UNIVERSIDADE EDUARDO MONDLANE

Faculdade de Ciências

Departamento de Matemática e Informática

Trabalho de Licenciatura em Estatística

Tratamento de Valores Omissos Num Conjunto de Dados

Autora: Fátima Bin A.A. Wilson

Maputo, Julho de 2010

UNIVERSIDADE EDUARDO MONDLANE

Faculdade de Ciências

Departamento de Matemática e Informática

Trabalho de Licenciatura em Estatística

Tratamento de Valores Omissos Num Conjunto de Dados

Autora: Fátima Bin A.A. Wilson

Supervisor: Dr. Alberto Mulenga

Maputo, Julho de 2010

DECLARAÇÃO DE HONRA

Declaro que este trabalho é fruto da minha própria investigação que não foi submetido para outro grau que não seja, **Licenciatura em Estatística** da Universidade Eduardo Mondlane**.**

Maputo, Julho de 2010

_____

(Fátima Bin Aly Amade Wilson)

DEDICATÓRIA

À toda família Wilson, em especial aos meus pais (em memória).

AGRADECIMENTOS

Em primeiro lugar agradecer à Deus, omnipotente e misericordioso, que tem me abençoado , iluminado, e me dado saúde e disposição durante esta longa caminhada.

Aos meus pais em memória, Aly Amade Wilson e Ancha Issufo Wilson, pelos ensinamentos que me transmitiram em vida.

A toda minha família em especial minha irmã, Mariamo Wilson que incansavelmente me apoiou nesta luta que finda com sucessos e meus primos.

Ao Dr. Mulenga, meu orientador, pela paciência, entrega, compreensão, prontidão e pelas sugestões e críticas valiosas.

A todos professores do curso de estatística em particular a Dra Rafica pelo profissionalismo e pelas contribuições durante os 4 anos de formação.

A todos meus amigos em especial, Rahma Issufo, Dalila Arune, Abiba Abdulcadre, Neide Botelho, Nancy, Benedita Alves e a sua família.

A todos colegas do DMI em especial do curso de estatística, Elisa, Luisa, Pangara, Calton, Pirolita, Langa, Eurico, Higino, Macuácua, Emília, Bata, Veríssimo.

A dra Felizarda pela disponibilidade no acesso aos dados, fornecidos pelas águas de Moçambique.

Aos funcionários do DMI em especial o Sr Eurico.

E a todos que directa ou indirectamente contribuíram para a concretização do presente trabalho.

Resumo

Qualquer estudo feito por amostragem está sujeito a vários erros. Os erros provenientes do próprio mecanismo de aleatorização utilizado, ditos erros de amostragem, sendo amostra apenas uma parte da população e os resultados obtidos destas amostras são apenas estimativas dos parâmetros da população, existirão provavelmente diferenças entre os dois valores e os erros de não – amostragem, que incluem todos restantes erros que podem surgir em qualquer dos passos do processo de estudo como por exemplo na definição do objectivo do estudo, no desenho do questionário ou mesmo por recusas ou não respostas, etc.

O presente trabalho tem como objectivo ilustrar o uso adequado de técnicas estatísticas para tratamento de informação com dados omissos, e para o efeito foram usadas duas bases de dados cujos tamanhos das amostras foram obtidos em função da informação disponível na base de dados.

Para o tratamento dos valores omissos foram usados os métodos de Imputação múltipla, Expectativa máxima e vários testes em particular o teste MCAR de Little para identificar o mecanismo de aleatoriedade da omissão de dados.

Os resultados mostram que o erro padrão obtido da média da amostra jornal depois da imputação é menor que o erro padrão da média antes da imputação dos valores omissos indicando assim que a amostra depois da imputação é mais representativa da população que provém. Os resultados mostram também que o modelo obtido da amostra agregado familiar classifica apenas 153 casos no universo de 613 observações enquanto que depois da imputação classifica todas observações. O grau de ajustamento do modelo depois da imputação é menor que o grau de ajustamento antes da imputação.

Concluindo-se que os valores omissos não só têm um impacto acentuado nos resultados, mas também pelo seu impacto prático na diminuição do tamanho da amostra disponível para análise, e quando atitudes correctivas não são aplicadas os resultados obtidos podem levar a uma conclusão errada sobre o comportamento do fenómeno na população.

Palavras Chaves: Omissão, Imputação, Aleatoriedade.

ÍNDICE

Lista de tabelas

Lista de figuras

I INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Na condução de um estudo é fundamental que se antecipem todos passos, da sua interdependência e possibilidade de em todos ocorrerem erros. A percepção e o controle desses erros é uma das tarefas mais árduas do técnico de investigação (Reis e Moreira 1992), os erros dividem-se em dois grandes grupos: erros de amostragem e erros de não - amostragem.

Os **erros de amostragem** resultam do próprio significado da amostra. Sendo amostra apenas uma parte da população e os resultados obtidos são apenas estimativas dos verdadeiros parâmetros da população, existirão provavelmente diferenças entre os dois valores. Porém essas diferenças são controláveis através da escolha de um processo de amostragem aleatório e do aumento da dimensão da amostra de tal modo que ela seja representativa da população em estudo e os resultados estejam associados a um grau de confiança elevado.

Segundo Reis e Moreira (1992) Os erros **não - amostrais** incluem todos os restantes erros que podem surgir em qualquer dos passos do processo de estudo e são bem mais difíceis de controlar, alguns dos erros que compõem este grupo são erro da definição dos objectivos do estudo, erro da definição do grupo alvo, número de recusas ou não - respostas, questionário incompleto ou inadequado, erros de prognóstico, influência do entrevistador, erros de processamento, erros de interpretação.

Estes erros podem ser agrupados de modo sistemático em: Recusas e não - respostas, erros de medida, erros introduzidos na introdução e processamento dos dados e na fase posterior de análise.

Define-se erros de **não - resposta** como a impossibilidade de medir as características de alguns dos casos ou indivíduos incluídos numa amostra (Reis e Moreira, 1992).

O problema de não - respostas divide-se em dois tipos: erros de não -respostas Globais e erros de não - respostas Parciais.

Designa-se Não - respostas Parciais - quando a matriz da amostra está incompleta, porque os campos referentes a algumas das variáveis, para alguns dos elementos da amostra não estão preenchidos enquanto que define-se não - respostas Globais quando a matriz – amostra de dimensão $\mathbf{n_r * q_r}$, onde n_r é o número de elementos da amostra para os quais foi observado pelo menos uma das q variáveis, existindo assim $n-n_r$ não -respostas Globais.

O presente trabalho toma como objecto de estudo não - respostas parciais, i.é, valores omissos (missing values) num conjunto de dados uma vez que este tipo de erros tem se verificado com frequência em muitos levantamentos de dados.

1.2 DEFINIÇÃO DO PROBLEMA

De acordo com Laaksonen (1996) a informação e a natureza de erros de não - respostas , em alguns Países, é parte integral da qualidade do sistema de levantamento de dados de instituições.

Segundo Reis e Moreira (1992) erros de não - resposta podem enviesar os resultados para dimensões e direcções muitas vezes desconhecidas, devido aos seus efeitos difíceis de controlar, são considerados como parte de erros mais importantes, sendo assim levantam-se as seguintes questões neste trabalho:

- Como evitar problemas de não –respostas no levantamento de dados?
- Como tratar problemas de não - respostas na análise dos dados ?
- Qual é o impacto que os erros de não - respostas têm sobre os resultados do estudo?

Nesta perspectiva, erros de não - respostas sendo erros de correcção difícil, constituem uma ameaça na qualidade de qualquer investigação, consequentemente na tomada de decisões, porque variáveis incluídas na análise são influenciadas pelos dados perdidos, dai a necessidade de entender e identificar a causa dos dados perdidos afim de tomar medidas correctivas apropriadas.

1.3 OBJECTIVOS DE ESTUDO

Geral

Ilustrar procedimentos para o tratamento de dados com valores omissos

Específicos:

• Identificar o mecanismo de omissão de dados;

• Ilustrar o tratamento de dados omissos usando uma técnica adequada;

• Avaliar o impacto que os valores omissos podem causar nos resultados do estudo.

1.4 LIMITAÇÕES

No que diz respeito à limitações no decorrer da investigação foram:

A dificuldade de obtenção de base de dados brutas para a aplicação das técnicas de imputação de dados omissos, e a escassez de fontes bibliográficas relacionadas com o tema, levando assim ao uso de fontes secundárias como a internet..

II REVISÃO DA LITERATURA

2.1 CONCEITOS E DEFINIÇÕES

Um problema comum em levantamento de dados é a ocorrência de dados omissos. A perda de dados é um grande desafio no planeamento e estudo para qualquer analista. Segundo Pestana e Gageiro (2004) quando a taxa da informação omissa excede 20% da informação, é necessário aplicar técnicas estatísticas de modo a minimizar o problema de viés. O desenvolvimento de técnicas estatísticas com vista a solucionar o problema de dados omissos constitui uma área bastante activa desde os anos 80.

No entanto o uso de técnicas inadequadas pode levar a conclusões erradas sobre o comportamento do fenómeno na população.

Definição 1- Segundo Macknight et. al (2007) o termo dados omissos significa quando existe algum tipo de omissão da informação acerca do fenómeno o qual está a analisar.

Definição 2. uma Função de máxima verossimilhança de uma amostra aleatória Independente e Identicamente Distribuída é definida como: $L(\theta) = \prod_1^n f(x_i \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) ... f(x_n \mid \theta)$, onde θ é o parâmetro de que depende a distribuição de probabilidade de x, e x_i são as variáveis da amostra.

