Working Paper nº 53

Modelos de Estimação em Contexto de Não Respostas

Osvaldo V. Caldeira

Working Paper nº 53

ISSN: 0872-895X
Depósito Legal nº: 90631/95

Osvaldo V. Caldeira
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Outubro 1996

Trabalho apresentado no âmbito do seminário de Estatística e Econometria, ano lectivo 1995/96 - coordenado pelo Professor Doutor Bento Murteira.

Modelos de Estimação em Contexto de Não Respostas

Osvaldo V. Caldeira*

28 de Fevereiro de 1996

Resumo

Neste texto apresenta-se algumas estratégias para se lidar com as não-respostas. Como veremos, o enviesamento cresce com a taxa de não-respostas. É difícil contudo, encontrar medidas objectivas do enviesamento, embora seja relativamente simples quantificar a extensão das não-respostas. Para minimizar os seus efeitos, apresentam-se estimadores que se apoiam em modelos de resposta.

Palavras-Chaves: não-resposta, medidas das não-respostas, enviesamento, amostragem em duas fases, distribuição das respostas, modelização das respostas, estimadores e variáveis auxiliares.

Abstract

We present some strategies for dealing with nonresponse. The bias increases with the rate of nonresponse. It is not easy to get measures of the bias, but it is simple to quantify the extend of the nonresponse. With the aid of response models, we present some estimators for dealing with nonresponse.

Key-Words: nonresponse, measures of nonresponse, bias, two-phase sampling, response distribution, response models, estimators and auxiliary variables.

* *Gabinete de Estudos do Instituto Nacional de Estatística e Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa. Comunicação apresentada no âmbito dos seminários organizados pelo Prof. Doutor Bento Murteira no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa.*

1. Introdução

(i) O problema das não-respostas existe em quase todos os inquéritos[1], contudo a sua extensão e os seus efeitos variam de inquérito para inquérito. Em regra, a qualidade das estimativas produzidas dependem bastante do tratamento que os seus responsáveis são capazes de implementar para minimizar os efeitos das não respostas. Tornou-se assim habitual em organismos credenciados de estatística, usar como indicador da boa realização do inquérito a taxa de respostas, o que só realça a importância do assunto. Contudo, como já referimos mesmo nos "melhores" inquéritos e não obstante todos os esforços, não é possível evitar as não-respostas, tornando-se por esse motivo numa questão incontornável dos inquéritos por amostragem. O tratamento das não-respostas visa controlar a **extensão** do fenómeno, assim como os seus **efeitos** na estimação.

(ii) Acresce a este facto que em inquéritos que se repetem no tempo (por exemplo: painel) a taxa de respostas poder variar consideravelmente ao longo do tempo, o que acrescenta à lista de problemas enunciados alguma dificuldade em comparar os resultados obtidos nas sucessivas ocasiões em que o inquérito é realizado, se não houver um tratamento adequado.

(iii) Para contornar os efeitos perniciosos das não respostas, várias técnicas têm sido desenvolvidas, que se podem classificar numa das seguintes famílias:

1- Subamostragem dos Não-Respondentes
2- Aleatorização de Respostas
3- Modelização do Mecanismo de Respostas
4- Imputação de Respostas

(iv) Ao longo deste texto, pretende-se **descrever o problema das não respostas**, fornecer medidas para **avaliar a extensão do fenómeno** e fornecer uma panorâmica das técnicas que podem ser utilizadas para se proceder à **estimação em contexto das não respostas**. Como

[1]A amostragem como processo para obter informação relativa a um universo, não é um processo exclusivo das ciências sociais e humanas onde em regra o uso do inquérito se tornou um instrumento privilegeado de observação. A amostragem também se realiza no contexto dos processos fisicos e industriais, como processo alternativo para a colheita de informação de um Universo. Mas é na verdade a possibilidade não-resposta, ou não-observação, que constitui um dos elementos distintivos da amostragem nestas ciências, uma vez que em muitos dos processos referidos (por exemplo: controle industrial da qualidade) raramente se depara com a impossibilidade de se observar a amostra gerada.

subproduto espera-se que o leitor fique na posse de um quadro, que permita avaliar conscienciosamente as situações de não-respostas e os respectivos modelos de tratamento. A modelização do mecanismo de resposta é o procedimento técnico que permite, em tempo de estimação, incorporar e minimizar as contrariedades do processo de resposta mediante o uso das potencialidades da amostragem em duas fases. Na verdade as não respostas podem ser encaradas como o remanescente de uma subamostragem. Dificilmente o problema será completamente resolvido, mas apenas consideravelmente minimizado, o que aliás é soberbamente sintetizados pela seguinte afirmação: *"...sampling practitioners do not believe that the nonresponse models on wich their adjustements are based hold exactly: they simply hope that they are improvements on the model of data missing at random over the total population implicitly assumed in estimating means with the 'do nothing' procedure."* *Kalton, G. in "Model in the Practice of Survey Sampling"-Int.Statist.Rev.51 (1983)*

2. Caracterização do Conceito de Não-Resposta

(i) Seleccionada uma amostra s, diz-se que nos encontramos perante um **problema de não-respostas**, desde que exista pelo menos um dos indivíduos da amostra[2], que não forneceu toda a informação (relativa às variáveis a observar por inquérito) que dele se pretendia.

(ii) O objectivo de um inquérito é observar um conjunto de variáveis consideradas de interesse, através de questões adequadamente formuladas para o efeito. Genericamente, pretende-se observar q variáveis, através de um inquérito, isto é:

$$\mathbf{y} = \left(y_1, \ldots, y_q\right) \qquad (2.1)$$

Os responsáveis do inquérito, solicitam a resposta às questões formuladas aos indivíduos que constituem a amostra. Se o indivíduo $k \in s$ responder a todas as perguntas, teremos um **vector de respostas completo** para esse indivíduo, isto é:

$$\mathbf{y}_k = \left(y_{1k}, \ldots, y_{qk}\right) \quad k \in s \qquad (2.2)$$

[2]Assume-se, obviamente que a base de amostragem caracteriza correctamente a população, isto é, não há erros de cobertura. Com efeito, a existência de fenómenos de natalidade e mortalidade em algumas populações, traduzem-se em erros de cobertura da base de amostragem, que levam por vezes a que na fase de recolha de informação, seja caracterizado como uma não-resposta, quando na verdade se trata de um problema de **sobre-cobertura** da base de amostragem.

(iii) Dizemos que o inquérito foi **completamente respondido** (*full response*) , se todos os inquiridos responderem a todas as questões colocadas, isto é:

$$\mathbf{y}_k = \left(y_{1k},\ldots,y_{jk},\ldots,y_{qk}\right) \quad \forall k \in s$$

y_{jk} – resposta do individuo k, à j - ésima pergunta do inquérito (2.3)

Neste caso, teremos uma **matriz de respostas de dimensão $n_S x q$ completamente preenchida**. Em todos os outros casos, estamos perante a existência de não-respostas, isto é, de casos em que a matriz de respostas de dimensão $n_S x q$ está **incompletamente preenchida**, ou dito de outro modo, com **valores em falta**[3].

(iv) As razões que dão origem às não-respostas, são muito diversas e afectam diferenciadamente os inquéritos consoante a entrevista seja conduzida directamente, por via telefónica ou por via postal:

1- Na inquirição directa ou telefónica[4], a recusa em responder a todas ou a algumas questões, a dificuldade em encontrar os inquiridos na residência, ou mesmo o desconhecimento da língua (ou, linguagem) do inquirido podem significar, à posteriori, a ausência de resposta por parte de alguns elementos da amostra.

