Ajuste de modelos lineares usando estimadores de regressão para amostras complexas

Renata Pacheco Nogueira Duarte[1]

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DE GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: **Estatística**
Orientador: **Prof. Dr. Pedro Luis do Nascimento Silva**

02 de setembro de 1999

---

[1] *Durante a elaboração deste trabalho a autora recebeu apoio financeiro do CNPq.*

Ajuste de modelos lineares usando estimadores de regressão para amostras complexas

Este exemplar corresponde à redação final devidamente corrigida e defendida por Renata Pacheco Nogueira Duarte *e aprovada pela comissão julgadora.*

São Paulo, 17 de janeiro de 2000.

Banca examinadora:

- Prof. Dr. Pedro Luis do Nascimento Silva
- Prof. Dr. Rinaldo Artes
- Prof. Dr. Wilton de Oliveira Bussab

Agradecimentos

Agradeço a todos que me apoiaram durante a realização deste trabalho, especialmente a Zelia Bianchini e a meus colegas do Departamento de Metodologia do Instituto Brasileiro de Geografia e Estatística.

Um obrigada especial ao meu orientador Pedro Luis do Nascimento Silva, a Sonia Albieri e ao professor Pedro Luiz Valls Pereira pela paciência, pela atenção e pela força que sempre me deram, mesmo nos momentos de maior dificuldade (e pelas inúmeras vezes que tiveram que ler esta dissertação!).

Meu muito obrigada a Antonio José Ribeiro Dias pelo seu indispensável suporte computacional.

Finalmente, obrigada a minha família por toda a paciência que tiveram comigo e pelo seu apoio incondicional.

ABSTRACT

Data from sample surveys are being used more and more in statistical modeling and analysis. Large part of the data available comes from complex sample surveys, where the hypothesis of iid (independent and identically distributed) observations, which facilitates the attainment of theoretical results, cannot be adopted in an automatic way.

Sampling Theory has developed estimators of population quantities, such as means, totals and ratios, which take in account the sample design used in the survey as well as the observation weights. However, the literature about estimation of more complex statistics, as those used in modeling and data analysis, is still scarce. For this reason, it is important to examine how the application of complex sampling designs can affect the utilization of data for estimation and model fitting.

Nascimento Silva(1996) investigated the utilization of auxiliary population information for estimation and fitting of parametric models, adopting the Pseudo Maximum Likelihood approach.

The main target of this thesis is to revisit the work of Nascimento Silva(1996), extending the simulation study in order to evaluate the performance of the variance estimators of distinct estimators of the coefficient of a linear regression model. Furthermore, several estimation methods for coefficients of the linear model were applied to sample data from the 1991 Brazilian Demographic Census, selected from a small area in Marilia, São Paulo, Brazil.

RESUMO

Dados provenientes de pesquisas por amostragem vêm sendo cada vez mais usados para modelagem e análise estatística. Grande parte dos dados disponíveis hoje em dia provém de pesquisas com desenho amostral complexo, onde a hipótese de observações iid (independentes e identicamente distribuídas), que tanto facilita a obtenção de resultados teóricos, não pode ser feita de forma automática.

A Teoria da Amostragem tem desenvolvido estimadores de quantidades populacionais, como médias, totais e razões, que levam em conta o peso e o desenho amostral utilizado na pesquisa. Mas ainda há relativamente pouca literatura que trate das estatísticas mais complexas, como as usadas nas modelagens e análises dos dados. Por esse motivo é importante examinar como o emprego de desenhos amostrais complexos pode afetar o aproveitamento dos dados para a estimação e ajuste de modelos.

Nascimento Silva(1996) investigou o aproveitamento de informações populacionais auxiliares para a estimação e ajuste de modelos paramétricos "regulares", empregando o método de Máxima Pseudo-Verossimilhança.

Um dos objetivos deste trabalho foi revisitar o de Nascimento Silva(1996), estendendo o estudo de simulação para avaliar o desempenho dos estimadores de variância de diferentes estimadores dos coeficientes de um modelo de regressão linear.

Além disso, aplicaram-se diversos métodos de estimação de coeficientes de um modelo linear a dados obtidos com a amostra do Censo Demográfico de 1991, para uma das áreas de ponderação do município de Marília (SP).

Sumário

1 - Introdução

Dados provenientes de pesquisas por amostragem vêm sendo cada vez mais usados para modelagem e análise estatística. Grande parte dos dados disponíveis hoje em dia provém de pesquisas com desenho amostral complexo, onde a hipótese de observações iid (independentes e identicamente distribuídas), que tanto facilita a obtenção de resultados teóricos, não pode ser feita. Por exemplo, um desenho amostral por conglomerados se caracteriza por maior homogeneidade entre os elementos de um conglomerado, o que vai contra a hipótese de independência dos erros amostrais.

A maioria dos métodos estatísticos foi desenvolvida sob a hipótese de observações independentes e identicamente distribuídas. A Teoria da Amostragem tem desenvolvido estimadores de quantidades populacionais, como médias, totais e razões, que levam em conta o peso e o desenho amostral utilizado na pesquisa. Mas ainda há relativamente pouca literatura que trate das estatísticas mais complexas, como as usadas nas modelagens e análises dos dados. Por esse motivo é importante examinar como o emprego de desenhos amostrais complexos (envolvendo estratificação, conglomeração, seleção com probabilidades desiguais, etc.) pode afetar o aproveitamento dos dados para a estimação e ajuste de modelos paramétricos.

Nascimento Silva(1996, cap.6) investigou o aproveitamento de informações populacionais auxiliares para a estimação e ajuste de modelos paramétricos "regulares", empregando o método de Máxima Pseudo-Verossimilhança (Binder, 1983; Skinner, Holt e Smith, 1989, p.84). Foi examinado o caso particular de modelos de regressão linear, tendo sido comparados diversos estimadores dos coeficientes do modelo. Foi feito, também, um estudo da consistência destes diferentes estimadores via simulação.

Um dos objetivos deste trabalho foi revisitar o de Nascimento Silva(1996, cap.6), explorando um aspecto que não foi ali examinado de forma exaustiva ou completa: o estudo de simulação para avaliar o desempenho dos estimadores de variância de diferentes estimadores dos coeficientes do modelo no exemplo estudado. Além disso, foram aplicados diversos métodos de estimação de coeficientes de um modelo linear a dados obtidos através da amostra do Censo Demográfico de 1991 (Instituto Brasileiro de Geografia e Estatística - IBGE), para uma das áreas de ponderação do município de Marília (SP).

Nesta aplicação são usadas diversas opções que um analista tem para a estimação destes coeficientes, a saber: (1) a utilização de um ajuste "ingênuo" obtido por Mínimos Quadrados Ordinários, sem levar em conta nem o plano amostral nem os pesos divulgados na pesquisa; (2) o uso de Mínimos Quadrados Ponderados, levando-se em conta os pesos divulgados na pesquisa, mas sem usar informações do desenho amostral (em geral, o método de Mínimos Quadrados

Ponderados é usado para "corrigir" o caso de modelos com erros heteroscedásticos, mas aqui os pesos não correspondem aos fatores derivados das variâncias não constantes dos erros; eles indicam o número de unidades da população que cada unidade da amostra representa.); (3) a estimação dos parâmetros do modelo em questão através de um "pacote estatístico" que utiliza métodos que levam em conta o desenho amostral e os pesos divulgados; e (4) a aplicação dos métodos de MPV (Máxima Pseudo-Verossimilhança). O objetivo da aplicação, além de apresentar diversas opções de estimação, foi mostrar como se obtêm estimativas diferentes dos coeficientes e, principalmente, das variâncias das estimativas destes coeficientes, quando se ignora ou não as informações auxiliares e do desenho.

No capítulo 2 são considerados alguns estimadores existentes na literatura que se preocupam com o uso da hipótese de dados iid para dados de pesquisas por amostra. No capítulo 3 é apresentado mais detalhadamente o método de Máxima Pseudo-Verossimilhança para estimação de modelos de regressão linear com dados amostrais complexos. Uma simulação realizada para a avaliar o desempenho dos estimadores de variância de diversos estimadores de coeficientes de regressão linear em dois desenhos amostrais diferentes (Amostragem Aleatória Simples e Amostragem Estratificada Simples com probabilidades de seleção desiguais) é descrita no capítulo 4. No capítulo 5 apresenta-se a aplicação de diversos métodos de estimação aos dados obtidos através da amostra do Censo Demográfico de 1991 em uma das áreas de ponderação de Marília. Finalmente, o capítulo 6 contém as principais conclusões deste trabalho. Para maior facilidade foi adicionado um glossário no final deste trabalho com as abreviaturas mais usadas.

2 - Estimação de Coeficientes de Regressão em Amostras Complexas

A análise de regressão é uma importante ferramenta para análise de dados, e seu uso é bem comum, não só por estatísticos, como também por um grande número de analistas de diversas áreas. Os modelos usuais de regressão linear exigem a hipótese de observações independentes e identicamente distribuídas. Porém, a maior parte dos dados disponíveis hoje em dia são provenientes de pesquisas por amostragem. Para que esta hipótese de observações iid fosse válida, o desenho amostral empregado deveria ser Amostragem Aleatória Simples com reposição (AASC), o que não acontece na prática. A maioria das pesquisas por amostragem utilizam desenhos amostrais complexos e, muitas vezes, os usuários de dados de uma pesquisa, como, por exemplo, a PNAD (Pesquisa Nacional por Amostra de Domicílios), desconhecem o desenho amostral da mesma e não o levam em conta na modelagem, o que pode levar a resultados incorretos na hora da estimação e causar distorções na análise. Nathan e Holt(1980), por exemplo, mostraram em seu artigo que os estimadores usuais de MQO (Mínimos Quadrados Ordinários) são assintoticamente viciados, exceto para a classe dos desenhos com equiprobabilidade de seleção para todos os membros, onde o estimador MQO dos coeficientes é assintoticamente não viciado.

Existe, em geral, muita discussão sobre qual o melhor desenho amostral para uma determinada pesquisa e qual o melhor estimador das quantidades populacionais dado um desenho amostral. Apesar de existir bastante literatura sobre a estimação de medidas descritivas que incorporem o desenho amostral usado na obtenção dos dados, existe pouca literatura sobre modelagem ou sobre análise de dados de pesquisas por amostragem, e ainda há pouca literatura sobre como incorporar o desenho amostral na análise de modelos lineares.

Kish e Frankel(1974) foram dos primeiros a tratar do tema e a iniciar uma discussão sobre o problema da estimação de estatísticas analíticas, tais como coeficientes de regressão e de correlação, com dados provenientes de amostras complexas. Eles classificaram os métodos de seleção de amostra e os parâmetros de interesse em três níveis cada, conforme pode ser visto no Quadro 2.1.

QUADRO 2.1

Classificação dos Métodos de Seleção da Amostra e dos Parâmetros de Interesse

Métodos de seleção dos elementos	Parâmetros de Interesse		
	1 Médias e totais de amostras completas	2 Médias e diferenças entre domínios	3 Estatísticas analíticas
A. Amostragem aleatória simples			
B. Amostragem estratificada		B2	B3
C. Amostragem complexa		C2	C3

A classificação dos problemas de inferência se deu pelo cruzamento dos métodos de seleção com os parâmetros de interesse. Não foram feitos comentários sobre problemas de inferência na linha A e coluna 1 do Quadro 2.1 por representarem áreas já muito discutidas em livros-texto da estatística clássica (linha A) e da teoria da amostragem (coluna 1). Sobre B2 e C2, os resultados apresentados no artigo eram simples e diretos, e já podem ser facilmente encontrados na literatura de amostragem (por exemplo, Särndal, Swensson e Wretman, 1992). Em relação às estatísticas analíticas sob uma amostra com desenho estratificado simples (B3) foram apresentadas algumas hipóteses razoáveis que estavam surgindo na época, com alguma justificativa empírica, que tornavam este problema mais simples. Mas o artigo se concentrou no problema da estimação de estatísticas analíticas com dados de amostras complexas, e apresentou vários resultados e sugestões sobre este problema, relativo à célula C3 do Quadro 2.1.

Segundo o artigo, pode-se aceitar a normalidade aproximada dos estimadores, na hipótese de que as amostras são grandes o suficiente, apesar da violação da hipótese de independência das observações. Porém, não prestar atenção na complexidade dos desenhos amostrais no cálculo dos erros padrões desses estimadores é que pode levar a problemas graves de vício na medição da precisão.

Pode-se dividir as estatísticas em dois níveis. Chamou-se de estatísticas de primeira ordem as estimativas dos parâmetros da distribuição da população que são o objetivo da pesquisa e não são tão afetadas pelo desenho amostral, por exemplo, médias e variâncias dos elementos, coeficientes de regressão e coeficientes de correlação. Por estatísticas de segunda ordem entende-se as medidas de variação amostral (variância, desvio padrão, coeficiente de variação (CV) e erro quadrático médio (EQM)) das estatísticas de primeira ordem. Elas estimam os

parâmetros de segunda ordem $\left(E\left\{\hat{\theta}-\theta\right\}^2\right)$, com base nos resultados agregados de amostras obtidas sob desenhos específicos, e são afetadas pelas correlações entre as observações amostrais, produzidas pelo desenho. Quando estamos interessados em estatísticas de primeira ordem, mesmo em amostras complexas, se sabemos apenas o peso das observações amostrais podemos chegar a estimativas pontuais "corretas" mesmo com os pacotes estatísticos usuais. Mas para estatísticas de segunda ordem podem ocorrer sérios erros se as informações sobre o desenho amostral não forem levadas em consideração nos seus cálculos. Assim, o principal foco da discussão deste artigo (Kish e Frankel, 1974) foram maneiras de obter estimativas mais corretas das estatísticas de segunda ordem que as estimativas "ingênuas", calculadas como se o desenho amostral fosse AASC (Amostragem Aleatória Simples com reposição).

Kish e Frankel(1974) usaram uma medida chamada Deff (do inglês Design Effect), ou efeito do desenho, que pode ser definido como a razão entre a variância do estimador sob o desenho amostral efetivamente adotado e a variância do estimador ignorando o plano amostral complexo, isto é, supondo que aquela mesma amostra fosse proveniente de AASC com o mesmo número de elementos (Kish, 1965, p.258; Pessoa e Nascimento Silva, 1998, p.54)). Esta medida, então, expressa o efeito que o uso de um plano amostral diferente de AASC tem na variância do estimador, ou ainda o erro que estaríamos cometendo ao usarmos um estimador "ingênuo" da variância do estimador sob um plano amostral diferente de AASC. Uma conclusão do artigo baseada num estudo empírico de simulação é que o deff para amostras complexas foi maior do que 1 em geral, ou seja, os erros padrões estimados com base em hipóteses de AASC tendem a subestimar os erros padrões de estatísticas analíticas em amostras complexas.

Também foi colocado em Kish e Frankel(1974) que as formas exatas de estatísticas complexas e, principalmente, das medidas de variação destas estatísticas, no caso de desenhos amostrais complexos, podem não ser nada simples de serem obtidas. As estatísticas de segunda ordem são mais complicadas que o cálculo das estatísticas de primeira ordem a que elas estão ligadas. Este problema tende a aumentar quando se trata de estatísticas multivariadas complexas, como coeficientes de regressão multivariada. Assim, neste artigo, houve pouco tratamento analítico explícito das propriedades amostrais dos coeficientes de regressão, e nenhum tratamento analítico explícito de suas propriedades/momentos de segunda ordem. Os autores deste artigo optaram por usar métodos "indiretos" para o cálculo das variâncias de diversos estimadores, inclusive de estimadores de coeficientes de regressão, com dados de diversos desenhos amostrais complexos. Os métodos "indiretos" usados foram o método de Linearização de Taylor, o método das Replicações Repetidas Balanceadas (BRR) e o método das Replicações

Repetidas Jackknife (JRR). Para compará-los, foi feito um estudo empírico, usando estes três métodos para se calcular a variância de várias estatísticas, entre elas médias e coeficientes de regressão múltipla, em três desenhos amostrais estratificados com conglomeração.

A principal conclusão a que os autores chegaram foi que, quando avaliados por vários critérios, nenhum desses 3 métodos se mostrou clara e consistentemente melhor ou pior que os outros. A escolha do método pode depender do custo relativo e da simplicidade, e isto varia muito com as situações e com as estatísticas de interesse. O método de Taylor pode ser melhor para estatísticas simples como razão de médias, e o BRR ou o JRR para estatísticas complexas como coeficientes de regressão múltipla. Os estimadores de variância baseados tanto nas aproximações de Taylor, como em BRR ou em JRR deram estimativas razoáveis da variância dos estimadores de coeficientes de regressão. Mas o mais importante deste trabalho foi destacar que o uso de estimadores "ingênuos" para os erros padrões pode levar a graves erros em desenhos amostrais complexos pois, em geral, estes estimadores subestimam os erros padrões verdadeiros dos estimadores.

Fuller(1975) também investigou o problema da estimação de equações de regressão para uma amostra selecionada de uma população finita, mas apenas para o caso de AAS sem reposição. Porém, neste artigo, este problema teve um tratamento analítico rigoroso e não foi usado nenhum método indireto de estimação, ao contrário de Kish e Frankel(1974). Em Fuller(1975) não só os momentos foram obtidos, como também a distribuição assintótica normal dos estimadores dos coeficientes num modelo de regressão.

A população finita foi tratada como uma amostra aleatória simples de uma população multivariada infinita, com os quatro primeiros momentos finitos, e foi suposto que a matriz de covariância desta população multivariada era positiva definida.

Para investigar o comportamento do estimador dos coeficientes de regressão de uma população finita é necessário fazer hipóteses sobre distribuição e/ou uso de aproximações para grandes amostras. Fuller investigou o comportamento, no limite, dos coeficientes estimados tanto quando o tamanho da amostra, quanto quando o tamanho da população aumentam.

Considere o modelo de regressão linear simples definido por $y = zB + e$, onde $z = (z_1' z_2' ... z_N')'$ é a matriz das variáveis independentes para os N elementos da população, $z_i = (z_{i0}, z_{i1}, ..., z_{ip})$, com $z_{i0} \equiv 1$, é o vetor de observações da i-ésima unidade da população, i=1,...,N, e é o vetor de erros, $y_{N\times 1}$ é o vetor da variável dependente e B é o vetor dos p+1 coeficientes de regressão da população finita.

O vetor dos (p+1) coeficientes de regressão da população infinita foi definido por:

$$\beta = Q^{-1}H \tag{2.1}$$

e o vetor dos coeficientes da população finita por:

$$B = Q_N^{-1}H_N \tag{2.2}$$

onde os elementos de Q, Q_N, H e H_N são dados respectivamente por:

$$q_{rs} = E\{z_{ir}z_{is}\} \tag{2.3}$$

$$q_{Nrs} = N^{-1}\sum_{i=1}^{N} z_{ir}z_{is} \tag{2.4}$$

$$h_s = E\{z_{is}y_i\} \tag{2.5}$$

$$h_{Ns} = N^{-1}\sum_{i=1}^{N} z_{is}y_i \tag{2.6}$$

e $z_{i0} \equiv 1$.

O estimador amostral de β, baseado numa Amostra Aleatória Simples de tamanho n, é dado por:

$$b = Q_n^{-1}H_n \tag{2.7}$$

onde o rs-ésimo elemento de Q_n é (r,s = 0,1,2,...,p):

$$q_{nrs} = n^{-1}\sum_{i=1}^{n} z_{ir}z_{is} \tag{2.8}$$

e o s-ésimo elemento de H_n é (s = 0,1,2,...,p):

$$h_{ns} = n^{-1}\sum_{i=1}^{n} z_{is}y_i . \tag{2.9}$$

Seja $\{\xi_\nu : \nu = 1,2,...\}$ uma seqüência de populações finitas, onde ξ_ν é uma amostra aleatória de tamanho N_ν, $N_\nu \geq N_{\nu-1}$, selecionada de uma população infinita (p+2)-dimensional. Assuma que a população infinita possui momento de 4ª ordem finito e uma matriz de covariância positiva definida. Selecione uma amostra aleatória simples sem reposição de tamanho n_ν da ν-ésima população finita, $\nu = 1,2,...$ e $n_\nu \geq n_{\nu+1}$. Defina $f_\nu = n_\nu / N_\nu$ e seja $\lim_{\nu\to\infty} f_\nu = f$, $0 \leq f < 1$. Sob estas condições, Fuller(1975) mostra no seu Teorema 1 que $n_\nu^{1/2}(b - B) \xrightarrow{\mathcal{L}} N\left(0,(1-f)Q^{-1}GQ^{-1}\right)$, quando $\nu \to \infty$, onde b é definido em (2.7), B é definido em (2.2), e o rs-ésimo elemento de G é $G_{rs} = E\{z_{ir}z_{is}e_i^2\}$, e $e_i = y_i - \sum_{r=0}^{p}\beta_r z_{ir}$.

Segue deste Teorema 1 que $n^{1/2}(b - \beta) \xrightarrow{\mathcal{L}} N\left(0, Q^{-1}GQ^{-1}\right)$. Assim, não se usaria a correção para a população finita para estimar o parâmetro da população infinita.

Seja

$$\hat{G} = \frac{1}{n-p-1}\sum_{i=1}^{n} \hat{d}_i' \hat{d}_i \qquad (2.10)$$

onde

$$\hat{d}_i' = z_i' \hat{e}_i \qquad (2.11)$$

$$\hat{e}_i = y_i - z_i b. \qquad (2.12)$$

Seja a seqüência de amostras e populações finitas que satisfazem as hipóteses do Teorema 1 e seja $\hat{G}$ como definido em (2.10). Pelo Teorema 2 do artigo, $p\lim\hat{G} = G$. Ou seja, com os dois teoremas o artigo fornece um método para se obter um estimador consistente da variância na ausência das hipóteses usuais dos modelos lineares, sem precisar recorrer a métodos indiretos de estimação, assim como uma distribuição assintótica e os momentos para os estimadores. Porém, como já foi dito, este artigo tratou apenas do caso de AAS sem reposição.

O Teorema 1 pode também ser estendido, pelo uso de TCL, para o coeficiente de regressão estimado para amostras estratificadas em 2 estágios.

Houve também artigos que trataram do problema de estimação dos coeficientes de regressão utilizando outras abordagens. Pfeffermann e Nathan(1977) e Pfeffermann e Nathan(1981) propuseram um método para estimação de parâmetros de um modelo de regressão em pesquisas com desenho amostral complexo, onde diferentes "grupos" da população apresentam diferentes relações de regressão, mas apenas uma parte dos grupos pode ser incluída na amostra. Ou seja, os grupos da população foram considerados mutuamente exclusivos e dentro de cada um destes grupos foi assumido o modelo:

$y_{ij} = z_{ij}\beta_i + e_{ij}$ ($j=1,...,M_i$ e $i=1,...,N$,), onde $E(e_{ij} \mid z_{ij}) = 0$, $E(e_{ij}\, e_{ij} \mid z_{ij}) = \sigma_i^2$ e $E(e_{ij}\, e_{kl} \mid z_{ij}, z_{kl})=0$ para $i\neq j$ ou $k\neq l$. Estes modelos foram restritos, por simplicidade, ao caso de z_{ij} e β_i serem escalares (regressão simples). Os autores trataram os coeficientes de regressão de cada grupo como variáveis aleatórias não-correlacionadas, com $E(\beta_i) = \beta$, e $Var(\beta_i) = \delta^2$ e $Cov(\beta_i,\beta_j) = 0$ para $i\neq j$. O parâmetro de interesse foi definido como sendo uma média ponderada destes coeficientes de regressão, ou seja, $\beta(\omega) = \sum_{i=1}^{N} \omega_i \beta_i$, com os ω_i ($i=1,...,N$) conhecidos e, sem perda de generalidade, consideraram que $\sum_{i=1}^{N} \omega_i = 1$. A abordagem adotada para obter o estimador ótimo $\hat{\beta}(\omega)$ foi a Bayesiana e este foi derivado sob as seguintes hipóteses: (1) distribuição normal para os erros de regressão dentro de cada grupo, com variâncias conhecidas, (2) distribuição a priori normal para os coeficientes de regressão, também com variância

conhecida, e (3) distribuição a priori localmente uniforme para a média destes coeficientes de regressão, ou seja $\beta_i \sim N(\beta,\delta^2)$ e $P(\beta)$ = constante. Esta última hipótese foi relaxada quando se investigou as propriedades do estimador.

Eles mostraram que o estimador de Bayes $\hat{\beta}(\omega)$, sob este modelo, é um estimador não viciado para $\beta(\omega)$ e o EQM de $\hat{\beta}(\omega)$ é igual à variância da distribuição a posteriori de $\beta(\omega)$. Pfeffermann e Nathan(1977) mostram, com seu Teorema 1, que sob o modelo descrito, o estimador de Bayes $\hat{\beta}(\omega)$ apresenta erro quadrático médio mínimo entre todos os estimadores lineares de $\beta(\omega)$, qualquer que seja a forma da distribuição a priori dos coeficientes de regressão dos grupos.

Denote por $\hat{\beta}_c(\omega)$ o estimador "clássico" de $\beta(\omega)$, o estimador obtido pela soma ponderada dos estimadores dos β_i, sendo que estes foram estimados por mínimos quadrados nos grupos que pertencem à amostra e, nos grupos que não pertencem, foram estimados por média simples dos estimadores dos β_i dos grupos representados na amostra. Este estimador clássico é também linear em y_{ij} e não-viciado. Foi mostrado, num corolário do artigo de Pfeffermann e Nathan(1977), que a perda pelo uso do estimador "clássico" ao invés do estimador ótimo (estimador de Bayes), $E_\xi\left[\hat{\beta}_c(\omega)-\beta(\omega)\right]^2 - E_\xi\left[\hat{\beta}(\omega)-\beta(\omega)\right]^2$, é maior do que zero.

Este artigo também mostra que todas as propriedades ótimas de $\hat{\beta}(\omega)$, que foram derivadas sob a hipótese de que os β_i são observações aleatórias não-correlacionadas de uma população com média β e variância δ^2, também podem ser estendidas para o caso onde os valores esperados dos β_i são funções lineares de valores conhecidos de uma variável adicional x. Assim, ao invés de se assumir que $E(\beta_i) = \beta$, e $Var(\beta_i) = \delta^2$ e $Cov(\beta_i,\beta_j) = 0$ para $i\neq j$ $(i=1,...,N)$, pode-se assumir que $E(\beta_i) = \beta+\alpha x_i$, $Var(\beta_i) = \delta^2$ e $Cov(\beta_i,\beta_j) = 0$ para $i\neq j$ $(i=1,...,N)$. Em Pfeffermann e Nathan(1981) foi mostrado que essas propriedades também podem ser estendidas ao caso onde $E(\beta_i) = \beta+\nu_i$, $Var(\beta_i) = \delta^2$ e $Cov(\beta_i,\beta_j) = 0$ para $i\neq j$ $(i=1,...,N)$.