Definição 3. Em estatística Imputação é a substituição de algum valor omisso ou um conjunto de valores omissos por um valor aproximado.

Definição 4. Dados omissos são valores não declarados num conjunto de dados.

A estratégia mais apropriada para o tratamento de dados incompletos não só depende dos mecanismos que os geram, mas também da taxa de não - respostas.

Quando se considera a dimensão de não - respostas para certas variáveis uma das interrogações que pode surgir é como analisar a informação incompleta e qual será o efeito destes ajustes nos procedimentos que se empregam.

Existem várias formas para distinguir o padrão de omissão de dados, que descrevem quais são os valores observados e quais são os valores omissos na matriz de dados, e o mecanismo de omissão de dados, no que diz respeito a relação entre a omissão e os valores das variáveis na matriz de dados, e as diferentes formas de omissão são: Omissão Univariada quando a omissão é limitada a uma única variável, Omissão Monótona cuja omissão é por atrito onde sujeitos desistem o fim do estudo e não retornam, Omissão Geral quando o caso prático em estudo tem valores omissos em observações particulares no questionário de sujeitos não respondentes e duas variáveis não observáveis em que a omissão ocorre em duas variáveis que não se observam conjuntamente, com uma larga quantidade de dados omissos.

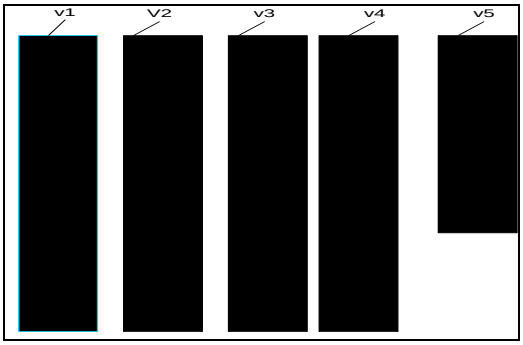

Figura 2.1. Omissão Univariada

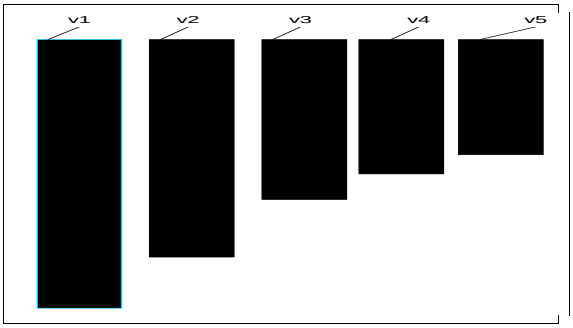

Figura 2.2. Omissão Monótona

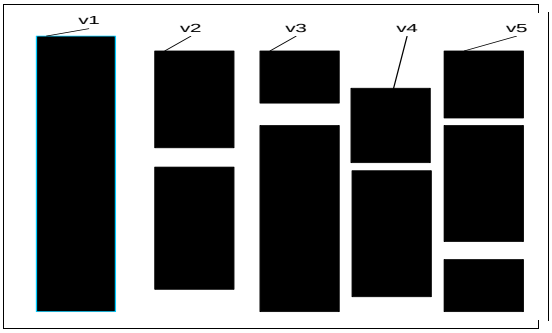

Figura 2.3. Omissão Geral

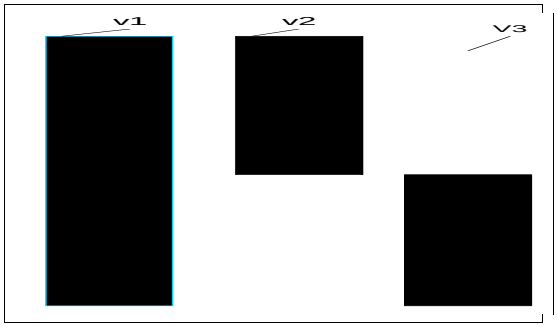

Figura 2.4. Duas variáveis não observáveis

Para solucionar o problema de analises estatísticas com dados incompletos é necessário, identificar em primeiro lugar, o mecanismo que descreve a distribuição dos valores omissos e sua implicação na inferência estatística.

Segundo Little e Rubin (1987) a classificação destes mecanismos se baseia na aleatoriedade com que se distribuem os valores omissos nos casos e nas variáveis. Estes autores definem três tipos de mecanismo nomeadamente: Processo de Omissão Completamente Aleatório (Missing

Completely At Random - MCAR), Processo de Omissão Aleatório (Missing At Random-MAR), e Processo de Omissão Não Aleatório (Missing Not At Random- MNAR).

O primeiro caso MCAR, aparece quando as unidades com dados completos são similares as de dados incompletos, i.é. os sujeitos com dados incompletos constituem uma amostra aleatória simples de todos os sujeitos que formam a amostra.

No segundo caso MAR, os sujeitos com informação completa diferem do resto. Os padrões dos dados omissos podem - se predizer a partir da informação contida em outras variáveis e não da variável que está incompleta.

No último caso MNAR, o padrão dos dados omissos não é aleatório e não se pode predizer a partir da informação contida noutras variáveis.

Seja **Z** uma matriz de dados representado por **n** casos de **k** variáveis aleatórias cuja distribuição conjunta depende do vector parâmetro θ. Z é só parcialmente observado quando é constituído por valores observados e valores omissos. A localização de valores omissos é feita na matriz R com a mesma dimensão de Z, onde R contém variáveis dicotómicas que indicam se cada valor em Z é observado ou omisso, i.é, R_{ij} =1 se Z_{ij} é omisso e R_{ij} =0 se Z_{ij} é observado.

De uma forma geral a distribuição de R é escrita da seguinte forma:

$$p(R|Z_{mis}, Z_{obs,} \Phi) \quad (2.1)$$

Se a omissão é do tipo MCAR ou não depende nem do valor observado e nem do valor omisso então a distribuição de R pode ser escrita como:

$$p(R| Z_{mis}, Z_{obs,} \Phi)= p(\Phi) \quad (2.2)$$

Se a omissão é do tipo MAR, i.é. depende do valor observado mas não do valor omisso então a distribuição de R pode ser simplificada e depender somente de Z_{obs} e de Φ:

$$p(R|Z_{mis}, Z_{obs,} \Phi) = p(Z_{obs,} \Phi) \quad (2.3)$$

Se a omissão é MNAR, então a distribuição de R não se pode escrever na sua forma simples.

2.2 TÉCNICAS ESTATÍSTICAS

2.2.1 Tratamento de observações somente com dados completos

Segundo Anderson et al (2005), a análise estatística de observações com dados completos é o tratamento mais simples e directo para lidar com dados perdidos que consiste em incluir somente as observações com dados completos, também conhecido como **abordagem de caso completo.** Uma desvantagem deste procedimento é que em muitas situações a amostra resultante é reduzida a um tamanho inadequado. A abordagem de caso completo é mais apropriada a bases de dados nos quais a extensão de dados perdidos é pequena, e a amostra é suficientemente grande para permitir a eliminação dos casos com dados perdidos, por outro lado as relações nos dados devem ser tão fortes para que não sejam afectadas por qualquer processo de dados perdidos.

Segundo Canizares et al (2003), este procedimento não facilita a análise por ser simples porque tem vários problemas entre eles a perda de informação que pode produzir um impacto negativo considerável na precisão e a eficiência dos estimadores. Os riscos podem ser graves quando o mecanismo que gera os dados omissos é MAR e não MCAR.

2.2.2 Eliminação de caso(s) e/ ou variáveis

Outro tratamento simples para dados perdidos é eliminar os casos ou variáveis com problemas. Neste tratamento, o investigador determina a extensão dos dados perdidos em cada caso e variável e só depois é que elimina os casos ou variáveis com níveis excessivos de dados omissos. Segundo Anderson et al (2005), em muitos casos, onde o padrão de dados perdidos não é aleatório, este procedimento pode ser a solução mais eficiente.

2.2.3 Métodos de atribuição simples

Uma terceira categoria de acções correctivas para lidar- se com dados perdidos é por meio de um entre os quatro métodos:

- método de substituição por um caso;
- método de substituição pela média;
- método de atribuição por uma carta marcada e

- método de atribuição por regressão.

Ainda segundo Anderson et al (2005) os **métodos de atribuição,** consistem em usar relações conhecidas que podem ser identificadas nos valores válidos da amostra para auxiliar na estimação dos valores perdidos.

a) Substituição por um caso - neste método, as observações com dados perdidos são substituídas por uma outra observação escolhida fora da amostra. Um exemplo comum é substituir uma família da amostra, que não pode ser contactada ou que tem extensos dados perdidos, por uma outra família que não esteja na amostra, de preferência que seja muito semelhante à observação original. Este método é comummente usado para substituir observações com dados completamente perdidos, apesar de também poder ser empregue para substituir observações com menos quantidade de dados perdidos.

b) Substituição pela média - este método é amplamente utilizado, a substituição pela média, troca os valores perdidos pelo valor médio da variável em análise, de salientar que a média é calculada com base em todas as respostas válidas. Dessa maneira, as respostas válidas da amostra são usadas para calcular o valor de substituição. O raciocínio deste tratamento é que a média é o melhor valor e único para substituição. Essa abordagem, apesar de muito usada, tem três desvantagens:

- Torna inválidas as estimativas da variância derivadas das fórmulas usuais por subestimar a verdadeira variância nos dados depois da substituição;
- A distribuição real de valores fica distorcida, quando os valores perdidos são substituídos pela média.
- O método comprime a correlação observada, pois todos os dados perdidos têm um único valor constante ”a média”.