2- Na inquirição por via postal, como é obvio, todas estas dificuldades se agravam com a fraca propensão (redução do estímulo à resposta) para responder da parte dos elementos da amostra. Em geral, em inquéritos semelhantes, a realização por via postal apresenta taxas de respostas inferiores.

---

[3]A designação de **missing values**, encontra-se bastante vulgarizada na língua inglesa. Informaticamente, os valores em falta são representados por "nulos" ou "brancos", com o significado óbvio, de valor a não processar (ou a processar com regras próprias), o que remove as dificuldades de representação matricial das respostas. De outro modo, encontrar-se-ia vedada a representação matricial, em casos de não-respostas, como facilmente se percebe. Mais exactamente, as regras de processamento dos valores em falta podem (e devem) ser fixados de acordo com princípios ou critérios de natureza estatística, o que se encontra facilitado nesta representação.

[4]A audiometria, recorre de dispositivos próprios, para o estudo de audiência de meios, contudo o suporte técnico de observação, são as "linhas de comunicação telefónica".

3- À posteriori o registo de dados revela-nos, por vezes, que por erro de digitação, por se tratar de um valor inaceitável ou ainda por ilegibilidade do valor, a resposta à questão não pode ser considerada e mais uma vez (objectivamente) nos encontramos perante uma não-resposta.

(v) Designemos o **conjunto de indivíduos que responderam a uma dada questão *j***, por

$$r_j = \left\{k : k \in s \wedge y_{jk} \text{ conhecido}\right\} \qquad (2.4)$$

o que nos permite definir no máximo q conjuntos (distintos) de respondentes (um para cada questão). Todos eles são obviamente subconjuntos da amostra s, isto é, o resultado de uma "subamostragem" em s.

(vi) A existência de não-respostas, coloca dificuldades à estimação dos parâmetros de interesse. Estas dificuldades derivam (sobretudo) de três razões essenciais:

1 - Em regra são **desconhecidos os mecanismos que dão origem às não-respostas**, ou melhor, como são gerados os conjuntos de respondentes, isto é:

$p_j\left(r_j \mid s\right)$ – probabilidade de r_j ser o conjunto dos respondentes à j - ésima questão do inquérito, para uma dada amostra s.

$p_j\left(\cdot \mid s\right)$ – lei de probabilidade desconhecida

Para contornar, as dificuldades apresentadas, são geralmente consideradas algumas hipóteses razoáveis de tratamento, a saber:

1ª Hipótese: Admite-se que o comportamento de cada indivíduo da amostra relativamente ao inquérito não **depende** dos outros indivíduos da amostra, isto é:

$$p_j(r_j|s) = p_{j1} \cdots p_{jk} \cdots p_{jK}$$

p_{jk} - probabilidade de o individuo $k \in s$ responder à j - ésima questão do inquérito (2.5)

2ª Hipótese: Admite-se que há indivíduos com **características semelhantes**, isto é, com características tais que dão origem a **comportamentos semelhantes perante um inquérito**. Esta hipótese é útil para a estimação da probabilidade de responder de cada indivíduo. Com maior generalidade, diga-se que se existe informação caracterizadora dos indivíduos então, por recurso a modelos, é possível estimar para cada um a propensão de responder ao inquérito.

2 - As inferências, isto é, a qualidade da estimação depende em grande medida, da **validade das hipóteses assumidas**. A formulação de hipóteses, é não só compreensível em face do problema, como também inevitável em inquéritos por amostragem, embora seguramente indesejável. A experiência dos responsáveis e a sua sensibilidade é fundamental, para a formulação de modelos credíveis, que a seu tempo serão testados e apoiarão a estimação.

3 - A estimação basear-se-á obviamente no conjunto dos respondentes, ou dizendo de outra maneira, a construção dos estimadores deve ter em conta eventuais cenários de não respostas (resultado de uma "subamostragem"). Neste contexto, a estimação é de uma complexidade acrescida, porque temos de nos confrontar com **um processo de amostragem, dentro do qual ocorre uma "subamostragem"**, o que a torna mais complexa, ainda que o mecanismo da "subamostragem" seja conhecido.

4 - O facto de haver usualmente **várias perguntas**, no mesmo inquérito, acresce à complexidade referida uma maior dimensão, uma vez que, o comportamento dos inquiridos (isto é, as não-respostas) podem diferenciar-se de questão para questão.

3. Extensão das Não Respostas.

(i) Em consequência das não respostas, o enviesamento aumenta (como adiante mostraremos) e uma vez que as hipóteses formuladas para o seu tratamento podem estar de algum modo

distorcidas a quantificação exacta dos seus efeitos, é em regra difícil. Sendo difícil a quantificação dos efeitos, não o é a avaliação da sua extensão, isto é, saber "quantos inquiridos" não responderam ou "quanta informação" se perdeu não é complicado. Por isso, em geral, a divulgação de estimativas é, muitas vezes, acompanhada de uma medida de não respostas para que os utilizadores possam ajuizar (subjectivamente) da credibilidade dos resultados divulgados. A importância de um indicador deste tipo reside ainda no facto de que tanto o enviesamento como a imprecisão do processo de estimação, em regra aumentarem, com a diminuição da taxa de respostas.

(ii) Neste contexto, convêm precisar, alguns conceitos relativos às não-respostas:

1 - O elemento k da amostra constitui um **não-respondente ao inquérito** (*unit nonresponse*), se se recusou responder a todas as questões do inquérito, isto é:

$$k \text{ é um não - respondente ao inquérito } \Leftrightarrow k \notin r_u$$
$$r_u = r_1 \cup \ldots \cup r_j \cup \ldots \cup r_q \text{ (unit response set)} \qquad (3.1)$$

2 - O elemento k da amostra constitui um **não respondente a questões** (*item nonresponse*), se se recusou a responder a alguma (a pelo menos uma) das questões colocadas, isto é:

$$k \text{ é um não - respondente a questões } \Leftrightarrow k \notin r_c$$
$$r_c = r_1 \cap \ldots \cap r_j \cap \ldots \cap r_q \text{ (item response set)} \qquad (3.2)$$

3 - O elemento k da amostra é um **não respondente à questão j** (*item j nonresponse*), se se recusou a responder a essa questão, isto é:

$$k \text{ é um não - respondente à questão j } \Leftrightarrow k \notin r_j$$
$$r_j \quad \text{(item } j \text{ response set)} \qquad (3.3)$$

4 - Ficam assim definidos os seguintes conjuntos notáveis:

$s - r_u$ - Conjunto de não - respondentes ao inquérito(unit nonresponse set)

$r_u - r_c$ - Conjunto de não - respondentes a questões (item nonresponse set) { dos que responderam ao inquérito }

$r_u - r_j$ - Conjunto de não - respondentes à questão j (item j nonresponse set) { dos que responderam ao inquérito }

Note-se que o conjunto de referência, é o conjunto r_u , embora nada obste a que se convencione um outro conjunto para esse efeito.