Tratou-se também, em Pfeffermann e Nathan(1977) e Pfeffermann e Nathan(1981), do problema das variâncias a priori serem desconhecidas, propondo-se um método para estimar a variância dos coeficientes de regressão pela maior solução de uma equação. Esta equação foi derivada sob as hipóteses de um modelo linear heteroscedástico, para a estimação dos β_i. O método garante um estimador não-negativo para a variância e que tende em probabilidade para o verdadeiro valor da variância. Foram usados os estimadores não viciados clássicos para os σ_i dentro dos grupos selecionados.

O principal problema da abordagem de Pfeffermann e Nathan(1977) e Pfeffermann e Nathan(1981) é que, como foi mostrado nos artigos, ela foi definida sob várias hipóteses (distribuição normal para os erros de regressão dentro de cada grupo, com variâncias conhecidas, distribuição a priori normal para os coeficientes de regressão com variância conhecida, e priori localmente uniforme para suas médias), o que restringe um pouco o seu uso. Além disso, a abordagem Bayesiana não foi seguida neste trabalho.

2.1 – Inclusão de informações auxiliares no modelo linear

Uma maneira de resolver o problema de ajuste de modelo linear em dados de amostras complexas seria incorporar informações populacionais auxiliares na modelagem. Essas informações auxiliares poderiam ser incorporadas de duas maneiras: na própria definição do modelo ou na estimação dos parâmetros.

Para a i-ésima observação, considere o vetor P×1 de dados $z_i = (z_{i1},\ldots,z_{iP})'$, que é o vetor das variáveis explicativas. Seja $y_i = (y_{i1},\ldots,y_{iA})'$ o vetor A×1 das variáveis dependentes. Denote por Y_i e Z_i as variáveis aleatórias e os vetores que geram y_i e z_i, para $i \in U$, onde $U = \{1,2,...,N\}$ é o conjunto de rótulos usados para identificar as unidades da população de interesse, também chamado de Universo. Seja x_i um vetor qx1 que contém as variáveis de desenho ou variáveis auxiliares para o elemento i e seja X_i o vetor aleatório que gera x_i. Então, se quisermos incorporar informações sobre o desenho amostral e/ou informações auxiliares, podemos fazer isso de duas maneiras.

a) Especifica-se um modelo supondo que a esperança condicional de Y_i dadas as variáveis explicativas Z_i e as variáveis auxiliares X_i é da forma:

$$E_M(Y_i \mid Z_i = z_i, X_i = x_i) = \beta_0 + z_i'\beta + x_i'\eta \qquad (2.13)$$

Ou seja, as variáveis x são diretamente incorporadas no modelo. Chamamos esta maneira de ABORDAGEM DESAGREGADA.

b) Especifica-se um modelo supondo que a esperança condicional de Y_i dadas as variáveis explicativas Z_i é da forma:

$$E_M(Y_i \mid Z_i = z_i) = \beta_0 + z_i'\beta \qquad (2.14)$$

Ou seja, as variáveis x não são diretamente incorporadas no modelo, mas podem ser consideradas na estimação dos parâmetros. Chamamos esta maneira de ABORDAGEM AGREGADA.

Quando consideramos o caso onde temos apenas uma variável explicativa e uma variável auxiliar, o ajuste de (2.13) recai num problema simples de estimação, e parece um modo direto

de resolver este problema. Apesar disso, Nascimento Silva(1996, cap.6) apresenta algumas considerações sobre dificuldades de se utilizar sempre (2.13) ao invés de (2.14). Uma delas é que nem sempre, em situações práticas, toda a informação auxiliar pode estar contida em uma única variável auxiliar. Em geral, são necessárias diversas variáveis, como indicadores de estratos e conglomerados, etc., o que torna o ajuste de (2.13) mais complexo. Outro problema é de como se pode deixar clara a divisão entre as variáveis Z e as variáveis auxiliares X. Nathan e Holt(1980) e Holt, Smith e Winter(1980) consideram como variáveis auxiliares apenas as que afetam a seleção amostral (variáveis de "desenho"). Mas pode existir um conjunto de variáveis que não são o interesse principal da pesquisa e não são variáveis de desenho, mas para as quais se possui algum tipo de informação populacional auxiliar que pode ser utilizada na estimação.

Smith(1981) faz uma discussão sobre as diversas maneiras existentes de ajustar os modelos dentro da abordagem agregada sugerida por (2.14). Existem 2 tipos de inferência: inferências descritivas e inferências analíticas. As inferências descritivas são aquelas em que o parâmetro de interesse é uma função conhecida de valores relacionados às N unidades da população finita como, por exemplo, o vetor B dos coeficientes da população finita, definido por (2.2), (2.4) e (2.6). O objetivo seria estimar uma dada propriedade ou uma dada população no momento em que se estava selecionando a amostra. Se as unidades fossem todas investigadas e não existissem erros de medida, não existiria incerteza nas inferências descritivas. Se o parâmetro de interesse não puder ser expresso como uma função de valores ligados às N unidades da população, então esta inferência será analítica, e o objetivo será estimar um parâmetro em outra população relacionada de alguma maneira com a população alvo. Esta outra população pode ser considerada como uma superpopulação que representa a esfera mais ampla de inferência. Estimação de um total da população finita é uma inferência descritiva; estimar coeficientes de um modelo econômico é, usualmente, uma inferência analítica.

Existem duas teorias bem desenvolvidas para fazer inferências descritivas. Uma é baseada na distribuição gerada pela amostragem aleatória, a p-distribuição p(s), que é a distribuição baseada no desenho. A segunda é baseada numa teoria que emprega um modelo estocástico para representar a estrutura da população e as inferências estarão baseadas na distribuição de probabilidades especificada no modelo, chamada de ξ-distribuição, ou distribuição baseada no modelo. Existem teorias híbridas, que combinam essas duas abordagens. Um problema com as inferências baseadas puramente no modelo é que não se leva em conta o desenho amostral da pesquisa na inferência.

Para as inferências analíticas, já que o parâmetro a ser estimado não é uma função dos valores de todas as unidades populacionais, é difícil formular uma teoria satisfatória de inferência

empregando apenas a distribuição baseada no desenho; precisa-se de uma hipótese que relacione a população finita com o parâmetro analítico de interesse. Em geral, se assume que os valores da população finita correspondem a uma amostra aleatória de uma superpopulação (infinita). Já que o tamanho N desta população finita é, em geral, bem grande, um estimador da população finita baseado nos N valores seria bem "próximo" do parâmetro desconhecido da superpopulação. Assim, considera-se o estimador da população finita como um parâmetro da população finita que poderá ser estimado usando-se a distribuição baseada no desenho.

É possível formular modelos de regressão para inferências descritivas ou analíticas, e para empregar tanto a distribuição sob o desenho quanto sob o modelo. Assim, pode-se dividir este problema em 3 casos, de acordo com o Quadro 2.2:

QUADRO 2.2

Parâmetros de Interesse	*Procedimento de Amostragem*	
	Em uma população finita	**Em um modelo de superpopulação**
Parâmetros de uma população finita	Caso 1: Teoria clássica de amostragem em população finita – inferências baseadas no desenho	Caso 2: Teoria de superpopulação para amostragem da população finita – inferências baseadas no modelo
Parâmetros de um modelo de superpopulação	Não definido	Caso 3: Inferências sobre os parâmetros da superpopulação - considerar amostragem em 2 estágios

Kish e Frankel(1974) estudaram o Caso 1, e Fuller(1975) estudou o Caso 2.

Uma das primeiras análises baseadas no Caso 3 foi de Pearson(1902). Os resultados de Pearson foram revisitados e desenvolvidos por vários artigos, inclusive por Holt, Smith e Winter(1980) e Nathan e Holt(1980).

2.2 – O estimador de Pearson e o de Pearson-ajustado

Vamos considerar a situação onde uma variável de desenho X é conhecida num estágio anterior à seleção para todos os membros de uma população finita. Após a amostragem, tem-se observações de Y, a variável dependente, e de Z, a variável independente da análise de regressão. Vamos considerar as duas possibilidades para o modelo de regressão em (2.13) e (2.14), re-escrevendo as equações particionando-se o vetor de parâmetros:

$$E\{Y \mid Z,X\} = \mu_Y + \beta_{YZ\cdot X}(Z-\mu_Z) + \beta_{YX\cdot Z}(X-\mu_X) \quad \text{(modelo desagregado)} \qquad (2.15)$$

$$E\{Y \mid Z\} = \mu_Y + \beta_{YZ}(Z-\mu_Z) \quad \text{(modelo agregado)} \qquad (2.16)$$

Em (2.15) (modelo desagregado) os estimadores não viciados dos parâmetros, baseados no modelo, são obtidos, por exemplo, por Mínimos Quadrados e a função do desenho amostral baseado em X seria somente a de melhorar as propriedades dos estimadores. A abordagem básica de Nathan e Holt (1980) e Holt, Smith e Winter(1980) é (2.16) (modelo agregado), onde X é usado no estágio de desenho, anterior à seleção da amostra, mas não explicitamente no modelo de regressão. Isto pode ser usado quando a variável de desenho X não explica a variável dependente. Neste caso, o parâmetro de regressão "marginal" β_{YZ} tem maior interesse que os parâmetros de regressão parciais $\beta_{YZ\cdot X}$ e $\beta_{YX\cdot Z}$.

O problema considerado, então, foi o de estimar o coeficiente de regressão linear de Y e Z quando dados de pesquisa por amostragem são obtidos por um desenho baseado em X, que é, em geral, relacionado a Y e Z. Ou seja, vamos considerar aqui a variável X como uma variável de desenho, que se conhece para toda a população finita antes de se retirar a amostra de Y e Z.

Nathan e Holt(1980) verificaram que o estimador usual de MQO da variância é, em geral, viciado, mesmo quando usado conjuntamente com um desenho com equiprobabilidade de seleção para todos os membros, com o qual o estimador de MQO dos coeficientes de regressão é não viciado neste caso. Foram então considerados estimadores alternativos que levem a estimativas não viciadas de coeficientes de regressão simples.

Uma segunda questão é em que nível se faz a inferência. O ponto de vista que foi usado em Nathan e Holt(1980) é que a regressão é descrita por um modelo de superpopulação, e que a população finita é vista como uma amostra desta superpopulação. O objetivo é o parâmetro β_{YZ} do modelo de superpopulação. Na prática, o que se tem é apenas uma amostra da população finita e as propriedades dos estimadores estarão relacionadas à distribuição de todas as possíveis amostras da população finita, consistente com o desenho usado. Se a população finita é considerada como uma amostra aleatória usualmente grande da superpopulação, então, B_{YZ} seria um bom estimador de β_{YZ}. Assim, qualquer estimador razoável de B_{YZ} de uma amostra da população finita espera-se que seja um estimador razoável de β_{YZ}.

Se o estimador considerado fosse o de MQO, ele seria viciado, exceto se a população tivesse uma estrutura especial ($\rho_{ZX} = 0$ ou $\rho_{YX\cdot Z} = 0$) ou se o desenho usado tivesse uma característica particular [$E(s_X^2) = \sigma_X^2$], o que é obedecido por alguns desenhos com equiprobabilidade de seleção em todos os membros. Quando se tem condições especiais da estrutura da população, o estimador de MQO da variância dos estimadores dos coeficientes é também não viciado. O mesmo não ocorre quando apenas a condição de balanceamento do desenho é satisfeita.

Em Nathan e Holt(1980) e Holt, Smith e Winter(1980), um estimador alternativo $\hat{\beta}_{YZ}$ foi considerado, e foi obtida sua variância assintótica. Sob as condições dadas, quando o estimador de MQO é assintoticamente não viciado, nenhum destes dois estimadores tem a variância uniformemente menor que o outro. Nestes artigos foi levada em conta uma população finita de tamanho N selecionada de uma superpopulação, e valores observados de X, independentes e identicamente distribuídos, com média μ_X e variância σ_X^2.

A hipótese do desenho é que X está relacionada de alguma maneira a Y e/ou a Z. Ao invés de adotarem hipóteses de normalidade trivariada, como fez Pearson(1902), Nathan e Holt(1980) adotaram apenas um conjunto de hipóteses de modelos lineares:

$$\begin{cases} y_i = \mu_Y + \beta_{YX}(x_i - \mu_X) + e_{Yi} \\ z_i = \mu_Z + \beta_{ZX}(x_i - \mu_X) + e_{Xi} \\ e_{Yi} = \beta_{YZ\cdot X}\, e_{Zi} + \eta_{Yi} \end{cases} \tag{2.17}$$

onde $E(e_{Zi} \mid X_i) = E(\eta_{Yi} \mid X_i) = E(e_{Zi}\eta_{Yi} \mid X_i) = 0$ (2.18)

$$\begin{aligned} E(e_{Zi}^2 \mid X_i) &= \sigma_{Z\cdot X}^2 \\ E(\eta_{Yi}^2 \mid X_i) &= \sigma_{Y\cdot ZX}^2 \end{aligned} \tag{2.19}$$

Estas condições são equivalentes aos valores esperados condicionais de Y e Z dado X, sendo linear em X e a matriz de covariância condicional de Y e Z, dado X, não depende de X. Além disso, foi assumida a independência condicional para diferentes unidades, isto é, que (η_{Yi}, e_{Zi}) e (η_{Yj}, e_{Zj}) são condicionalmente independentes dado X.

Uma amostra s foi selecionada da população finita por um desenho amostral com tamanho fixo n. O desenho pode ser baseado nos valores de X conhecidos para toda a população. Assim, X pode servir como uma variável de tamanho para amostragem proporcional ao tamanho ou como a variável de agrupamento para estratificação ou amostragem por conglomerados. O parâmetro de interesse é o coeficiente de regressão da superpopulação de Y em Z, β_{YZ}, ou seja,

$$\beta_{YZ} = \sigma_{YZ}\sigma_{ZZ}^{-1}. \tag{2.20}$$

Os resultados a seguir não dependem da existência desta relação linear.

$$\text{Sejam} \begin{cases} \overline{Y} = \sum_{i=1}^{N} y_i / N, \quad S_Y^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2 / N \\ \overline{Z} = \sum_{i=1}^{N} z_i / N, \quad S_Z^2 = \sum_{i=1}^{N} (z_i - \overline{Z})^2 / N \\ \overline{X} = \sum_{i=1}^{N} x_i / N, \quad S_X^2 = \sum_{i=1}^{N} (x_i - \overline{X})^2 / N \end{cases} \tag{2.21}$$

As estatísticas amostrais $\bar{y}$, $\bar{z}$, $\bar{x}$, s_Y^2, s_Z^2, s_X^2, s_{YX}, s_{YZ}, s_{ZX}, são definidas como

$\bar{y} = \sum_{i\in s} y_i / n$, $\bar{z} = \sum_{i\in s} z_i / n$, $\bar{x} = \sum_{i\in s} x_i / n$, $s_Y^2 = \sum_{i\in s} (y_i - \bar{y})^2 / (n-1)$, $s_Z^2 = \sum_{i\in s} (z_i - \bar{z})^2 / (n-1)$, $s_X^2 = \sum_{i\in s} (x_i - \bar{x})^2 / (n-1)$, $s_{YZ} = \sum_{i\in s} (y_i - \bar{y})(z_i - \bar{z}) / (n-1)$, $s_{YX} = \sum_{i\in s} (y_i - \bar{y})(x_i - \bar{x}) / (n-1)$, e $s_{ZX} = \sum_{i\in s} (z_i - \bar{z})(x_i - \bar{x}) / (n-1)$. Os parâmetros μ_Y, μ_Z, μ_X, σ_Y^2, σ_Z^2, σ_X^2, σ_{YX}, σ_{ZX} e σ_{YZ}, são definidos como o usual.

Foi considerada uma seqüência de populações e de desenhos amostrais apropriados, tal que tanto n quanto N tendam ao infinito, com n < N. O valor esperado e a variância condicionais sob o modelo, dado X foram representados por $E_\xi(\bullet|X,s)$ e $V_\xi(\bullet|X,s)$ e se denotou por $E_\pi(\bullet|D)$ e $V_\pi(\bullet|D)$ os valores esperados condicionais dados os valores na população finita de D'=(Y',Z',X'), gerado pelo desenho amostral p(s). Finalmente, foram usados os valores esperados sob a distribuição conjunta do modelo e do desenho, denotados por $E(\bullet)$ e $V(\bullet)$.

Considerou-se que além das variáveis de regressão observadas na pesquisa, os valores das variáveis de desenho eram conhecidos para todas as unidades da população finita na matriz $X_{N\times q}$. Esta matriz possui toda a informação numérica que é usada para determinar o desenho amostral p(s), escrito, então, como p(s|X). O desenho possui informação sobre as variáveis de pesquisa através de sua relação com as variáveis de desenho. Numa regressão, tanto os valores da variável dependente Y quanto das variáveis independentes Z podem ser relacionados a X. A proposta do artigo é que se leve este relacionamento em conta para inferências.

Dada uma amostra de n unidades selecionada com o desenho p(s|X), foi derivado o estimador de máxima verossimilhança de β_{YZ}. Se $(n-1)s_{jk}$ é a soma de quadrados ou a soma de produtos cruzados na amostra correspondente a σ_{jk}, (com j, k podendo ser substituídos por Y, Z ou X) e NS_X^2 é a soma de quadrados ou a soma de produtos cruzados baseada em todas as N observações, então o estimador de máxima verossimilhança de β_{YZ}, conhecido como estimador de Pearson é (Holt, Smith e Winter; 1980)

$$\hat{\beta}_{YZ} = \left\{ s_{YZ} + s_{YX} s_X^{-2} \left(S_X^2 s_X^{-2} - I \right) s_{XZ} \right\} \left\{ s_Z^2 + s_{ZX} s_X^{-2} \left(S_X^2 s_X^{-2} - I \right) s_{XZ} \right\}^{-1} \qquad (2.22)$$

Holt, Smith e Winter(1980) apresentam também o resultado derivado por DeMets e Halperin(1977) para o caso especial de regressão onde Y, Z e X são vetores unitários. Assim (2.22) se reduz a

$$\hat{\beta}_{YZ} = \left\{ s_{YZ} + \frac{s_{YX} s_{ZX}}{s_X^2} \left(\frac{S_X^2}{s_X^2} - 1 \right) \right\} \Bigg/ \left\{ s_Z^2 + \frac{s_{ZX}^2}{s_X^2} \left(\frac{S_X^2}{s_X^2} - 1 \right) \right\}. \qquad (2.23)$$

Em Pfeffermann e Holmes(1985) também é apresentado um estimador da constante do modelo neste caso onde Y, Z e X são vetores unitários:

$$\hat{\beta}_{0YZ} = \hat{\mu}_Y - \hat{\beta}_{YZ}\hat{\mu}_Z \qquad (2.24)$$

onde $\hat{\mu}_Y = \bar{y} + b_{YX}(\overline{X} - \bar{x})$, $\hat{\mu}_Z = \bar{z} + b_{ZX}(\overline{X} - \bar{x})$, $b_{YX} = s_{YX}/s_X^2$ e $b_{ZX} = s_{ZX}/s_X^2$.

Este estimador é assintoticamente não viciado para β_{YZ} e pode ser usado sem qualquer hipótese sobre distribuição, sendo também assintoticamente não viciado sob as hipóteses mais fracas (2.17).

Sob as hipóteses (2.17) e (2.18), a variância de $\hat{\beta}_{YZ}$ foi obtida por Nathan e Holt(1980)

$$V\left[\hat{\beta}_{YZ}\right] = \left(1-\rho_{YZ}^2\right)\frac{\sigma_Y^2}{n\sigma_Z^2}\left\{\left(1-\rho_{YX\cdot Z}^2\right)\left[1+\rho_{ZX}^2\left(\frac{1}{Q-1}\right)\right]+\frac{1}{Q}\rho_{YX\cdot Z}^2\left(1-2\rho_{ZX}^2\right)^2+\right.$$
$$\left.\left(1-\rho_{ZX}^2\right)\rho_{ZX}^2\rho_{YX\cdot Z}^2\left[\frac{\mu_4(z\mid x)}{\sigma_{Z.X}^4}-1+\frac{n}{N}\left(\frac{\mu_4(x)}{\sigma_X^4}-1\right)\right]\right\} \qquad (2.25)$$

onde $\mu_4(z \mid x) = E_\xi(e_Z^4 \mid x)$, $\mu_4(x) = E_\xi(x^4)$ e $Q = E(s_X^2)/\sigma_X^2$. O estimador de $V\left[\hat{\beta}_{YZ}\right]$, $\hat{V}\left[\hat{\beta}_{YZ}\right]$, no caso de Y, Z e X serem vetores unitários, é dado por (Smith, 1981):

$$\hat{V}\left[\hat{\beta}_{YZ}\right] = \frac{\hat{\sigma}_{Y.ZX}\left\{s_Z^2\frac{s_{ZX}}{s_X^2}\left(\frac{S_X^2}{s_X^2}-1\right)^2+2\frac{s_{ZX}}{s_X^2}\left(\frac{S_X^2}{s_X^2}-1\right)\right\}}{n\left\{s_Z^2+\frac{s_{ZX}}{s_X^2}\left(\frac{S_X^2}{s_X^2}-1\right)\right\}^2} \qquad (2.26)$$

onde $\hat{\sigma}_{Y.ZX} = s_Y^2 - \frac{\left(s_{YX}s_Z^2 + s_{YZ}s_X^2 - 2(s_{YZ}s_{YX}s_{ZX})^{1/2}\right)}{\left(s_Z^2s_X^2 - s_{ZX}\right)}$.

Nathan e Holt (1980) mostram que, se a amostragem balanceada pela variável do desenho X for empregada (veja Royal e Herson,1973) com o balanceamento sendo feito nos primeiros dois momentos, então os estimadores de MQO e de Pearson coincidem. Assim, o estimador de MQO será assintoticamente não viciado nesta classe de desenhos. A amostragem balanceada permite o uso do estimador de MQO, mas o custo é que desenhos mais eficientes empregando o estimador de Pearson são excluídos. Este estimador de β_{YZ} é completamente independente do desenho amostral, embora suas propriedades dependam do desenho. Nathan e Holt(1980) apresentam também um outro estimador cujas propriedades serão úteis especialmente para estes dois casos:

- Embora os valores de X sejam usados para a seleção amostral, pode ser impossível obter valores de X para toda a população. Neste caso, S_X^2 pode não estar disponível para a estimação. Isto pode acontecer, por exemplo, quando a amostragem é baseada numa categorização de valores de X (por exemplo, numa estratificação).

- As hipóteses do modelo (2.17), necessárias para o estimador de Pearson ser assintoticamente não viciado, podem não ser verificadas na prática e, assim, seriam necessários estimadores robustos para os desvios do modelo.

Este novo estimador parte do princípio de que um modo óbvio de utilizar as informações do desenho amostral é baseado nos estimadores de médias, variâncias e covariâncias amostrais ponderadas, onde os pesos são o inverso das probabilidades de inclusão na amostra π_i. Portanto, só foram considerados desenhos amostrais probabilísticos, isto é, desenhos que, com probabilidade 1, têm todas as probabilidades de inclusão positivas. As médias, variâncias e covariâncias amostrais ponderadas são, então, definidas similarmente às estatísticas amostrais ponderadas, como

$$
\begin{aligned}
&\bar{y}^{*} = \sum_{k\in s} y_k / N\pi_k, \quad \bar{z}^{*} = \sum_{k\in s} z_k / N\pi_k, \quad \bar{x}^{*} = \sum_{k\in s} x_k / N\pi_k, \\
&s^{*}_{YZ} = \sum_{k\in s} y_k z_k / N\pi_k - \bar{y}^{*}\bar{z}^{*} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right), \\
&s^{*}_{YX} = \sum_{k\in s} y_k x_k / N\pi_k - \bar{y}^{*}\bar{x}^{*} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right), \\
&s^{*}_{ZX} = \sum_{k\in s} z_k x_k / N\pi_k - \bar{z}^{*}\bar{x}^{*} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right), \\
&s^{*2}_{Y} = \sum_{k\in s} y_k^2 / N\pi_k - \bar{y}^{*2} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right), \\
&s^{*2}_{Z} = \sum_{k\in s} z_k^2 / N\pi_k - \bar{z}^{*2} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right), \\
&s^{*2}_{X} = \sum_{k\in s} x_k^2 / N\pi_k - \bar{x}^{*2} \Big/ \left(\sum_{k\in s} 1\ N\pi_k\right).
\end{aligned}
\tag{2.27}
$$

Deste modo, este novo estimador, chamado de "Pearson-ajustado" será obtido substituindo-se as estatísticas ponderadas acima em suas versões não ponderadas em (2.23).

As variâncias dos estimadores ponderados são assintoticamente iguais aos valores esperados, sob o modelo, das variâncias condicionais sob o desenho, dados Y, Z e X, isto é, .

$$V\left(\hat{\beta}^{*}_{YZ}\right) = E_{\xi}\left[V_p\left(\hat{\beta}^{*}_{YZ} \mid X\right)\right] + O\left(N^{-1}\right) \tag{2.28}$$

Nascimento Silva(1996, pp.147) obteve a variância assintótica sob AAS do estimador de Pearson para o caso onde Y, Z e X são vetores, ou seja:

$$V\left(\hat{\beta}_{YZ}\right) = \frac{1-f}{n} S_Z^{-2}\left[S_{Z^2e^2} + S_r - S_e + 2A_{XZ}A_{Xe}\left(S_{X^2e} - S_{X^2Ze}\right)\right] + O\left(n^{-3/2}\right) \tag{2.29}$$

onde $S_e = N^{-1}\sum_{i=1}^{N} e_i^2$, $S_r = N^{-1}\sum_{i=1}^{N}\left(r_i^2 - \overline{R}\right)^2$, $r_i = e_i - A_{XZ}A_{Xe}\left(x_i - \overline{X}\right)^2$, $\overline{R} = N^{-1}\sum_{i=1}^{N} r_i$,

$A_{XZ} = S_X^{-1}S_{XZ}$, $A_{Xe} = S_X^{-1}S_{Xe}$, $S_{Xe} = N^{-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)e_i$,

$S_{X^2Ze} = N^{-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2\left(z_i - \overline{Z}\right)e_i$ e $S_{X^2e} = N^{-1}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 e_i$.

Pfeffermann e Holmes (1985) seguiram a investigação iniciada em Holt, Smith e Winter (1980) e Nathan e Holt(1980), e compararam o desempenho de métodos baseados no modelo com métodos baseados no desenho, para análise de regressão de dados de pesquisas por amostragem, onde algumas das hipóteses do modelo são violadas. A principal conclusão de Holt, Smith e Winter(1980) e Nathan e Holt(1980) foi que o método de MQO não é apropriado para inferências sobre modelos de regressão e que se deve levar em conta a informação da amostra usada. Dois procedimentos foram, então, investigados por Pfeffermann e Holmes(1985): o primeiro, baseado na inferência pela teoria convencional de amostragem probabilística e o outro baseado em inferência na teoria de máxima verossimilhança, utilizando-se os valores conhecidos das variáveis de desenho em toda a população.

