UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

# PRESENÇA DE DADOS *MISSING* EM MODELOS DE REGRESSÃO LOGÍSTICA

**Natália Manduca Ferreira**

# UNIVERSIDADE FEDERAL DE SÃO CARLOS

## CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

## PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA


# PRESENÇA DE DADOS *MISSING* EM MODELOS DE REGRESSÃO LOGÍSTICA


**Natália Manduca Ferreira**

**Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz**


Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.


São Carlos

Março/2009

**Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar**

F383pd

Ferreira, Natália Manduca.
Presença de dados missing em modelos de regressão logística / Natália Manduca Ferreira. -- São Carlos : UFSCar, 2009.
93 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Regressão logística. 2. Ausência de dados (Estatística). 3. Estimadores. I. Título.

CDD: 519.5 (20ª)

# Resumo

Neste trabalho apresentamos um estudo detalhado do modelo de regressão logística na presença de valores *missing* nas covariáveis considerando as técnicas Caso Completo, Imputação pela Média e Caso Completo Corrigido. Um novo método, denotado EMVGM, dado pela combinação entre os estimadores de Caso Completo e os estimadores obtidos via Máxima Verossimilhança com uso da Quadratura Gaussiana, é sugerido.

No desenvolvimento do estudo são realizadas simulações para a verificação do desempenho dos estimadores de máxima verossimilhança obtidos em cada técnica citada acima. A avaliação mostra que a qualidade dos parâmetros estimados obtidos por meio de cada técnica varia de acordo com o tamanho da amostra e com o número de dados *missing* e que, em geral, o estimador sugerido, EMVGM, apresenta os melhores estimadores levando em conta as métricas variância estimada, vício estimado e erro quadrático médio estimado.

# Abstract

In this work we present a detailed study of the logistic regression model with missing data in the independent variables. Several techniques are considered such as Complete Case, Mean Imputation and Corrected Complete Case. We present a new estimator, denoted EMVGM, given by the combination between the Complete Case estimator and the ML-estimator with the use of Gaussian quadrature. A simulation study is carried out to evaluate the performance of the ML-estimators obtained in each technique above mentioned. In general, the alternative estimador, EMVGM, presents a better performance taking into account the variance, the bias and the mean quadratic error.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

A inferência estatística em conjunto de dados com ausência de informação (dados *missing*, faltantes, incompletos ou não observados) é, segundo Little (1992), uma importante área de pesquisa. Os dados *missing* são geralmente encontrados em situações reais, por razões de acidente, falta de informação, informações errôneas ou até mesmo por conveniência. Entre alguns casos mais comuns, podemos citar:

- Em uma pesquisa de campo, uma pessoa pode recusar-se a responder determinada pergunta. Conseqüentemente, a resposta referente à pergunta não respondida é o dado *missing*.

- Em um experimento industrial, alguns resultados podem ser *missing* por causa de um certo acidente mecânico não esperado. Neste caso, é natural tratar os dados que não são observados como sendo *missing*, visto que seriam observados se o acidente não tivesse ocorrido (Little *et al.*, 1987).

- Uma pesquisa familiar com muitas variáveis socioeconômicas realizada em um certo período e a mesma pesquisa realizada com as mesmas famílias em um período subseqüente. É provável a presença de muitos dados *missing* no último conjunto de dados coletados, já que algumas famílias podem não ser localizadas na pesquisa seguinte, como exemplificado em Rubin (1976).

As diferentes causas que nos levam a obter os dados *missing* são importantes na escolha da análise a ser feita e na interpretação dos dados. Enquanto a maioria das

análises de dados ignoram as causas dos dados *missing*, assumindo-os como acidentais, a literatura estatística discute, embora que pouco, as causas da ocorrência destes dados faltantes, assumindo-os como intencionais. Neste caso, o processo que causa esses dados é geralmente considerado explícito. Um exemplo de método que pode criar *missing* intencional é a análise robusta, onde os *outliers* podem ser descartados ou tidos como *missing*.