(iii) Um caso de grande valor pedagógico[5], é o caso em que os inquiridos se comportam do mesmo modo em relação a todas as questões, isto é:

$$r = r_1 = \ldots = r_j = \ldots = r_q \Rightarrow r_u - r_j = \varnothing \qquad (3.4)$$

(iv) Em face do exposto, duas medidas da capacidade de obter a informação desejada são utilizadas habitualmente, a saber:

1- A probabilidade de um indivíduo escolhido aleatoriamente[6], de entre aqueles que integram a amostra ter respondido[7] ao inquérito (realizado em condições conhecidas). Trata-se obviamente de uma probabilidade condicional, isto é, dado que o indivíduo foi seleccionado para a amostra. A sua estimação à posteriori, resulta de:

$$P(k \in r|s) = \sum_{k \in s} p_{ks} \underbrace{P(k \text{ ser selececcionado} | k \in s)}_{\frac{1}{n_s}} = \frac{1}{n_s} \sum_{k \in s} p_{ks}$$

(Teor. das Prob. Totais)

$$p_{r|s} = \frac{1}{n_s} \sum_{k \in s} y_k = \frac{n_r}{n_s} \qquad y_k = \begin{cases} 0 \text{ se não respondeu} \\ 1 \text{ se respondeu} \end{cases} \qquad (3.5)$$

$$E\{p_{r|s}|s\} = \frac{1}{n_s} \sum_{k \in s} E\{y_k|s\} = \frac{1}{n_s} \sum_{k \in s} p_{ks} = P(k \in r|s)$$

Como se percebe as características do desenho de amostragem, no espírito desta medida, não influenciam o comportamento dos inquiridos, daí que as taxas de

[5] Que será bastante usado nesta exposição, em parte, para simplificar a exposição.

[6] Sendo a selecção aleatória efectuada de modo a garantir para todos os indivíduos a equiprobabilidade, ou seja, para efeitos de controle da qualidade da inquirição (respostas), todos os indivíduos da amostra tem a mesma importância. Trata-se de estimar a probabilidade de um indivíduo escolhido "ao acaso" na amostra, vir a responder ao inquérito.

[7] A distinção entre respondentes a uma questão determinada, a todo o questionário, ou a pelo menos parte do questionário, só será considerada quando a distinção for relevante para o desenvolvimento teórico. A adaptação das formulas para casos diversos, faz-se sem qualquer dificuldade.

resposta seriam supostamente homogéneas. Em concreto, admite que a perda de informação é indiferenciada, isto é, não se relaciona com o desenho de amostragem razão que leva à consideração da amostra como grupo de referência homogéneo.

2- A probabilidade de num desenho de amostragem, um indivíduo seleccionado aleatoriamente[8] na população, pertencer ao grupo dos que não se recusam (ou não recusariam) a responder ao inquérito. De modo análogo, esta probabilidade pode ser "à posteriori" estimada, com base na teoria de amostragem, e em particular através do clássico estimador π de Horvitz-Thompson:

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \qquad E\left[\hat{t}_\pi\right] = t = \sum_k y_k \tag{3.6}$$

π_k – probabilidade de $k \in s$ pertencer à amostra

assim podemos derivar:

$$P(k \in r) = \sum_{k \in U} p_{ks} \underbrace{P(k \text{ ser seleccionado})}_{\frac{1}{N}} = \frac{1}{N} \sum_{k \in U} p_{ks}$$

(Teor. das Prob. Totais)

$$p_r = \frac{1}{N} \sum_{k \in U} \frac{y_k I_k}{\pi_k} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{N} \sum_{k \in r} \frac{1}{\pi_k} \tag{3.7}$$

$$y_k = \begin{cases} 0 \text{ se não respondeu} \\ 1 \text{ se respondeu} \end{cases}$$

[8] A selecção aleatória é efectuada admitindo probabilidades iguais, isto é, para efeitos de controle da qualidade da inquirição (desenho + respostas), todos os indivíduos da população tem a mesma importância. Trata-se de estimar a probabilidade de um indivíduo escolhido "ao acaso" na População não se recusar a responder ao inquérito.

$$E\{p_r\} = \frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}E\left\{\overbrace{\underbrace{E\{y_k|k\in s\}}_{p_{ks}} I_k}^{\text{independentes}}\right\}$$
$$= \frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}p_{ks}\underbrace{E\{I_k\}}_{\pi_k} = P(k\in r)$$

Se atendermos ao facto de que em desenhos de dimensão variável, existe um estimador alternativo (baseado no ratio) mais eficiente que o estimador π, então podemos modificar o estimador para aumentar a sua eficiência, isto é:

$$p_r = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}} = \frac{\sum_{k\in r}\frac{1}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}} \qquad (3.8)$$

De notar que, a taxa de resposta assim calculada não depende do desenho de amostragem. Note-se apenas, o facto de esta taxa assentar o seu calculo, na presunção de que o comportamento (probabilidade de responder) do inquirido não se altera de amostra para amostra (requisito de independência de p_{ks} e I_k), ou seja, que as condições de inquirição permanecem inalteradas de amostra para amostra.

3- Se houver uma medida da quantidade de informação (variável x) proporcionada por cada indivíduo, podemos de modo análogo estimar a percentagem de informação recolhida, isto é:

$$p_{xr|s} = \frac{\sum_{k\in r}x_k}{\sum_{k\in s}x_k} \qquad p_{xr} = \frac{\sum_{k\in r}\frac{x_k}{\pi_k}}{\sum_{k\in s}\frac{x_k}{\pi_k}} \qquad (3.9)$$

onde, a primeira formula representa a percentagem de informação recolhida, de entre aquela que se pretendia recolher, isto é, no contexto da amostra. A segunda formula representa a percentagem da informação caracterizadora[9] do universo recolhida, de entre a informação necessária para a caracterização adequada/desejada do Universo, isto é, o ponto de referência é o Universo.

A distinção entre medidas condicionadas e não condicionadas das respostas, esbate-se no caso de se tratarem de desenhos de amostragem com iguais probabilidades de inclusão.

(v) As medidas de não-respostas, são adequadas para caracterizar a capacidade de recolha de informação, mas dizem pouco sobre as suas causas e sobretudo acerca dos seus efeitos na estimação, que como já se referiu é um complicado problema metodológico.

4. Não-Respostas: Avaliação Empírica das Consequências

(i) Uma boa estratégia de amostragem para um dado inquérito, deve prever a existência de não respostas e conceber estratégias destinadas ao seu tratamento. Em particular, as acções a empreender podem ser de três tipos, consoante a fase em que se processa a intervenção, a saber:

1-Na fase de **planeamento**, são definidas as acções a empreender de forma a minimizar as não respostas, isto é, pretendendo reduzi-las a níveis insignificantes e/ou a minimizar o enviesamento. Além de algumas recomendações de "bom senso"[10], como a insistência junto do inquirido e o cuidado com a extensão do questionário, a **aleatorização das respostas** [Warner(1965)] é um dos processos que poderá ser utilizado para esse fim, sobretudo nos casos em que os inquiridos revelam relutância em responder às questões colocadas. Na verdade, um mecanismo aleatório cripta a resposta do inquirido de forma a impedir a sua descodificação, mas com a delicadeza necessária para não impedir a estimação.

---

[9] O valor da informação, depende também do desenho de amostragem utilizado, o que justifica a ponderação da medida de informação. O desenho de amostragem, não só determina o modo de selecção da amostra, como também **condiciona a escolha dos estimadores** adequados à caracterização do universo.

[10] Apesar disso, nem sempre óbvias, conforme se constata da prática de algumas instituições com responsabilidade nesta área.