Os resultados relatados na literatura sobre a questão de se escolher uma determinada abordagem para se tratar a informação do desenho num ajuste não são conclusivos, levando em conta que eles são derivados sob modelos onde a máxima verossimilhança parece apropriada. Pfeffermann e Holmes(1985) complementaram o estudo de Nathan e Holt(1980) e Holt, Smith e Winter(1980) examinando o desempenho de vários métodos de inferência em situações onde as hipóteses do modelo de trabalho que motivam o uso da máxima verossimilhança em (2.17) podem não ser adequadas. Os resultados indicam a sensibilidade do procedimento de máxima verossimilhança de Pearson, com respeito à especificação correta das relações entre as variáveis de regressão e as variáveis de desenho. Assim, Pfeffermann e Holmes(1985) sugerem que a distribuição das probabilidades sob o desenho não pode ser ignorada no processo de inferência. As hipóteses do modelo devem ser empregadas na construção de estimadores, mas, em geral, os estimadores baseados nos modelos devem ser modificados para ficarem consistentes com o vetor do "censo" dos coeficientes (vetor dos coeficientes obtido com toda a população finita).

Pfeffermann e Holmes(1985) também utilizaram a abordagem de superpopulação descrita em Smith(1981), ou seja, eles assumiram que a população finita é uma amostra aleatória de uma superpopulação infinita, tal que as observações das variáveis dependentes e independentes seguem o modelo linear clássico. A motivação para a abordagem dos autores é que as hipóteses do modelo, e em particular a hipótese de resíduos independentes inerente ao modelo clássico de

regressão, são seriamente violadas nos desenhos amostrais práticos, que envolvem estratificação e conglomeração.

Considere a população finita $U=\{1,...,N\}$ e seja (Y_i,Z_i,X_i) uma realização de uma variável aleatória trivariada (Y,Z,X) associada a cada indivíduo $i\in U$. A variável X é a variável de desenho, que é conhecida para toda a população. As variáveis Y e Z são, respectivamente, a variável dependente e a variável explicadora, que só são observadas para uma amostra *S* definida pelas variáveis indicadoras $d_1,...,d_N$ onde $d_i=1$, se $i\in S$ e $d_i=0$ caso contrário, $\sum d_i = n$, $i=1,...,N$. Para apenas uma variável de desenho e uma variável explicadora, assumiu-se que o desenho é condicionalmente não-informativo, ou seja, $f(Y_i,Z_i|X_i,d_i)= f(Y_i,Z_i|X_i)$. Todas as informações da amostra estão contidas em Y, tal que dados os valores da variável de desenho, a distribuição conjunta de Y e Z é a mesma se a unidade pertencer ou não à amostra.

O objetivo é estimar os coeficientes $(\beta_{0YZ}, \beta_{YZ})$ da regressão linear de Z em Y, definidos por

$$\begin{aligned} \beta_{YZ} &= \mathrm{cov}(Y,Z)/\,\mathrm{var}(Z) = \sigma_{YZ}/\sigma_{ZZ} \\ \beta_{0YZ} &= \mu_Y - \beta_{YZ}\mu_Z \\ \mu_Y &= \mathrm{E}(Y) \\ \mu_Z &= \mathrm{E}(Z) \end{aligned} \qquad (2.30)$$

Os momentos de (2.30) se referem ao modelo de "superpopulação". Em Holt, Smith Winter(1980) foi apresentado o estimador de máxima verossimilhança de Pearson para este modelo da superpopulação, usando-se a hipótese de distribuição conjunta trivariada normal. Nathan e Holt(1980) mostraram que esta hipótese pode ser relaxada e os estimadores continuarem sendo consistentes se as restrições em (2.17) forem levadas em conta, além de restrições relacionadas às variâncias condicionais das estimativas amostrais das variâncias e covariâncias tenderem a zero quando o tamanho da amostra tender a infinito, ou seja, $V(s_{ab}|X,S)=O(n^{-1})$, onde *a* representa a variável Y ou a variável Z e *b* representa a variável Z ou a variável X. Essa "relaxamento" de hipóteses é importante, já que as variáveis nem sempre têm distribuição normal.

Pfeffermann e Holmes(1985) estudaram a robustez deste estimador de máxima verossimilhança, através da comparação do desempenho com outros estimadores sob dois conjuntos de hipóteses que diferem ligeiramente de (2.17), mas que possuem informação extra. Pfeffermann e Holmes(1985) fizeram uma avaliação do vício dos estimadores (2.23) e (2.24) sob um destes outros conjuntos de hipóteses.

Neste artigo foi apresentado, ainda, um estimador alternativo, baseado no desenho. A principal vantagem deste tipo de estimador é que suas principais propriedades amostrais são geralmente independentes das relações entre as variáveis de regressão e as variáveis de desenho, já que foi mostrado no artigo que variando-se ligeiramente o conjunto de hipóteses em (2.17) pode-se introduzir vício no estimador. O principal interesse deste método é a estimação dos parâmetros da população finita (B_{0YZ},B_{YZ}) (chamados de "census coefficients", isto é, coeficientes do censo) análogos dos parâmetros da superpopulação (β_{0YZ},β_{YZ}). Eles são definidos por:

$$B_{YZ} = \frac{\sum_{i=1}^{N} y_i z_i / N - \overline{YZ}}{\sum_{i=1}^{N} z_i^2 / N - \overline{Z}^2}$$
$$B_{0YZ} = \overline{Y} - B_{YZ}\overline{Z} \qquad (2.31)$$
$$\text{onde} \quad \overline{Y} = \sum_{i=1}^{N} y_i / N \quad \text{e} \quad \overline{Z} = \sum_{i=1}^{N} z_i / N.$$

Sejam $E_p(\bullet)$ e $V_p(\bullet)$ o valor esperado e a variância em relação à distribuição das probabilidades $p(S)$ e sejam $E_0(\bullet)$ e $V_0(\bullet)$ os momentos correspondentes com relação à distribuição conjunta sobre $p(S)$ e sob o modelo (chamada de "pξ-distribution"), que é a distribuição sobre todas as possíveis populações e sobre todas as possíveis amostras de uma dada população. Então, para qualquer estimador b_{0YZ}, tal que $E_p(b_{0YZ}) = B_{0YZ} + O(n^{-1})$ e $V_\xi(B_{0YZ}) = O(N^{-1})$, segue-se que:

$$E_0(b_{0YZ}) = \beta_{0YZ} + O(n^{-1}) \quad \text{e}$$
$$V_0(b_{0YZ}) = E_\xi\{V_p(b_{0YZ})\} + O(N^{-1}) \quad . \qquad (2.32)$$

Assim, b_{0YZ} é aproximadamente não viciado sob o modelo e sob o desenho para o parâmetro da superpopulação β_{0YZ}, com a variância global sendo aproximadamente igual ao valor esperado sob o modelo da variância de b_{0YZ} sob o desenho.

Este processo de inferência difere do de Kish e Frankel(1974) mencionado anteriormente por este último não assumir o modelo de superpopulação. Ou seja, em Kish e Frankel(1974) os coeficientes da população finita eram o objetivo da estimação, e aqui são apenas um passo intermediário para se fazer inferências sobre β_{0YZ} e β_{YZ}. Portanto, em relação ao Quadro 2.2, o processo de inferência de Kish e Frankel(1974) pertenceria ao Caso 1 e o de Pfeffermann e Holmes(1985) pertenceria ao Caso 3.

Pfeffermann e Holmes(1985) propuseram o uso de dois estimadores diferentes, ambos baseados no desenho:

(a) Estimadores "ponderados pelas probabilidades de inclusão na amostra", $\{b_{0W}, b_W\}$, obtidos pela substituição da média populacional nas expressões de B_{0YZ} e B_{YZ} pelo seus estimadores de Horwitz-Thompson, da forma $\sum_{i\in s} y_i/\pi_i$, tal que as variâncias condicionais em D sob o desenho amostral podem ser escritas explicitamente, onde π_i são as probabilidades de inclusão da unidade i na amostra, $i \in U$.

(b) Estimadores "ponderados pelas probabilidades" ajustados propostos para o estudo por Fuller(1975). Estes estimadores utilizam as diferenças conhecidas entre os momentos da amostra e os momentos da população finita das variáveis de desenho, com a finalidade de se conseguir um ganho na eficiência dos estimadores "ponderados pelas probabilidades" citados em (a). No caso onde Z e X forem vetores unitários, estes estimadores ajustados são definidos como:

$$\begin{bmatrix} B_{0YZ} \\ B_{YZ} \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^{N} Z_i \\ \sum_{i=1}^{N} Z_i & \sum_{i=1}^{N} Z_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} Y_i Z_i \end{bmatrix};$$

$A_i' = [1, Z_i, Z_i^2, Y_i, Y_i Z_i]$, tal que $(B_{0YZ}, B_{YZ}) = g(A)$, $A = \sum_{i=1}^{N} A_i$. Seja $X_i' = [X_i - \bar{X}, (X_i - \bar{X})^2 - \hat{\sigma}_X^2]$, tal que $\sum_{i=1}^{N} X_i = 0$. Os estimadores $\{\hat{B}_{0W}(F), \hat{B}_W(F)\}$ são definidos como $g(\hat{A})$, onde $\hat{A}$ é o estimador ponderado de regressão de A definido por:

$$\hat{A} = \sum_{i\in s}(A_i/\pi_i) - \sum_{i\in s}(1/\pi_i)A_i(X_i - \bar{X}_\pi)' \left\{\sum_{i\in s}(1/\pi_i)(X_i - \bar{X}_\pi)(X_i - \bar{X}_\pi)'\right\}^{-1} \times \sum_{i\in s}(1/\pi_i)X_i \qquad (2.33)$$

onde $\bar{X}_\pi = (1/N)\sum_{i\in s}(X_i/\pi_i)$ e π_i são as probabilidades de inclusão do indivíduo i na amostra. Os estimadores $\{b_{0W}(F), b_W(F)\}$ "corrigem" os estimadores $[\{b_{0W}, b_W\} = g\{\sum_{i\in s}(A_i/\pi_i)\}]$ por uma função de regressão das diferenças entre as estatísticas amostrais e populacionais das variáveis de desenho X.

Pfeffermann e Holmes(1985) fizeram um estudo de simulação para observar a magnitude do vício do estimador em (2.23) quando as hipóteses em (2.17) são violadas e para comparar o estimador de Pearson com os estimadores "ponderados pelas probabilidades" ajustados propostos no artigo. A principal conclusão deste artigo foi que as inferências baseadas no modelo propostas por Nathan e Holt(1980) e Holt, Smith e Winter(1980) são mais eficientes se o modelo estiver corretamente especificado, mas podem levar a conclusões equivocadas se uma das

hipóteses em (2.17) for violada, embora a diferença de eficiência não tenha sido grande em muitos casos. Identificar corretamente o modelo, na prática, é bastante complexo. O procedimento baseado no desenho, ao contrário, pareceu ser mais robusto para situações práticas.

Pfeffermann e Holmes(1985) observaram que a modelagem da relação entre as variáveis de regressão e as de desenho faz surgir uma "grande e possivelmente mais eficiente família de estimadores", que utilizam tanto a modelagem usual quanto as informações sobre o desenho amostral. Um exemplo desta estratégia, citado por Pfeffermann e Holmes(1985) foi o estimador ponderado de máxima verossimilhança considerado por Nathan e Holt (1980). Mas esta estratégia também é a base do estimador de Máxima Pseudo-Verossimilhança, que é o foco deste trabalho e que está descrito detalhadamente na Seção 3.

3 - Método da Máxima Pseudo-Verossimilhança

3.1 - Introdução

No capítulo anterior foi mostrada a discussão feita por Pfeffermann e Holmes(1985), onde foi mostrada uma nova estratégia de estimação baseada no desenho, com modelos de superpopulação, e que levam em conta tanto informações do desenho quanto estimação por máxima verossimilhança. Neste capítulo, veremos um desenvolvimento deste estimador discutido inicialmente por Pfeffermann e Holmes(1985).

A maneira utilizada aqui para incorporar as informações sobre o desenho amostral e/ou informações auxiliares foi a abordagem agregada, através do Método de Máxima Pseudo-Verossimilhança (MPV), que é aplicado a modelos paramétricos regulares quando são usados dados de pesquisas por amostra.

3.2 - Método da Máxima Pseudo-Verossimilhança

Considere o vetor P×1 de dados $z_i = (z_{i1},\ldots,z_{iP})'$, que é o vetor das variáveis explicativas. Seja x_i um vetor q×1 que contém as variáveis de desenho ou variáveis auxiliares para o elemento i e sejam Zi e Xi os vetores aleatórios que geram z_i e x_i. Seja $y_i = (y_{i1},\ldots,y_{iA})'$ um vetor A×1 das variáveis de pesquisa do elemento i, gerado por um vetor aleatório Y_i, para $i \in U$. Vamos assumir que $Y_1,\ldots,Y_N$ são iid com densidade $f(y;\theta)$, tal que:

$$f(y_1,\ldots,y_N;\theta) = \prod_{i\in U} f(y_i;\theta), \tag{3.1}$$

onde $\theta = (\theta_1,\ldots,\theta_K)'$ é o vetor K×1 dos parâmetros desconhecidos. Se considerarmos todos os elementos da população finita U, as funções de verossimilhança e de log-verossimilhança do "Censo" serão dadas por:

$$l_U(\theta) = \prod_{i\in U} f(y_i;\theta) \qquad \text{e} \tag{3.2}$$

$$L_U(\theta) = \sum_{i\in U} \log[f(y_i;\theta)]. \tag{3.3}$$

Igualando-se a zero as derivadas parciais de $L_U(\theta)$ em relação a cada componente de θ, temos $\sum_{i\in U} u_i(\theta) = \underset{\sim}{0}$, onde $u_i(\theta) = \partial[\log(f(y_i;\theta))]/\partial\theta$ é o vetor K×1 dos escores do elemento i, $i \in U$. Resolvendo-se este sistema, encontraremos θ_U, que é o estimador de máxima verossimilhança de θ no caso de um "Censo". Mas podemos também interpretar θ_U como uma quantidade desconhecida da população finita para a qual se quer fazer inferências com base em informações de uma amostra. Sob certas condições de regularidade (que são satisfeitas, por

exemplo, pelos Modelos Lineares Generalizados usuais; veja Cox e Hinkley, 1974, pp.281). tem-se que $\theta_U - \theta = o_p(1)$.

Seja $T(\theta) = \sum_{i \in U} u_i(\theta)$ a soma dos escores, que é um vetor de totais populacionais. Para estimar este vetor de totais, pode-se usar um estimador linear ponderado $\hat{T}(\theta)$ da forma $\sum_{i \in s} w_i u_i(\theta)$, onde w_i são pesos definidos apropriadamente. A abordagem proposta por Binder(1983) é que o estimador de Máxima Pseudo-Verossimilhança $\hat{\theta}_{MPV}$ será a solução das equações de Pseudo-Verossimilhança dadas por $\hat{T}(\hat{\theta}_{MPV}) = \sum_{i \in s} w_i u_i(\hat{\theta}_{MPV}) = \underset{\sim}{0}$.

Através da linearização de Taylor podemos chegar à variância (do desenho) do estimador $\hat{\theta}_{MPV}$ e a seu estimador correspondente:

$$V_p(\hat{\theta}_{MPV}) = [J(\theta_U)]^{-1} V_p\left[\sum_{i \in s} w_i u_i(\theta_U)\right][J(\theta_U)]^{-1} \quad \text{e} \tag{3.4}$$

$$\hat{V}(\hat{\theta}_{MPV}) = [\hat{J}(\hat{\theta}_{MPV})]^{-1} \hat{V}\left[\sum_{i \in s} w_i u_i(\hat{\theta}_{MPV})\right][\hat{J}(\hat{\theta}_{MPV})]^{-1}, \text{ onde} \tag{3.5}$$

$$J(\theta_U) = \left.\frac{\partial T(\theta)}{\partial \theta}\right|_{\theta=\theta_U} = \sum_{i \in U} \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta=\theta_U} \quad \text{e} \tag{3.6}$$

$$\hat{J}(\hat{\theta}_{MPV}) = \left.\frac{\partial \hat{T}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}} = \sum_{i \in s} w_i \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}}. \tag{3.7}$$

$V_p\left[\sum_{i \in s} w_i u_i(\theta_U)\right]$ é a matriz de variância (do desenho) do estimador do total populacional dos escores e $\hat{V}\left[\sum_{i \in s} w_i u_i(\hat{\theta}_{MPV})\right]$ é um estimador consistente para esta variância.

Muitos modelos paramétricos com dados provenientes de pesquisas usando muitos desenhos amostrais diferentes podem ser ajustados resolvendo-se as equações de Pseudo-Verossimilhança, desde que sejam satisfeitas algumas condições de regularidade, que foram discutidas por Binder(1983, App.), e incluem:

c1) tamanho da amostra (n) grande;

c2) estimadores de totais de funções das variáveis de pesquisa baseados no desenho, assintoticamente não viciados (sob o plano amostral) e normalmente distribuídos;

c3) existência de estimador consistente para a variância assintótica (do desenho) de estimadores de totais dos escores (funções das variáveis de pesquisa);

c4) condições padrão de regularidade dos modelos paramétricos que levam às equações de verossimilhança do Censo.

Os estimadores de MPV não serão únicos, já que existem diversas maneiras de se definir os pesos w_i. Serão vistos, nas duas próximas subseções, os estimadores de MPV quando usados dois tipos diferentes de peso.

Em resumo, os passos de um procedimento padrão para ajustarmos um modelo paramétrico regular $f(y;\theta)$ pelo método da Máxima Pseudo-Verossimilhança seriam:

1. Resolver $\sum_{i\in s} w_i u_i(\theta) = \underset{\sim}{0}$ e calcular o estimador pontual $\hat{\theta}_{MPV}$ do parâmetro θ no modelo $f(y;\theta)$;
2. Calcular a matriz de variância estimada

$$\hat{V}(\hat{\theta}_{MPV}) = [\hat{j}(\hat{\theta}_{MPV})]^{-1} \hat{V}\left[\sum_{i\in s} w_i u_i(\hat{\theta}_{MPV})\right][\hat{j}(\hat{\theta}_{MPV})]^{-1}; \quad \text{e}$$

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

Vantagens de usar o procedimento de MPV

1. Proporciona estimativas para a variância dos estimadores dos parâmetros que levam em conta o desenho, que são razoavelmente simples de calcular e são consistentes sob "condições fracas" no desenho amostral e na especificação do modelo. Mesmo quando o estimador pontual de MPV coincidir com o estimador de máxima verossimilhança ordinário, a estimativa da variância obtida pelo procedimento de MPV pode ser preferível aos estimadores ordinários da variância baseados em modelos obtidos ignorando-se o desenho amostral.
2. Robustez, no sentido de que em muitos casos a quantidade θ_U da população finita permanece um "alvo" válido para inferência, mesmo se o modelo especificado por $f(y;\theta)$ não proporcionar uma descrição adequada para a distribuição das variáveis de pesquisa.

Este procedimento requer conhecimento de informações detalhadas sobre os elementos da amostra, assim como sobre sua pertinência a estratos e conglomerados (ou unidades primárias de amostragem) e suas probabilidades de inclusão.

Desvantagens do método de MPV

1. Não se conhece as propriedades dos estimadores para pequenas amostras. Isto pode não ser um problema muito grande em análises que usam os dados de pesquisas que, por exemplo, são feitas pelas agências oficiais de estatística, desde que em tais análises se utilize a amostra

inteira, ou, no caso de domínios estudados separadamente, que se use uma amostra suficientemente grande.

2. O fato de não se poder utilizar diagnóstico e outros procedimentos da inferência clássica, tais como gráficos de resíduos e testes estatísticos de Razão de Verossimilhança.

3.3 - Método da Máxima Pseudo-Verossimilhança com pesos π

Já foi visto que os estimadores de MPV não serão únicos, pois existem diversas maneiras de se definir os pesos w_i. Os pesos que geralmente se usa, e que veremos nesta seção, são o inverso da probabilidade de inclusão do indivíduo i, ou seja, $w_i = \pi_i^{-1}$, para $i \in s$, que formam um vetor de pesos amostrais dado por $w_\pi = \Pi_s^{-1} 1_s$, onde Π_s^{-1} é uma matriz diagonal com os elementos da diagonal iguais ao inverso da probabilidade de inclusão para cada indivíduo na amostra.

Substituindo w_i por π_i^{-1} nas equações de Pseudo-Verossimilhança, temos que o estimador de Máxima Pseudo-Verossimilhança $\hat{\theta}_\pi$ neste caso será a solução de $\sum_{i \in s} \pi_i^{-1} u_i(\theta) = \underset{\sim}{0}$. Com resultados apresentados em Godambe e Thompson(1986) chega-se a que o estimador $\hat{\theta}_\pi$ obtido pela solução destas equações é um estimador ótimo para θ (ou θ_U), segundo a abordagem de equações de estimação (tradução de "estimating equations approach"), defendida pelos autores.

Substituindo w_i por π_i^{-1} e $\hat{\theta}_{MPV}$ por $\hat{\theta}_\pi$ em (3.4) e (3.5), teremos:

$$V_p\left(\hat{\theta}_\pi\right) = \left[J(\theta_U)\right]^{-1} V_p\left[\sum_{i \in s} \pi_i^{-1} u_i(\theta_U)\right] \left[J(\theta_U)\right]^{-1} \qquad \text{e} \tag{3.8}$$

$$\hat{V}\left(\hat{\theta}_\pi\right) = \left[\hat{J}\left(\hat{\theta}_\pi\right)\right]^{-1} \hat{V}_\pi\left[\sum_{i \in s} \pi_i^{-1} u_i\left(\hat{\theta}_\pi\right)\right] \left[\hat{J}\left(\hat{\theta}_\pi\right)\right]^{-1}, \tag{3.9}$$

onde $J(\theta_U)$ é dado por (3.6),

$$V_p\left[\sum_{i \in s} \pi_i^{-1} u_i(\theta_U)\right] = \sum_{i \in U} \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \left[u_i(\theta_U)\right]\left[u_j(\theta_U)\right]' \tag{3.10}$$

$$\hat{J}\left(\hat{\theta}_\pi\right) = \sum_{i \in s} \pi_i^{-1} \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta = \hat{\theta}_\pi} \tag{3.11}$$

$$\hat{V}_\pi\left[\sum_{i \in s} \pi_i^{-1} u_i\left(\hat{\theta}_\pi\right)\right] = \sum_{i \in s} \sum_{j \in s} \left(\pi_i^{-1} \pi_j^{-1} - \pi_{ij}^{-1}\right)\left[u_i\left(\hat{\theta}_\pi\right)\right]\left[u_j\left(\hat{\theta}_\pi\right)\right]' \tag{3.12}$$

Sob as condições de regularidade (c1) a (c4), o estimador de variância em (3.9) é um estimador consistente (sob o desenho) para (3.8).

Assim, o procedimento padrão para ajustar um modelo paramétrico regular $f(y;\theta)$ pelo método da Máxima Pseudo-Verossimilhança utilizando-se $w_i = \pi_i^{-1}$, consistiria em substituir $\hat{\theta}_{MPV}$ por $\hat{\theta}_\pi$ e $\hat{V}(\hat{\theta}_{MPV})$ por $\hat{V}(\hat{\theta}_\pi)$no passo 3) da seção anterior.

O ajuste de modelos por este procedimento é simples e envolve, basicamente, estimação de vetores de totais da população e suas matrizes de variância correspondentes por estimadores tradicionais baseados no desenho. Isto facilitou o desenvolvimento de sistemas que utilizam esta técnica para o ajuste de várias classes de modelos (regressões linear e logística, modelos log-lineares para tabelas de contingência, etc) a dados de pesquisas por amostra, como, por exemplo, o SUDAAN (SUrvey DAta ANalysis), o Stata, o Cenvar, o PC Carp, entre outros (ver Pessoa e Nascimento Silva, 1998, capítulo 10, para maiores detalhes).

Sob probabilidades iguais de seleção, os pesos π_i^{-1} serão constantes e o estimador pontual $\hat{\theta}_\pi$ será idêntico ao estimador de máxima verossimilhança ordinário em uma amostra de observações iid com distribuição $f(y;\theta)$, mas o mesmo não é verdade em se tratando da variância do estimador. A igualdade da variância só se verifica quando for empregada amostragem aleatória simples com reposição.

3.4 - MPV com pesos do estimador de regressão

Utiliza-se freqüentemente os pesos π_i^{-1} como padrão pela sua simplicidade. Mas, quando existe informação sobre variáveis auxiliares para toda a população e tais variáveis são relacionadas com as variáveis de pesquisa, estes estimadores (com os pesos π_i^{-1}) podem não ser os mais adequados. Um outro tipo de estimador bem conhecido para totais é o estimador de regressão. Tal estimador é motivado supondo que existe uma relação de regressão linear entre a variável de pesquisa e as variáveis auxiliares x[2]. Então, o estimador de regressão para o total de uma variável de pesquisa pode ser expresso como uma combinação linear ponderada dos valores amostrais, com o vetor de pesos dado por $w_{R\pi} = \Pi_s^{-1} g_s$,

$$g_s = \underset{\sim}{1} + V_s^{-1} x_s^* \left(x_s^{*\prime} V_s^{-1} \Pi_s^{-1} x_s^* \right)^{-1} \left(X^* - \hat{X}_\pi^* \right), \qquad (3.13)$$

onde $x_s^* = [1_s : x_s]$, $x_s = [x_{i_1}, \ldots, x_{i_n}]'$, x_i é a matriz das informações auxiliares para os elementos da amostra, V_s^{-1} é uma matriz conhecida onde se pode incluir a estrutura de variância dos dados,

---

[2] Estas variáveis x são variáveis relacionadas com as variáveis do modelo e que são conhecidas para toda a população, mas não se tem interesse de que elas entrem diretamente no modelo.

no caso de heteroscedasticidade, $X^* = \sum_{i\in U} x_i^* = x_U^{*\prime} 1_U$, $x_U^* = [1_U : x_U]$, $x_U = [x_1, \ldots, x_N]'$, x_U é a matriz das informações auxiliares para os elementos da população, e $\hat{X}_\pi^* = \sum_{i\in s} x_i^* \pi_i^{-1} = x_s^{*\prime} \Pi_s^{-1} 1_s$. Então, Nascimento Silva(1996, cap.6) desenvolveu os estimadores de MPV utilizando os pesos de regressão $w_i^R = \pi_i^{-1} g_{si}$, onde g_{si} é o i-ésimo elemento de g_{si}, ao invés de π_i^{-1}.

O uso de estimadores para os coeficientes de regressão e para suas variâncias que levem em conta os pesos de regressão ao invés dos pesos π é interessante porque, em algumas pesquisas, os únicos pesos divulgados podem ser os pesos w_i^R e o analista pode não ter acesso a informações que o ajudem a reconstruir os pesos π_i. Os pesos de regressão são usados em amostragem como uma maneira de compensar não-resposta, quando se tem informações de uma variável para toda a população.