De salientar que este método tem a vantagem de ser fácil de implementar e fornecer todos os casos com informação completa.

c) Atribuição por carta marcada - neste método substitui-se os valores perdidos por um valor constante obtido de fontes externas ou estudos anteriores. Pela natureza o método é semelhante ao método de substituição pela média, diferindo apenas na fonte do valor de substituição. A atribuição por carta marcada tem as mesmas desvantagens do método de substituição pela média, e o investigador neste método deve certificar se o valor para a substituição de fonte externa é mais válido do que um valor gerado internamente, como a média. Este método pode fornecer ao investigador a opção de substituir os dados perdidos por um outro valor que pode ser considerado mais adequado do que a média da amostra.

d) Atribuição por regressão - neste método, a análise de regressão é usada para estimar os valores perdidos de uma variável com base em sua relação com outras variáveis na base de dados. Apesar do recurso de usar relações já existentes na amostra como a base de estimação, o método tem várias desvantagens:

- Reforça as relações já existentes nos dados. À medida que o emprego desse método aumenta, os dados resultantes se tornam mais característicos da amostra e menos generalizáveis.

- A menos que termos estocásticos sejam acrescentados aos valores estimados, a variância da distribuição é subestimada;

- Este método pressupõe que a variável com dados perdidos tenha correlações substanciais com as outras variáveis. Se essas correlações não forem suficientes para produzir uma estimativa significativa, então outros métodos, como a substituição pela média, atribuição por um caso são preferíveis. O método de atribuições por regressão se mantém promissor quando níveis moderados de dados perdidos, amplamente dispersos, existem e as relações entre variáveis são suficientemente estabelecidas, de tal modo que o investigador está confiante de que o uso desse método não influenciará a generalidade dos resultados (Anderson et al, 20005).

2.2.4 Método de atribuição múltipla.

Autores como Rubin (1987), Allison (2002), Schafer e Graham (2002), afirmam que o método de atribuição múltipla tem se tornado o método estatístico mais elogiado para a manipulação de

dados omissos. Assim de acordo com Rubin (1987) citado por Macknigt et. al (2007) o método fornece estimativas generalizadas e recupera a variância crítica populacional para inferência estatística.

Para Macknight et al (2007), este método é facilmente implementado porque fornece manualmente ou pelo uso de pacotes estatísticos estimativas de confiança incluindo erros padrão, e permite a obtenção de informação omissa e o impacto nos parâmetros de estimação, podendo ser aplicado em diferentes situações de dados omissos.

Segundo Anderson et al (2005) o procedimento de atribuição múltipla é uma combinação de diversos métodos. Este tratamento possui dois ou mais métodos de atribuição que são usados para derivar uma estimativa composta (geralmente a média das várias estimativas) para o valor perdido. O raciocínio dessa abordagem fundamenta - se no princípio de que o uso de tratamentos múltiplos minimiza as preocupações específicas com qualquer método particular e que a composição final tem sido a melhor estimativa possível.

2.2.5 Procedimentos baseados em Modelos de distribuição Teórica

Segundo Anderson et al (2005), este conjunto de procedimentos incorpora explicitamente os dados perdidos na análise, quer seja por um processo especificamente planeado para estimação de dados perdidos, seja como uma parte da análise multivariada padrão. Constituem este grupo o método de máxima verossimilhança e o método de expectativa máxima.

a) Método de Máxima Verossimilhança

Os procedimentos de Máxima Verosimilhança (MV) são considerados como estimadores mais

robustos. Estes métodos são vantajosos por produzirem estimativas não enviesada (as) em amostras grandes; são também eficientes por produzirem pequenos erros padrão.

Outra vantagem do uso de procedimentos de Máxima Verossimilhança para a estimação de parâmetros é a boa manipulação dos dados omissos, especialmente quando o mecanismo de omissão de dados é MAR.

b) Método de Expectativa Máxima

Segundo Little e Rubin (2002) o método de Expectativa Máxima é um método geral de obtenção de estimativas de Máxima Verossimilhança quando os dados são omissos.

A vantagem dos algoritmos de Máxima Verossimilhança e Expectativa Máxima para manipular dados omissos está fundamentada nas suas desejáveis propriedades de estimação quando o mecanismo de omissão é MAR ou MNAR.

2.2.6 Métodos de ajustamento

a) Ajustamento com variáveis dicotómicas

Geralmente usado em análise de regressão quando uma variável predictora é omissa. O método consiste na criação de duas variáveis: uma variável dicotómica e outra variável que faz réplica de valores observados e substitui os valores omissos com uma constante. Embora pareça ser um método razoável para a manipulação de dados omissos, Jones (1996) e Allison (2002) mostraram que o procedimento geralmente produz estimativas enviesadas independentemente do mecanismo gerador de dados omissos.

b) Procedimentos com coeficientes de ponderação

Um método para endereçamento do problema associado com procedimentos de eliminação (pequenas amostras, diminuição do poder estatístico) é o procedimento com coeficientes de ponderação de casos ou parâmetros baseados nos dados observados.

Depois da eliminação dos casos incompletos, os dados completos restantes são ponderados como se a sua distribuição aproximasse a amostra completa ou a distribuição populacional. Os coeficientes para ponderações são empregues para corrigir a variabilidade populacional ou erros padrão associados aos parâmetros. Para derivar coeficientes pesos adequados a probabilidade de cada resposta possível da variável com dados omissos é estimada a partir dos dados observados.

Schefer e Graham (2002), mostraram, que os coeficientes de ponderação podiam eliminar o erro de viés associado à taxas de respostas diferenciais para a variável usada na estimação das probabilidades de respostas, entretanto não corrige os erros de viés relacionados com variáveis não usadas.

Procedimentos com coeficientes de ponderação são uma opção viável especificamente, em situações quando os padrões de dados omissos são monótonos ou em análises univariadas.

Embora estes métodos não necessitem de um modelo de distribuição fundamental como os procedimentos baseados em modelos com distribuição teórica, a aplicação de probabilidade de respostas pode tornar-se absolutamente embaraçoso o procedimento sobretudo quando múltiplas variáveis tiverem diferentes probabilidades de respostas ou quando muitos padrões de dados omissos estão presentes. Por isso, os métodos são raramente uma opção adequada para muitos investigadores em Ciências Sociais a menos que os padrões de dados omissos sejam poucos e as probabilidades de respostas são conhecidas e relativamente uniformes dentro das variáveis.

2.2.7 Cadeias de Markov e simulação de Monte Carlo

Uma limitação dos métodos de Máxima Verossimilhança é que eles são embaraçados pela suposição da distribuição teórica, por exemplo a distribuição normal multivariada. O procedimento baseado nas Cadeias de Markov e simulação de Monte Carlo (CMMC) promete óptimas flexibilidades quando as distribuições fundamentais não são conhecidas.

As CMMC foram desenvolvidas para investigar o equilíbrio de distribuição de interacção de moléculas, e em estatística os processos de CMMC tem uma boa qualidade de estimativas mesmo sob condições impróprias das bases de dados incluindo situações quando a base de dados tem valores omissos e quando a distribuição fundamental não satisfaz os pressupostos de procedimentos de Máxima Verossimilhança. O processo é caracterizado como Bayesiano onde o objectivo final é obter uma distribuição de probabilidades chamada distribuição posterior que pode ser usada para as próximas estimativas. Segundo Gill (2002), a distribuição posterior é a distribuição de parâmetros desconhecidos depois da observação dos dados e do uso da informação obtida dos dados para o modelo estatístico.

Os métodos padrão de Monte Carlo são usados para gerar valores simulados independentemente um do outro de acordo com uma distribuição de probabilidade previamente escolhida. Os métodos de CMMC geram valores em cadeias de Markov, o qual é a sequência de valores aleatórios cujas probabilidades dependem só de valores no passo antecedente. A vantagem dos procedimentos de CMMC sobre procedimentos de Máxima Verossimilhança é que eles são eficientes porque permitem a análise de dados para estimativas quando as distribuições

fundamentais não são conhecidas e nem se aproxima a distribuição normal. Pelo uso de pacotes estatísticos estes procedimentos tendem a oferecer soluções óptimas mesmo para o mais complicado problema de dados omissos.

2.2.8- Regressão Logística

Segundo Hosmer e Lemeshow.(1989) a regressão logística é uma técnica estatística multivariada que consiste em perceber o que diferencia dois grupos de casos, isto é, o que diferencia dois níveis de uma variável dependente dicotómica, com base num conjunto de variáveis independentes.

De acordo com Gujarat (2006) a regressão logística possui uma variável dependente de carácter não - métrica que é inserida através do uso de variáveis dicotómicas, que tomam o valor "0" para indicar ausência de um atributo e "1" para indicar a presença do tal atributo.

Na regressão logística importa saber se um evento ocorreu para de seguida usar um valor dicotómico como sendo a variável dependente. A partir desse valor, o procedimento prevê a sua estimativa da probabilidade de que o evento ocorrerá ou não, normalmente usa-se a seguinte regra de decisão: se a probabilidade prevista for maior que 0.5, então a previsão será sim, caso contrário não.

a) Interpretação dos coeficientes da regressão logística

O procedimento que calcula os coeficientes da regressão logística, compara a probabilidade de um evento ocorrer com a de um não ocorrer. A razão de desigualdade pode ser expressa como a probabilidade de um evento ocorrer e é dada pela expressão:

$$\pi = \frac{1}{1 + e^{(\beta 0 + \beta 1 X1 + \beta 2 X2 \ldots + \beta i Xi)}} \qquad (2.4)$$

Onde Xi são variáveis independentes inclusas no modelo, β_i são os coeficientes das variáveis exógenas inclusas no modelo.

b) **Ajuste do modelo logístico**

Segundo Anderson et al (2005), ao se fazer uma estimação de máxima verossimilhança deve-se ter em conta o ajuste da curva logística aos dados da amostra, o que pode ajudar em grande medida a decidir quais variáveis devem fazer parte do modelo. A medida geral, para o efeito, é dada pelo valor da verossimilhança (-2Log(verossimilhança)). O valor mínimo para -2LL é zero que representa um ajuste perfeito com verossimilhança igual a 1.