2-Na fase de **recolha**, perante a constatação de que existem não-respostas, poder-se-á proceder a uma **subamostragem de não respondentes** [Hansen e Hurwitz (1946)] , uma vez que uma parcela da estimativa desejada continua desconhecida. Trata-se de um procedimento para controlar/dominar o processo de não -respostas que possibilita a estimação não-enviesada, recorrendo à estimação por subamostragem da parcela desconhecida por efeito das não-respostas. Apesar da elegância da técnica usada, dever-se-á alertar para o seu elevado custo de implementação, sobretudo quando se trata de inquéritos com alguma dimensão, uma vez que se exige o conhecimento da matriz de resposta completa da subamostra.

3-Na fase de **estimação**, perante a necessidade de produzir resultados, o conhecimento do **mecanismo** (aleatório) **das não-respostas** permite-nos contornar o enviesamento proveniente das não-respostas, mediante a consideração das probabilidades com que cada inquirido apresenta responde ao inquérito. A **modelização do mecanismo desconhecido de resposta** permite em função de informação auxiliar, proceder à estimação das referidas probabilidades e reduzir o enviesamento do estimador. Em certo sentido a **imputação**, é um processo semelhante, na medida em que, através do uso de informação auxiliar relevante (isto é, relacionada com a variável desconhecida e baseando nas semelhanças entre indivíduos), poder-se-á contribuir para a produção de "bons" valores a imputar (estimativas para indivíduos). É a apresentação destes mecanismos que visa a presente exposição.

(ii) As atitudes de **"bom senso"** que referimos, dirigem-se aos factores que são responsáveis pelas não-respostas, dos quais se devem referir:

1 - A selecção, formação e supervisionamento dos entrevistadores são actividades relevantes, que devem ser alvos do maior cuidado, uma vez que, em geral, se encontra na mão dos entrevistadores, a possibilidade de se conseguir a colaboração do entrevistado.

2.- As questões, isto é, a forma como elas são colocadas e a própria extensão do questionário, devem ser avaliadas de forma que o nível de não respostas, se mantenha dentro de limites aceitáveis. Paralelamente[11] a escolha do método de recolha de informação , isto é, por via postal, por telefone ou por entrevista directa, deve ser ponderada no contexto das disponibilidades orçamentais, uma vez que dela derivam diferentes taxas de respostas.

3 - A relação de confiança estabelecida entre a instituição responsável pela operação e os inquiridos é fundamental, uma vez que dela resultará nomeadamente a convicção de que não será realizado qualquer uso abusivo da informação fornecida. Se assim suceder, reduzir-se-ão as razões que muitas vezes levam os inquiridos a recusar a sua colaboração.

Não há duvidas, do ponto vista lógico, de que a forma como as instituições encaram as não-respostas, é muito importante para a qualidade da informação produzida. No nosso país, esta preocupação tem estado cada vez mais presente em algumas instituições com responsabilidades na produção estatística. Mas importa referir, que nem sempre a força da razão tem sido suficiente para a mudança de atitude de muitas instituições. Dalenius (1976), a este propósito, apresenta o caso de duas instituições responsáveis pela condução de um mesmo inquérito, na mesma região, em que uma delas teve apenas 10% de não-respostas, enquanto outra teve uma taxa de não-respostas de 50%. A diferença, resultou - segundo Dalenius - da diferença de atitude, que as duas instituições tinham perante as não-respostas.

(iii) A argumentação apresentada sugere que os efeitos das não-respostas não se limitam a uma redução da dimensão da amostra, isto é, à redução da eficiência. Pelo contrário, modificam outras propriedades dos estimadores e em especial o enviesamento. A concepção de uma estratégia de amostragem credível exige a modificação do estimador, para contornar as dificuldades apresentadas e sobretudo para o tornar mais robusto. Sugerimos que as estimativas produzidas por esses estimadores - na ignorância do mecanismo das respostas - afastam-se, em média, significativamente do parâmetro a estimar. Adiante, justificar-se-á esta ideia empírica com argumentos formais. Por agora, limitar-nos-emos a ilustrar com dois casos para reforçar a

[11] Mas tendo em consideração as características do questionário, nomeadamente a extensão.

ideia de que as não-respostas devem ser encaradas com cuidado. Uma ideia básica, para reduzir as não-respostas, é a de tentar mais do que uma insistência no contacto[12] com os inquiridos em falta, o que - em regra - contribui significativamente para a redução das não-respostas. Ao mesmo tempo, o acompanhamento das insistências permite-nos aperceber melhor os efeitos das não-respostas. Com efeito, vejamos:

1- P. Rao (1983) cita os dados de Hilgard e Payne (1944), onde se pretendia conhecer a percentagem de agregados familiares com filhos de idade inferior a 2 anos. No primeiro contacto responderam 63.5% dos inquiridos (que se encontravam na residência), dos quais 17.2% tinham crianças com idade inferior a 2 anos. Das pessoas que responderam ao segundo contacto (22.2%), só 9.5% tinham crianças nas idades referidas. Os restantes agregados responderam ao terceiro contacto, mas destes apenas 6.2% tinham crianças nas idades referidas. A explicação óbvia, parece ser a redução de mobilidade a que se encontram sujeitos os agregados com crianças com idades inferiores a 2 anos, o que facilita o contacto pelos entrevistadores no local de residência.

	Nº de Insistências			Não	
	1	2	>2	Respostas	Total
Percentagem de Respostas	63.5%	22.2%	14.3%	0%	100%
Percentagem de Famílias com crianças com menos de 2 anos	17,2%	9.5%	6.2%	-	13.9%

2- Um outro exemplo, é dado por Danermark e Swenson(1987), relativo a um inquérito realizado a estudantes com idades superiores a 16 anos em estabelecimentos de ensino, com o objectivo de saber a percentagem de estudantes que já haviam experimentado drogas leves (marijuana e haxixe). No primeiro contacto foram inquiridos 89.6% dos indivíduos da amostra dos quais apenas 4.9% tinham experimentado drogas leves. No segundo contacto foram

[12] Em Portugal vulgarizou-se, em alguns organismos com responsabilidade na produção estatistica, a designação de insistência.

inquiridos mais 8.9%, dos quais 14.9% já haviam experimentado drogas leves. A relação (e não causalidade) existente entre o consumo de drogas leves e algum absentismo escolar, será eventualmente a razão da discrepância entre os resultados provenientes dos dois contactos.

	Nº de Insistências		Não Respostas	Total
	1	>1		
Percentagem de Respostas	89.6%	8.9%	1.5%	100%
Perc. de Estudantes que já experi- mentaram Haxixe e/ou Marijuana	4.9%	14.9%	?	?

Como dissemos, as não-respostas podem subavaliar ou sobreavaliar os parâmetros, conforme foi evidenciado nos exemplos precedentes. Podia-se, sem grande dificuldade, acrescentar outros exemplos, evidenciadores das consequências das não respostas. As suas consequências não devem ser menosprezadas, pois delas resultam enviesamentos nos parâmetros e justificam pelo menos um esforço de insistência nos contactos. Em inquéritos, por via postal, a realização de duas a três insistências é mesmo uma norma aceite em inquéritos que prezam alguns requisitos de qualidade.

5. Utilização do Mecanismo de Respostas para a Estimação

Noção de Amostragem Quase-Probabilística

(i) A modelização do mecanismo das respostas, assenta na ideia de que se fosse conhecida a propensão (probabilidade) com que cada elemento da amostra responde ao inquérito, então poder-se-ia construir um estimador para os parâmetros da população, tirando partido da amostragem em duas fases.