O estimador $\hat{\theta}_R$ é definido pela solução do sistema de equações: $\hat{T}(\hat{\theta}_R) = \sum_{i\in s} \pi_i^{-1} g_{si} u_i(\hat{\theta}_R) = \sum_{i\in s} w_i^R u_i(\hat{\theta}_R) = \underset{\sim}{0}$. Se $u_i(\theta)$ for considerado um vetor de variáveis de pesquisa, o estimador de regressão do total $\hat{T}_R(\hat{\theta}_R)$ usado para calcular o estimador de MPV de regressão $\hat{\theta}_R$ pode ser interpretado como motivado pelo modelo de trabalho:

$$\begin{aligned} E_M(u_i(\theta) \mid X_i = x_i) &= \gamma' x_i^* \\ V_M(u_i(\theta) \mid X_i = x_i) &= \nu_i \Sigma_U \end{aligned}, \qquad (3.14)$$

onde γ é uma matriz $(Q+1)\times K$ de parâmetros desconhecidos, Σ_U é uma matriz de dispersão $K\times K$ desconhecida e os ν_i são escalares conhecidos, para $i\in U$. Isto implica que o estimador MPV de regressão $\hat{\theta}_R$ pode ser mais eficiente que o estimador MPV padrão $\hat{\theta}_\pi$ quando alguma variação nos escores $u_i(\theta)$ puder ser explicada por uma combinação linear das variáveis auxiliares em x_i.

Para termos um procedimento de ajuste do modelo de MPV baseado na estimação de regressão, basta substituir $\hat{\theta}_{MPV}$ e $\hat{V}(\hat{\theta}_{MPV})$ por $\hat{\theta}_R$ e $\hat{V}_g(\hat{\theta}_R)$ respectivamente nos passos 1, 2 e 3 do procedimento padrão de MPV, onde $\hat{\theta}_R$ é a solução para, $\hat{T}_R(\hat{\theta}_R) = \sum_{i\in s} \pi_i^{-1} g_{si} u_i(\hat{\theta}_R) = \sum_{i\in s} w_i^R u_i(\hat{\theta}_R) = \underset{\sim}{0}$, e

$$V_p(\hat{\theta}_R) = [J(\theta_U)]^{-1} V_p\left[\sum_{i\in s} \pi_i^{-1} g_{si} u_i(\theta_U)\right][J(\theta_U)]^{-1}, \qquad (3.15)$$

$$\hat{V}_g(\hat{\theta}_R) = [\hat{J}(\hat{\theta}_R)]^{-1} \hat{V}_g\left[\sum_{i\in s} \pi_i^{-1} g_{si} u_i(\hat{\theta}_R)\right][\hat{J}(\hat{\theta}_R)]^{-1}, \qquad (3.16)$$

com $J(\theta_U)$ dado por (3.6), $\hat{J}(\hat{\theta}_R) = \left.\frac{\partial \hat{T}_R(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_R} = \sum_{i\in s} \pi_i^{-1} g_{si} \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_R}$,

$$V_p\left[\sum_{i\in s} \pi_i^{-1} g_{si} u_i(\theta_U)\right] = \sum_{i\in U}\sum_{j\in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} [e_i(\theta_U)][e_j(\theta_U)]', \quad \text{e} \tag{3.17}$$

$$\hat{V}_g\left[\sum_{i\in s} \pi_i^{-1} g_{si} u_i(\hat{\theta}_R)\right] = \sum_{i\in s}\sum_{j\in s} (\pi_i^{-1}\pi_j^{-1} - \pi_{ij}^{-1})[g_{si}\hat{e}_i(\hat{\theta}_R)][g_{sj}\hat{e}_j(\hat{\theta}_R)]', \quad \text{onde} \tag{3.18}$$

$e_i(\theta_U) = u_i(\theta_U) - G' x_i^*$, $\quad G = \left(x_U^{*\prime} V_U^{-1} x_U^*\right)^{-1} x_U^{*\prime} V_U^{-1} h(\theta_U)$, $\quad \hat{e}_i(\hat{\theta}_R) = u_i(\hat{\theta}_R) - \hat{G}' x_i^*$,

$\hat{G} = \left(x_s^{*\prime} V_s^{-1} \Pi_s^{-1} x_s^*\right)^{-1} x_s^{*\prime} V_s^{-1} h(\hat{\theta}_R)$, com $x_U^{*\prime} V_U^{-1} x_U^*$ e $x_s^{*\prime} V_s^{-1} \Pi_s^{-1} x_s^*$ não singulares,

$h(\theta_U) = [u_1(\theta_U), \cdots, u_N(\theta_U)]'$ e $h(\hat{\theta}_R) = [u_{i_1}(\hat{\theta}_R), \cdots, u_{i_n}(\hat{\theta}_R)]'$, e com V_U^{-1} sendo uma matriz conhecida para toda a população finita onde se pode incluir a estrutura de variância dos dados, no caso de heteroscedasticidade.

O uso dos pesos de regressão w_i^R é justificável ao menos em dois casos:

1. Os pesos w_i^R podem ser os únicos disponíveis, na pesquisa, para o analista. Tais pesos são freqüentemente adotados para melhorar a precisão ou fazer calibração nos totais conhecidos das variáveis auxiliares durante a pesquisa. O analista pode não ter acesso às probabilidades de seleção π_i nem a informações detalhadas sobre as variáveis auxiliares que permita a reconstrução destas probabilidades.
2. No caso em que tanto os pesos de regressão w_i^R como os pesos ordinários π_i^{-1} estiverem disponíveis, pode-se preferir os pesos de regressão devido à utilização dos estimadores de regressão para compensar a não-resposta, por exemplo, com o uso da pós-estratificação. Seria interessante a reformulação do modelo levando em conta o mecanismo de não-resposta, para se evitar o vício na estimação de θ via $\hat{\theta}_\pi$. Mas isto pode não ser possível se o analista não tiver acesso às informações detalhadas para modelar e considerar o mecanismo de não-resposta.

O uso dos pesos de regressão pode ser melhor do que o uso dos pesos ordinários pois levam a um estimador $\hat{\theta}_R$ que pode ser mais eficiente do que o estimador $\hat{\theta}_\pi$. Isto pode acontecer mesmo quando os pesos de regressão foram calculados por razões que não tem a ver com a estimação de θ.

3.5 - MPV para ajuste de modelos de regressão linear simples

Considere o modelo de regressão linear relacionando uma variável resposta simples y com um vetor P×1 de variáveis explicativas z, ou seja, assuma que $(Y_i, Z_i')'$, i ∈ U, são vetores aleatórios normais iid tais que

$$f(y_i \mid z_i;\beta^*,\sigma_e)=(2\pi\sigma_e)^{-1/2}\exp\left[-\left(y_i - z_i^{*\prime}\beta^*\right)^2 \Big/ 2\sigma_e\right] \tag{3.19}$$

onde $\beta^* = (\beta_0,\beta')' = \left(\beta_0,(\beta_1,\cdots,\beta_P)'\right)'$ e $\sigma_e > 0$ são os parâmetros desconhecidos do modelo, e $z_i^* = (1, z_{i1},\ldots,z_{iP})'$ para i ∈ U. Nascimento Silva(1996, cap.6) desenvolveu o MPV aplicado a este modelo.

As funções dos escores para β^* e σ_e correspondentes ao modelo (3.19) são:

$$\partial log\left[f(y_i \mid z_i;\beta^*,\sigma_e)\right]/\partial\beta^* = z_i^*\left(y_i - z_i^{*\prime}\beta^*\right)/\sigma_e \propto z_i^*\left(y_i - z_i^{*\prime}\beta^*\right) = u_i(\beta^*) \tag{3.20}$$

$$\partial log\left[f(y_i \mid z_i;\beta^*,\sigma_e)\right]/\partial\sigma_e = \left[\left(y_i - z_i^{*\prime}\beta^*\right)^2 - \sigma_e\right]\Big/2\sigma_e^2 \propto \left(y_i - z_i^{*\prime}\beta^*\right)^2 - \sigma_e = u_i(\sigma_e) \tag{3.21}$$

Chamaremos de B^* e S_e os estimadores de Máxima Verossimilhança para β^* e σ_e no caso de um Censo, obtidos mediante a solução das equações de verossimilhança do Censo $\sum_{i\in U} u_i(\theta) = 0$ para $z_U' z_U$ não-singular:

$$B^* = (z_U' z_U)^{-1} z_U' y_U \tag{3.22}$$

$$S_e = N^{-1}\left(y_U - z_U B^*\right)'\left(y_U - z_U B^*\right) \tag{3.23}$$

onde $z_U = (z_1^*,\ldots,z_N^*)'$ e $y_U = (y_1^*,\ldots,y_N^*)'$.

Quando temos apenas uma amostra desta população, usam-se pesos para chegarmos aos estimadores de Máxima Pseudo-Verossimilhança de β^* e σ_e (ou de B^* e S_e). Se os pesos usados satisfizerem às condições de regularidade c1) a c4), podemos resolver as equações de pseudo-verossimilhança $\sum_{i\in s} w_i u_i\left(\hat{\theta}_{MPV}\right) = 0$ correspondentes ao modelo (3.19) de regressão linear. Os estimadores pontuais dos parâmetros são (assumindo daqui para frente que $z_s' W_s z_s$ é não-singular):

$$b_w^* = (z_s' W_s z_s)^{-1} z_s' W_s y_s \tag{3.24}$$

$$\begin{aligned} s_e^w &= (1_s' W_s 1_s)^{-1} (y_s - z_s b_w^*)' W_s (y_s - z_s b_w^*) = \\ &= (1_s' W_s 1_s)^{-1} y_s' \left[W_s - W_s z_s (z_s' W_s z_s)^{-1} z_s' W_s \right] y_s \end{aligned} \tag{3.25}$$

onde z_s e y_s são análogos, para a amostra, a z_U e y_U, respectivamente, $W_s = \text{diag}\left[(w_{i_1}, \ldots, w_{i_n})\right]$ é uma matriz n×n de pesos para os elementos da amostra na diagonal principal, e b_w^* e s_e^w são os estimadores de MPV de β^* e σ_e obtidos com os pesos W, respectivamente.

Podemos observar que, quando o desenho da amostra dá a todos os elementos da amostra pesos iguais, ou seja, $w_i = \overline{w}$ e $W_s = \overline{w} I_n$, os estimadores não dependerão de $\overline{w}$ e os estimadores serão iguais aos estimadores de mínimos quadrados ordinários (ou de máxima verossimilhança sob normalidade).

Para os estimadores de β^*, que serão os de interesse daqui para frente, usando (3.20), e considerando que

$$e_i = y_i - z_i^{*\prime} \beta^*, \text{ para } i \in U, \tag{3.26}$$

temos que (3.6) e (3.7) ficam:

$$J(B^*) = \sum_{i \in U} \partial z_i^* e_i / \partial \beta^* \big|_{\beta^* = B^*} = -z_U' z_U \tag{3.27}$$

$$J(b_w^*) = \sum_{i \in s} w_i \, \partial z_i^* e_i / \partial \beta^* \big|_{\beta^* = b_w^*} = -z_s' W_s z_s \tag{3.28}$$

Assim, substituindo-se o (3.27) e (3.28) em (3.4) e (3.5), obtemos as expressões de variância assintótica do estimador e do estimador desta variância:

$$V_p(b_w^*) = [z_U' z_U]^{-1} V_p\left[\sum_{i \in s} w_i z_i^* e_i \right] [z_U' z_U]^{-1} \quad \text{e} \tag{3.29}$$

$$\hat{V}(b_w^*) = [z_s' W_s z_s]^{-1} \hat{V}\left[\sum_{i \in s} w_i z_i^* \hat{e}_i \right] [z_s' W_s z_s]^{-1}, \text{ onde} \tag{3.30}$$

$\hat{e}_i = y_i - z_i^{*\prime} b_w^*$ para $i \in s$.

3.5.1 - Método da Máxima Pseudo-Verossimilhança com pesos π

No caso de considerarmos os pesos $w_i = \pi_i^{-1}$ com o modelo de regressão linear simples, os estimadores pontuais serão:

$$b_\pi^* = \left(z_s' \Pi_s^{-1} z_s\right)^{-1} z_s' \Pi_s^{-1} y_s, \quad \text{e} \tag{3.31}$$

$$s_e^\pi = \left(1_s' \Pi_s^{-1} 1_s\right)^{-1} \left(y_s - z_s b_\pi^*\right)' \Pi_s^{-1} \left(y_s - z_s b_\pi^*\right). \tag{3.32}$$

As expressões da variância assintótica do *estimador* de b_π^* e do estimador consistente desta variância serão:

$$V\left(b_\pi^*\right) = \left[z_U' z_U\right]^{-1} V_p\left[\sum_{i\in s} \pi_i^{-1} z_i^* e_i\right] \left[z_U' z_U\right]^{-1} \tag{3.33}$$

$$\hat{V}\left(b_\pi^*\right) = \left[z_s' \Pi_s^{-1} z_s\right]^{-1} \hat{V}\left[\sum_{i\in s} \pi_i^{-1} z_i^* \hat{e}_i^\pi\right] \left[z_s' \Pi_s^{-1} z_s\right]^{-1}, \text{ onde} \tag{3.34}$$

$\hat{e}_i^\pi = y_i - z_i^{*\prime} b_\pi^*$ para $i \in s$, e

$$V_p\left[\sum_{i\in s} \pi_i^{-1} z_i^* e_i\right] = \sum_{i\in U}\sum_{j\in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} e_i z_i^* z_j^{*\prime} e_j, \quad \text{e} \tag{3.35}$$

$$\hat{V}_\pi\left[\sum_{i\in s} \pi_i^{-1} z_i^* \hat{e}_i^\pi\right] = \sum_{i\in s}\sum_{j\in s} \left(\pi_i^{-1}\pi_j^{-1} - \pi_{ij}^{-1}\right) \hat{e}_i^\pi z_i^* z_j^{*\prime} \hat{e}_j^\pi. \tag{3.36}$$

3.5.1.1 - Método da Máxima Pseudo-Verossimilhança com pesos π sob AAS

Sob AAS, a probabilidade π_i de inclusão do elemento i é igual a $f = n/N$ para qualquer $i \in U$. As probabilidades conjuntas π_{ij} de inclusão são dadas por:

$$\pi_{ij} = \begin{cases} n/N & \text{se } i = j \\ n(n-1)/(N(N-1)) & \text{se } i \neq j \end{cases} \quad \text{para } i, j \in U. \tag{3.37}$$

Além disto,

$\Pi_s^{-1} = \frac{N}{n} I_n$, onde I_n é a matriz identidade de tamanho n.

Como já foi visto anteriormente, neste caso os estimadores pontuais dos parâmetros têm a mesma forma dos estimadores de mínimos quadrados ordinários com observações iid:

$$b_\pi^* = \left(z_s' \Pi_s^{-1} z_s\right)^{-1} z_s' \Pi_s^{-1} y_s = \left(z_s' z_s\right)^{-1} z_s' y = \left[\sum_{i\in s}\left(z_i^* z_i^{*\prime}\right)\right]^{-1} \sum_{i\in s} z_i^* y_i = b^* \tag{3.38}$$

Definindo $u_i = z_i^* e_i$, $u_i^\pi = z_i^* e_i^\pi$, $\bar{u}_s^\pi = n^{-1}\sum_{i\in s} u_i^\pi$, $S_u = N^{-1}\sum_{i\in U} u_i u_i'$ e $\hat{S}_u = (n-1)^{-1}\sum_{i\in s}\left(u_i^\pi - \bar{u}_s^\pi\right)\left(u_i^\pi - \bar{u}_s^\pi\right)'$, as expressões da variância assintótica e do seu estimador são obtidas substituindo-se

$$V_p\left[\sum_{i\in s} \pi_i^{-1} z_i^* e_i\right] = N^2 \frac{1-f}{n} S_u \quad \text{e} \tag{3.39}$$

$$\hat{V}_{\pi}\left[\sum_{i\in s}\pi_i^{-1}z_i^{*}\hat{e}_i^{\pi}\right]=N^2\frac{1-f}{n}\hat{S}_u \tag{3.40}$$

em (3.33) e (3.34) acima, que ficam iguais a

$$V_p\left(b_{\pi}^{*}\right)=\left[z_U'z_U\right]^{-1}N^2\frac{1-f}{n}S_u\left[z_U'z_U\right]^{-1}=\left[\sum_{i=1}^{N}z_i^{*}z_i^{*'}\right]^{-1}N^2\frac{1-f}{n}S_u\left[\sum_{i=1}^{N}z_i^{*}z_i^{*'}\right]^{-1} \quad \text{e} \tag{3.41}$$

$$\hat{V}\left(b_{\pi}^{*}\right)=\left[\sum_{i\in s}z_i^{*}z_i^{*'}\right]^{-1}N^2\frac{1-f}{n}\hat{S}_u\left[\sum_{i\in s}z_i^{*}z_i^{*'}\right]^{-1}. \tag{3.42}$$

3.5.1.2 - Método da Máxima Pseudo-Verossimilhança com pesos π sob Amostragem Estratificada

Sob Amostragem Estratificada Simples (AES) sem reposição, as probabilidades π_i de inclusão de cada elemento na amostra são iguais a $f_h = n_h/N_h$ para i pertencente ao h-ésimo estrato. As probabilidades conjuntas π_{ij} ficam:

$$\pi_{ij}=\begin{cases}n_h/N_h & \text{se } i=j \text{ e } i,j\in h\\ n_hn_k/(N_hN_k) & \text{se } i\neq j \text{ e } h\neq k\\ n_h(n_k-1)/(N_h(N_k-1)) & \text{se } i\neq j,\ h=k,\ i\in \text{estrato h e } j\in \text{estrato k}\end{cases} \tag{3.43}$$

para $i,j\in U$.

As estimativas pontuais, neste caso, ficam:

$$b_{\pi}^{*}=\left(z_s'\Pi_s^{-1}z_s\right)^{-1}z_s'\Pi_s^{-1}y_s=\left(\sum_{h=1}^{H}\sum_{i\in s}\frac{N_h}{n_h}z_{hi}^{*}z_{hi}^{*'}\right)^{-1}\left(\sum_{h=1}^{H}\sum_{i\in s}\frac{N_h}{n_h}z_{hi}^{*}y_{hi}\right). \tag{3.44}$$

Definindo $u_i=z_i^{*}e_i$, $u_i^{\pi}=z_i^{*}e_i^{\pi}$, $\overline{U}_h=N_h^{-1}\sum_{\substack{i=1\\ i\in h}}^{N_h}u_i$, $S_{uh}=N_h^{-1}\sum_{i\in U}^{i\in h}\left(u_i-\overline{U}_h\right)\left(u_i-\overline{U}_h\right)'$

$\overline{u}_{sh}^{\pi}=n_h^{-1}\sum_{\substack{i=1\\ i\in h}}^{n_h}u_i^{\pi}$ e $\hat{S}_{uh}=\left(n-1\right)^{-1}\sum_{i\in s}^{i\in h}\left(u_{i_i}^{\pi}-\overline{u}_{sh}^{\pi}\right)\left(u_{i_i}^{\pi}-\overline{u}_{sh}^{\pi}\right)'$, as expressões da variância assintótica

e do seu estimador são obtidas substituindo-se

$$V_p\left[\sum_{i\in s}\pi_i^{-1}z_i^{*}e_i\right]=\sum_{h=1}^{H}N_h^2\frac{1-f_h}{n_n}S_{uh} \quad \text{e} \tag{3.45}$$

$$\hat{V}_{\pi}\left[\sum_{i\in s}\pi_i^{-1}z_i^{*}\hat{e}_i^{\pi}\right]=\sum_{h=1}^{H}N_h{}^2\frac{1-f_h}{n}\hat{S}_{uh} \tag{3.46}$$

em (3.33) e (3.34) acima, que ficam iguais a

$$V_p\left(b_\pi^*\right)=\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h} z_{hi}^* z_{hi}^{*\prime}\right)^{-1}\left(\sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{uh}\right)\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h} z_{hi}^* z_{hi}^{*\prime}\right)^{-1} \tag{3.47}$$

$$\hat{V}_p\left(b_\pi^*\right)=\left(\sum_{h=1}^{H}\sum_{i\in s} \frac{N_h}{n_h} z_{hi}^* z_{hi}^{*\prime}\right)^{-1}\left(\sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} \hat{S}_{uh}\right)\left(\sum_{h=1}^{H}\sum_{i\in s} \frac{N_h}{n_h} z_{hi}^* z_{hi}^{*\prime}\right)^{-1} \tag{3.48}$$

3.5.2 - MPV com pesos do estimador de regressão

No caso de usarmos os pesos de regressão, os estimadores de máxima pseudo-verossimilhança ficam (Nascimento Silva, 1996, cap.6):

$$b_R^* = \left(z_s' W_s^R z_s\right)^{-1} z_s' W_s^R y_s \tag{3.49}$$

$$s_e^R = \left(1_s' W_s^R 1_s\right)^{-1}\left(y_s - z_s b_R^*\right)' W_s^R \left(y_s - z_s b_R^*\right) \tag{3.50}$$

onde $W_s^R = \mathrm{diag}\left(w_{R\pi}\right)$ e $w_{R\pi} = \Pi_s^{-1} g_s$, e a expressão da variância assintótica do estimador de b_π^* e do estimador consistente desta variância serão:

$$V_p\left(b_R^*\right)=\left[z_U' z_U\right]^{-1} V_p\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* e_i\right]\left[z_U' z_U\right]^{-1} \qquad \text{e} \tag{3.51}$$

$$\hat{V}_g\left(b_R^*\right)=\left[z_s' W_s^R z_s\right]^{-1} \hat{V}_g\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* \hat{e}_i^R\right]\left[z_s' W_s^R z_s\right]^{-1}, \text{ onde} \tag{3.52}$$

$\hat{e}_i^R = y_i - z_i^{*\prime} b_R^*$ para $i \in s$ e

$$V_p\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* e_i\right] = \sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left(z_i^* e_i - G' x_i^*\right)\left(z_j^* e_j - G' x_j^*\right)' \qquad \text{e} \tag{3.53}$$

$$\hat{V}_g\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* \hat{e}_i^R\right] = \sum_{i\in s}\sum_{j\in s}\left(\pi_i^{-1}\pi_j^{-1} - \pi_{ij}^{-1}\right) g_{si}\left(z_i^* \hat{e}_i^R - \hat{G}' x_i^*\right)\left(z_j^* \hat{e}_j^R - \hat{G}' x_j^*\right) g_{sj}. \tag{3.54}$$

3.5.2.1 - MPV com pesos do estimador de regressão sob AAS

Por simplicidade, vamos assumir que a regressão dos escores nas variáveis auxiliares é homoscedástica ($\nu_i = 1, \forall i \in U$) para obter os resultados desta seção.

Como já foi visto anteriormente, neste caso os estimadores pontuais dos parâmetros ficam iguais a:

$$b_R^* = \left(z_s' W_s^R z_s\right)^{-1} z_s' W_s^R y_s = \left(\sum_{i\in s} g_{si} z_i^* z_i^{*\prime}\right)^{-1}\left(\sum_{i\in s} g_{si} z_i^* y_i\right) \tag{3.56}$$

onde $W_s^R = \mathrm{diag}\left(w_{R\pi}\right)$ e $w_{R\pi} = \Pi_s^{-1} g_s$.

Definindo $r_i = z_i^* e_i - G'x_i^*$, $r_i^R = z_i^* \hat{e}_i^R - \hat{G}'x_i^*$, $\bar{r}_s^R = n^{-1}\sum_{i\in s} r_i^R$, $\bar{R} = N^{-1}\sum_{i\in U} r_i^R$, $S_r = N^{-1}\sum_{i\in U}(r_i - \bar{R})(r_i - \bar{R})'$ e $\hat{S}_r = (n-1)^{-1}\sum_{i\in s}(r_i^R - \bar{r}_s^R)(r_i^R - \bar{r}_s^R)'$, as expressões da variância assintótica e do seu estimador são obtidas substituindo-se

$$V_p\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* e_i\right] = N^2\frac{1-f}{n}S_r \quad \text{e} \tag{3.57}$$

$$\hat{V}_g\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* \hat{e}_i^R\right] = N^2\frac{1-f}{n}\hat{S}_r. \tag{3.58}$$

em (3.51) e (3.52) acima.

3.5.2.2 - MPV com pesos do estimador de regressão sob Amostragem Estratificada

Vamos assumir que a regressão dos escores nas variáveis auxiliares é homoscedástica ($\nu_i = 1, \forall i \in U$).

Neste caso os estimadores pontuais dos parâmetros são:

$$b_R^* = (z_s' W_s^R z_s)^{-1} z_s' W_s^R y_s = \left(\sum_{h=1}^{H}\sum_{i\in h\, i\in s}\frac{N_h}{n_h} g_{shi} z_{hi}^* z_{hi}^{*\prime}\right)^{-1}\left(\sum_{h=1}^{H}\sum_{i\in h\, i\in s}\frac{N_h}{n_h} g_{shi} z_{hi}^* y_{hi}\right) \tag{3.59}$$

onde $W_s^R = \text{diag}(w_{R\pi})$ e $w_{R\pi} = \Pi_s^{-1} g_s$, e o índice h em g_{si}, z_i^* e y_{hi} indica que este elemento de z_s, g_s e y_s pertencem ao estrato h.

Definindo: $r_i = z_i^* e_i - G'x_i^*$, $r_i^R = z_i^* \hat{e}_i^R - \hat{G}'x_i^*$, $\bar{r}_h^R = n_h^{-1}\sum_{i\in s\, i\in h} r_i^R$, $\bar{R}_h = N_h^{-1}\sum_{i\in U\, i\in h} r_i^R$, $S_{rh} = N_h^{-1}\sum_{i\in U\, i\in h}(r_i - \bar{R}_h)(r_i - \bar{R}_h)'$ e $\hat{S}_{rh} = (n_h - 1)^{-1}\sum_{i\in s\, i\in h}(r_i^R - \bar{r}_{sh}^R)(r_i^R - \bar{r}_{sh}^R)'$, as expressões da variância assintótica e do seu estimador são obtidas substituindo-se

$$V_p\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* e_i\right] = \sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}S_{rh} \quad \text{e} \tag{3.60}$$

$$\hat{V}_g\left[\sum_{i\in s}\pi_i^{-1} g_{si} z_i^* \hat{e}_i^R\right] = \sum_{h=1}^{H} N_h^2\frac{1-f_h}{n_h}\hat{S}_{rh}. \tag{3.61}$$

em (3.51) e (3.52) acima.

4 - Estudo de simulação dos estimadores pontuais e de variância

4.1 - Introdução

Nos capítulos 2 e 3 foram examinados diversos estimadores existentes na literatura que permitem levar em conta o desenho da amostra na hora de se ajustar um modelo de regressão. Em especial, foi considerado o método da Máxima Pseudo-Verossimilhança (Binder, 1983; Skinner, Holt e Smith, 1989, p.84) com o uso de pesos de regressão (Nascimento Silva, 1996, cap.6). O objetivo deste capítulo é comparar o desempenho do estimador de variância dos coeficientes estimados obtido por este método com outros estimadores já existentes na literatura, através de um estudo de simulação.

A população cujos dados foram utilizados aqui é a mesma do exemplo apresentado em Nascimento Silva(1996, cap.6), onde foi feita uma simulação para se comparar diversos estimadores pontuais dos coeficientes de modelos de regressão linear. Neste capítulo, iremos utilizar simulação de amostras desta população para comparar os diversos estimadores da variância dos coeficientes do modelo de regressão linear.