Os métodos formais para medir a redução do valor de verossimilhança com o aumento de variáveis independentes na equação são: O teste de Qui - quadrado, o R^2 de Cox e Snell que não pode alcançar o valor máximo de 1, o R^2 de Nagelkerke que representa a modificação do Cox e Snell (varia de 0 à 1) e o R^2 Pseudo é dado pela equação:

$$R^2 logit = \frac{-2LLnulo - (-2LLmodelo)}{-2LLnulo} \quad (2.5)$$

2.2-9 Erro Padrão

Segundo Pestana e Gageiro (2005**),** o erro padrão é importante pois o investigador recorre normalmente a amostra para inferir sobre a população. O seu valor é obtido pelo quociente entre o desvio padrão da amostra e a raiz quadrada da dimensão da amostra conforme a equação:

$$EP = \frac{s}{\sqrt{n}} \quad (2.6)$$

O erro padrão é o desvio padrão das médias amostrais, como tal é uma medida que indica o quão a amostra é representativa da população.

2.3 ETAPAS PARA O TRATAMENTO DE VALORES OMISSOS

Para o tratamento de valores omissos segundo Rubin (1987) e Anderson et al (2005) confere a verificação dos seguintes passos:

1) Cálculo da percentagem de dados omissos em cada variável;
2) Cálculo da percentagem de dados omissos por caso;
3) Cálculo da percentagem conjunta de dados omissos;
4) Cálculo das correlação entre as variáveis.

5) Identificar o mecanismo que descreve os dados omissos se é MCAR, MAR ou MNAR; Nesta etapa faz-se o teste para avaliar a aleatoriedade comparando as observações com e sem dados omissos para cada variável com relação às outras variáveis, o segundo teste para a aleatoriedade envolve o uso de correlações entre variáveis dicotómicas, e o teste final é um teste geral dos dados perdidos para detectar o mecanismo MCAR, este teste faz uma comparação do verdadeiro padrão de dados perdidos com o que se esperaria se os dados perdidos fossem distribuídos totalmente ao acaso,

6) Identificação da melhor técnica de tratamento de valores omissos. Nesta etapa é escolhida a técnica de imputação múltipla porque esta técnica produz melhores resultados quando o mecanismo de omissão de dados for MCAR em comparação com outras técnicas sem deixar de lado o método de Máxima verossimilhança porque se o mecanismo for MAR produz resultados menos tendenciosos .

III MATERIAL E MÉTODOS

3.1 MATERIAL

Os dados foram fornecidos pela empresa Águas de Moçambique e pela Ca Computers. Os tamanhos das amostras usadas neste trabalho foram obtidas em função da informação disponível na base de dados.

Com a finalidade de ilustrar a aplicação dos procedimentos ou técnicas para o tratamento de dados com valores omissos descritos no capítulo 2, foram usadas duas bases de dados correspondentes a duas amostras daqui em diante: uma amostra de 613 observações fornecida pelas águas de Moçambique, esta amostra tem nove variáveis das quais 3 são qualitativas e as restantes são quantitativas referentes ao mês de Outubro de 2009 conforme mostra a Tabela 3.1.

Tabela 3.1: Lista de variáveis na base de dados referentes ao agregado familiar (2009)

Código	Nome das variáveis
Y	Despesa da família com água no mês de Outubro
X1	Despesa da família com electricidade no mês de Outubro
X2	Despesa da família com transporte no mês de Outubro
X3	Despesa da família com renda no mês de Outubro
X4	Consumo de água no mês de Outubro
X5	Valor de investimentos feitos pela família no ano 2008
X6	Género de chefe da família
X7	Número de homens no agregado familiar
X8	Número de mulheres no agregado familiar
X9	Frequência de saída de água em horas por dia.

Outra amostra fornecida pela Ca Computers é composta por 100 observações. Esta amostra tem

duas variáveis quantitativas e duas qualitativas, que são a idade do indivíduo, o tempo em minutos despendido a ler jornal, o sexo do indivíduo e o tipo de jornal preferido respectivamente.

Para o processamento dos dados foram usados os seguintes softwares: SPSS versão 14 e 17, e foram compilados usando o Microsoft Word.

3.2 MÉTODOS

3.2.1 Estatística Descritiva e testes aplicados

Como em todos estudos estatísticos é fundamental o uso da estatística descritiva para produzir as medidas de tendência central, medidas de dispersão e análise gráfica. Para a qualidade dos dados é necessário fazer o teste da normalidade das variáveis através do teste de Kolmogorov Smirnov cujas hipótese são:

H_0: X = N (0,1)

H_a: X $\neq$ N(0,1)

Regra de Decisão: No presente trabalho todos testes usados, são analisados a 5% ou seja α = 0.05. Se o sig associado ao valor do teste for menor que α = 0.05 rejeita-se a hipótese nula. Se o valor do sig associada ao valor do teste for maior que α não se rejeita a hipótese nula.

Para a detecção de valores aberrantes (*outliers*), recomenda-se muitas vezes a análise gráfica recorrendo ao uso de caixa de bigodes (*boxplots*). Para além do teste de normalidade e correcção de dados aberrantes, em estatística também é necessário ter um resumo das estatísticas descritivas de cada variável que consta na base de dados. Assim, o número de casos válidos, de casos perdidos e as suas respectivas percentagens é exemplo de algumas estatísticas descritivas de interesse.

A disposição gráfica de dados omissos, é uma das etapas da análise preliminar, esta etapa serve para identificar os casos com dados omissos nas variáveis, i.é, o número de valores omissos em cada caso nas respectivas variáveis. Uma outra etapa importante nesta análise é o Padrão de dados omissos que consta na matriz dos dados. Nesta etapa retrata-se os padrões para identificar a variável com mais dados omissos.

A identificação do mecanismo de aleatoriedade de dados omissos, é mais uma fase do processo onde se faz um exame empírico dos padrões de dados perdidos, para determinar se estes estão distribuídos ao acaso pelos casos e pelas variáveis ou não.

O segundo teste é usado para analisar as correlações entre as variáveis dicotómicas. Estas variáveis são formadas pela substituição de valores válidos pelo valor "1" e de valores perdidos pelo valor "0". As correlações resultantes entre as variáveis dicotómicas indicam o quanto os dados perdidos estão relacionados em pares de variáveis, sendo baixas correlações apontarem uma fraca associação entre o processo de dados perdidos para as duas variáveis.

O último teste que é necessário fazer de aleatoriedade é o teste geral para detectar MCAR. Este é um teste multivariado conhecido como teste MCAR de Little, que consiste em testar as correlações das variáveis. Para tal faz-se uma comparação do verdadeiro Padrão de dados perdidos com o que se esperaria se os dados perdidos fossem distribuídos totalmente ao acaso, cujas hipóteses do teste são:

H_0: Os dados omissos são completamente aleatórios.

H_a: Os dados omissos não são completamente aleatórios.

A escolha da técnica para o tratamento dos dados omissos é baseada na aleatoriedade dos dados perdidos, i.é, uma vez identificado o mecanismo gerador dos dados omissos, segue-se a escolha da técnica para o tratamento dos dados omissos.

Segundo Anderson et al (2005), se o investigador detectar que o processo de dados perdidos é MCAR pode-se aplicar os diferentes métodos de atribuição, se o Padrão de dados perdidos é MAR ou MNAR, recomenda-se apenas uma acção correctiva o tratamento de modelagem ou o uso de procedimentos baseados em modelos.

3.2.2 Método de Imputação Múltipla

A aplicação da imputação múltipla segue os seguintes passos:

- A estimativa combinada - é a média das várias estimativas conforme a equação:

$$Q = \frac{1}{m}\sum_{i=1}^{m} Qj \qquad (2.7)$$

Onde Q_j é o parâmetro estimado em cada imputação. Este parâmetro pode ser qualquer medida escalar como a média, correlação, coeficiente da recta de regressão, e j = 1, 2, 3, …,*m*., sendo *m* o número de imputações.