Quando queremos estimar parâmetros de regressão tendo valores *missing* nas variáveis de entrada (covariáveis), uma solução é usar a análise de caso completo, onde todos os casos incompletos são simplesmente descartados. Isto torna, na maioria das vezes, o estimador ineficiente. Outro método consiste em imputar (completar) valores no lugar dos dados não observados e então tratar o conjunto de valores como se fosse completo, porém, alguns dos métodos de imputação não levam em conta a incerteza dos dados adicionados, podendo gerar erro na estimação, como afirmou Didelez (2002). A escolha mais razoável é usar a imputação múltipla que é auxiliada por métodos de simulação, como Monte Carlo via cadeia de Markov, por exemplo.

Este trabalho trata de um estudo detalhado do modelo de regressão logística na presença de valores *missing* nas variáveis de entrada considerando as técnicas Caso Completo, Imputação pela Média e Caso Completo Corrigido. Além disso, sugerimos um novo estimador, denotado EMVGM, que é dado pela combinação entre o estimador de Caso Completo e o estimador de Máxima Verossimilhança com uso da Quadratura Gaussiana. Apresentamos também um estudo sobre o desempenho dos estimadores de máxima verossimilhança obtidos em cada uma das técnicas citadas acima.

O desenvolvimento da dissertação é dado por cinco capitulos. Neste Capítulo vimos uma breve introdução sobre dados faltantes. No Capítulo 2 apresentamos alguns modelos e mecanismos existentes para se trabalhar quando tem-se ausência de informação. No Capítulo 3, apresentamos o modelo de regressão logística especificando detalhadamente as variáveis adotadas no estudo. No Capítulo 4, propomos a aplicação do método de estimação dos parâmetros por máxima verossimilhança levando em conta as técnicas Caso Completo, Caso Completo Corrigido, Imputação pela Média e EMVGM. No Capítulo 5 comparamos, por meio de simulação, as performances dos diferentes estimadores, obtidos através das técnicas apresentadas no Capítulo 4. E, finalmente, no Capítulo 6 mostramos

a conclusão deste trabalho. Para obtenção dos resultados aqui presentes, foram utilizados os Softwares SAS 9.0, R 2.7.1 e Maple 11. No apêndice disponibilizamos todos os códigos utilizados.

# Capítulo 2

# Dados *Missing*

Conforme dito na Introdução, o problema de dados *missing* surge freqüentemente na prática. Ao contrário dos dados presentes em textos ilustrativos, os dados reais quase sempre são incompletos.

Muitos *Softwares* estatísticos permitem a identificação de dados não observados criando códigos especiais para eles na matriz de dados. No *Software* SAS, por exemplo, cada valor *missing* é indicado por um ponto (.), por *default*. Porém, no presente texto, indicamos a ausência de informação com um traço ($-$). Segue exemplo:

**Exemplo 2.1.** Considere as variáveis $X_1, X_2, X_3, X_4$ e $X_5$ definidas como quantidade de cigarro consumido por dia, idade (em anos), taxa de gordura no sangue (em hdl), hereditariedade (1, se há casos da doença coronária na família e 0, caso contrário) e realização regular de atividade física (1, se sim e 0, caso contrário), respectivamente. Seja $Y$ a variável de interesse codificada em 0 quando há ausência da doença coronária e 1, quando há presença da doença.

Observe os dados abaixo:

TABELA 2.1: Presença ou não de doença coronária em 6 pacientes

| Paciente | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | - | 170 | 0 | 1 |
| 2 | 0 | 3 | 38 | 200 | 0 | 1 |
| 3 | 1 | - | 51 | 210 | - | - |
| 4 | 1 | 1 | - | 350 | 1 | 1 |
| 5 | 1 | 4 | 40 | 430 | 0 | 0 |
| 6 | 0 | 2 | - | - | - | 0 |

Alguns pacotes estatísticos (SAS e R, por exemplo) excluem, por *default*, os indivíduos que apresentam valores *missing* em qualquer variável envolvida na análise. A Tabela 2.1 nos mostra que os pacientes (casos, unidades ou indivíduos) $1, 3, 4$ e $6$ são descartados em uma possível análise por apresentarem variáveis com dados *missing*.