(ii) Como atrás referimos, admitem-se duas hipóteses razoáveis, de forma a permitir a modelização do mecanismo das respostas, a saber:

1- A independência entre os indivíduos na decisão de responder ou não ao inquérito.

2- Que indivíduos com características semelhantes apresentam comportamentos semelhantes em relação ao inquérito.

(ii) Num primeiro momento, considere-se apenas a primeira hipótese e vamos supor conhecidas as probabilidades com que os indivíduos se dispõem a fornecer a informação solicitada. Neste caso, de acordo com a teoria da amostragem em duas fases[13], teremos como estimador condicional do total π^* a substituir o estimador π de Horwitz-Thompson:

$$\hat{t}_{\pi^*} = \sum_{k \in r} \frac{y_k}{\pi_k \theta_k} \quad \text{a substituir } \hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \text{ com } E\left[\hat{t}_\pi\right] = t \qquad (5.1)$$

$P(k \in s) = \pi_k$ - probabilidade do indivíduo k da população pertencer à amostra .

$P(k \in r|s) = \theta_k$ - probabilidade do indivíduo k da amostra s responder ao inquérito.

donde:

$$E_{RD}\left[\hat{t}_{\pi^*}|s\right] = \hat{t}_\pi \Rightarrow E\left[\hat{t}_{\pi^*}\right] = t \qquad (5.2)$$

(iii) O facto de a resposta (inclusão dos indivíduos na segunda fase) ser independente de indivíduo para indivíduo, aponta para uma amostragem de Poisson na segunda fase. Neste caso, como sabemos, existe um estimador mais eficiente não-informativo do que o estimador π, conhecido como o estimador alternativo, inspirado no estimador do ratio e aproximadamente não-enviesado, cuja expressão é:

$$\hat{t}_1 = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k \theta_k}}{\sum_{k \in r} \frac{1}{\pi_k \theta_k}} = N \bar{y} \qquad (5.3)$$

(iv) O caso mais simples, acontece, se todos os indivíduos da amostra apresentam a mesma probabilidade de responder. Neste caso o desenho de Poisson, reduz-se a um desenho de Bernoulli e então temos:

[13] O uso desta notação, neste contexto pretende reforçar a ideia de que as probabilidades de inclusão da 2ª fase, são um parâmetro desconhecido θ a estimar. Não se considerará, por razões de oportunidade, as probabilidades de inclusão de 2ª ordem, assim como a estimação da variância a elas associada. A sigla RD, é usada como abreviatura de Response Distribution.

$$\theta_k = \theta \Rightarrow \hat{t}_1 = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k \theta}}{\sum_{k \in r} \frac{1}{\pi_k \theta}} = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k}}{\sum_{k \in r} \frac{1}{\pi_k}} \tag{5.4}$$

Neste caso, admite-se implicitamente que o mecanismo de resposta não influência a estimação, como se percebe da leitura do estimador. O único efeito situar-se-á a nível da variância, que aumenta com a redução da dimensão da amostra. Formalmente, tudo está correcto, **mas convenhamos que o modelo apresenta uma fraca adesão à realidade**.

(v) No caso do modelo de Bernoulli coincidir com o verdadeiro mecanismo de respostas, não haverá inconveniente em o utilizar. O enviesamento do estimador resultará apenas da sua inspiração no ratio e que, como se sabe, em amostras com alguma dimensão é aproximadamente nulo. Os problemas põem-se naturalmente, quando o modelo se afasta da realidade. Dois modelos podem ser então, usados para análise:

1- O verdadeiro modelo é de Poisson. Neste caso, o enviesamento será:

$$B\left[\hat{t}_1\right] = E\left[\hat{t}_1\right] - t \cong N \frac{\sum_k y_k \theta_k}{\sum_k \theta_k} - t = \frac{N \overline{y\theta}_U}{\bar{\theta}_U} - \frac{N \bar{y}_U \bar{\theta}_U}{\bar{\theta}_U}$$

$$\text{Como} \quad S_{y\theta U} = \frac{\sum_k \left(y_k - \bar{y}_U\right)\left(\theta_k - \bar{\theta}_U\right)}{N-1} \qquad R_{y\theta U} = \frac{S_{y\theta U}}{S_{yU} S_{\theta U}}$$

$$B\left[\hat{t}_1\right] = (N-1)\frac{S_{y\theta U}}{\bar{\theta}_U} = \frac{t(N-1) R_{y\theta U} cv_{yU} cv_{\theta U}}{N}$$

donde:

$$RB\left[\hat{t}_1\right] = \frac{B\left[\hat{t}_1\right]}{t} \cong R_{y\theta U} cv_{yU} cv_{\theta U}$$

isto é

$RB\left[\hat{t}_1\right]$ depende de $R_{y\theta U}$

O que mostra que o problema do enviesamento é tanto mais importante, quanto maior for a **relação entre os valores da variável e a probabilidade de responder**.

2- O mecanismo de resposta é deterministico, isto é, a população é estratificada em dois grupos, um que responde sempre (com probabilidade 1) e outro que nunca responde (com probabilidade 0). O enviesamento, será então:

U_1, U_2 - Estractos de respondentes e não - respondentes com dimensões N_1, N_2 respectivamente

$$B\left[\hat{t}_1\right] \cong N_2(\bar{y}_{U_1} - \bar{y}_{U_2})$$

O que mostra, neste caso, que o enviesamento aumenta com a **diferença das médias dos estratos** e com a **dimensão do estrato dos não-respondentes**. Este resultado, alerta-nos para a necessidade de conceber estratégias, que reduzam o peso do estrato de não-respondentes.

Se na teoria base da amostragem, a ênfase é colocada na redução da dispersão, por motivos óbvios a presença das não-respostas, **recoloca o problema do enviesamento**. A ideia é obviamente, de procurar conhecer o mecanismo de respostas e em face do conhecimento deste mecanismo eliminar ou pelo menos reduzir o enviesamento.

(vi) Durante muitos anos, o **modelo deterministico** foi a base para o tratamento das não respostas, seguindo uma sugestão de Cochran(1977). Assente neste modelo, nas décadas de 60 e 70, muitos trabalhos de tratamento de não respostas foram desenvolvidos nos Institutos de Estatística. Mas como confessava Cochran, a ideia baseava-se numa supersimplificação da realidade. Ainda é esta ideia, que remotamente inspira, o processo de subamostragem dos não-respondentes. O inconveniente deste processo - como referimos - é o seu custo e a sua difícil aplicação a amostragens de grande dimensão. São estes inconvenientes, que nos remetem para a pesquisa de outras soluções. Da sua exploração, ficam duas ideias importantes para a redução do enviesamento, a **necessidade de reduzir a dimensão do estrato das não respostas** e que o **enviesamento é proporcional à diferença das médias dos estratos**[14].

[14]Se existir uma associação (correlação) entre o valor da variável de interesse e a resposta (ou não) ao inquérito, então o modelo deterministico, permite dizer que a estimação efectuada através dos respondentes é enviesada. Esta constatação, é retomada no modelo aleatório.