- Variância dentro das imputações:

$$V = \frac{1}{m}\sum_{i=1}^{m} Vj \qquad (2.8)$$

Onde v_j são as variâncias de cada imputação;

- Variância entre as imputações

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(Qj - Q)^2 \qquad (2.9)$$

- A taxa de informação omissa é calculada a partir da equação:

$$\gamma = \frac{r+\frac{2}{(gl+3)}}{r+1}$$

r representa o aumento relativo na variância devido a não - respostas, seu valor é calculado pela equação:

$$r = \left(\frac{\left(1+\frac{1}{m}\right)B}{V}\right) \qquad (2.11)$$

- Eficiência de imputação múltipla (IM) para parâmetros de estimação com dados omissos

$$Eficiencia = \frac{1}{1+\gamma} \qquad (2.12)$$

- A variância combinada também designada variância total é dada por:

$$T = V + (1 + \frac{1}{m})B \qquad (2.13)$$

O erro padrão global eficiente para reportar o modelo final para o teste de significância é dado por: $\sqrt{T}$

- O intervalo de confiança da estimativa combinada Q é dado por :

$$Q \pm 1.96\left(\sqrt{T}\right) \tag{2.14}$$

- Finalmente o t - value similar ao t - test de Students é calculado por:

$$t(gl) = \left(\frac{Q}{\sqrt{T}}\right) \tag{2.15}$$

Os graus de liberdade são calculados pela equação:

$$gl = (m-1)\left\{\frac{1+mV}{\left(((m+1))B\right)}\right\}^2 \tag{2.16}$$

3.2.3- Método de Expectativa Máxima

Este procedimento consiste em dois passos: Expectativa e Maximização, estes passos repetem-se múltiplas vezes em um processo de iteração que eventualmente converge em estimadores de Máxima Verossimilhança. As principais etapas do método são:

Etapa 1 Imputar valores omissos usando estimativas de Máxima Verossimilhança;

Etapa 2. Gerar estimativas da média, variância e covariância baseados na etapa 1;

Etapa 3. Reimputar valores baseados nas estimativas obtidas na etapa 2;

Etapa 4. Reestimar parâmetros baseados nos dados reimputados da etapa 3.

O processo continua até que a etapa final convirja em uma solução que difere muito pouco da solução anterior.

Para avaliar o impacto que os valores omissos têm sobre os resultados, os dados devem ser analisados antes e depois do tratamento, para tal são aplicadas técnicas estatísticas tais como: a Regressão Logística e análise do erro padrão da amostra**.**

Teste Qui - quadrado

Segundo Pestana e Gageiro (2005), na escala ordinal quando se tem mais de duas categorias e o tamanho da amostra for maior que 30 a variável pode ter tratamento qualitativo e pode- se

aplicar o teste Qui - quadrado X^2, para verificar a relação de dependência entre as variáveis que podem ou não ser relações de casualidade. O teste é também aplicado para verificar se a variável em causa segue uma distribuição específica (por exemplo Uniforme, de Poisson, etc).

As hipóteses do teste Qui - quadrado para verificar se uma variável segue a distribuição Uniforme são formuladas da seguinte maneira:

H_0: A distribuição da variável é uniforme;

H_a: A distribuição da variável não é uniforme;

3.3 MODELO LOGÍSTICO

3.3.1 Estimação do modelo logístico

Para a obtenção das estimativas mais prováveis para os coeficientes usa - se o procedimento iterativo de Máxima Verossimilhança, o qual consiste no uso do valor de verossimilhança ao calcular a medida de ajuste geral do modelo. Antes de estimar o modelo a usar para efeitos de interpretação do fenómeno, determina-se o modelo base que fornece padrões para comparação.

No presente trabalho usou – se o método Stepwise for Ward ” critério de redução da razão do logaritmo de verossimilhança” que indica a ordem de inclusão das variáveis independentes no modelo. Neste método entra primeiro a variável que apresentar o maior coeficiente de Wald de entre todas as variáveis que ainda se encontram fora do modelo.

3.3.2 Significância estatística dos coeficientes

Para interpretar os coeficientes obtidos no modelo ajustado, é preciso verificar se os mesmos têm significância estatística, para tal usa-se o teste de Wald, cujas hipóteses são:

Ho : Os coeficientes do modelo são iguais à zero;

Ha : Existe pelo menos um coeficiente do modelo diferente de zero;

3.3.3 Testes do ajuste do modelo logístico estimado

- O valor de -2LL (verossimilhança), onde valores menores indicam um bom ajuste;

- O valor de R^2 de Nagelkerke;

- Estatística de Hosmer e Lemeshow, e as hipóteses a testar são:

H_0: As classificações observadas são iguais as previstas

H_a: As classificações observadas são diferentes das classificações previstas

3.4 ANÁLISE DO ERRO PADRÃO DE MÉDIAS

Para a comparação de médias é analisado o erro padrão das médias da variável em estudo, e para o efeito as médias são analisadas antes e depois da imputação dos valores omissos para verificar a sua eficiência.

Um erro padrão grande significa que existe muita variabilidade entre as médias de amostras diferentes e por isso a média global pode não ser representativa da população. Um erro padrão pequeno significa que as médias de diferentes amostras são semelhantes entre si, e portanto semelhantes à da população de onde provém, neste caso a média da amostra é representativa da população.

IV RESULTADOS E DISCUSSÃO

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Tabela 4.1: Estatística descritiva das variáveis (Agregado Familiar)

Variável	N		Media	Mediana	Moda	Desvio padrão	Mínimo	Máximo
	Valido	Omisso						
X4	326	287	9.83	8.00	5.00(a)	5.05	1.00	20.00
X2	601	12	604.22	450.00	300.00	698.34	20.00	10,800.00
X5	406	207	9,597.22	5,150.00	3,000.00	11,780.44	1.00	85,000.00
Y	469	144	148.98	150.00	15.00(a)	142.66	10.00	1,500.00
X1	261	352	251.42	200.00	200.00	180.95	30.00	2,000.00
X8	612	1	2.50	2.00	2.00	1.26	0.00	7.00
X7	611	2	2.25	2.00	2.00	1.31	0.00	9.00

(a) Multiple modes exist. The smallest value is show

As estatísticas da Tabela 4.1 mostram que o consumo de água (x4) no período em estudo em média foi cerca de 9.83 m^3, com uma dispersão no consumo entre as famílias de 5.05m^3. O número médio de homens que vivem em cada agregado familiar (x7) é cerca de 2 homens com uma dispersão de número de homens de aproximadamente um homem, sendo que o número mínimo e máximo de homens é cerca de 0 e 9 respectivamente. A tabela 4.1 mostra também que as despesas da família (Y) foram em média de 149 meticais com uma dispersão das despesas de cerca de 143 meticais indicando uma dispersão muito grande de despesas entre as famílias, o valor das despesas de água com maior frequência é de 15 meticais, mais de 50% das famílias têm despesas superiores a 150 meticais e 50% das famílias têm despesas abaixo de 150 meticais, o valor máximo e mínimo das despesas com água são 10 meticais e 1500 meticais respectivamente.

A Tabela 4.2 mostra o teste de normalidade das variáveis através do teste de Kolmogorov-Smirnov, podendo assim se afirmar que o tempo gasto a ler jornal está normalmente distribuído com média zero e variância constante enquanto que a variável idade dos inqueridos não tem distribuição normal.

Tabela 4.2: Teste de normalidade das variáveis Idade e Tempo despendido em minutos..

One-Sample Kolmogorov-Smirnov Test

		Idade do inquirido	Temp em minutos dispendido
N		83	79
Normal Parameters[a,b]	Mean	38.33	75.93
	Std. Deviation	15.237	36.028
Most Extreme Differences	Absolute	.312	.146
	Positive	.312	.146
	Negative	-.189	-.101
Kolmogorov-Smirnov Z		2.841	1.293
Asymp. Sig. (2-tailed)		.000	.070

a. Test distribution is Normal.

b. Calculated from data.

4.2 CÁLCULO DA PERCENTAGEM DE OMISSÃO EM CADA VARIÁVEL

A Tabela 4.3 contém as estatísticas descritivas das observações com casos válidos, incluindo a percentagem de casos perdidos em cada variável. A média do tempo gasto a ler jornal é de 75.93 minutos com uma dispersão de tempo de 36.03 minutos, e a idade média dos inquiridos é de 38 anos com uma dispersão de idade de 15 anos. A percentagem de dados perdidos para cada variável varia desde 17% para a variável idade até a um máximo de 24% dos casos para a variável Jornal.

Tabela 4.3: Resumo das estatísticas de dados do pré-teste

Variável	Omissão		N válido	Média	Desvio padrão
	N	Percentagem			
Tipo de Jornal	24	24.0%	76		
Tempo despendido em minutos	21	21.0%	79	75.93	36.03
Idade do inquirido	17	17.0%	83	38.33	15.24

A Tabela 4.4 mostra as estatísticas descritivas das observações com casos válidos, incluindo a percentagem de casos perdidos em cada variável. A percentagem de dados omissos varia desde 0.2% para a variável número de mulheres em cada agregado familiar até um máximo de 57.4% dos casos na variável despesas com electricidade.

Tabela 4.4: Resumo das estatísticas de dados do pré-teste

Variável	N	Média	Desvio padrão	Omissão		Nr. de extremos	
				Contagem	Percentagem	Baixo	Alto
X4	326	9.83	5.05	287.00	46.80	0.00	0.00
X5	406	9,597.22	11,780.44	207.00	33.80	0.00	48.00
X3	613	41.19	640.23	0.00	0.00	0.00	0.00
X2	601	604.22	698.34	12.00	2.00	0.00	52.00
Y	469	148.98	142.66	144.00	23.50	0.00	14.00
X1	261	251.42	180.95	352.00	57.40	0.00	8.00
X8	612	2.50	1.26	1.00	0.20	5.00	44.00
X7	611	2.25	1.31	2.00	0.30	0.00	11.00
X6	613			0.00	0.00		

4.3 ILUSTRAÇÃO DO PADRÃO DE DADOS OMISSOS

Tabela 4.5 : Padrão de dados omissos da base de dados 1 (Agregado familiar)

Número de casos	Padrão de omissão									Complete if ...[b]
	X3	X6	X8	X7	X2	Y	X4	X1	X5	
145										145
82								X		227
51							X	X		308
30							X			175
27						X	X			210
8						X				153
50						X	X	X		397
41						X	X	X	X	599
62							X	X	X	461
49								X	X	302
26									X	171
16							X		X	217

b. Number of complete cases if variables missing in that pattern (marked with X) are not used

A Tabela 4.5 mostra os padrões de dados perdidos na base de dados Agregado Familiar, sendo que o padrão predominante é o de dados perdidos para a variável despesas com electricidade e foi encontrado em 82 casos e o segundo padrão foi encontrado em 62 casos nas variáveis despesas com electricidade, consumo de água e investimento, o terceiro padrão encontrado em 51 casos nas variáveis consumo de água e despesas com electricidade, e os outros padrões foram encontrados em menos de 50 casos.