4.2 - Descrição da População

A população de simulação aqui utilizada é composta por 452 fazendas de ovelhas da Austrália, com rebanhos entre 200 e 50.000 cabeças. As variáveis de interesse são:

QUADRO 4.1
Variáveis de interesse

LROV	Logaritmo da renda proveniente da criação de ovelhas
LRTOT	Logaritmo da renda total da fazenda
LLA	Logaritmo da produção de lã
NOV	Número de ovelhas da fazenda
LNOV	Logaritmo do número de ovelhas da fazenda
NEQOV	Número equivalente em ovelhas. É uma medida global das operações da fazenda, e é definida por: NEQOV=NOV+8×NGCOR+12×AREAPL, onde NGCOR denota a quantidade de gado de corte da fazenda e AREAPL denota a área destinada à plantação.
LNEQOV	Logaritmo de NEQOV.

O estimador de mínimos quadrados da população (considerado como o "parâmetro verdadeiro") foi calculado para dois modelos lineares da forma:

$E_M(Y_i|Z_i=z_i) = \beta_0 + z_i \beta_1$

$V_M(Y_i|Z_i=z_i) = \sigma_e$

onde as combinações de variáveis dependentes e independentes utilizadas foram:

QUADRO 4.2
Modelos de regressão considerados

Modelo	Variável dependente (Y)	Variável independente (Z)
(1)	LROV	LLA
(2)	LRTOT	LROV

Pelos gráficos 4.1 e 4.2 a seguir, podemos notar que as variáveis usadas no primeiro modelo parecem ter uma relação linear, embora os dados tenham uma grande variabilidade. Já para o segundo modelo, apesar da variabilidade ser menor, os dados apresentam uma estrutura que possivelmente não será bem explicada por um modelo linear simples: a variável LRTOT só pode ser maior ou igual a LROV, pela própria definição dessas duas variáveis. Esse caso poderia ser explicado por um outro tipo de modelo, como regressão monotônica, que não será tratado aqui. Este caso foi considerado aqui como uma regressão linear que teria uma má-especificação parecida com heteroscedasticidade, com as variâncias decrescentes com a variável Z. Isto não é exatamente uma simulação, com uma estrutura de heteroscedasticidade controlada, mas optou-se por este caminho pois é um problema que aparece na prática.

GRÁFICO 4.1
Diagrama de dispersão das variáveis do Modelo 1:
Logaritmo da renda proveniente da criação de ovelhas versus
o logaritmo da produção de lã

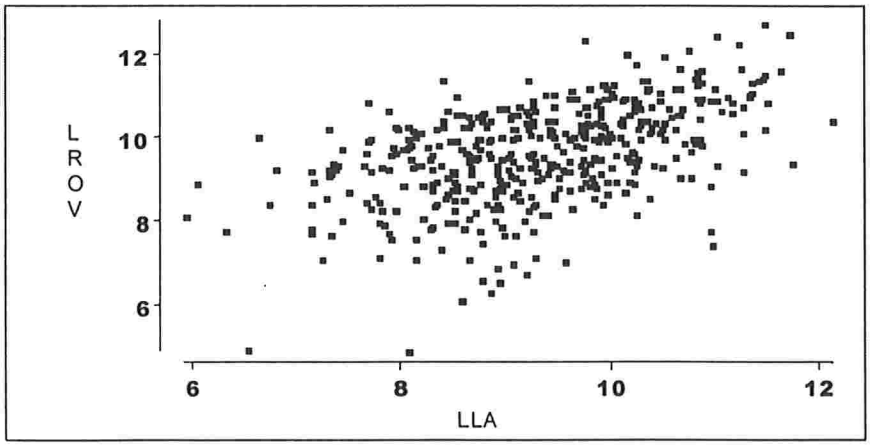
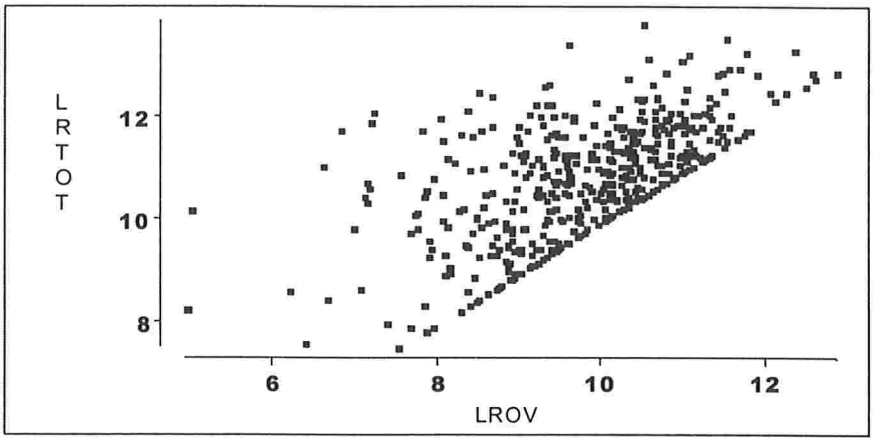


GRÁFICO 4.2
Diagrama de dispersão das variáveis do Modelo 2: Logaritmo da renda total da fazenda versus o logaritmo da renda proveniente da criação de ovelhas



Ajustando-se os dois modelos, por MQO, aos dados da população (com 452 observações), temos:

TABELA 4.1
Ajuste dos modelos de regressão aos dados da população

Modelo	Parâmetro	Estimativa	Desvio Padrão
(1)	β_0	4,4618	0,4091
Y: LROV	β_1	0,5649	0,0436
Z: LLA	σ_e	1,0227	--
	R^2	0,2721	--
(2)	β_0	5,0671	0,3563
Y: LRTOT	β_1	0,5951	0,0363
Z: LROV	σ_e	0,8349	--
	R^2	0,3734	--

Ao se analisar a Tabela 4.1 e os gráficos 4.3 e 4.4 dos valores preditos contra resíduos destes modelos, apresentados a seguir, podemos verificar que o baixo valor do R^2 do primeiro modelo mostra que ele não está sendo suficiente para explicar bem a variabilidade dos dados, embora os resíduos não apresentem estrutura nesse caso. Por outro lado, os resíduos do modelo 2 têm uma certa estrutura de heteroscedasticidade, indicando que o modelo realmente não está bem especificado. Este resultado já era esperado pela própria definição das variáveis deste modelo, pois o menor valor que a variável LRTOT pode assumir é igual a LROV. Logo, é de se

esperar que, com a amostragem, as estimativas dos parâmetros deste modelo também sofram influência desta má especificação.

GRÁFICO 4.3
Valores preditos contra resíduos do Modelo 1

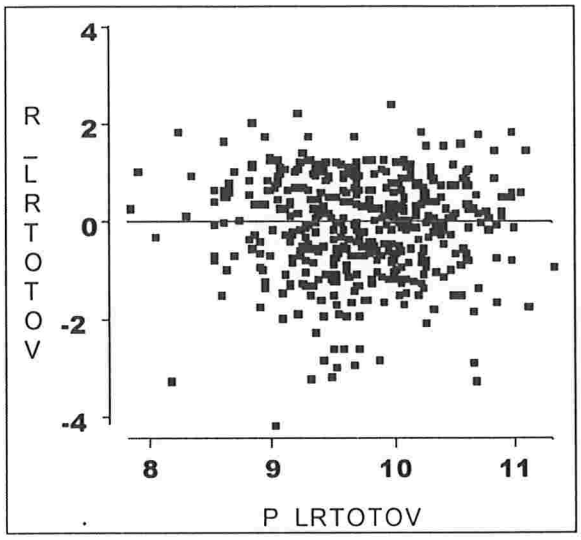
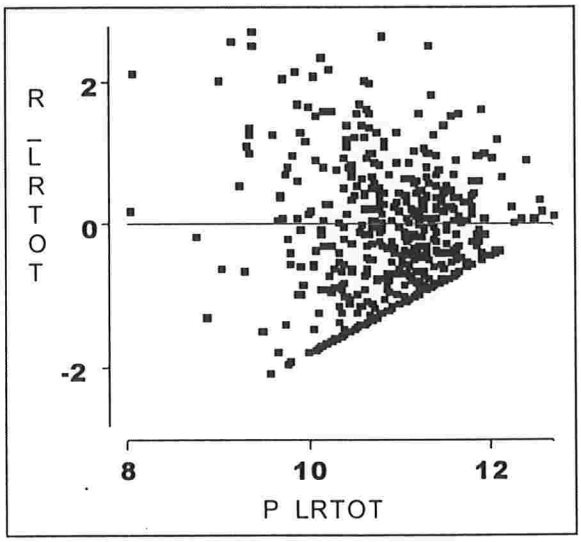


GRÁFICO 4.4
Valores preditos contra resíduos do Modelo 2



4.3 - Descrição da Simulação

A simulação[3] feita consistiu em retirar 1000 amostras de tamanho 50 da população, com dois desenhos amostrais diferentes: Amostragem Aleatória Simples e Amostragem Estratificada Simples. Em cada amostra, foram calculadas estimativas dos parâmetros dos modelos 1 e 2 pelos métodos MQO, MPV-π, MPV-R e Pearson, com 4 opções de variáveis auxiliares consideradas para cada modelo: LNOV, LNEQOV, NOV e NEQOV. Para cada amostra, foram obtidas, portanto, 4 estimativas dos parâmetros de interesse para cada uma das 8 situações a seguir:

QUADRO 4.3
Situações geradas pelos modelos 1 e 2 com as opções de variáveis auxiliares

SITUAÇÕES	Y	Z	X
1	LROV	LLA	LNOV
2	LROV	LLA	LNEQOV
3	LROV	LLA	NOV
4	LROV	LLA	NEQOV
5	LRTOT	LROV	LNOV
6	LRTOT	LROV	LNEQOV
7	LRTOT	LROV	NOV
8	LRTOT	LROV	NEQOV

Para este estudo foram considerados os dois modelos descritos na seção 4.2 para observar o desempenho dos diversos estimadores em duas situações distintas: um caso onde o modelo linear é indicado e um outro onde o modelo tem um problema de má especificação (heteroscedasticidade dos resíduos). Da mesma maneira, foram utilizadas 4 opções de variáveis auxiliares para se observar o quanto os estimadores são influenciados pela escolha da variável auxiliar. Assim, veremos nos gráficos 4.5 e 4.6 a seguir a relação de cada variável auxiliar com as variáveis de cada um dos modelos considerados.

[3] Toda a simulação e o cálculo dos estimadores foi feita através de programas em SAS - Statistical Analysis System.

GRÁFICO 4.5

Diagramas de dispersão das variáveis Y e Z do modelo 1 contra as variáveis auxiliares X

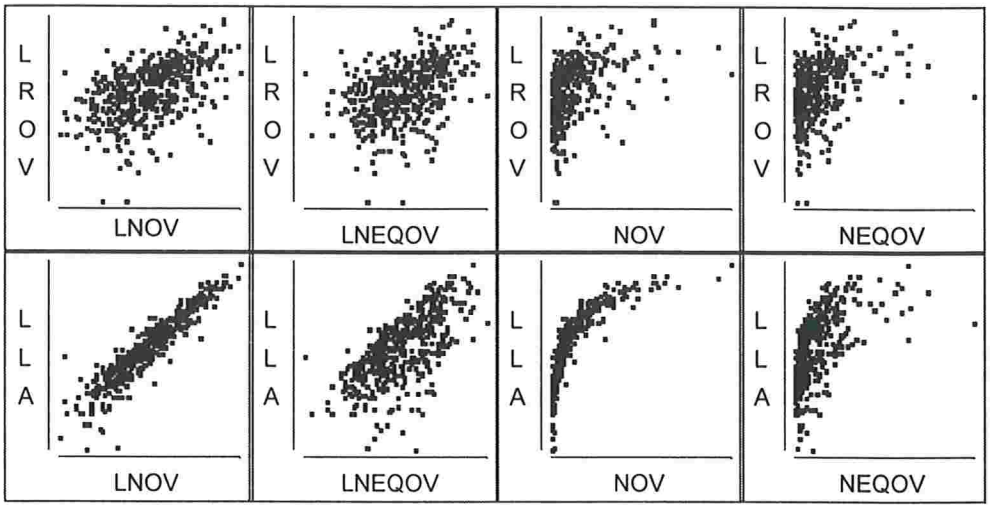
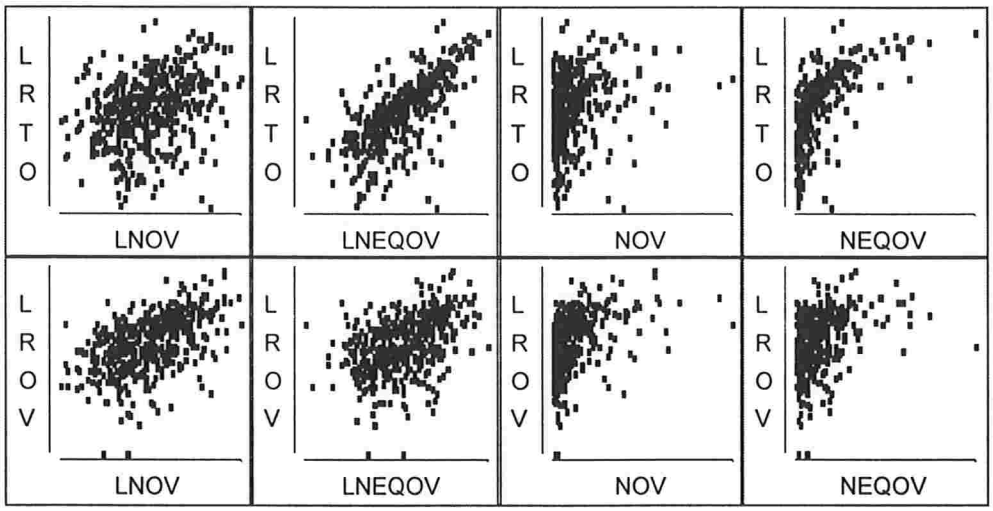


GRÁFICO 4.6

Diagramas de dispersão das variáveis Y e Z do modelo 2 contra as variáveis auxiliares X



Analisando os gráficos 4.5 e 4.6 podemos notar que:

- No modelo 1 (LROV explicado por LLA), as variáveis auxiliares com transformação logarítmica (LNOV e LNEQOV) têm relação linear tanto com a variável explicativa (LLA) como com a variável resposta (LROV) deste modelo, sendo esta relação mais forte com a variável explicativa LLA, o que deve resultar em maior eficiência para estimadores que incorporem essa informação auxiliar;
- No modelo 2 (LRTOT explicado por LROV), as variáveis auxiliares com transformação logarítmica (LNOV e LNEQOV), não apresentam uma relação linear forte com a variável

explicativa deste modelo (LROV), o que indica que seu uso no cálculo dos estimadores considerando informação auxiliar não deve resultar em acréscimo de eficiência;

- Já as variáveis auxiliares sem transformação logarítmica (NOV e NEQOV) não apresentam relação linear com qualquer das variáveis explicativa ou resposta de nenhum dos dois modelos, o que deve implicar em problemas para estimadores considerando uma destas como variável auxiliar.

4.3.1 - Estimação pontual dos coeficientes

Com a finalidade de fazer comparações sobre a eficiência dos quatro métodos considerados para estimação pontual dos coeficientes, foram calculadas as variâncias dos estimadores por duas maneiras diferentes. A primeira consistiu em utilizar as fórmulas assintóticas (vide seções 3.5.1 e 3.5.2) aplicadas aos dados da população completa, exceção feita para o caso do coeficiente B_0 para o estimador de Pearson, para o qual não está disponível na literatura a expressão assintótica da variância. A segunda consistiu em estimar por simulação o Erro Quadrático Médio, usando as estimativas amostrais e o "parâmetro verdadeiro" (parâmetro estimado considerando a população completa), ou seja:

$$\hat{EQM}\left(b_{jk}\right) = \frac{\sum_{i=1}^{1000}\left(b_{jk}^{i} - B_k\right)^2}{1000}, \text{ onde:}$$

$k = 0,1$ é o índice do coeficiente do modelo que está sendo considerado;

B_k é o estimador de mínimos quadrados do β_k usando toda a população (valores na Tabela 4.1);

j indica o tipo de estimador usado; j=MQO para o estimador de Mínimos Quadrados Ordinários, j=π para o estimador de MPV com pesos π, j=R para o estimador de MPV com pesos de regressão e j=P para o estimador de Pearson;

b_{jk}^{i} é o estimador j de B_k na i-ésima amostra da população .

O estimador do EQM pode ser decomposto na soma de duas parcelas: uma relativa à variância e outra relativa ao quadrado do vício, sendo:

$$Var\left(b_{jk}\right) = \left(\frac{\sum_{i=1}^{1000}\left(b_{jk}^{i} - \bar{b}_{jk}\right)^2}{1000} \right)$$

$$Vicio\left(b_{jk}\right) = \bar{b}_{jk} - B_k ;$$

onde $\bar{b}_{jk} = \frac{\sum_{i=1}^{1000} b_{jk}^{i}}{1000}$.

O EQM de cada tipo de estimador foi estimado por simulação para se ter uma idéia de como cada método funcionava para cada um dos dois modelos, um deles com um visível problema de especificação. O EQM e a variância assintótica são medidas de dispersão dos estimadores dos parâmetros, e o esperado é que fiquem próximas uma da outra, para o caso de estimadores assintoticamente não viciados e amostras grandes. Como nos métodos de Pearson e da Máxima Pseudo-Verossimilhança com pesos de regressão são feitas hipóteses mais "fortes" que nos outros dois métodos de estimação (MQO e MPV com pesos π), mesmo apresentando um vício estimado pequeno, na prática o EQM pode não estar tão próximo da variância assintótica.

4.3.2 - Estimação da variância dos estimadores dos coeficientes

Como já foi dito, o objetivo primordial deste capítulo é comparar, através de simulação, o desempenho dos estimadores da variância dos estimadores alternativos para coeficientes do modelo linear. Assim, foram calculadas as estimativas das variâncias de b_0 e b_1 por cada um dos quatro métodos usados na subseção 4.3.1 e descritos nas seções 2.2 (Pearson) e 3.5. Depois, tirou-se a média, em cada método, das estimativas de variância, isto é, foram encontrados os valores médios das variâncias estimadas para os estimadores de β_0 e β_1 pelos quatro métodos estudados, definidos por:

$\bar{\hat{V}}(b_{jk}) = \frac{\sum_{i=1}^{1000} \left(\hat{V}(b_{jk}^{i}) \right)}{1000}$, onde $\hat{V}(b_{jk}^{i})$ é o estimador da variância do estimador b_{jk}^{i} do coeficiente β_k, estimada pelo método j, na amostra i.

Foi usada, então, a variância assintótica apresentada na subseção anterior como valor de comparação, para termos idéia de quanto os estimadores da variância dos diversos métodos estudados aqui se aproximam da "verdadeira variância".

Para que se tenha idéia do comportamento dos estimadores da variância dos coeficientes estimados por cada método, foi feito um teste t para se saber se estes estimadores, em média, são iguais às variâncias calculadas pelas fórmulas assintóticas correspondentes apresentadas no capítulo anterior. Ou seja, iremos testar:

H_0: média de $\hat{V}(b_{jk}^{i}) = V(b_{jk})$

×

H_1: média de $\hat{V}(b_{jk}^{i}) \neq V(b_{jk})$,

onde o valor de $V(b_{jk})$, a variância assintótica calculada com os dados da população completa, é uma constante conhecida. Para fazer este teste foi calculada também uma estimativa de simulação da variância dos estimadores de variância dos coeficientes estimados, ou seja,

$$\hat{V}\left(\hat{V}\left(b_{jk}\right)\right) = \frac{\sum_{i=1}^{1000}\left(\hat{V}\left(b_{jk}^{i}\right) - \overline{\hat{V}}\left(b_{jk}\right)\right)^{2}}{1000}.$$

4.4 – Resultados das Simulações para Amostragem Aleatória Simples

Nesta seção serão apresentados os resultados para as 8 situações de simulação descritas na seção anterior, para os 4 métodos de estimação, usando-se 1000 amostras de tamanho 50 selecionadas através de Amostragem Aleatória Simples. Este é um desenho bem simples e ignorável, que não fere as hipóteses de observações iid dos modelos usuais.

4.4.1 - Estimação pontual dos coeficientes

Na Tabela 4.2 a seguir estão os resultados obtidos para a estimação de B_0, para as 8 situações de simulação acima descritas, para os 4 métodos de estimação. Estes resultados incluem, para cada estimador, a variância calculada através das fórmulas assintóticas, o EQM e o vício estimados por simulação, e a razão entre a variância calculada pela fórmula assintótica e o EQM estimado, conforme descritos na seção 4.3.1. A Tabela 4.3 apresenta as mesmas medidas, mas em relação à estimação de B_1.

TABELA 4.2

Medidas da variação de estimadores de B_0

AMOSTRAGEM ALEATÓRIA SIMPLES

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Variância Assintótica[4]							
			LNOV	1,5121	1,5121	1,5120	--
			LNEQOV	1,5121	1,5121	1,5120	--
1	LROV	LLA	NOV	1,5121	1,5121	1,5100	--
			NEQOV	1,5121	1,5121	1,5067	--
			LNOV	1,6234	1,6234	1,6229	--
			LNEQOV	1,6234	1,6234	1,6231	--
2	LRTOT	LROV	NOV	1,6234	1,6234	1,6120	--
			NEQOV	1,6234	1,6234	1,6226	--
Erro Quadrático Médio[5]							
			LNOV	1,5878	1,5878	1,6454	1,6003
			LNEQOV	1,5878	1,5878	1,6193	1,5969
1	LROV	LLA	NOV	1,5878	1,5878	1,6638	1,7009
			NEQOV	1,5878	1,5878	1,6153	1,6731
			LNOV	1,7195	1,7195	1,7406	1,7435
			LNEQOV	1,7195	1,7195	1,7234	1,6283
2	LRTOT	LROV	NOV	1,7195	1,7195	1,7473	1,7593
			NEQOV	1,7195	1,7195	1,7759	2,1218
Vício(× 1000)							
			LNOV	4,3899	4,3899	4,8936	3,3595
			LNEQOV	4,3899	4,3899	5,4028	2,3546
1	LROV	LLA	NOV	4,3899	4,3899	1,6604	0,7446
			NEQOV	4,3899	4,3899	2,9446	0,1151
			LNOV	5,3318	5,3318	4,2681	6,5870
			LNEQOV	5,3318	5,3318	3,5252	10,4972
2	LRTOT	LROV	NOV	5,3318	5,3318	9,0839	10,9998
			NEQOV	5,3318	5,3318	0,1961	112,2673
Razão EQM/Variância Assintótica							
			LNOV	1,0501	1,0501	1,0882	--
			LNEQOV	1,0501	1,0501	1,0710	--
1	LROV	LLA	NOV	1,0501	1,0501	1,1019	--
			NEQOV	1,0501	1,0501	1,0721	--
			LNOV	1,0592	1,0592	1,0725	--
			LNEQOV	1,0592	1,0592	1,0618	--
2	LRTOT	LROV	NOV	1,0592	1,0592	1,0839	--
			NEQOV	1,0592	1,0592	1,0945	--

[4] Nas Tabelas 4.2 e 4.3 foram usadas as fórmulas (3.41) para o MPV-π e (3.51) e (3.57) para o MPV-R. Na Tabela 4.3 foi usada também a fórmula (2.25) para o estimador de Pearson.

[5] Como parâmetros verdadeiros foram utilizados os valores da Tabela 4.1, e as estimativas dos parâmetros nas Tabela 4.2 e 4.3 foram calculadas usando-se (3.38) para o MPV-π e (3.56) para o MPV-R em cada uma das 1000 amostras. Na Tabela 4.3 as estimativas para Pearson foram calculadas usando-se a fórmula (2.23) em cada uma das 1000 amostras.

TABELA 4.3
Medidas da variação de estimadores de B_1
AMOSTRAGEM ALEATÓRIA SIMPLES

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Variância Assintótica (× 100)							
			LNOV	1,6687	1,6687	1,6686	1,6691
			LNEQOV	1,6687	1,6687	1,6686	1,6667
1	LROV	LLA	NOV	1,6687	1,6687	1,6659	1,6776
			NEQOV	1,6687	1,6687	1,6609	1,6924
			LNOV	1,5508	1,5508	1,5502	1,5510
			LNEQOV	1,5508	1,5508	1,5507	1,4569
2	LRTOT	LROV	NOV	1,5508	1,5508	1,5379	1,5475
			NEQOV	1,5508	1,5508	1,5447	1,8145
Erro Quadrático Médio (× 100)							
			LNOV	1,7472	1,7472	1,8148	1,7618
			LNEQOV	1,7472	1,7472	1,7837	1,7569
1	LROV	LLA	NOV	1,7472	1,7472	1,8414	1,8856
			NEQOV	1,7472	1,7472	1,7797	1,8484
			LNOV	1,6606	1,6606	1,6820	1,6854
			LNEQOV	1,6606	1,6606	1,6814	1,5767
2	LRTOT	LROV	NOV	1,6606	1,6606	1,6866	1,7107
			NEQOV	1,6606	1,6606	1,7012	2,1107
Vício($\times 10^5$)							
			LNOV	5,1202	5,1202	5,6835	3,9073
			LNEQOV	5,1202	5,1202	6,6197	2,9955
1	LROV	LLA	NOV	5,1202	5,1202	1,8159	0,8125
			NEQOV	5,1202	5,1202	3,2254	0,1390
			LNOV	5,0482	5,0482	4,0950	6,4407
			LNEQOV	5,0482	5,0482	2,7678	9,4688
2	LRTOT	LROV	NOV	5,0482	5,0482	9,1124	10,9985
			NEQOV	5,0482	5,0482	0,4042	120,6968
Razão EQM/Variância Assintótica							
			LNOV	1,0470	1,0470	1,0876	1,0556
			LNEQOV	1,0470	1,0470	1,0690	1,0541
1	LROV	LLA	NOV	1,0470	1,0470	1,1053	1,1240
			NEQOV	1,0470	1,0470	1,0715	1,0922
			LNOV	1,0708	1,0708	1,0850	1,0867
			LNEQOV	1,0708	1,0708	1,0843	1,0822
2	LRTOT	LROV	NOV	1,0708	1,0708	1,0967	1,1054
			NEQOV	1,0708	1,0708	1,1013	1,1632

As fórmulas da variância assintótica e das estimativas pontuais dos parâmetros para os métodos de MQO e MPV-π, sob AAS, são iguais. Assim, todos os valores para estes dois métodos são iguais nas duas tabelas.

Podemos observar nas Tabelas 4.2 e 4.3 que a variância assintótica calculada foi bem próxima para todos os estimadores. O uso da variável auxiliar não contribuiu muito para que a variância diminuísse, embora as variâncias assintóticas calculadas pelo método de MPV-R tenham sido ligeiramente menores em quase todos os casos, exceto para os estimadores de B_1,

quando foi usada a variável auxiliar LNEQOV no modelo 2. Neste caso, o estimador com menor variância assintótica foi o de Pearson.