(vii) Uma representação mais fiel da realidade, assenta na hipótese de que o processo de respostas é aleatório e não deterministico. Como mostramos, a ideia de que todos indivíduos possuem o mesmo mecanismo aleatório de resposta, conduz-nos à amostragem de Bernoulli na segunda fase. É a versão mais simples do modelo aleatório. Neste caso a estimação, como vimos, considera que das não-respostas apenas resulta uma redução da dimensão da amostra. Em termos práticos, equivale isto a dizer, que as não-respostas mantém inalteradas as propriedades do não-enviesamento, o que em geral, não é aceitável. Sociologicamente, corresponde a assumir que não existem características individuais (ou diferenças de comportamento), que aumentem ou reduzam a propensão a responder e sobretudo que essa propensão não se encontra correlacionada com a variável de interesse. Da exploração efectuada, há que reter a ideia de que o enviesamento resulta, da associação (correlação) que pode existir entre o mecanismo de respostas e os valores da variável, o que pode comprometer os valores observados, sobrevalorizando o perfil dos respondentes.

(viii) Em qualquer dos casos considerados, o verdadeiro mecanismo de respostas é na prática desconhecido. As alternativas, são "impor" um dos modelos sugeridos ou em face da informação disponível, estimar as probabilidades de respostas. A estimação parece a solução mais sensata e que mais se aproximará do verdadeiro modelo. É esta a abordagem, que adiante desenvolveremos. Importa no entanto reconhecer, que o uso desta abordagem afasta-nos dos conceitos da amostragem probabilistica, uma vez que se desconhece o desenho de amostragem que dá origem à "amostra observada". Autores existem que sugerem a designação de **"amostragem quase-probabilistica"**[15], para os métodos em que o desenho de subamostragem só se torna conhecido por recurso à estimação.

(ix) Ao contrário do que à primeira vista se poderia julgar, como já se percebe, **a estimação no contexto das não-respostas não desvaloriza o papel da teoria da amostragem, mas antes exige a concepção de modelos robustos e resistentes**. O seu desenvolvimento assenta obviamente, na amostragem em duas fases. Em todo o caso, o problema do enviesamento[16]

---

[15] Os trabalhos de Oh e Scheuren(1983) e Plateck e Gray (1983) seguem esta abordagem, não apenas para a resolução do problema das não-respostas, mas visando também a resolução de outros erros não-amostrais.

[16] É o que refere, Kalton(1983) e Kalton e Kasprzyk(1986), ao considerarem que o conhecimento do verdadeiro mecanismo de resposta depende não só da validade das hipóteses formuladas, como também dos procedimentos de estimação.

nunca se encontrará completamente resolvido, mas apenas substancialmente reduzido, isto é, quando muito minimizado, face à informação disponível.

(x) Geralmente, mas não necessariamente, a estimação do mecanismo de resposta é precedido da imputação de "bons valores" para os indivíduos que responderam apenas parcialmente ao inquérito, isto é, apenas a algumas questões do inquérito. Nestes casos, a imputação será obviamente uma estratégia a considerar.

6. Modelização do Mecanismo de Respostas

Uma Estratégia Realizável de Estimação: Grupos de Resposta Homogéneo

(i) A sofisticação estatística, nem sempre é compatível, com a necessidade de produzir estimativas rápidas, o que aconselha em muitos casos o uso técnicas robustas mas simultaneamente simples de estimação. Neste sentido, a estratificação dos elementos da amostra, em grupos que apresentem comportamentos homogéneos de resposta, torna-se um modelo simples, mas poderoso de estimação. O modelo RHG (response homogeneity group) é caracterizado do seguinte modo:

1. A amostra é repartida em H_s grupos distintos, com comportamentos hómogeneos de resposta:

$$s = s_1 \cup \ldots \cup s_h \cup \ldots s_{H_s} \quad s_h \cap s_{h'} = \varnothing \; \forall h \neq h'$$

2. As probabilidades de inclusão são uniformes em cada grupo, isto é:

$$\pi_{k|s} = P(k \in r|s) = \theta_{hs} \quad \forall k \in s_h$$

Supõem-se agora - como se percebe - que quem respondeu ao inquérito, o fez a todas as questões do inquérito. Esta suposição, não retira a generalidade ao modelo e é geralmente contornada como referimos através da imputação. O objectivo é sobretudo de se reduzir a complexidade da estimação.

(ii) Deve-se notar, em primeiro lugar que para efeitos de estratificação da amostra, se usa não apenas a informação prévia à amostragem, mas também a informação relativa à amostra concreta (da 1ª fase) na qual se inclui a informação relativa às condições em se processou a

recolha de informação, e que é susceptível de ter influenciado o mecanismo de resposta. Em particular, diga-se que:

1. A formação de grupos, pode variar de amostra para amostra[17], de acordo com a informação disponível para a caracterização do mecanismo de resposta.

2. A estratificação, pode (e deve) levar em conta os meios e as dificuldades ocorridas no "trabalho de campo", como o número de entrevistadores e/ou o seu perfil, que se traduzem no processo de resposta, ou seja, as condições de inquirição.

O sucesso do modelo, assenta na capacidade e astúcia do estaticista, em formar grupos com mecanismos de resposta homogéneos. Assim a identificação dos factores que condicionam as respostas, desempenha um papel determinante, nas estratégias de amostragem quase-probabilistica.

(iii) No caso de o modelo ser verdadeiro, as probabilidades de inclusão, poderão ser estimadas, com base nas frequências observadas em cada grupo homogéneo. Neste caso, sob a hipótese de validade do modelo, há a considerar:

1. Em cada um dos H_s grupos distintos, foram solicitadas respostas a:

$$|s_h| = n_h \text{ inquiridos e houve } |r_h| = m_h \text{ respondentes} \qquad (6.3)$$
$$\mathbf{m} = \left(m_1,\ldots,m_h,\ldots,m_{H_s}\right)$$

2. As probabilidades de inclusão são estimadas de acordo com o modelo e com as frequências observadas:

$$\pi_{k|s,m} = P(k \in r|s,\mathbf{m}) = \frac{m_h}{n_h} = f_h \qquad (6.4)$$
$$\forall k \in s_h \text{ pois } E_{ARD}(m_h) = n_h\theta_{hs} \qquad n_h \geq 1$$

[17] Pode variar de amostra para amostra, mas não numa mesma amostra sob condições dadas de inquirição, isto é, as condições de inquirição ficam determinadas antes do processo de resposta.

As probabilidades de inclusão da 2ª fase são agora condicionadas à validade do modelo, isto é, à distribuição das respostas assumida (*ARD*). Este modelo foi detalhadamente estudado por Särndal e Swenson (1987). A sua utilização pode ser feita, com ou sem o uso de variáveis auxiliares na fase de estimação, como adiante veremos.

(iv) Um modelo, com inspiração algo semelhante, deve-se a Oh e Scheuren (1983) conhecido, como Modelo de Subpopulações com Mecanismos de Resposta Uniforme. A grande diferença em relação ao modelo preconizado, reside no facto de **se realizar a estratificação da amostra, num momento prévio à amostragem**, isto é, sem atender à amostra particular e às condições de realização do inquérito.

(v) Autores de inspiração Bayesiana[18], efectuam a distinção entre **mecanismo de resposta ignorável e não-ignorável**, pretendendo com isso notar que se o mecanismo de resposta é não-correlacionado com a variável de interesse, então o conhecimento do mecanismo é irrelevante para a estimação. Não se pondo em causa, a ideia atrás referida, que o mecanismo de resposta só é relevante quando correlacionado com a variável de interesse, julga-se problemático assumir "à priori" a sua irrelevância, isto é, sem atender a informação que o confirme.