A Tabela 4.6 mostra os padrões de dados perdido na base de dados 2, sendo que o padrão predominante é o de dados perdidos para a variável idade encontrado em 14 casos e o segundo encontrado em 13 casos na variável horas, o terceiro encontrado em 7 casos na variável Horas e jornal.

Tabela 4.6: Padrões de dados omissos das variáveis da base de dados Jornal.

Número de Casos	Padrões de dados perdidos			
	Idade	Horas	Jornal	
48				48
14	X			62
13		X		61
7		X	X	83

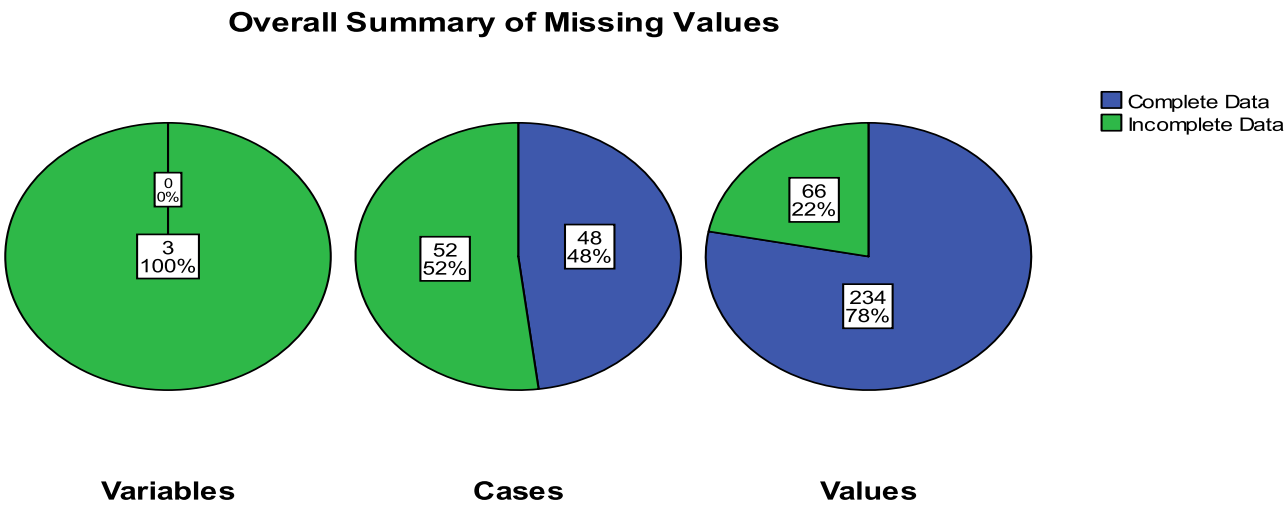

Figura 4.1: Distribuição geral de dados omissos.

A Figura 4.1 mostra que todas variáveis têm valores omissos, por outro lado a figura mostra que dos 100 casos 52 tem pelo menos um valor omisso numa das variáveis e finalmente dos 300 valores 66 são valores omissos.

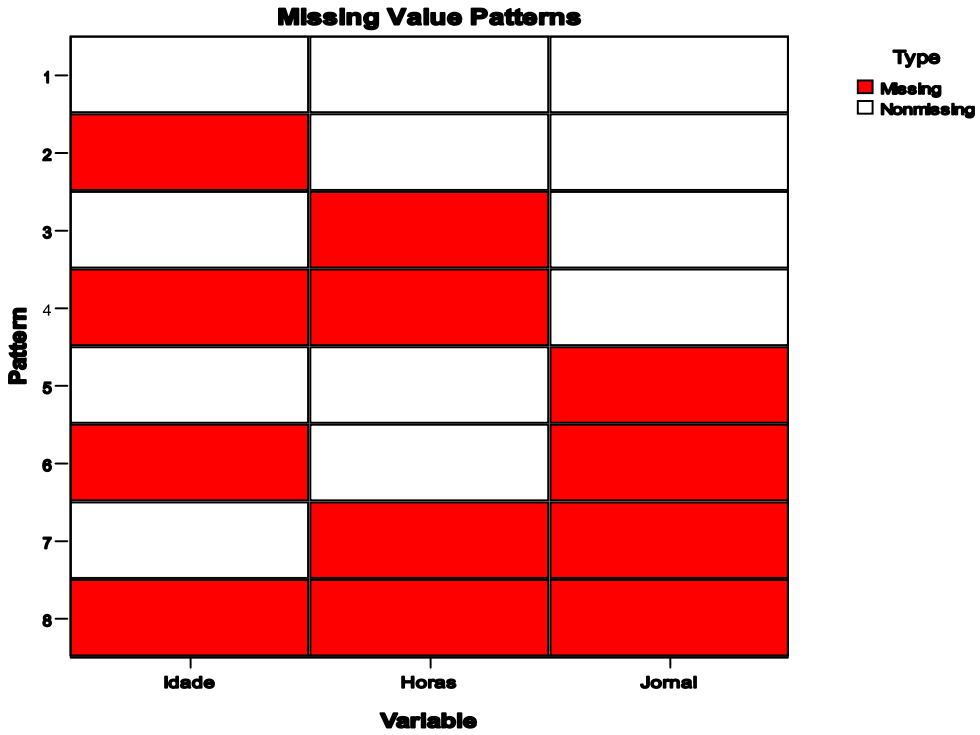

Figura 4.2: Padrão de valores omissos

A Figura 4.2 mostra os padrões de omissão de dados para cada variável, de salientar que, cada padrão corresponde ao grupo de caso com o mesmo padrão de dados completos e incompletos. Por exemplo o padrão 4 representa casos que têm valores omissos na variável horas e idade.

4.4 CÁLCULO DA CORRELAÇÃO ENTRE VARIÁVEIS E IDENTIFICAÇÃO DO MECANISMO GERADOR DE DADOS OMISSOS

Cada célula da Tabela 4.7, contém 7 valores: o valor de t, o número de graus de liberdade, significância do valor de t, o número de observações válidas e observações perdidas, as médias das observações válidas e observações perdidas.

Tabela 4.7: Avaliação da aleatoriedade de dados perdidos via comparação de grupos de observações com dados perdidos versus dados válidos

		Idade	Jornal	Horas
Idade	T	.	-1.4	.4
	Df	.	23.8	22.9
	P(2-tail)	.	.168	.706
	# Present	83	61	63
	# Missing	0	15	16
	Mean(Present)	38.33	1.82	76.72
	Mean(Missing)	.	2.13	72.80
Jornal	T	-.3	.	1.4
	Df	76.2	.	45.7
	P(2-tail)	.801	.	.164
	# Present	61	76	62
	# Missing	22	0	17
	Mean(Present)	38.15	1.88	78.15
	Mean(Missing)	38.85	.	67.83
Horas	T	1.8	.1	.
	Df	79.8	18.8	.
	P(2-tail)	.077	.907	.
	# Present	63	62	79
	# Missing	20	14	0
	Mean(Present)	39.37	1.89	75.93
	Mean(Missing)	35.08	1.86	.

Na Tabela 4.7 ilustra-se o exame empírico dos padrões de dados perdidos, para verificar se estes estão distribuídos ao acaso pelos casos e pelas variáveis. Os valores de p - value do teste t mostram que não são significativos, indicando que os padrões de dados perdidos estão distribuídos ao acaso entre os casos e variáveis.

Tabela 4.8: Avaliação da aleatoriedade de dados perdidos usando a correlação de variáveis dicotómicas e teste multivariado para detectar MCAR(Jornal)

	Horas	Jornal	Idade
Horas	1.000		
Jornal	-0.377	1.000	
Idade	0.083	-0.161	1.000

a. Little's MCAR test: Chi-Square = 4.080, DF = 9, Sig. = .906

Na Tabela 4.8 o nível de significância do teste MCAR é de 0.906, indicando que o processo de dados perdidos pode ser considerado MCAR para qualquer erro do tipo I do analista.

Tabela 4.9 Avaliação da aleatoriedade de dados perdidos usando a correlação de variáveis dicotómicas e teste multivariado para detectar MCAR (Agregado familiar)

	X7	X8	Y	X3	X5	X2	X1	X4
X7	1							
X8	.027	1						
Y	.107	.161	1					
X3	.067	.039	.336	1				
X5	.128	.142	.232	-.043	1			
X2	.106	.025	.118	.065	.253	1		
X1	.057	.131	.447	.187	.335	.235	1	
X4	-.021	.033	.088	.036	.023	-.058	.041	1

a. Little's MCAR test: Chi-Square = 433.201, DF = 131, Sig. = .000

Na Tabela 4.9 o nível de significância do teste MCAR é de 0.000, indicando que o processo de dados perdidos não pode ser considerado MCAR para qualquer erro do tipo I do analista.