O EQM estimado foi um pouco maior para os estimadores que utilizaram variáveis auxiliares, exceto para o caso do modelo 2 quando a variável auxiliar foi LNEQOV, pois neste caso o estimador com menor EQM foi o de Pearson. O estimador de Pearson se mostrou sensível à má especificação do modelo. Para o modelo 1, foi o método que apresentou menor vício, mas foi o que apresentou maior vício no modelo 2. Ele também foi sensível à escolha de uma variável auxiliar que não tenha relação linear com Z e com Y, pois apesar dos vícios para os estimadores de B_0 como o de B_1 serem pequenos, quando a variável auxiliar foi NEQOV, o vício do estimador de Pearson foi de cerca de 10 vezes o vício dos outros estimadores. Como os vícios estimados para todos os estimadores foram pequenos (na Tabela 4.3 eles estão, inclusive, multiplicados por 10^5), as conclusões baseadas no EQM ou na a variância assintótica são idênticas.

Os resultados apresentados nas Tabelas 4.2 e 4.3 mostram que os estimadores pontuais dos coeficientes B_0 e B_1 sob AAS se comportam como o esperado com base nas fórmulas assintóticas, já que os vícios são desprezíveis e as razões entre o EQM estimado e a variância assintótica são próximas de 1. Os métodos que utilizam alguma variável auxiliar (MPV-R e Pearson) têm essa razão variando mais que os que ignoram a informação auxiliar (MQO e MPV-π). Isso deve ter acontecido porque estes últimos são métodos onde as hipóteses feitas são mais fortes. Além disso, o uso de variáveis auxiliares não contribuiu para uma maior eficiência dos estimadores, pois os vícios foram bem pequenos e os EQM estimados foram maiores para os métodos que utilizaram alguma variável auxiliar (MPV-R e Pearson) do que para os outros que ignoraram informações auxiliares (MQO e MPV-π). O valor da razão EQM/Variância Assintótica no modelo 2, que é mal especificado, é bem mais distante de 1 para os dois primeiros métodos do que para os dois últimos, principalmente com a variável auxiliar NEQOV.

4.4.2 – Estimação da variância dos estimadores dos coeficientes

As Tabelas 4.4 e 4.5 a seguir contém os valores médios das estimativas das variâncias para os estimadores de B_0 e B_1 pelos quatro métodos estudados, além da estimativa de simulação da variância dos estimadores de variância dos coeficientes estimados e os p-valores do teste t, conforme descritos na seção 4.3.2. Além disto, foram também calculadas as razões entre a média dos estimadores de variância dos coeficientes estimados e a variância calculada pelas fórmulas assintóticas, para cada método e cada modelo, com o objetivo de se examinar o

desempenho de cada estimador da variância. Essas razões também se encontram nas Tabelas 4.4 e 4.5.

TABELA 4.4
Resultados da simulação dos estimadores de variância dos estimadores do coeficiente B_0
AMOSTRAGEM ALEATÓRIA SIMPLES

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Média das estimativas de variância[6]							
			LNOV	1,5378	1,4385	1,5940	--
			LNEQOV	1,5378	1,4385	1,5813	--
1	LROV	LLA	NOV	1,5378	1,4385	1,6086	--
			NEQOV	1,5378	1,4385	1,5898	--
			LNOV	1,1898	1,5338	1,5177	--
			LNEQOV	1,1898	1,5338	1,5159	--
2	LRTOT	LROV	NOV	1,1898	1,5338	1,5273	--
			NEQOV	1,1898	1,5338	1,5229	--
Variância das estimativas de variância (×10)							
			LNOV	1,7982	4,0252	1,0996	--
			LNEQOV	1,7982	4,0252	0,9555	--
1	LROV	LLA	NOV	1,7982	4,0252	1,1911	--
			NEQOV	1,7982	4,0252	1,0067	--
			LNOV	1,0059	4,3789	1,3367	--
			LNEQOV	1,0059	4,3789	1,3333	--
2	LRTOT	LROV	NOV	1,0059	4,3789	1,4268	--
			NEQOV	1,0059	4,3789	1,3859	--
P-valor do Teste T (em %)							
			LNOV	5,57	0,03	0,00	--
			LNEQOV	5,57	0,03	0,00	--
1	LROV	LLA	NOV	5,57	0,03	0,00	--
			NEQOV	5,57	0,03	0,00	--
			LNOV	0,00	0,00	0,00	--
			LNEQOV	0,00	0,00	0,00	--
2	LRTOT	LROV	NOV	0,00	0,00	0,00	--
			NEQOV	0,00	0,00	0,00	--
Razão: Médias das estimativas de variância/ Variância assintótica							
			LNOV	1,0170	0,9513	1,0542	--
			LNEQOV	1,0170	0,9513	1,0458	--
1	LROV	LLA	NOV	1,0170	0,9513	1,0653	--
			NEQOV	1,0170	0,9513	1,0552	--
			LNOV	0,7329	0,9448	0,9352	--
			LNEQOV	0,7329	0,9448	0,9339	--
2	LRTOT	LROV	NOV	0,7329	0,9448	0,9474	--
			NEQOV	0,7329	0,9448	0,9386	--

[6] As estimativas de variâncias para cada uma das 1000 amostras foram calculadas pelas fórmulas (3.42) para MPV-π e (3.52) e (3.58) para MPV-R, para as Tabelas 4.4 e 4.5. As estimativas de Pearson das variâncias para cada uma das 1000 amostras na Tabela 4.5 foram calculadas usando-se a fórmula (2.26).

TABELA 4.5
Resultados da simulação dos estimadores de variância dos estimadores do coeficiente B_1
AMOSTRAGEM ALEATÓRIA SIMPLES

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Média das estimativas de variância (×100)							
			LNOV	1,7401	1,5857	1,8042	1,6672
			LNEQOV	1,7401	1,5857	1,7881	1,6719
1	LROV	LLA	NOV	1,7401	1,5857	1,8257	1,7399
			NEQOV	1,7401	1,5857	1,8018	1,7479
			LNOV	1,2376	1,4663	1,5753	1,1907
			LNEQOV	1,2376	1,4663	1,5733	0,6879
2	LRTOT	LROV	NOV	1,2376	1,4663	1,5882	1,2388
			NEQOV	1,2376	1,4663	1,5828	0,8247
Variância das estimativas de variância ($\times 10^5$)							
			LNOV	2,3008	4,7519	1,3582	2,1974
			LNEQOV	2,3008	4,7519	1,1751	2,2107
1	LROV	LLA	NOV	2,3008	4,7519	1,5354	2,4723
			NEQOV	2,3008	4,7519	1,2762	2,8210
			LNOV	1,1094	4,1385	1,3330	1,0378
			LNEQOV	1,1094	4,1385	1,3312	0,5161
2	LRTOT	LROV	NOV	1,1094	4,1385	1,4462	1,2375
			NEQOV	1,1094	4,1385	1,4068	0,5542
P-valor do Teste T (em %)							
			LNOV	0,00	0,01	0,00	90,03
			LNEQOV	0,00	0,01	0,00	72,79
1	LROV	LLA	NOV	0,00	0,01	0,00	0,01
			NEQOV	0,00	0,01	0,00	0,10
			LNOV	0,00	0,00	3,02	0,00
			LNEQOV	0,00	0,00	5,01	0,00
2	LRTOT	LROV	NOV	0,00	0,00	0,00	0,00
			NEQOV	0,00	0,00	0,14	0,00
Razão: Médias das estimativas de variância / Variância assintótica							
			LNOV	1,0428	0,9503	1,0813	0,9989
			LNEQOV	1,0428	0,9503	1,0717	1,0031
1	LROV	LLA	NOV	1,0428	0,9503	1,0959	1,0371
			NEQOV	1,0428	0,9503	1,0848	1,0328
			LNOV	0,7980	0,9456	1,0162	0,7677
			LNEQOV	0,7980	0,9456	1,0146	0,4722
2	LRTOT	LROV	NOV	0,7980	0,9456	1,0327	0,8005
			NEQOV	0,7980	0,9456	1,0247	0,4545

Podemos observar, tanto pelos p-valores da Tabela 4.4 como da 4.5, que não se pode dizer que as variâncias estimadas são, em média, iguais às calculadas pelas fórmulas assintóticas, ao nível de 5% de significância. As exceções a essa conclusão geral são os valores obtidos para o modelo 1 pelo método de MQO, para as estimativas relativas a B_0 (Tabela 4.4) e os obtidos pelo método de Pearson com as variáveis auxiliares LNOV e LNEQO, para as estimativas

relativas a B_1 (Tabela 4.5), e para o modelo 2, os valores obtidos pelo método de MPV-R com a variável auxiliar LNEQO, para as estimativas relativas a B_1 (Tabela 4.5).

Observando a razão entre a média das estimativas das variâncias dos coeficientes estimados e a variância assintótica, podemos observar que, na Tabela 4.4, para o modelo 1, a que mais próxima ficou de 1 foi a referente ao método de MQO. Na Tabela 4.5, para o modelo 1, as duas mais próximas de 1 foram as referentes ao método de Pearson e ao de MQO. Em compensação, para o modelo 2, que estava mal-especificado, as razões mais distantes de 1 foram, justamente, as relativas a estes dois métodos. Isso mostra que, quando o modelo está bem especificado e estamos trabalhando apenas sob AAS (plano amostral ignorável), podemos não ter problemas em utilizar MQO, mas quando, como neste exemplo, usamos este método num modelo mal-especificado, a média das variâncias estimadas ficou mais de 20% abaixo da variância assintótica tanto no caso relativo a B_0 quanto ao relativo a B_1.

O método de Pearson, apesar de só ter sido usado para estimativas relativas a B_1, também funcionou muito bem para o modelo 1, neste caso sob AAS. Em compensação, no modelo 2, as médias das variâncias estimadas chegaram a mais de 50% abaixo da variância assintótica, quando utilizadas as variáveis auxiliares LNEQOV e NEQOV.

As estimativas calculadas pelos métodos de MPV, tanto com pesos π, quanto com pesos de regressão, apresentaram desempenho mais "robusto" quanto à aderência do estimador de variância às variâncias assintóticas, porém invariante à boa ou má especificação do modelo. Ou seja, as razões apresentadas nas Tabelas 4.4 e 4.5 relativas a estes dois modelos não se distanciaram de 1 por mais de 10%. O MPV-R funcionou especialmente bem no modelo 2, com problemas de especificação. É interessante observar também que a variância das estimativas de variância é maior para o MPV-π, que para o MPV-R, independente do modelo utilizado.

4.5 - Resultados da Simulação para Amostragem Estratificada Simples

Em geral, as pesquisas por amostragem possuem um desenho amostral mais complexo que AAS. Assim, repetiu-se o procedimento de simulação apresentado na seção 4.3, utilizando-se agora Amostragem Estratificada, com seleção aleatória simples sem reposição dentro dos estratos, para a seleção das 1000 amostras. Usou-se o mesmo exemplo de Nascimento Silva(1996, cap.6), onde as 452 fazendas foram divididas em 2 estratos:

Estrato 1: Fazendas "pequenas", com menos de 10.000 ovelhas. Havia 407 fazendas pertencentes a este estrato.

Estrato 2: Fazendas "grandes", com mais de 10.000 ovelhas. Havia 45 fazendas pertencentes a este estrato.

Foram selecionadas 25 fazendas em cada estrato, por AAS sem reposição. Utilizou-se este desenho, inclusive, por se parecer com o desenho da aplicação feita no capítulo seguinte, com dados do Censo Demográfico de 1991[7]. Trata-se de um desenho não ignorável, devido ao emprego de frações amostrais desiguais nos estratos.

Serão apresentados a seguir os resultados da simulação da mesma maneira que na seção anterior, agora para este novo desenho.

4.5.1 - Estimação pontual dos coeficientes

Na Tabela 4.6 a seguir estão os resultados obtidos para a estimação de B_0, para as 8 situações de simulação acima descritas, para os 4 métodos de estimação. Estes resultados incluem, para cada estimador, a variância calculada através das fórmulas assintóticas, o EQM e o vício estimados por simulação, e a razão entre a variância calculada pela fórmula assintótica e o EQM estimado, conforme descritos na seção 4.3.1. A Tabela 4.7 apresenta as mesmas medidas, mas em relação à estimação de B_1.

[7] Na verdade, o desenho usado no Censo Demográfico inclui seleção sistemática dentro dos estratos, mas usou-se AAS sem reposição por simplificação.

TABELA 4.6
Medidas da variação de estimadores de B_0
Amostragem Estratificada Simples

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Variância Assintótica[8]							
			LNOV	1,5121	2,2172	2,2012	--
			LNEQOV	1,5121	2,2172	2,2065	--
1	LROV	LLA	NOV	1,5121	2,2172	2,2099	--
			NEQOV	1,5121	2,2172	2,2094	--
			LNOV	1,6234	2,8167	2,7878	--
			LNEQOV	1,6234	2,8167	2,7207	--
2	LRTOT	LROV	NOV	1,6234	2,8167	2,8004	--
			NEQOV	1,6234	2,8167	2,8297	--
Erro Quadrático Médio[9]							
			LNOV	1,3825	2,2955	2,4157	1,4330
			LNEQOV	1,3825	2,2955	2,4032	1,4156
1	LROV	LLA	NOV	1,3825	2,2955	2,3181	1,7573
			NEQOV	1,3825	2,2955	2,2685	1,4905
			LNOV	2,9453	3,0114	3,0811	3,0292
			LNEQOV	2,9453	3,0114	2,9603	2,9593
2	LRTOT	LROV	NOV	2,9453	3,0114	3,0037	3,1807
			NEQOV	2,9453	3,0114	3,0446	1,9941
Vício(× 1000)							
			LNOV	28,7494	0,7558	0,7415	30,2536
			LNEQOV	28,7494	0,7558	0,4684	33,2112
1	LROV	LLA	NOV	28,7494	0,7558	1,4209	12,9243
			NEQOV	28,7494	0,7558	1,8663	21,3102
			LNOV	1464,4180	26,4226	28,7188	1528,7322
			LNEQOV	1464,4180	26,4226	25,7578	1643,6808
2	LRTOT	LROV	NOV	1464,4180	26,4226	28,6560	1609,3206
			NEQOV	1464,4180	26,4226	29,4895	518,5421
Razão EQM/Variância Assintótica							
			LNOV	0,9143	1,0353	1,0974	--
			LNEQOV	0,9143	1,0353	1,0891	--
1	LROV	LLA	NOV	0,9143	1,0353	1,0489	--
			NEQOV	0,9143	1,0353	1,0267	--
			LNOV	1,8142	1,0691	1,1052	--
			LNEQOV	1,8142	1,0691	1,0881	--
2	LRTOT	LROV	NOV	1,8142	1,0691	1,0726	--
			NEQOV	1,8142	1,0691	1,0759	--

[8] Nas Tabelas 4.6 e 4.7 foram usadas as fórmulas (3.47) para o MPV-π e (3.51) e (3.60) para o MPV-R. Na Tabela 4.7 foi usada também a fórmula (2.25) para o estimador de Pearson.

[9] Como parâmetros verdadeiros foram utilizados os valores da Tabela 4.1, e as estimativas dos parâmetros nas Tabela 4.6 e 4.7 foram calculadas usando-se (3.44) para o MPV-π e (3.59) para o MPV-R em cada uma das 1000 amostras. Na Tabela 4.7 as estimativas para Pearson foram calculadas usando-se a fórmula (2.23) em cada uma das 1000 amostras.

TABELA 4.7
Medidas da variação de estimadores de B_1
Amostragem Estratificada Simples

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Variância Assintótica (× 100)							
			LNOV	1,6687	2,2974	2,2787	1,6691
			LNEQOV	1,6687	2,2974	2,2830	1,6667
1	LROV	LLA	NOV	1,6687	2,2974	2,2880	1,6776
			NEQOV	1,6687	2,2974	2,2857	1,6924
			LNOV	1,5508	2,6634	2,6310	1,5510
			LNEQOV	1,5508	2,6634	2,6236	1,4569
2	LRTOT	LROV	NOV	1,5508	2,6634	2,6443	1,5475
			NEQOV	1,5508	2,6634	2,6841	1,8145
Erro Quadrático Médio (× 100)							
			LNOV	1,2600	2,3722	2,5013	1,3322
			LNEQOV	1,2600	2,3722	2,4872	1,2829
1	LROV	LLA	NOV	1,2600	2,3722	2,3918	1,7123
			NEQOV	1,2600	2,3722	2,3362	1,4309
			LNOV	2,6965	2,8216	2,8878	2,8924
			LNEQOV	2,6965	2,8216	2,8388	2,3397
2	LRTOT	LROV	NOV	2,6965	2,8216	2,8093	3,0301
			NEQOV	2,6965	2,8216	2,8658	1,5715
Vício($\times 10^5$)							
			LNOV	32,0163	1,0784	1,0248	33,8929
			LNEQOV	32,0163	1,0784	0,6057	20,2778
1	LROV	LLA	NOV	32,0163	1,0784	1,9657	12,7148
			NEQOV	32,0163	1,0784	2,7229	49,1543
			LNOV	1436,9834	25,5000	27,6780	1584,4186
			LNEQOV	1436,9834	25,5000	22,0959	1153,4396
2	LRTOT	LROV	NOV	1436,9834	25,5000	27,9391	1639,1990
			NEQOV	1436,9834	25,5000	28,1264	258,3495
Razão EQM/ Variância Assintótica							
			LNOV	0,7551	1,0325	1,0977	0,7982
			LNEQOV	0,7551	1,0325	1,0894	0,7697
1	LROV	LLA	NOV	0,7551	1,0325	1,0453	1,0207
			NEQOV	0,7551	1,0325	1,0221	0,8455
			LNOV	1,7388	1,0594	1,0976	1,8649
			LNEQOV	1,7388	1,0594	1,0820	1,6059
2	LRTOT	LROV	NOV	1,7388	1,0594	1,0624	1,9580
			NEQOV	1,7388	1,0594	1,0677	0,8661

Para o modelo 1, que foi bem especificado, o EQM estimado dos estimadores de métodos baseados no modelo, sem informação sobre o desenho (MQO e Pearson) é quase a metade do obtido para os estimadores do método de MPV, embora esses EQMs sejam muito pequenos. Mas para o modelo 2, com problema de má especificação, o EQM estimado é bem próximo para todos os estimadores, com exceção ao de Pearson, usando a variável auxiliar NEQOV, que foi bem menor que os outros.

Os vícios estimados foram pequenos em relação à ordem de grandeza do EQM estimado, visto que estão multiplicados por 10^3 na Tabela 4.6 e por 10^5 na Tabela 4.7. Apesar disso, para o

modelo 1, os estimadores que usaram as informações do desenho amostral (MPV-π e MPV-R) apresentaram um vício cerca de 30 vezes menor que o dos outros dois. Já para o modelo 2, que apresenta má especificação, os vícios foram bem maiores, principalmente para os estimadores de MQO e de Pearson, mais de 50 vezes maior que os do MPV-π e do MPV-R. Estes dois últimos apresentaram vícios bem próximos. O MPV-π apresentou, em geral, um vício ligeiramente menor que o do MPV-R, exceto quando a variável auxiliar usada foi LNEQOV.

Podemos observar nas Tabelas 4.6 e 4.7 que a variância assintótica calculada pelos dois métodos de MPV foram bem próximas do EQM estimado, sendo este último, no máximo, 10% maior que a variância assintótica, o que já era de se esperar pois os vícios estimados destes estimadores foram pequenos. Ou seja, os estimadores pontuais dos coeficientes B_0 e B_1 se comportam como o esperado com base nas fórmulas assintóticas, já que os vícios são desprezíveis e as razões entre o EQM estimado e a variância assintótica são próximas de 1 para os estimadores de MPV-R e MPV-π. Os métodos que não utilizam informações sobre o desenho amostral (MQO e Pearson) têm essa razão variando mais que os que ignoram estas informações (MPV-π e MPV-R). O EQM estimado chega a ser cerca de 95% maior que a variância assintótica para o estimador de Pearson e cerca de 80% maior que a variância assintótica do MQO para o modelo com má especificação. Para este modelo, o uso do EQM para a comparação do desempenho dos estimadores pode levar a conclusões diferentes das obtidas pela variância assintótica. Para o modelo 1, estes estimadores apresentam uma variância assintótica ligeiramente maior que o EQM estimado, mas isto deve ter ocorrido porque estes dois valores foram calculados por maneiras diferentes: a variância assintótica foi calculada com os dados de toda a população finita e o EQM foi calculado pela média entre todas as amostras da distância entre a estimativa amostral e o parâmetro "verdadeiro" (Tabela 4.1).

O uso de variáveis auxiliares e informações sobre o desenho amostral contribuiu para a redução dos vícios dos estimadores dos coeficientes, embora o EQM fique ligeiramente maior para esses estimadores quando comparados com o estimador MQO.

4.5.2 – Estimação da variância dos estimadores dos coeficientes

As Tabelas 4.8 e 4.9 a seguir contém os valores médios das variâncias estimadas para os estimadores de B_0 e B_1 pelos quatro métodos estudados, além da estimativa de simulação da variância dos estimadores de variância dos coeficientes estimados, os p-valores do teste t, conforme descritos na seção 4.3.2, além das razões entre a média dos estimadores de variância dos coeficientes estimados e a variância calculada pelas fórmulas assintóticas, para cada método e cada modelo.

TABELA 4.8
Resultados da simulação dos estimadores de variância dos estimadores do coeficiente B_0
Amostragem Estratificada Simples

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Média das estimativas de variância[10]							
			LNOV	1,6021	2,0509	2,0764	--
			LNEQOV	1,6021	2,0509	2,0353	--
1	LROV	LLA	NOV	1,6021	2,0509	2,0549	--
			NEQOV	1,6021	2,0509	1,9961	--
			LNOV	1,0754	2,3594	2,3188	--
			LNEQOV	1,0754	2,3594	2,2773	--
2	LRTOT	LROV	NOV	1,0754	2,3594	2,3344	--
			NEQOV	1,0754	2,3594	2,3245	--
Variância das estimativas de variância (×10)							
			LNOV	1,7158	19,7129	27,7469	--
			LNEQOV	1,7158	19,7129	21,2119	--
1	LROV	LLA	NOV	1,7158	19,7129	21,3666	--
			NEQOV	1,7158	19,7129	19,5196	--
			LNOV	0,5934	18,2937	18,7839	--
			LNEQOV	0,5934	18,2937	19,1752	--
2	LRTOT	LROV	NOV	0,5934	18,2937	18,0640	--
			NEQOV	0,5934	18,2937	18,3841	--
P-valor do Teste T (em %)							
			LNOV	0,00	0,02	1,80	--
			LNEQOV	0,00	0,02	0,02	--
1	LROV	LLA	NOV	0,00	0,02	0,08	--
			NEQOV	0,00	0,02	0,00	--
			LNOV	0,00	0,00	0,00	--
			LNEQOV	0,00	0,00	0,00	--
2	LRTOT	LROV	NOV	0,00	0,00	0,00	--
			NEQOV	0,00	0,00	0,00	--
Razão: Médias das estimativas de variância/ Variância assintótica							
			LNOV	1,0595	0,9250	0,9433	--
			LNEQOV	1,0595	0,9250	0,9224	--
1	LROV	LLA	NOV	1,0595	0,9250	0,9298	--
			NEQOV	1,0595	0,9250	0,9035	--
			LNOV	0,6624	0,8376	0,8317	--
			LNEQOV	0,6624	0,8376	0,8370	--
2	LRTOT	LROV	NOV	0,6624	0,8376	0,8336	--
			NEQOV	0,6624	0,8376	0,8215	--

[10] As estimativas de variâncias para cada uma das 1000 amostras foram calculadas pelas fórmulas (3.48) para MPV-π e (3.52) e (3.61) para MPV-R. para as Tabelas 4.8 e 4.9. As estimativas de Pearson das variâncias para cada uma das 1000 amostras na Tabela 4.9 foram calculadas usando-se a fórmula (2.26).

Tabela 4.9
Resultados da simulação dos estimadores de variância dos estimadores do coeficiente B_1
AMOSTRAGEM ESTRATIFICADA SIMPLES

Modelo	Y	Z	X	MQO	MPV-π	MPV-R	Pearson
Média das estimativas de variância (×100)							
			LNOV	1,5374	2,1096	2,1415	1,4656
			LNEQOV	1,5374	2,1096	2,0995	1,4435
1	LROV	LLA	NOV	1,5374	2,1096	2,1105	1,8534
			NEQOV	1,5374	2,1096	2,0418	1,5227
			LNOV	1,0190	2,2188	2,1786	0,9788
			LNEQOV	1,0190	2,2188	2,1924	0,7833
2	LRTOT	LROV	NOV	1,0190	2,2188	2,1898	1,0606
			NEQOV	1,0190	2,2188	2,1904	0,7784
Variância das estimativas de variância ($\times 10^5$)							
			LNOV	1,4897	19,7912	29,2304	1,3064
			LNEQOV	1,4897	19,7912	21,7388	1,3237
1	LROV	LLA	NOV	1,4897	19,7912	21,5186	2,1175
			NEQOV	1,4897	19,7912	19,7123	1,4128
			LNOV	0,5332	16,5161	16,9806	0,4903
			LNEQOV	0,5332	16,5161	18,6805	0,3393
2	LRTOT	LROV	NOV	0,5332	16,5161	16,1458	0,5643
			NEQOV	0,5332	16,5161	17,1577	0,3278
P-valor do Teste T (em %)							
			LNOV	0,00	0,00	1,13	0,00
			LNEQOV	0,00	0,00	0,01	0,00
1	LROV	LLA	NOV	0,00	0,00	0,01	0,00
			NEQOV	0,00	0,00	0,00	0,00
			LNOV	0,00	0,00	0,00	0,00
			LNEQOV	0,00	0,00	0,00	0,00
2	LRTOT	LROV	NOV	0,00	0,00	0,00	0,00
			NEQOV	0,00	0,00	0,00	0,00
Razão: Médias das estimativas de variância / Variância assintótica							
			LNOV	0,9213	0,9182	0,9398	0,8781
			LNEQOV	0,9213	0,9182	0,9196	0,8661
1	LROV	LLA	NOV	0,9213	0,9182	0,9224	1,1048
			NEQOV	0,9213	0,9182	0,8933	0,8997
			LNOV	0,6571	0,8331	0,8280	0,6311
			LNEQOV	0,6571	0,8331	0,8357	0,5376
2	LRTOT	LROV	NOV	0,6571	0,8331	0,8281	0,6854
			NEQOV	0,6571	0,8331	0,8161	0,4290

Podemos observar, tanto pelos p-valores da Tabela 4.8 como da 4.9, que não se pode dizer que as variâncias estimadas são, em média, iguais às calculadas pelas fórmulas assintóticas, ao nível de 5% de significância.

Observando a razão entre a média das estimativas das variâncias dos coeficientes estimados e a variância assintótica, podemos observar que as estimativas calculadas pelos métodos de MPV, tanto com pesos π, quanto com pesos de regressão, apresentaram desempenho, mais "robusto" quanto à aderência do estimador de variância às variâncias assintóticas, porém

invariante à boa ou má especificação do modelo. Ou seja, as razões apresentadas nas Tabelas 4.8 e 4.9 relativas a estes dois modelos não se distanciaram de 1 por mais de 20%. Já os outros dois métodos apresentaram esta razão distante de 1 para o modelo 2, com problemas de especificação. Estes estimadores apresentaram esta razão próxima de 0,67, ou seja, as variâncias assintóticas são quase 50% maiores do que essa média das estimativas das variâncias dos coeficientes estimados.