Esta estratégia de "descarte", apesar de ser facilmente implementada e, às vezes, satisfatória quando se tem poucos *missing*, é geralmente inapropriada, pois perdem-se possíveis informações relevantes observadas nos indivíduos excluídos. Se houver uma grande diferença entre os casos completos e os incompletos, onde por caso completo entende-se, no presente parágrafo, todos os casos presentes na amostra inicial e casos incompletos, a amostra de tamanho reduzida (já sem os indivíduos com dados *missing*), a inferência estatística baseada nos casos completos pode ser viciada, já que há uma perda considerável de dados.

Com isso, o objetivo deste trabalho é descrever, testar e comparar técnicas que sejam mais apropriadas para trabalhar com conjuntos onde há dados *missing* em Modelos de Regressão Logística.

Segundo Little *et al.* (1987), a estrutura de dados mais simples é a amostra aleatória univariada onde tem-se unidades *missing* ou não. Seja $x_i$ o i-ésimo valor da variável $X$ e suponha que para uma amostra aleatória simples de tamanho $n$, $x_1, x_2, ..., x_m$ são observadas e $x_{m+1}, ..., x_n$ são *missing*, com $m < n$. Uma consequência óbvia é a redução do tamanho amostral de $n$ para $m$, podendo-se fazer as mesmas inferências na amostra reduzida (tamanho $m$) que seria feita na amostra original (tamanho $n$).

Por exemplo, se assumimos que os dados são normalmente distribuídos, a média é estimada pela média amostral das unidades correspondentes e a estimativa da variância é dada por $S^2/m$, onde $S^2$ é a variância amostral das unidades correspondentes. Ao fazer isto, estamos ignorando o mecanismo que causou os valores *missing* (Little *et al.*, 1987).

Para se trabalhar com dados faltantes, precisamos identificar qual modelo de dados *missing* estamos analisando e qual mecanismo a ser adotado.

## 2.1 Modelos e Mecanismos *Missing*

Nesta seção, apresentamos quatro modelos para valores *missing* que, segundo Little (1992), foram classificados em: Valor *missing* univariado, Monotonia de valor *missing*, Modelo especial e Modelo geral.

Na sequência, definimos também três mecanismos para dados incompletos.

### 2.1.1 Modelos de valores *missing*

Para o estudo dos modelos abaixo, consideramos as variáveis aleatórias $X_1, X_2, ..., X_p$ (com presença de *missing* ou não) e $Y$ a variável resposta (de interesse). Nos quatro modelos são apresentadas Figuras como forma de ilustração.

**i)** *Valor missing univariado*

Temos valores *missing* univariados quando os valores faltantes aparecem em apenas uma das variáveis estudadas. A Figura 2.1, por exemplo, mostra-nos que todas as variáveis, exceto $X_1$, são completamente observadas.

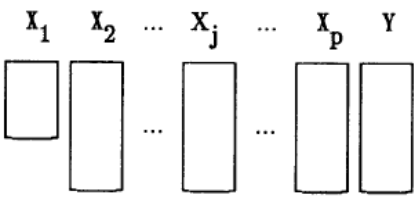


FIGURA 2.1: Modelo de valor *missing* univariado (Fonte: Little, 1992)

No Exemplo 2.1, se as variáveis $X_2, X_3, X_4$ e $X_5$ fossem completamente observadas, teríamos o modelo de valor univariado, onde apenas a variável $X_1$ teria dados *missing*.