4.5 RESULTADOS ANTES E DEPOIS DA IMPUTAÇÃO DA BASE JORNAL

Tabela 4.10: O erro padrão do tempo dispendido em minutos

Base de dados original	**N**	**Média**	**Erro padrão**
Expresso	25	100.28	8.692
Semanário	19	61.93	6.424
Independente	18	64.52	6.240
Total	62	78.15	4.931
Base de dados imputada	**N**	**Media**	**Erro padrão**
Expresso	40	92.43	6.133
Semanário	30	59.06	4.831
Independente	30	65.82	4.836
Total	100	74.28	3.415

A Tabela 4.10 mostra que o erro padrão da média de categorias diferentes de jornal antes da imputação múltipla assim como do total é maior (4.931) relativamente ao erro da média depois do tratamento de valores omissos (3.415), evidenciando que a média dos dados depois do tratamento de dados omissos é mais representativa da população de onde provém.

Resultados de despesas de água antes da Imputação

Tabela 4.11: Resultados do modelo antes da inclusão das variáveis

		B	S.E.	Wald	GL	Sig.	Exp(B)
Passo 0	Constante	-.013	.162	.007	1	.936	.987

O modelo estima que a constante seja de -0.013.

Tabela 4.12: Classificação dos casos antes da imputação

Observado			Preditos		
			despesas variáveis		Percentagem
			Baixo	Alto	correcta
Passo 0	despesas variáveis	Baixa	77	0	100.0
		Altas	76	0	.0
Percentagem total					50.3

(a) Constant is included in the model

Quando se inclui apenas a constante, cada agregado é colocado numa das categorias da variável dependente despesas variáveis. Assim, Da Tabela 4.12, pode – se notar que 77 agregados que exprimem despesas baixas e 76 que exprimem despesas altas e a previsão de despesas baixas é de 50.3% enquanto que a previsão de despesas altas é cerca de 49.7% dos casos. Globalmente o modelo só com a constante classifica apenas 50.3% dos casos.

Tabela 4.13: O teste de Hosmer e Lameshow

Passos	Qui - Quadrado	Graus de liberdade	Sig.
1	55.852	7	.000
2	6.042	8	.643
3	7.442	8	.490

A hipótese nula do teste de Hosmer e Lemeshow, afirma que não existem diferenças significativas entre os valores observados e os previstos. Como o valor de sig = 0.490 maior que 0.05 no terceiro passo, não se rejeita a hipótese nula, o que indica que o modelo se ajusta bem aos dados.

Tabela 4.14: Teste de ajustamento do modelo

Passos	-2log likelihood	Cox e Snell R Square	Nagelkerke R Square
1	202.765[a]	.059	.079
2	196.465[b]	.097	.129
3	189.316[c]	.138	.184

A Tabela 4.14 mostra o teste de Nagelkerke para o ajustamento do modelo. O R^2 de Nagelkerke indica que 18.4% das variações das despesas são justificadas pelo modelo.

Resultados de despesas de água depois da imputação dos dados

Tabela 4.15 : Resultados do modelo antes da inclusão das variáveis

		B	S.E.	Wald	GL	Sig.	Exp(B)
Passo 0	Constante	-.710	.086	68.336	1	.000	.491

Tabela 4.16: Classificações dos casos depois da imputação

Observado			Preditos		
			despesas variáveis		Percentagem correcta
			Baixo	Alto	
Passo 0	despesas variáveis	Baixa	411	0	100.0
		Altas	202	0	.0
	Percentagem total				67.0

(a) Constant is included in the model

No modelo incluindo apenas a constante existem 411 agregados que exprimem despesas baixas e 202 que exprimem despesas altas e a previsão de despesas baixas estará correcta em 67% dos casos enquanto que a previsão das despesas altas estará certa em 33% dos casos e o modelo só com a constante classifica 67% dos casos, porém a constante não é estatisticamente significativa.

Tabela 4.17: O teste de Hosmer e Lemeshow

Passos	Qui - Quadrado	Graus de liberdade	Sig.
1	30.958	8	.000
2	14.251	8	.075
3	9.009	8	.342
4	12.351	8	.136
5	11.154	8	.193

A hipótese nula do teste de Hosmer e Lemeshow, afirma que não existem diferenças significativas entre os valores observados e os previstos. E como o valor de sig = 0.193 superior a 5%, não se rejeita a hipótese nula, o que indica que o modelo se ajusta bem aos dados.

Tabela 4.18 : Teste de ajustamento do modelo

Passos	-2log likelihood	Cox e Snell R Square	Nagelkerke R Square
1	743.613[a]	.053	.074
2	729.889[a]	.074	.103
3	719.897[a]	.089	.124
4	710.259[a]	.103	.144
5	706.339[a]	.109	.152

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

O R^2 de Nagelkerke da Tabela 4.18 indica que 15.2% da variação das despesas é justificada pelo modelo, neste caso o grau de ajustamento é mais baixo relativamente ao modelo dos dados antes do tratamento com uma diferença de 3.2% de ajustamento.

Tabela 4.19 : Teste dos coeficientes do modelo

Passos		Qui - Quadrado	Graus de Liberdade	Sig.
1	Step	33.479	1	.000
	Block	33.479	1	.000
	Model	33.479	1	.000
2	Step	13.723	1	.000
	Block	47.202	2	.000
	Model	47.202	2	.000
3	Step	9.992	1	.002
	Block	57.195	3	.000
	Model	57.195	3	.000
4	Step	9.638	1	.002
	Block	66.833	4	.000
	Model	66.833	4	.000
5	Step	3.919	1	.048
	Block	70.750	5	.000
	Model	70.752	5	.000

O valor de Qui - quadrado da Tabela 4.19 passo 5, é de 70.752 com sig = 0.000 menor que o valor do nível de significação escolhido para o teste (5%). Portanto, isto significa que esta diferença é estatisticamente significativa, mostrando que o modelo com as variáveis exógenas é melhor para prever as despesas de água do agregado familiar do que aquele que tinha apenas a constante.

Tabela 4.20: Variáveis inclusas na equação do modelo

Passos		B	S.E.	Wald	GL	Sig.	Exp(B)
1[a]	investimentos	.000	.000	28.438	1	.000	1.000
	Constante	-1.210	.128	89.065	1	.000	.298
2[b]	consumágua	.089	.024	13.383	1	.000	1.093
	investimentos	.000	.000	27.672	1	.000	1.000
	Constante	-2.097	.282	55.438	1	.000	.123
3[c]	consumágua	.093	.025	14.180	1	.000	1.097
	investimentos	.000	.000	23.394	1	.000	1.000
	homemcasa	.221	.071	9.755	1	.002	1.247
	Constante	-2.603	.332	61.541	1	.000	.074
4[d]	consumágua	.091	.025	13.351	1	.000	1.095
	investimentos	.000	.000	19.850	1	.000	1.000
	homemcasa	.224	.072	9.792	1	.002	1.251
	Mulhercasa	.224	.072	9.576	1	.002	1.251
	Constante	-3.129	.381	67.297	1	.000	.044
5[e]	consumágua	.096	.025	14.550	1	.000	1.100
	investimentos	.000	.000	15.583	1	.000	1.000
	homemcasa	.215	.072	8.966	1	.003	1.240
	Mulhercasa	.230	.073	10.018	1	.002	1.259
	desptransport	.000	.000	3.218	1	.007	1.000
	Constante	-3.312	.398	69.325	1	.000	.036

A Tabela 4.20 mostra que os coeficientes de todas variáveis inclusas no modelo são estatisticamente significativas incluindo a constante, diferentemente dos resultados antes da imputação dos valores omissos alguns coeficientes das variáveis inclusas não são significativas. (Tabela em anexo).

V CONCLUSÕES E RECOMENDAÇÕES

5.1 CONCLUSÕES

A partir dos resultados obtidos no presente trabalho pode-se concluir que:

1) O uso de técnicas estatísticas para o tratamento de informação com uma taxa acima de 20% de dados omissos é imprescindível.

2) A identificação do mecanismo que leva a omissão de dados é indispensável à medida que a aplicação de técnicas inadequadas pode levar a conclusões erradas sobre o comportamento do fenómeno na população.

3) O método de Imputação Múltipla é o mais recomendado se o mecanismo é MCAR, e o método de Expectativa Máxima é mais recomendado se o mecanismo é MAR ou MNAR .

4) Quando atitudes correctivas não são aplicadas os dados omissos num conjunto de dados têm um impacto muito acentuado, não apenas pelas suas tendências ocultas sobre os resultados, mas também pelo seu impacto na diminuição do tamanho da amostra disponível para análise.

5.2 RECOMENDAÇÕES

Para estudos futuros relacionados com o tratamento de valores omissos num conjunto de dados recomenda-se :

- O uso de softwares mais completos tais como: SAS, STATA, R, entre outros;

- A realização de um inquérito para melhor compreensão do problema de não - respostas;

REFERÊNCIA BIBLIOGRÁFICA

Allison, P.D. (2002). Missing data. Thousand Oaks, CA: Sage.

Anderson, R.E , J.F Hair. R.L. Tathan W.C. Black (2005), Análise Multivariada de Dados, 5ª edição. Bookman Porto Alegre;

Canizares, et al (2003). Datos incompletos: uma Mirada critica para su manejo en studios Sanitarios. Maio;

Gill, J. (2002). Bayesian methods: A social and behavioral sciences approach. New York: Chapman e Hall;

Gujarat, D.N (2006). Econometría básica, 4ª edição (tradução) .São Paulo Editora Campus;

Hosmer, David and Stanley Lemeshow.(1989). Applied Logistic Regression. John Wiley and . Sons, Inc

Jones, M.P.(1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association, 91, 222-230.