É interessante observar também que a variância das estimativas de variância, em geral, foi maior para o MPV-R que para o MPV-π, independente do modelo utilizado. Os métodos que não utilizaram as informações do desenho (MQO e Pearson) apresentaram esta variância bem menor que os outros dois.

4.6 – Conclusões

Neste capítulo foi apresentado um estudo de simulação para comparar o desempenho de quatro estimadores de coeficientes de regressão e respectivos estimadores de variância. Este estudo foi iniciado por Nascimento Silva(1996, cap.6), apenas para a comparação entre as estimativas pontuais dos parâmetros.

Quando se utilizou uma AAS como método de seleção das amostras, viu-se que o uso de variáveis auxiliares não contribuiu muito para a aumentar a eficiência ou diminuir o vício dos estimadores. O MQO funcionou bem, com variância e vício pequenos. As variâncias assintóticas ficaram, em geral, próximas dos EQMs estimados.

Tanto as médias quanto as variâncias das estimativas de variância ficaram bem próximas para todos os métodos, para o modelo bem ajustado. Para o outro modelo, estas médias e variâncias foram um pouco mais heterogêneas.

Foi observado claramente que o estimador de Pearson foi muito sensível à má especificação do modelo e à escolha inadequada das variáveis auxiliares.

Quando se utilizou um outro desenho amostral não ignorável, o AES, viu-se que os estimadores que não levam em conta as informações do desenho apresentaram um vício bem maior (MQO e Pearson), principalmente para o modelo mal especificado, embora a eficiência (em termos de EQM) dos diversos estimadores seja parecida. Neste caso, o uso do MPV-π ou do MPV-R implica numa eficiência um pouco menor, mas sem vício.

As médias e variâncias das estimativas de variância foram maiores para os dois métodos de MPV. A razão entre a média das estimativas de variância e a variância assintótica do estimador de Pearson, no modelo 2, apresenta uma grande distância de 1. Para os métodos de MPV esta razão parece variar menos.

5 - Uma aplicação do MPV-R aos dados do Censo Demográfico de 1991

No capítulo anterior os diversos estimadores foram comparados através de um estudo de simulação. Neste capítulo o objetivo é comparar estimativas pontuais e de precisão obtidas por vários métodos, para o ajuste de um modelo de regressão aplicado a dados do município de Marília (SP), retirados da amostra Censo Demográfico de 1991, realizado pelo IBGE.

Este capítulo será dividido em três partes. Na primeira será descrito o desenho amostral utilizado pelo Censo Demográfico de 1991. Na segunda, descreveremos o modelo a ser ajustado. Finalmente, na terceira serão apresentados os resultados dos ajustes.

5.1 - Descrição do desenho amostral

No Censo Demográfico de 1991, como em todos desde 1960, foi utilizada amostragem na coleta dos dados, para a aplicação de dois tipos diferentes de questionário. Os domicílios que não foram selecionados para a amostra responderam a um questionário básico (chamado de CD 1.01), mais simples, que continha perguntas investigadas para toda a população, como sexo, idade, parentesco em relação ao chefe de domicílio e alfabetização de todas as pessoas do domicílio e, para os chefes e moradores individuais em domicílios coletivos, o rendimento mensal bruto e informações sobre instrução, como a última série concluída e o grau de instrução. Os domicílios selecionados para a amostra responderam a um outro questionário bem mais longo (chamado de CD 1.02), que além das perguntas contidas no CD 1.01, tinha outras mais detalhadas sobre características dos domicílios e das pessoas, como religião, escolaridade, fecundidade, mão-de-obra, migração, etc.[11]

O desenho amostral usado consistiu em seleção sistemática com eqüiprobabilidade, dentro de cada setor censitário, de uma amostra de domicílios particulares e famílias ou componentes de grupos conviventes recenseados em domicílios coletivos, com uma fração amostral constante em cada município. Para municípios com população estimada[12] superior a 15.000 habitantes, como era o caso de Marília, usou-se uma fração amostral de 10%; para os demais municípios usou-se a fração amostral de 20%.

---

[11] Para os domicílios selecionados na amostra, foi criado um questionário básico (CD 1.01) a partir do CD 1.02, através da transcrição manual de informações pelo próprio recenseador, e feita uma apuração independente para cada um dos dois questionários. Ver Albieri e Bianchini(1993) e Censo Demográfico de 1991, CD 1.07, Manual do Recenseador (instrumento de coleta).

[12] Estimativas de população para o ano de 1991, baseadas nas projeções independentes realizadas pelo Departamento de População e Indicadores Sociais da Diretoria de Pesquisas.

Aqui, por simplicidade, trabalhou-se com uma aproximação deste desenho, ou seja, considerou-se nesta aplicação que os dados eram provenientes de uma amostra estratificada simples sem reposição. Os estratos são os setores censitários e as probabilidades de inclusão π_{ij}, dos elementos i e j na amostra, ónde i pertence a um estrato h e j pertence a um estrato k, são definidas por:

$$\pi_{ij} = \begin{cases} (n_h \times n_k)/(N_h \times N_k) & \text{se } i \neq j \text{ e } k \neq h \\ (n_h \times (n_h-1))/(N_h \times (N_h-1)) & \text{se } i \neq j \text{ e } k = h \\ n_h/N_h & \text{se } i = j \end{cases}$$

onde n_h é o numero de elementos do estrato h na amostra e N_h é o número de elementos do estrato h na população.

A matriz Π_s será uma matriz diagonal, com os elementos da diagonal iguais aos π_{ii}, como definidos acima.

5.1.1 - A Obtenção dos Pesos Divulgados na Amostra do Censo Demográfico de 1991

Nos microdados correspondentes aos questionários da amostra do Censo Demográfico de 1991 existe uma variável PESO, que é utilizada para a expansão da amostra. Esta variável não foi obtida simplesmente pelo inverso das probabilidades de inclusão. Esses pesos foram obtidos através de um método de estimação denominado, no IBGE, de MQG2 (Mínimos Quadrados Generalizados em duas etapas) (Nascimento Silva, Bianchini e Albieri; 1993). Esse método é baseado em um procedimento multivariado de estimação por mínimos quadrados generalizados em duas etapas e foi desenvolvido pelos técnicos do *Statistics Canada* e aplicado na expansão dos Censos de População Canadense de 1991 e 1996.

O MQG2 foi concebido de maneira a produzir pesos que, quando usados para calcular as estimativas de totais de algumas características básicas (ou seja, as características comuns ao CD 1.01 e ao CD 1.02), fornecem estimativas que se igualam aos valores conhecidos para toda a população dos totais destas características, divulgados como resultados do universo. Sua aplicação consiste em determinar os pesos das unidades amostrais baseados no ajuste de um modelo linear sujeito a restrições, que são justamente as condições que buscam igualar as estimativas aos valores conhecidos do universo para este conjunto de variáveis. O resultado desta metodologia é um peso ajustado para cada unidade domiciliar da amostra.

Um aspecto importante deste método é a escolha das variáveis auxiliares usadas neste processo de ajuste. Como a metodologia de ajuste de um modelo linear multivariado envolve cálculos com matrizes, inclusive inversão, as restrições definidas, que dão origem a cada coluna dessas matrizes, devem satisfazer algumas condições essenciais. A principal delas é a de que as

restrições não podem ser linearmente dependentes. Além disso, considera-se também o caso dessas restrições serem quase linearmente dependentes, pois afetariam a estabilidade da solução do modelo. Outras duas condições importantes para o ajuste deste modelo dizem respeito à significância estatística. Quando uma restrição não atinge um número mínimo de unidades domiciliares, fixado considerando a fração amostral usada naquele município, esta restrição é considerada rara. E quando uma restrição definida pode causar um peso extremo (muito grande ou muito pequeno), considera-se essa restrição geradora de peso extremo.

Assim, o programa de ajuste do modelo possui também procedimentos para eliminação das restrições linearmente dependentes e quase linearmente dependentes, raras e geradoras de pesos extremos. A eliminação de restrições pode implicar no fato de não se ter garantia da consistência desejada para as variáveis correspondentes às restrições eliminadas.

Esta metodologia da expansão da amostra foi aplicada separada e independentemente para cada uma das áreas de ponderação, que são formadas por um conjunto de setores. Desta maneira, apesar de se entrar com o mesmo conjunto grande de restrições em todas as áreas, durante o processo houve eliminações de restrições de forma independente para cada área. Por este motivo, e por se resolver usar as variáveis correspondentes às restrições não eliminadas como variáveis auxiliares do método de MPV com pesos de regressão, trabalhou-se, nesta aplicação, com apenas uma das cinco áreas de ponderação de Marília. O conjunto de variáveis auxiliares usado está descrito no Quadro 5.1.

QUADRO 5.1
Variáveis Auxiliares

TPES	Total de pessoas no domicílio
MURB	Número de mulheres do domicílio, se este domicílio pertence ao setor urbano.[13]
HCHEFE	HCHEFE= 1 se o chefe do domicílio (ou individual) for do sexo masculino, 0 caso contrário.
NID0A4	Número de pessoas do domicílio, com 0 a 4 anos completos.
NID5A9	Número de pessoas do domicílio, com 5 a 9 anos completos.
NID10A14	Número de pessoas do domicílio, com 10 a 14 anos completos.
NID15A19	Número de pessoas do domicílio, com 15 a 19 anos completos.
NID20A24	Número de pessoas do domicílio, com 20 a 24 anos completos.
NID25A29	Número de pessoas do domicílio, com 25 a 29 anos completos.
NID30A34	Número de pessoas do domicílio, com 30 a 34 anos completos.
NID35A39	Número de pessoas do domicílio, com 35 a39 anos completos.
NID40A44	Número de pessoas do domicílio, com 40 a 44 anos completos.
NID45A49	Número de pessoas do domicílio, com 45 a 49 anos completos.
NID50A59	Número de pessoas do domicílio, com 50 a 59 anos completos.
NID60A69	Número de pessoas do domicílio, com 60 a 69 anos completos.
NHID0A4	Número de pessoas do sexo masculino do domicílio, com 0 a 4 anos completos.
NHID5A9	Número de pessoas do sexo masculino do domicílio, com 5 a 9 anos completos.
NHID10A14	Número de pessoas do sexo masculino do domicílio, com 10 a 14 anos completos.
NHID15A19	Número de pessoas do sexo masculino domicílio, com 15 a 19 anos completos.
NHID25A29	Número de pessoas do sexo masculino do domicílio, com 25 a 29 anos completos.
NHID30A34	Número de pessoas do sexo masculino do domicílio, com 30 a 34 anos completos.
MOR3	MOR3= 1 se o domicílio tiver 3 moradores, 0 caso contrário.
MOR4	MOR4= 1 se o domicílio tiver 4 moradores, 0 caso contrário.
MOR5	MOR5= 1 se o domicílio tiver 5 moradores, 0 caso contrário.

Uma das restrições que não foi eliminada pelo MQG2, mas que também não foi usada como variável auxiliar, foi uma chamada de UNIDADES DOMICILIARES, que consistia em cada domicílio ter o valor 1 para esta variável. Mas como é colocado um vetor unitário nas

[13] Como a área de ponderação escolhida pertence a um setor urbano, esta variável corresponde simplesmente ao número de mulheres do domicílio.

matrizes x_s^* e x_u^* usadas na "correção" dos pesos π_i^{-1}, haveria o problema de matrizes não inversíveis na estimação pelo método MPV-R

Este método de obtenção dos pesos se assemelha a uso de um estimador de regressão, com as mesmas variáveis auxiliares, embora seja mais complexo.

5.2 - Descrição do modelo

Como já foi dito no início deste capítulo, foram utilizados dados do município de Marília provenientes da amostra do Censo Demográfico de 1991, apenas para uma área de ponderação urbana. As variáveis de interesse do modelo são:

QUADRO 5.2
Variáveis de interesse

RCHEFE	Renda do chefe do domicílio no CD 1.02
LRCHEFE	LRCHEFE=Log(RCHEFE+1)
HOMEM	HOMEM= 1 se o chefe do domicílio ou individual for do sexo masculino. 0 caso contrário.
IDADE	Idade em anos do chefe do domicílio ou individual no CD 1.02
IDADE2	IDADE2= IDADE ^{2}
ANOSEST	Número de anos completos de estudo do chefe do domicílio ou individual.
ADMIN	ADMIN= 1 se o chefe do domicílio ou individual possui ocupação administrativa. 0 caso contrário.
TECART	TECART= 1 se o chefe do domicílio ou individual possui ocupação na área técnica, científica, artística e assemelhadas. 0 caso contrário.
INDCONST	INDCONST= 1 se o chefe do domicílio ou individual possui ocupação na área de indústrias de transformação e construção civil.. 0 caso contrário.
COMERC	COMERC= 1 se o chefe do domicílio ou individual possui ocupação na área do comércio ou atividades auxiliares. 0 caso contrário.
TRANSP	TRANSP= 1 se o chefe do domicílio ou individual possui ocupação na área de transportes e comunicações. 0 caso contrário.

A variável dependente do modelo de regressão a ser ajustado é LRCHEFE, e as variáveis independentes consideradas foram HOMEM, IDADE, IDADE2, ANOSEST, ADMIN, TECART, INDCONST, COMERC e TRANSP.

Não foram considerados os chefes de domicílio que não responderam a alguma das perguntas relacionadas com as variáveis consideradas no modelo descrito, dando um total de 772 chefes de domicílio na amostra para a área de ponderação considerada.

Nos gráficos 5.1 a 5.8 podemos ver a relação da variável dependente LRCHEFE e cada uma das outras variáveis independentes. Podemos notar, indicado pelas setas, um ponto bastante destacado dos demais, que provavelmente dará problema no ajuste do modelo. Pode-se notar também que todas as variáveis parecem ter poder explicativo em relação à variável dependente. Os boxplots foram calculados levando-se em conta os pesos de MQG2 dos microdados da amostra, ou seja, usando-se os pesos truncados como variável de freqüência de cada observação na hora de se fazer os boxplot no SAS.

A variável idade demonstra uma leve tendência quadrática. Desta maneira IDADE2 também foi incluída no modelo.

GRÁFICO 5.1
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável HOMEM

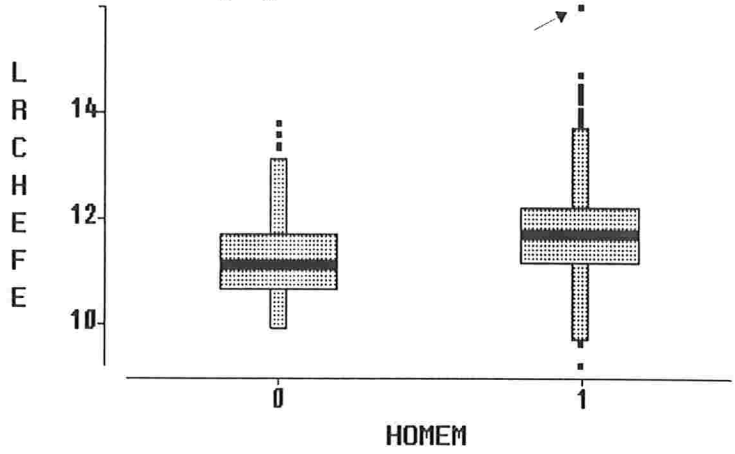


GRÁFICO 5.2
Gráfico de Dispersão do Logaritmo da renda do chefe como função da variável IDADE

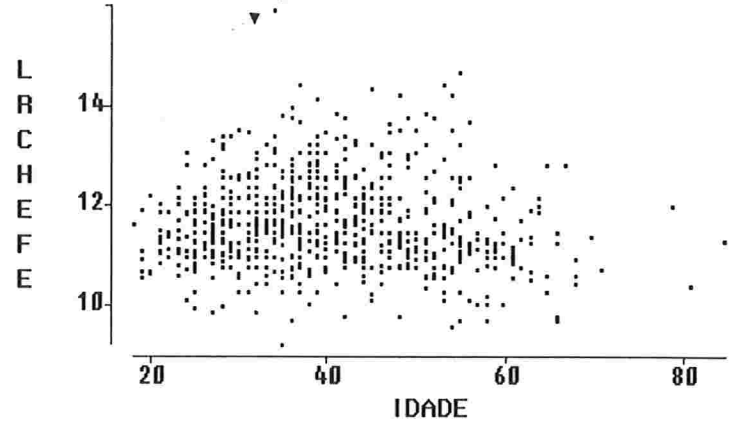


GRÁFICO 5.3
Gráfico de Dispersão do Logaritmo da renda do chefe como função da variável Anos de Estudo

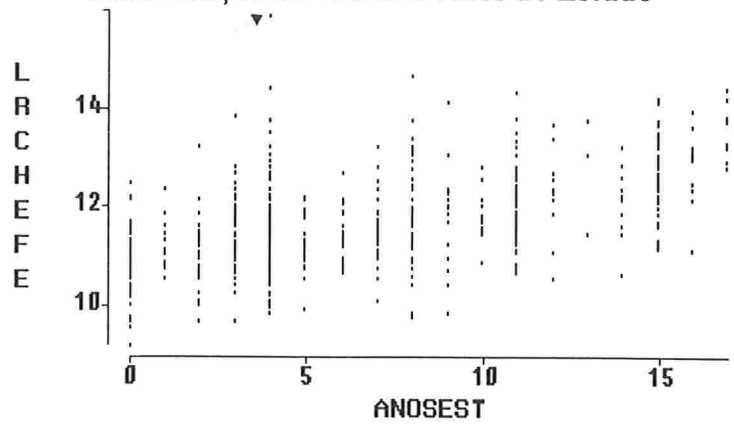


GRÁFICO 5.4
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável ADMIN

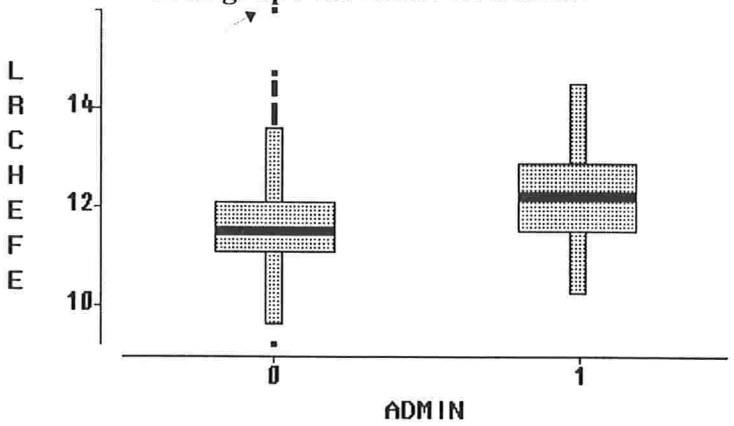


GRÁFICO 5.5
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável TECART

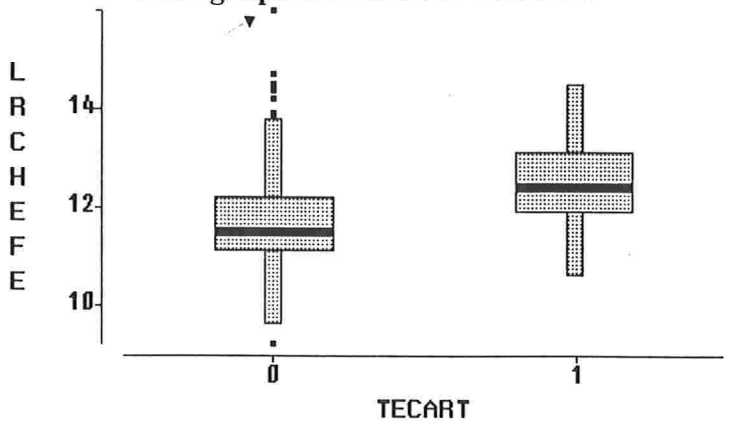


GRÁFICO 5.6
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável INDCONST

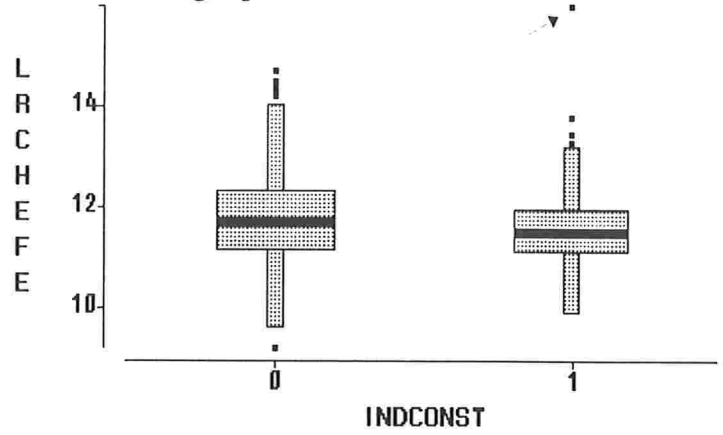


GRÁFICO 5.7
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável COMERC

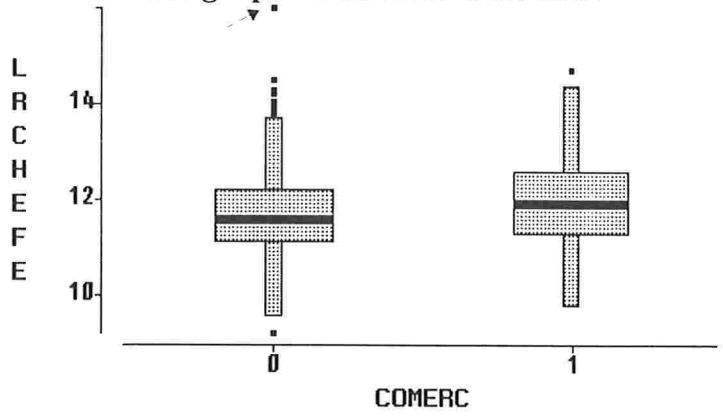
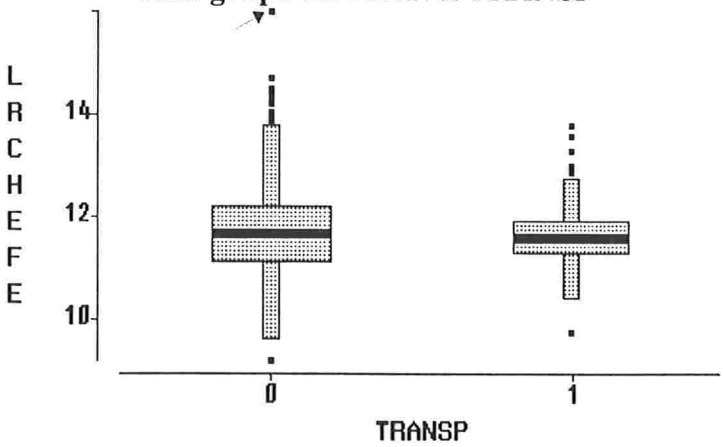


GRÁFICO 5.8
Boxplots do Logaritmo da Renda do Chefe dentro de cada grupo da variável TRANSP



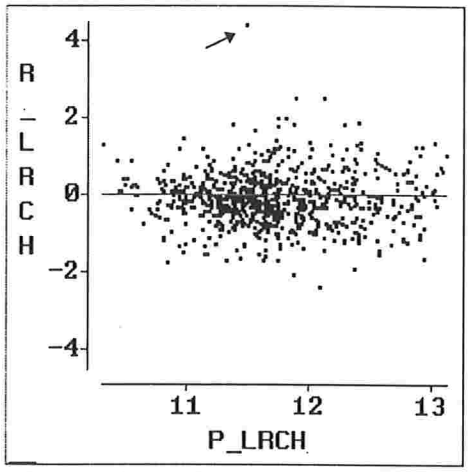
Um primeiro ajuste foi feito, usando apenas os pesos de MQG2 e o método dos Mínimos Quadrados Ponderados, apenas para se observar se o modelo era adequado. Os resultados estão na Tabela 5.1 a seguir. Esse ajuste usou as informações que comumente um analista usaria para os dados do Censo Demográfico: um software estatístico para o ajuste de um modelo de

regressão linear (no caso, o *SAS - Statistical Analysis System*), e os pesos que são disponibilizados pelo IBGE junto com os microdados da amostra.

TABELA 5.1
Ajuste do Modelo usando os Pesos de MQG2 e Mínimos Quadrados Ponderados

Modelo	Parâmetro	Estimativa	Desvio Padrão Estimado	P-valor do Teste t
Y: LRCHEFE				
Intercept	β_0	0,8737	0,2826	0,0001
Z: HOMEM	β_1	0,4076	0,0850	0,0001
Z: IDADE	β_2	0,0690	0,0128	0,0001
Z: IDADE2	β_3	-0,0007	0,0001	0,0001
Z: ANOSEST	β_4	0,0974	0,0068	0,0001
Z: ADMIN	β_5	0,5154	0,0820	0,0001
Z: TECART	β_6	0,5038	0,1196	0,0001
Z: INDCONST	β_7	0,2770	0,0670	0,0001
Z: COMERC	β_8	0,3652	0,0826	0,0001
Z: TRANSP	β_9	0,3608	0,0936	0,0001
	R^2	0,4088		

GRÁFICO 5.9
Resíduos do Ajuste por Mínimos Quadrados Ponderados e Pesos do MQG2



Pode-se ver pela coluna de p-valores do teste t da Tabela 5.1 que os parâmetros parecem ser significativamente diferentes de zero, mostrando a importância deles na explicação da variável dependente. Apesar disso, o R^2 não é muito alto, embora os resíduos não pareçam ter estrutura. Pelo Gráfico 5.9 pode-se notar que os resíduos estão aleatoriamente distribuídos, mas aquele ponto que se destacava nos gráficos anteriores gerou um valor discrepante no gráfico de resíduos (marcado pela seta). Assim, vamos considerar nos ajustes, daqui em diante, os dados amostrais usados, sem incluir este ponto, ficando a amostra, então, com 771 dados de chefes de domicílios ou individuais, não considerando nenhum dado com não resposta.

5.3 - Comparações entre os ajustes

Com o objetivo de comparar os resultados de diversos métodos de ajuste, alguns levando em conta a estrutura da amostra e outros não, iremos ajustar um modelo com as variáveis

descritas no Quadro 5.2, onde LRCHEFE é a variável dependente e HOMEM, IDADE, IDADE2, ANOSEST, ADMIN, TECART, INDCONST, COMERC e TRANSP as variáveis independentes, a 771 dados para chefes de domicílios. Estes dados, como já foi dito, foram retirados do Censo Demográfico de 1991, para uma das 5 áreas de ponderação do município de Marília, sendo esta área na parte urbana. Nestes dados foram retirados todos os que apresentaram não resposta para alguma variável, além de um outlier.