(vi) Estes resultados (6.4 e 5..3) , permitem construir os estimadores condicionais não enviesados[19] e determinar as suas propriedades, usando a técnica de amostragem estratificada de Bernoulli na 2ªfase , a saber:

$$\hat{t}_{c\pi^*} = \sum_{h=1}^{H_s} \sum_{k \in r} \frac{\overset{v}{y}_k}{\pi_{k|s,m}} = \sum_{h=1}^{H_s} f_h^{-1} \sum_{k \in r} \overset{v}{y}_k \qquad (6.5)$$

(vii) Importa notar, que tudo assenta na hipótese de que o modelo assumido é coincidente com o verdadeiro mecanismo de resposta, caso em que, o estimador[20] é não enviesado, pois:

$$\begin{aligned} E\left[\hat{t}_{c\pi^*}|s\right] &= E_m E_{RD}\left[\hat{t}_{c\pi^*}|s,m\right] = E_m E_{RD}\left[\sum_{h=1}^{H_s} f_h^{-1} \sum_{k \in r} \overset{v}{y}_k |s,m\right] = \\ &= E_m\left[\sum_{h=1}^{H_s} \sum_{k \in s} \overset{v}{y}_k |s\right] = \hat{t}_\pi \end{aligned} \qquad (6.6)$$

[18] É o caso dos estimadores apresentados por Rubin (1976, 1983 e 1987) e Little e Rubin (1987).

[19] Se o modelo assumido for verdadeiro.

[20] Refira-se, que neste caso a notação é pouco clara e confunde o estimador e a estimativa.

Que tem o efeito de corrigir, a dimensão esperada da amostra através da dimensão da amostra observada.

(viii) Este estimador quando aplicado a desenhos usuais - para efeitos de computação - apresenta expressões menos elegantes, mas que por serem o resultado de um trabalho adequado, reduzem a complexidade computacional. Por exemplo, no caso do desenho SI (amostragem aleatória simples sem reposição na 1ª fase), temos:

$$\hat{t}_{c\pi} = N\left[\frac{\sum_{h=1}^{H_s} n_h \bar{y}_{r_h}}{\underbrace{n}_{\bar{y}_U}}\right] = N\hat{\bar{y}}_U \qquad (6.7)$$

(ix) Uma recomendação efectuada, sempre que se trabalha em contexto das não respostas, é a de se usar sempre que possível estimadores informativos, como o da diferença ou o da regressão. A grande vantagem de uma estratégia assim definida é de aumentar não só a eficiência, como também a resistência a percalços de amostragem, mesmo que o modelo assumido não seja verdadeiro. Tal resulta do facto de nesses casos ser possível a produção de previsores, para cada elemento da amostra e/ou população, que contribuem para a redução da incerteza. Estas ideias foram exploradas por Särndal e Swenson (1987) e Bethlehem(1988), que sugerem o uso de informação auxiliar relevante quando ela existe. Os estimadores informativos sugeridos, deverão ser modificados de forma a contemplarem a modelização das não-respostas. Os estimadores informativos[ver Särndal (1992 - Cap. 9)], terão então a seguinte expressão[21]:

$$\hat{t}_{c\pi} = \overbrace{\sum_{k\in U}\hat{y}_{1k}}^{\text{pré-estimativa}} + \sum_{h=1}^{H_s}\left[\overbrace{\sum_{k\in s_h}\frac{y_k-\hat{y}_{1k}}{\pi_k}}^{\text{erro de pré-estimativa}} + \overbrace{f_h^{-1}\sum_{k\in r_h}\frac{y_k-\hat{y}_k}{\pi_k}}^{\text{erro de não-resposta}}\right]$$

com $\hat{y}_k$ e $\hat{y}_{1k}$, referidos à informação auxiliar da amostra e do Universo respectivamente

$\hat{y}_k = f_s(\mathbf{x}'_k)$ $\quad k \in s$ - forte previsor de y_k

$\mathbf{x}'_k$ - vector com n variáveis auxiliares

$\hat{y}_{1k} = f_U(\mathbf{x}'_{1k})$ $\quad k \in U$ - fraco previsor de y_k

$\mathbf{x}'_{1k}$ - vector com n_1 ($\leq$ n) variáveis auxiliares

[21] Substituindo $\pi_{k|s}$ pelas estimativas obtidas, através da formação de grupos homogéneos de resposta.

que apresenta duas formas particulares, para os casos em que existe apenas informação auxiliar recolhida previamente à amostragem (isto é, relativa ao Universo) ou apenas informação auxiliar recolhida após a amostragem (isto é, relativa à Amostra):

Se apenas existir informação auxiliar a nível da amostra:

$$\hat{t}_{c\pi^*} = \sum_{h=1}^{H_s}\left[\sum_{k\in s_h}\frac{\hat{y}_k}{\pi_k} + f_h^{-1}\sum_{k\in r_h}\frac{y_k - \hat{y}_k}{\pi_k}\right]$$

Se apenas existir informação auxiliar a nível da população:

$$\hat{t}_{c\pi^*} = \sum_{k\in U}\hat{y}_{1k} + \sum_{h=1}^{H_s}\left[f_h^{-1}\sum_{k\in r_h}\frac{y_k - \hat{y}_{1k}}{\pi_k}\right]$$

7. Modelização do Mecanismo de Respostas

Avaliação através da Simulação de Monte Carlo

(i) É importante voltar a referir, que se está a trabalhar com um modelo de respostas assumido que este poderá não coincidir com o verdadeiro mecanismo de resposta, o que na prática, significa que a modelização do mecanismo de resposta não elimina completamente o enviesamento. Estudos de simulação, usando o método de Monte Carlo, foram realizados por Särndal e Swenson (1987) e mostraram que o uso de mecanismos de resposta "não-exactamente" identificados, eram preferíveis a nenhuma modelização, quer em termos de enviesamento, quer ainda em termos da probabilidade de cobertura[22]. Por outro lado, mostraram ainda que as qualidades de não enviesamento e a probabilidade de cobertura melhoram se na fase de estimação se usar um estimador informativo.

(ii) A abordagem sugerida, estima as probabilidades de resposta admitindo homogeneidade de comportamentos no interior de cada grupo. Supõe-se, é claro que os grupos formados tem dimensões, que tornam possível a sua estimação. Uma diferente abordagem, consiste na estimação das probabilidades individuais de resposta (desiguais), com base numa abordagem paramétrica (Cassel, Särndal e Wretman 1983) ou não-paramétrica (Giommi 1988), usando informação auxiliar, correlacionada com a variável de interesse. A abordagem de Greenlees,

[22] Admitindo que o modelo possui alguma verdade.

Reece e Zieshang (1982) não só usa a variável de auxiliar, como também a própria variável de interesse. Seguindo uma sugestão de Gorieroux(1989) - embora com uma metodologia diferente Caldeira (1995) propôs o uso dos modelos GLM como forma de unificar as diferentes abordagens apresentadas.