Laaksonen ,S. (1996). Preface : Nonresponse – An Essential Indicator of Survey Quality, in International Perspectives on Nonresponse, Proceedings of the Sixth International Workshop on Household Survey Nonresponse, Statistics Finland, Helsinki.

Little, R.J.A. e Rubin, D.B (1987), Statistical with Missing data, New York; John Wiley e Sons;

Little, R.J.A. e Rubin, D.B (2002) Statistical with Missing data, 2nd Edition:

McKnight, P.E et al (2007), Missing Data, a gentle intruduction the Guilford press, 1st . edition new york london;

Pestana, M.H e Gageiro, J.N (2004), Analise de Dados para Ciências Sociais (com a Complementaridade do SPSS), 4ª edição;

Pestana, M. H. e Gageiro, J.N. (2005) , Descobrindo a Regressão (com a Complementaridade do SPSS), 1ª edição Lisboa;

Reis, E e Moreira,R. (1992), Pesquisa de Mercado **,** 1ª edição;

Reis, et al (1999)**,** Estatística Aplicada, 3ª edição;

Rubin, D.B. (1987) Multiple imputation for non-response in Survey. New York; Wiley

Shaffer, J.L. (1997) Analysis of Incomplete multivariate data. London: Chapman and Hall;

Schafer, J.L. e Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

BIBLIOGRÁFICA NÃO CITADA

Carpenter, James (2009), Statistical modelling with missing data using multiple imputation November 17;

Cochran, W. G. (1953), Técnicas de Amostragem, 1ª edição. São Paulo, Editora Brasileira;

Rubin, D. (1996), Multiple imputation after 18 years. *Journal of the American Statistical Association*, 91:473–490,

Hippel, Von Paul T. (2003), Data analysis with missing values, May 2;

Ming-Yu Fan, PhD (2008) Missing Data Analysis – Multiple Imputation, April 30;

Jerilee Grandy (1998), Nonresponse bias in the gre background questionnaire, GRE Board . Professional Report No. 85-6PETS Research Report 88-51 December This report;

Nunes et al (2009), Cadernos de Saúde Pública vol.25.no.2 Rio de Janeiro, Fevereiro;

Royston, P. (2005), Multiple Imputation of Missing Values: Update. The Stata Journal .Vol. 5 No. 2, pp. 188-201

Anexos

Tabela A.1: Resultados da imputação

Imputation Method		Fully Conditional Specification
Fully Conditional Specification Method Iterations		10
Dependent Variables	Imputed	Horas,Jornal,Idade
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	Sexo
Imputation Sequence		Horas,Jornal,Idade,Sexo

Tabela A.2: Especificação da imputação

Imputation Method	Fully Conditional Specification
Number of Imputations	5
Model for Scale Variables	Predictive Mean Matching
Interactions Included in Models	(none)
Maximum Percentage of Missing Values	30.0%
Maximum Number of Parameters in Imputation Model	100

Tabela A.3: Estatística descritiva da variável hora

Data	Imputation	N	Mean		Minimum	Maximum
Original Data		79	75.93	36.028	30.00	180.00
Imputed Values	1	21	73.35	24.367	30.00	105.00
	2	21	63.78	34.778	30.00	180.00
	3	21	68.30	25.293	30.00	120.00
	4	21	67.65	24.867	30.00	120.00
	5	21	64.13	22.866	35.00	105.00
Complete Data After Imputation	1	100	75.39	33.819	30.00	180.00
	2	100	73.38	35.941	30.00	180.00
	3	100	74.33	34.083	30.00	180.00
	4	100	74.19	34.045	30.00	180.00
	5	100	73.45	33.935	30.00	180.00

Tabela A.4: Estatística descritiva da variável idade

Data	Imputation	N	Mean	Std. Deviation	Minimum	Maximum
Original Data		83	38.33	15.237	24.00	99.00
Imputed Values	1	17	38.92	17.412	28.00	99.00
	2	17	37.81	16.381	28.00	99.00
	3	17	36.74	9.023	28.00	54.00
	4	17	38.54	16.842	28.00	99.00
	5	17	42.39	22.243	28.00	99.00
Complete Data After Imputation	1	100	38.43	15.535	24.00	99.00
	2	100	38.24	15.352	24.00	99.00
	3	100	38.06	14.346	24.00	99.00
	4	100	38.37	15.432	24.00	99.00
	5	100	39.02	16.571	24.00	99.00

Tabela A.5: Estatística descritiva da variável jornal

Data	Imputation	Category	N	Percent
Original Data		1	32	42.1
		2	22	28.9
		3	22	28.9
Imputed Values	1	1	11	47.8
		2	4	17.4
		3	8	34.8
	2	1	10	43.5
		2	7	30.4
		3	6	26.1
	3	1	7	30.4
		2	8	34.8
		3	8	34.8
	4	1	6	26.1
		2	4	17.4
		3	13	56.5
	5	1	8	34.8
		2	12	52.2
		3	3	13.0
Complete Data After Imputation	1	1	43	43.4
		2	26	26.3
		3	30	30.3
	2	1	42	42.4
		2	29	29.3
		3	28	28.3
	3	1	39	39.4
		2	30	30.3
		3	30	30.3
	4	1	38	38.4
		2	26	26.3

		3	35	35.4
	5	1	40	40.4
		2	34	34.3
		3	25	25.3

Tabela A.6: Tipo de Jornal

	Observed N	Expected N	Residual
Expresso	31	25.3	5.7
Semanário	23	25.3	-2.3
Independente	22	25.3	-3.3
Total	76		

Tabela A. 7: O teste Qui-quadrado

	Tipo de Jornal
Chi-Square(a)	1.921
Df	2
Asymp. Sig.	.383

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.3.

Tabela A .8 : Resumo das estimativas dos desvios padrões

	mulhercasa	homemcasa	Despelectric	Despágua	investimentos	consumágua	desptransporte
All Values	1.257	1.310	180.9516	142.6622	11,780.4419	5.0544	698.3393
EM	1.257	1.310	188.2178	142.1657	11,766.0660	5.0488	698.7236

Tabela A. 9: Resumo das estimativas das médias

	mulhercasa	homemcasa	despelectric	despágua	investimentos	consumágua	desptransporte
All Values	2.50	2.25	251.425	148.985	9,597.224	9.828	604.218
EM	2.50	2.25	240.650	146.811	9,460.681	9.745	604.999

Tabela A.10 : Casos processados depois da imputação

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	613	100.0
	Missing Cases	0	.0
	Total	613	100.0
Unselected Cases		0	.0
Total		613	100.0

a. If weight is in effect, see classification table for the total number of cases.

Tabela A.11 : Tabela das classificações antes da imputação

			Previstos		
			despesas variaveis		Percentagem
	Observados		Baixas	Altas	Correct
Passo 1	despesas variaveis	Baixas	48	29	62.3
		Altas	38	38	50.0
	Overall Percentage				56.2
Passo 2	despesas variaveis	Baixas	52	25	67.5
		Altas	31	45	59.2
	Overall Percentage				63.4
Passo 3	despesas variaveis	Baixas	53	24	68.8
		Altas	28	48	63.2
	Overall Percentage				66.0

a. The cut value is .500

Tabela A.12: Variáveis na equação antes da imputação

		B	S.E.	Wald	GL	Sig.	Exp(B)
Passo 1[a]	consumágua	.099	.033	8.841	1	.003	1.104
	Constant	-1.009	.372	7.341	1	.007	.365
Passo [b]	consumágua	.101	.034	8.714	1	.003	1.106
	homemcasa	.326	.136	5.765	1	.016	1.385
	Constant	-1.848	.526	12.323	1	.000	.158
Passo [c]	consumágua	.097	.035	7.657	1	.006	1.102
	despelectric	.003	.001	5.972	1	.105	1.003
	homemcasa	.335	.139	5.819	1	.016	1.398
	Constant	-2.731	.666	16.794	1	.000	.065

Tabela A.13 : Tabela das iterações depois da imputação

			Coefficients			
Iteração		-2 Log likelihood	Constant	Consumágua	homemcasa	despelectric
Passo 1	1	202.777	-.976	.095		
	2	202.765	-1.009	.099		
	3	202.765	-1.009	.099		
Passo 2	1	196.596	-1.676	.093	.287	
	2	196.465	-1.844	.101	.324	
	3	196.465	-1.848	.101	.326	
	4	196.465	-1.848	.101	.326	
Passo 3	1	191.348	-2.002	.083	.277	.002
	2	189.363	-2.625	.095	.329	.003
	3	189.316	-2.729	.097	.335	.003
	4	189.316	-2.731	.097	.335	.003
	5	189.316	-2.731	.097	.335	.003

Tabela A. **14**: Tabela das classificações depois da imputação

			Previstos		
			desaspas variáveis		Percentagem
	Observado		despesas baixas	despesas altas	Correcta
Passo 1	desaspes variaveis	despesas baixas	389	22	94.6
		despesas altas	175	27	13.4
	Overall Percentage				67.9
Passo 2	desaspes variaveis	despesas baixas	390	21	94.9
		despesas altas	168	34	16.8
	Overall Percentage				69.2
Passo 3	desaspes variaveis	despesas baixas	386	25	93.9
		despesas altas	160	42	20.8
	Overall Percentage				69.8
Passo 4	desaspes variaveis	despesas baixas	384	27	93.4
		despesas altas	152	50	24.8
	Overall Percentage				70.8
Passo 5	desaspes variaveis	despesas baixas	383	28	93.2
		despesas altas	147	55	27.2
	Overall Percentage				71.5

a. The cut value is .500

Figura A.1: ILustração de outliers