Foram feitos ajustes usando-se os seguintes métodos:

1 - Mínimos Quadrados Ordinários (MQO);

2 - Mínimos Quadrados Ponderados(MQP), usando-se os pesos de MQG2 divulgados na pesquisa;

3 - Ajuste dos Mínimos Quadrados Ponderados usando-se o SUDAAN, que leva em conta o desenho amostral;

4 - Método da Máxima Pseudo-Verossimilhança usando-se os pesos π (probabilidades de inclusão de cada indivíduo na amostra) (MPV-π); e

5 - Método da Máxima Pseudo-Verossimilhança usando-se os pesos de regressão (MPV-R).[14]

O SUDAAN faz o ajuste pontual dos parâmetros pelo MPV-π, bastando entrar com o vetor de pesos igual à diagonal da matriz Π_s^{-1}, e utiliza uma aproximação de Taylor para o cálculo das estimativas das variâncias dos parâmetros estimados.

Para o ajuste do MPV-R, foram usadas as 23 variáveis do Quadro 5.1 como variáveis auxiliares em x^*_U e x^*_s, da população e da amostra, respectivamente, em (3.13).

As estimativas pontuais dos parâmetros estão na Tabela 5.2 a seguir.

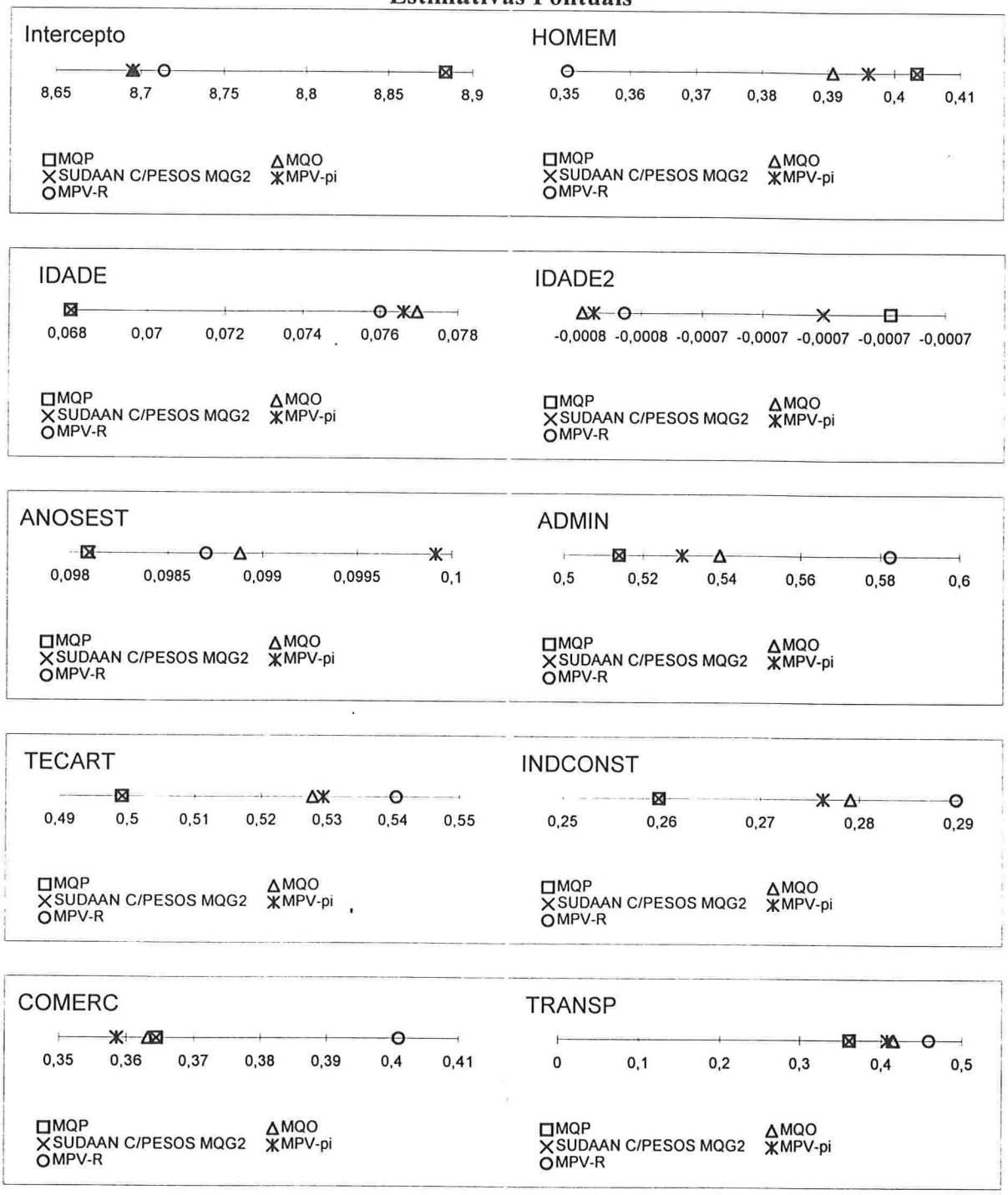
Tabela 5.2
Estimativas pontuais calculadas por diversos métodos

Parâmetros	MQO	MQP	SUDAAN c/ pesos de MQG2	MPV-π	MPV-R
Intercepto	8,696	8,885	8,885	8,696	8,714
HOMEM (×10)	3,908	4,035	4,035	3,961	3,504
IDADE (×10^2)	7,696	6,809	6,810	7,662	7,602
IDADE2 (×10^4)	-7,794	-6,780	-7,000	-7,755	-7,653
ANOSEST(×10^2)	9,888	9,809	9,810	9,992	9,870
ADMIN (×10)	5,393	5,141	5,141	5,299	5,825
TECART (×10)	5,278	4,993	4,993	5,294	5,405
INDCONST (×10)	2,791	2,597	2,597	2,763	2,898
COMERC (×10)	3,631	3,643	3,643	3,587	4,011
TRANSP (×10)	4,155	3,613	3,613	4,077	4,597

[14] Essa maneira daria os mesmos valores de estimadores que o ajuste de Mínimos Quadrados Ponderados usando-se o SUDAAN. Isto não ocorreu na prática porque os pesos de MQG2 têm mais ajustes que os pesos de regressão usuais, aplicados no MPV-R. Além disso, estes últimos só usam os dados que entraram no ajuste, ao contrário do MQG2. Estes pesos de regressão foram recalculados para se estudar o caso onde os pesos aos quais o usuário tem acesso não são os de regressão, embora tenha acesso a variáveis auxiliares que permitam o cálculo dos pesos de regressão.

Podemos comparar melhor estas estimativas no Gráfico 5.10 a seguir, onde foi feito um gráfico para cada parâmetro com as cinco estimativas.

GRÁFICO 5.10
Estimativas Pontuais



Em quase todas as variáveis, os correspondentes parâmetros estimados por métodos que usaram os pesos de MQG2 foram os que ficaram mais distantes do MPV-R, com exceção de ANOSEST e COMERC. Para estas duas variáveis, o que ficou mais distante do MPV-R foi o MPV-π.

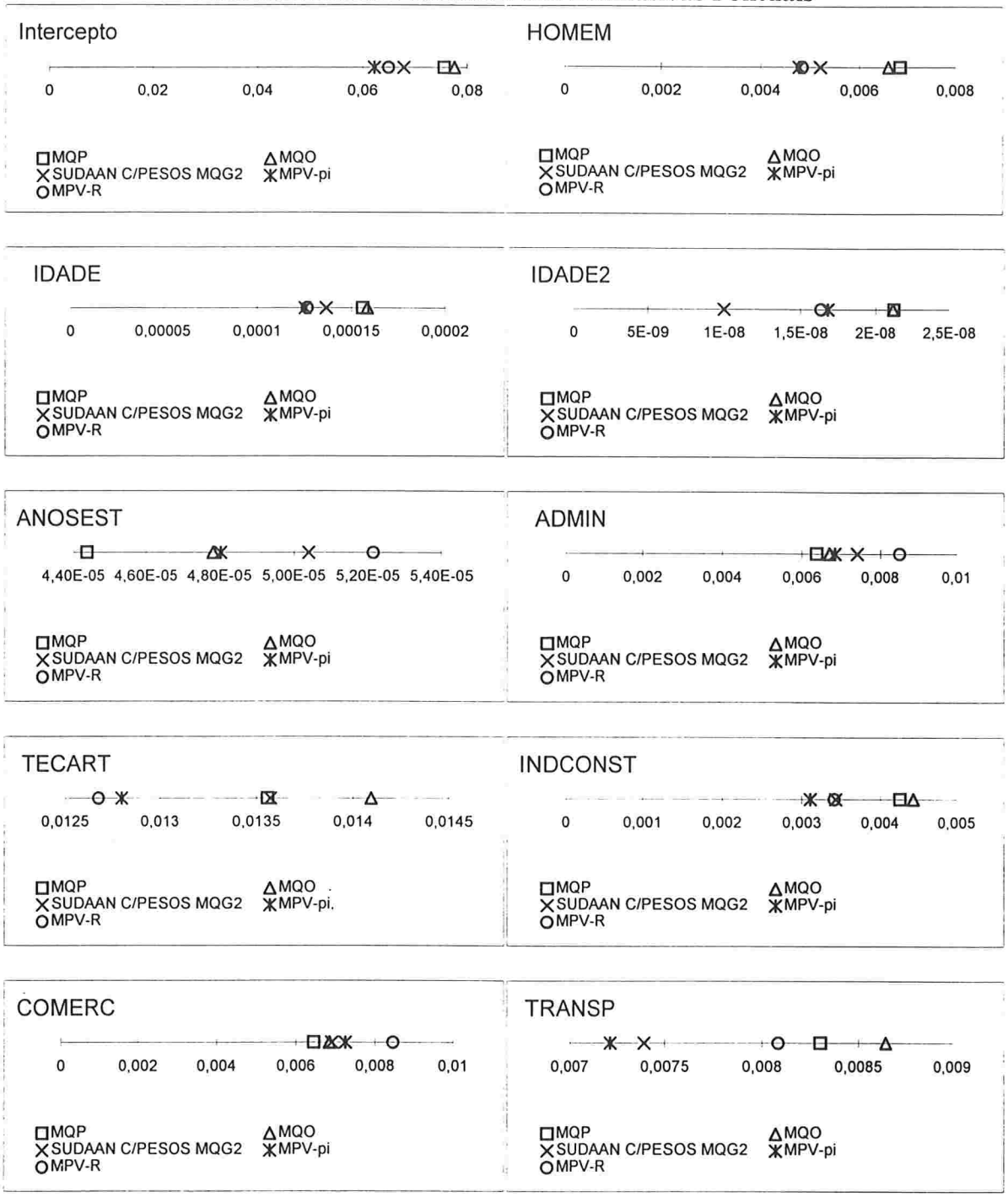
A diferença dos métodos que usaram pesos de MQG2 deve ter ocorrido devido ao procedimento de calibração para os totais do universo, usado para a obtenção dos pesos de MQG2. Além disso, neste procedimento foram utilizados todos os indivíduos, ao contrário desta aplicação que está sendo feita, onde foram excluídos os indivíduos que não eram chefes, os que apresentavam problema de não resposta, além de um outlier. As estimativas pontuais do MQP e do SUDAAN com pesos MQG2 deveriam ser exatamente iguais, as diferenças ocorridas foram por problemas de arredondamento dos sistemas usados.

As estimativas das variâncias das estimativas pontuais dos parâmetros estão na Tabela 5.3 e no Gráfico 5.11 a seguir.

TABELA 5.3
Estimativas das variâncias das estimativas pontuais calculadas por diversos métodos

Parâmetros	MQO	MQP	SUDAAN c/ pesos de MQG2	MPV-π	MPV-R
Intercepto($\times 10^2$)	7,768	7,573	6,791	6,224	6,500
HOMEM ($\times 10^3$)	6,605	6,860	5,213	4,778	4,848
IDADE ($\times 10^4$)	1,584	1,561	1,369	1,254	1,268
IDADE2 ($\times 10^8$)	2,121	2,127	1,000	1,681	1,639
ANOSEST($\times 10^5$)	4,783	4,437	5,041	4,803	5,215
ADMIN ($\times 10^3$)	6,708	6,385	7,413	6,845	8,525
TECART ($\times 10^2$)	1,409	1,356	1,357	1,280	1,268
INDCONST ($\times 10^3$)	4,437	4,264	3,434	3,112	3,416
COMERC ($\times 10^3$)	6,865	6,468	6,922	7,261	8,470
TRANSP ($\times 10^3$)	8,645	8,303	7,396	7,219	8,085

GRÁFICO 5.11
Estimativas das Variâncias das Estimativas Pontuais



Em quase todas as variáveis, as correspondentes estimativas das variâncias das estimativas pontuais por métodos que não utilizaram as informações do desenho da amostra foram os que ficaram mais distantes do MPV-R, com exceção das variáveis IDADE2 e TRANSP. Em IDADE2 a estimativa que ficou mais distante da do método MPV-R foi a calculada pelo SUDAAN, e em TRANSP foi pelo MPV-π.

Com exceção de ANOSEST, ADMIN e COMERC, as variâncias estimadas por métodos que não usaram o desenho amostral ficaram quase sempre maiores que as outras.

Nas Tabelas 5.4 e 5.5 a seguir podemos ver a distância percentual (em módulo) de cada estimativa (pontual e de variância) em relação aos respectivos valores calculados por MQO, onde não se leva em conta nem o desenho amostral, nem nenhuma ponderação. Ou seja, se uma estimativa calculada por um dos métodos for a e a estimativa correspondente calculada por MQO for b, o módulo da distância percentual encontrada foi $100\times|b-a|/|b|$.

TABELA 5.4
Diferenças percentuais em relação às estimativas pontuais obtidas por MQO

Variáveis	MQP	SUDAAN c/ pesos de MQG2	MPV-π	MPV-R
INTERCEP	2,17	2,17	0,00	0,21
HOMEM	3,24	3,25	1,36	10,32
IDADE	11,53	11,51	0,45	1,23
IDADE2	13,01	10,18	0,50	1,81
ANOSEST	0,80	0,79	1,05	0,18
ADMIN	4,68	4,68	1,76	8,00
TECART	5,41	5,40	0,30	2,41
INDCONST	6,95	6,96	0,99	3,83
COMERC	0,33	0,32	1,23	10,44
TRANSP	13,04	13,04	1,86	10,66
Diferença Média	6,12	5,83	0,95	4,91

TABELA 5.5
Diferenças percentuais em relação às estimativas de variâncias das estimativas pontuais obtidas por MQO

Variáveis	MQP	SUDAAN c/ pesos de MQG2	MPV-π	MPV-R
INTERCEP	2,51	12,58	19,88	16,32
HOMEM	3,85	21,08	27,66	26,60
IDADE	1,42	13,57	20,80	19,92
IDADE2	0,26	52,86	20,75	22,73
ANOSEST	7,22	5,40	0,43	9,04
ADMIN	4,81	10,52	2,05	27,09
TECART	3,81	3,71	9,17	10,05
INDCONST	3,90	22,61	29,87	23,02
COMERC	5,78	0,84	5,78	23,38
TRANSP	3,95	14,44	16,49	6,48
Diferença Média	3,75	15,76	15,29	18,46

Pode-se ver, pela Tabela 5.4, que os métodos que levam em conta o desenho da amostra não chegaram a estimativas pontuais muito diferentes, em média, das obtidas por métodos que não levam em conta o desenho. O MPV-π, que não utiliza pesos muito elaborados, como os de MQG2 e os de regressão, é o que menos difere das estimativas pontuais de MQO. Os que obtiveram estimativas pontuais mais afastadas das de MQO foram os métodos que utilizaram os

pesos de MQG2, talvez por causa do método de calibração pelos totais do universo. Ou seja, na diferença entre as estimativas pontuais entre os diversos métodos, o que mais pesou foi o uso de informações auxiliares da população, e não informações sobre desenho amostral.

Ao contrário das estimativas pontuais, o módulo das diferenças percentuais das estimativas de variância em relação às obtidas por MQO foram, em média, bem maiores para os métodos que usaram informação sobre o desenho da amostra, ou seja, esta informação foi a que pesou mais na diferença entre as estimativas de variâncias pelos diversos métodos.

Podemos concluir que, no caso de Amostragem Estratificada Simples sem reposição, não levar em conta informações auxiliares da população, como o uso dos pesos divulgados pelas pesquisas, pode influir na estimação pontual dos coeficientes. Por outro lado, ignorar no ajuste do modelo informações sobre o desenho amostral utilizado na pesquisa, como a estratificação por setor e as diferentes frações amostrais observadas, pode levar a diferenças nas estimativas das variâncias dos coeficientes estimados.

6 - Conclusões

Este trabalho tratou do problema de estimação de regressão linear com dados provenientes de amostras complexas, considerando em particular o método de MPV com pesos de regressão. Este método incorpora informações tanto do desenho quanto de variáveis auxiliares, como médias populacionais conhecidas.

Foram feitas simulações para verificar o desempenho dos estimadores dos coeficientes e das variâncias dos coeficientes, tanto para o caso de AAS como para o de Amostragem Estratificada Simples. Com isto, comparou-se estes estimadores com o de MQO, mais usual, e o de Pearson, que é um estimador baseado no modelo, mas que também incorpora informações de variáveis auxiliares.

Por estas simulações, verificou-se que, para AAS, o estimador de MQO teve um desempenho bom e um vício pequeno, mesmo quando o modelo estava mal-especificado. O estimador de Pearson se mostrou sensível à má especificação do modelo e a uma escolha ruim de variável auxiliar. Em geral, todos os estimadores tiveram um desempenho razoável e um vício pequeno, e, portanto, o uso de variáveis auxiliares em modelos como MPV-R e Pearson não implica em uma grande melhora de precisão ou em uma sensível diminuição do vício.

Em termos dos estimadores das variâncias, verificou-se que os estimadores de MPV-R foram os mais "robustos", pois apresentaram menor variabilidade e menor distância entre a média dos estimadores de variância e a variância assintótica calculada com os dados populacionais. Essa distância foi ligeiramente menor para o MPV-π, mas a variabilidade dos estimadores das variâncias deste método foi 4 vezes maior que a do MPV-R, mesmo para o modelo mal especificado e variáveis auxiliares mal escolhidas. Para o modelo com problemas de especificação, os estimadores de variância que apresentaram menor variância foram os de Pearson e o do MQO, que são baseados no modelo. Ao contrário, quando o modelo era bem especificado, quem apresentou menor variação foi o MPV-R.

Para Amostragem Estratificada Simples, o vício foi maior para os métodos que não usaram as informações do desenho amostral (MQO e Pearson) do que para os que usaram. Quando o modelo era mal especificado, essa diferença foi ainda maior. De novo, Pearson se mostrou mais sensível que o MPV-R à má especificação do modelo e a uma escolha errada da variável auxiliar. Quando o modelo estava bem especificado, Pearson teve um melhor desempenho, mas um vício maior que os MPVs.

Pode-se concluir que quando o modelo está mal especificado, os estimadores que incorporam os pesos do desenho são mais indicados, pois o desempenho ficou mais ou menos no

mesmo nível de EQM, mas o vício foi bem menor. Além do MPV-R ser mais robusto que o de Pearson, ele não exige que as variáveis auxiliares sejam normais multivariadas, e nem contínuas.

Quanto aos estimadores de variância dos coeficientes estimados, podemos ver que a variabilidade dos estimadores de variância pelo método de MQO e Pearson foi bem menor que a variabilidade destes estimadores pelo método da Máxima Pseudo-Verossimilhança. Em compensação, a distância entre a média das estimativas de variância e a variância assintótica desses estimadores é bem maior, no modelo mal especificado, para os estimadores de MQO e Pearson, chegando a haver uma superestimação desta variância assintótica de quase 50%.

Também foi feita, neste trabalho, uma aplicação do método do MPV no ajuste de um modelo linear aos dados de uma área de ponderação de Marília (SP) com dados do Censo Demográfico de 1991. Considerou-se a AES como desenho amostral usado para a obtenção destes dados, e vários métodos de estimação. As estimativas pontuais dos métodos que não usaram os pesos de MQG2, divulgados na pesquisa e ajustados para o total da população pelas variáveis auxiliares usadas aqui, ficaram bem próximas. Já para as estimativas das variâncias das estimativas pontuais, o componente de vício apareceu. Os estimadores que não usaram as informações amostrais ficaram mais distantes dos outros. Escolhendo-se o estimador mais usual, o MQO, para ser o termo de comparação, obteve-se que os estimadores tiveram uma distância média dos coeficientes estimados do MQO de 5% apenas. Mas em relação às estimativas de variância dos estimadores pontuais, vimos que os métodos que incluem informações de desenho na estimação ficam entre 15% e 18,5%, em média, diferentes dos obtidos pelo MQO, mostrando como o uso dos métodos usuais de estimação com desenhos amostrais diferentes de AAS, mais complexos, causa mudanças na estimação da variância.

O método a ser escolhido por um analista para o ajuste de um modelo linear a dados de pesquisa por amostra depende das informações a que ele tiver acesso. Se não tiver acesso a nenhuma informação sobre o desenho, e quiser ver a relação de regressão entre um conjunto de variáveis, pode até ajustar o modelo por MQO, embora deva ter em mente que a não utilização de informações sobre o desenho amostral da pesquisa poderá acarretar estimadores viciados tanto para os coeficientes como para as variâncias dos coeficientes. Se o analista só tiver informações sobre os pesos, mas não tiver acesso a nenhum detalhe sobre o desenho utilizado, pode usar MQP, que embora possa chegar a estimativas "corretas" dos coeficientes, levará a estimativas viciadas das variâncias dos coeficientes. Se o analista tiver informações sobre o desenho utilizado e sobre as probabilidades de inclusão, poderá usar o MPV-π, que já foi incluído em alguns sistemas como o SUDAAN. No caso do analista que tiver acesso a informações sobre o desenho utilizado e aos pesos de regressão, ou que tiver acesso aos valores de variáveis auxiliares

para toda a população e quiser usar alguma compensação de não resposta, pode-se usar o MPV-R, que ainda não foi incluído em nenhum sistema estatístico padrão. Apesar disso, não é difícil usar o MPV-R mediante aplicações sucessivas de sistemas como o SUDAAN, devidamente programado para calcular variâncias de resíduos.

7 - Glossário de siglas usadas no texto

AAS	Amostragem Aleatória Simples
AASC	Amostragem Aleatória Simples com reposição
AES	Amostragem Estratificada Simples
BRR	Método das Replicações Repetidas Balanceadas, método "indireto" de estimação
Cov	Covariância
CV	Coeficiente de Variação
E	Valor esperado
EQM	Erro Quadrático Médio
JRR	Método das Replicações Repetidas Jackknife, método "indireto" de estimação
MPV	Máxima Pseudo-Verossimilhança
MPV-R	Máxima Pseudo-Verossimilhança com pesos de regressão.
MPV-π	Máxima Pseudo-Verossimilhança com pesos iguais ao inverso da probabilidade de inclusão de cada elemento na amostra.
MQG2	Mínimos Quadrados Generalizados em 2 estágios, método de obtenção do ajuste dos pesos usado no Censo Demográfico de 1991.
MQO	Mínimos Quadrados Ordinários
MQP	Mínimos Quadrados Ponderados
Pearson	Estimador de Pearson (ver capítulo 2)
PNAD	Pesquisa Nacional por Amostra de Domicílios, pesquisa demográfica realizada pelo IBGE anualmente, exceto em ano de Censo Demográfico.
SAS	Statistical Analysis System, sistema estatístico
SUDAAN	Survey Data Analysis, sistema estatístico
V	Variância

8 – Bibliografia

ALBIERI, S. e BIANCHINI, Z.M.(1993). *Censo Demográfico de 1991: Sobre a independência da apuração do CD 1.01 e do CD 1.02.* Rio de Janeiro: IBGE, Divisão de Metodologia, 20pp.

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

CENSO DEMOGRÁFICO DE 1991. CD1.07 – Manual do Recenseador (instrumento de coleta).

COX, D.R. e HINKLEY, D.V. (1974). *Theoretical Statistics.* Londres: Chapman and Hall.

DEMETS, D. e HALPERIN, M. (1977). Estimation of a simple regression coefficient in samples arising from a sub-sampling procedure. *Biometrics*, 33, pp. 47-56.

FULLER, W. A. (1975) Regression Analysis for Sample Survey. *Sankhyã: The Indian Journal of Statistics.*, Volume 37, Series C, Parte 3, pp 117-132.

GODAMBE, V.P.; THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 2, pp. 127-138.

HOLT, D; SMITH, T.M.F; E WINTER, P.D.(1980). Regression Analysis of Data from Complex Surveys. *J.R.Statist.Soc.* A, 143, Part4, pp.474-487.

KISH, L. (1965). *Survey sampling.* Nova Iorque: John Wiley.

KISH, L. e FRANKEL, M.R. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society*, Serie B, 36, pp. 1-37.

NASCIMENTO SILVA, P.L.D. (1996). *Utilizing auxiliary information for estimation and analysis in sample surveys.* Southampton: University of Southampton, Department of Social Statistics, Tese de Doutorado.

NASCIMENTO SILVA, P.L.D.; BIANCHINI, Z.M. e ALBIERI, S. (1993). *Uma proposta de metodologia para a expansão da amostra do Censo Demográfico de 1991.* Rio de Janeiro: IBGE, Texto para Dicussão nº 62, 106 pp.

NATHAN, G. e HOLT, D. (1980). The Effect of Survey Design on Regression Analysis. *Journal of Royal Statistical Society* B., 42, nº 3, pp.377 a 386.

PEARSON, K. (1902). On the influence of natural selection on the variability and correlation of organs, *Phil.Trans.Roy.Soc.* A, 200, pp.1-66.

PESSOA, D.G.C. e NASCIMENTO SILVA, P.L.D. (1998). *Análise de Dados Amostrais Complexos.* Caxambu: Associação Brasileira de Estatística (ABE), Mini-Curso do 13º SINAPE.

PFEFFERMANN, D. e HOLMES, D.J. (1985). Robustness Considerations in the Choice of a Method of Inference for Regression Analysis of Survey Data. *J.R.Statist. Soc. A*, 148, Part 3, pp268-278.

PFEFFERMANN, D. e NATHAN, G. (1977). *Analysis of Data From Complex Samples.* Viena: Proceedings of the 41ª Session of the ISI. XLVII, livro 3, pp. 21-42.

PFEFFERMANN, D. e NATHAN, G. (1981). Regression Analysis of Data from a Cluster Sample. *Journal of the American Statistical Association.*, Volume 76, nº 375, pp.681-689, Theory and Methods Section.

ROYALL, R. e HERSON, J. (1973). Robust estimation in finite populations. *J.Amer.Statist.Ass.*, 68, pp.880-893.

SÄRNDAL, C.E.; SWENSSON, B. e WRETMAN, J.H. (1992). *Model Assisted Survey Sampling.* Nova Iorque: Springer-Verlag.

SKINNER, C.J.; HOLT, D. e SMITH, T.M.F(1989). *Analysis of complex surveys.* Chichester: John Willey & Sons.

SMITH, T.M.F. (1981). Regression analysis for complex surveys. In Krewski, D; Platek, R.; Rao, J.N.K., *Current topics in survey sampling*. Nova Iorque: Academic Press, pp 267-292.