(iii) Para realizar a simulação de Monte Carlo, Särndal e Swensson consideraram uma população de N=1227 famílias suecas, das quais havia informação relativa ao rendimento disponível (y) e ao seu rendimento fiscal (x) as quais apresentam um coeficiente de correlação de 83%. De acordo com o rendimento fiscal foram repartidos por 4 estratos com diferentes probabilidades de resposta. Em cada estrato os seus elementos tinham uma probabilidade de resposta (θ_h) uniforme (hipótese de trabalho para a simulação). A probabilidade de resposta apresentava, de acordo com o modelo, um coeficiente de correlação de 44% com o rendimento disponível. Apesar de baixa, esta correlação tem efeitos graves no enviesamento conforme adiante mostraremos. As características da população de referência são sintetizada no seguinte quadro

Estratos	N_h	θ_h	$\bar{y}_h$	$\bar{x}_h$
U_1	373	0.45	4.18	1.68
U_2	303	0.65	5.65	2.25
U_3	280	0.75	6.73	2.54
U_4	271	0.90	8.31	2.79

(iv) Foi gerada uma amostra aleatória simples (SI) de 400 elementos da população das 1227 famílias. Esta amostra de 400 elementos, simulava responder ao inquérito aleatoriamente, de acordo com o modelo atrás descrito. Foram gerados assim 1000 conjuntos de respondentes, cada um deles a suportar uma estimativa, para um dado estimador.

(v) Foram considerados, para efeitos de estimação, três hipóteses de modelização do comportamento dos respondentes, saber:

1. Conhece-se o comportamento dos inquiridos, com diferenciação perfeita dos 4 estratos atrás referidos. Hipótese desejável mas irrealista, adiante designado como Modelo Perfeito.

2. Conhece-se estatisticamente o comportamento dos inquiridos, com diferenciação imperfeita em 2 estratos, $U_1 + U_2$ e $U_3 + U_4$ [23]. Hipótese mais realista, adiante designado como Modelo Imperfeito.

3. Desconhece-se o comportamento dos inquiridos e não se realiza esforço para a percepção de comportamentos diferenciados, isto é, supõem-se identicos. Hipótese pouco realista, adiante designado como Modelo Ausente.

(iv) Foram ainda considerados três estimadores $\hat{t}_A$, $\hat{t}_B$ e $\hat{t}_C$, o primeiro dos quais não usa informação auxiliar e os outros usam-na com base nos modelos de regressão.

(v) Com base nos estimadores apresentados e nas várias hipóteses de modelização do mecanismo de resposta, procedeu-se à apreciação estatística do enviesamento relativo, que é nulo se o estimador for não-enviesado.

Enviesamento Relativo Estimado (%): $B\left(\hat{t}\right) / t$

Modelo	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$
Perfeito	0%	0%	0%
Imperfeito	1.5%	0.3%	0.5%
Ausente	6.5%	1.7%	1.6%

(vi) Procedeu-se ainda à apreciação da taxa de cobertura dos Intervalos de Confiança Nominais a 95% , conforme consta do quadro. Esperava-se que 95% dos intervalos aleatórios estimados, por amostragem, contivessem o parametro a estimar. Contudo a simulação reservou-nos algumas surpresas.

Taxa de Cobertura do Intervalo de Confiança

Modelo	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$
Perfeito	95.2%	95.5%	95.6%
Imperfeito	93.3%	95.6%	95.6%
Ausente	46.3%	92.6%	93.9%

[23] O sinal + tem o significado de reunião na Algebra de Conjuntos

(vi) A **leitura dos quadros em coluna** revela que a modelização do mecanismo de resposta, reduz o enviesamento e aumenta a taxa de cobertura dos Intervalos de Confiança Nominais. Quanto mais perfeita for a modelização do mecanismo de resposta, maior será a qualidade das estimativas produzidas, como aliás era esperado. Mas o que é importante, uma vez que a modelização perfeita não é possível no tratamento realista das não-respostas, é o facto de **que a modelização mesmo imperfeita trás importantes ganhos de qualidade na estimação**, qualquer que seja o estimador usado.

(vii) A **leitura dos quadro em linha** evidência uma redução do enviesamento e um aumento da taxa de cobertura se se fizer uso da informação auxiliar, no caso através do estimador da regressão. Em resultado, se **o uso de informação auxiliar relevante contribui para tornar robusta a estratégia de estimação** é obvio que melhor solução será obtida **combinando a modelização do mecanismo de resposta com o uso de informação auxiliar relevante**.

Bibliografia

Bethlehem, D.R. 1988. Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics 4.*

Caldeira, O. 1995. Potencialidades da amostragem em duas fases com aplicação ao desenvolvimento de modelos de tratamento de não respostas Tese de Mestrado. *UNL-ISEGI:Lisboa.*

Cochran, W.G. 1977. Sampling Techniques. *New York: Wiley*

Dalenius, T. 1976. Bortfallsproblemet vid statistiska undersökningar. *Marknadsvetande 4.*

Danermark, B. e Swensson, B. 1987. Measuring drug use among Swedish adolescents: randomized response versus anonymous questionnaires. *Journal of Official Statistics 3.*

Giommi, A. 1988. Nonparametric methods for estimating individual response. *Survey Methodology 13.*

Gorieroux, C. 1989. Econometrie des variables qualitatives. *Paris: Economica.*

Greenlees, J.S., Reece, W.S. e Ziesh, K.D. 1982. Inputation of missing values when the probabities of response depends on the variable being imputed. *Journ. of Amer. Statist. Association 77.*

Hansen, M.H. e Hurwitz, W.N. 1946. The problem of non-response in sample surveys. *Journal of American Statistical Association 41.*

Hilgard, E.R. e Payne, S.L. 1944. Those not at home: riddle for pollsters. *Public Opinion Quaterly 8.*

Horwitz, D.G. e Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association 47.*

Kalton, G. 1983. Models in practice of survey sampling. *International Statistical Review 51.*

Kalton, G. e Kasprzyk, D. 1986. The treatment of missing survey data. *Survey Methodology 12.*

Little, R.J.A. Rubin, D.B. 1987. Statistical analysis with missing data. *New York:Wiley*

Oh, H.L. Scheuren, F.J. 1983. Weighting adjustment for unit nonresponse (*vol 2*) Panel on Incomplete Data in Sample Survey (*vol 1*) *Incomplete Data in Sample Surveys, Vol. 1e 2. New York: Academic Press*

Plateck,R. e Gray, G.B.	1983.	Inputation methodology. *Incomplete Data in Sample Surveys, Vol. 2. New York: Academic Press*
Rao, P.S.R.S.	1983.	Callbacks, follow-ups, and repeated telephone calls *Incomplete Data in Sample Surveys, Vol. 2. New York: Academic Press*
Rubin, D.B.	1976.	Inference and Missing Data. *Biometrika, 63.*
Rubin, D.B:	1983.	Conceptual issues in the presence of non-response. *Incomplete Data in Sample Surveys, Vol. 2. New York: Academic Press*
Rubin, D.B.	1987.	Multiple inputation for nonresponse in surveys. *New York: Wiley*
Sarndal, C.E. e Swenson, B.	1987.	A general view of estimation for two phases of selection with applications to two phase sampling and nonresponse. *International Statistical Review 55.*
Cassel, C.M., Sarndal, C.E. e Wretman, J.H.	1983.	Some uses of statistical models in connection with the nonresponse problem. *Incomplete Data in Sample Surveys, Vol. 3. New York: Academic Press*
Sarndal, C.E. , Swenson, B. e Wretman, J.	1992.	Model assisted survey sampling. *New York: Springer-Verlag.*
Warner, S.L.	1965.	Randomized response: a survey technique for eliminating evasive answer bias. *Journal of American Statistical Association, 57.*