**ii)** *Monotonia de valores missing*

Neste modelo, as colunas são arranjadas de modo que $X_{j+1}$ é observado para todos os casos onde $X_j$ é observado, $j = 1, 2, ..., p$. Observe a figura:

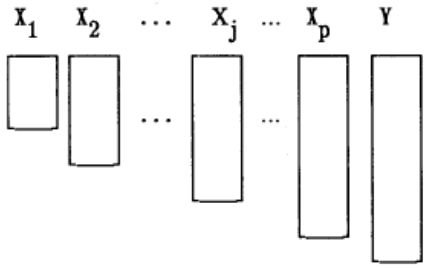


FIGURA 2.2: Modelo de valor *missing* monótono (Fonte: Little, 1992)

Considerando as mesmas variáveis do Exemplo 2.1, teríamos um modelo de monotonia de valores *missing* se os dados estivessem, por exemplo, apresentados da seguinte maneira:

| i | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 24 | 170 | 0 | 1 |
| 2 | 0 | 3 | 38 | 200 | 0 | 1 |
| 3 | 1 | - | 51 | 210 | 1 | 0 |
| 4 | 1 | - | - | 350 | 1 | 1 |
| 5 | 1 | - | - | - | 0 | 0 |
| 6 | 0 | - | - | - | - | 0 |

Note que o indivíduo $x_{55}$ é observado, já que $x_{45}$ é observado. O mesmo raciocínio é aplicado a todos os outros $x_{ji}$, onde $i = 1, ..., 6$ e $j = 1, ..., 5$.

**iii)** *Modelo Especial*

O modelo especial ocorre quando duas variáveis nunca são observadas simultaneamente. Ou seja, se considerarmos três variáveis $X_1, X_2$ e $X_3$, sendo $X_1$ e $X_2$ variáveis incompletas, teremos a seguinte disposição dos dados:

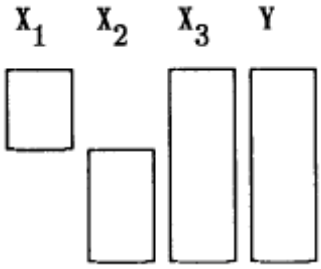


FIGURA 2.3: Modelo especial de valor *missing* (Fonte: Little, 1992)

Tomando apenas as variáveis $X_1$ e $X_2$ do Exemplo 2.1, teríamos um modelo especial se os dados fossem apresentados como segue

| i | $Y$ | $X_1$ | $X_2$ |
|---|---|---|---|
| 1 | 0 | 1 | - |
| 2 | 0 | 3 | - |
| 3 | 1 | - | 51 |
| 4 | 1 | 1 | - |
| 5 | 1 | - | 40 |
| 6 | 0 | 2 | - |

onde, se $x_{11}$ é observado, $x_{21}$ é *missing* e, se $x_{13}$ é *missing*, $x_{23}$ é observado. O mesmo aplicando-se a todos os outros $x_{ji}$, onde $i = 1, ..., 6$ e $j = 1, 2$.

**iv)** *Modelo geral*

Este modelo não apresenta estrutura especial, ou seja, os dados podem estar dispostos de qualquer maneira. Veja a Figura a seguir:

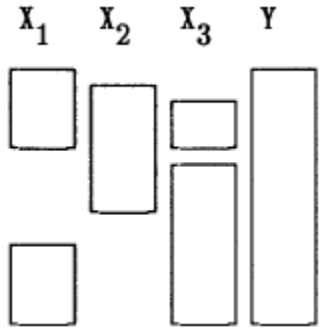


FIGURA 2.4: Modelo geral de valor *missing* (Fonte: Little, 1992)

Como exemplo, temos os dados apresentados inicialmente no Exemplo 2.1.

### 2.1.2 Mecanismos de valores *missing*

Existem três tipos de mecanismos: *Missing completely at random* (MCAR), *Missing at random* (MAR) e Missing não ignorável (MNI). Estes têm como objetivo verificar se os dados (*missing* ou não) estão relacionados aos valores observados.

**Exemplo 2.2.** Considere a situação ilustrada na Figura 2.1, onde todas as variáveis são completamente observadas, exceto $X_1$. Podemos ter os seguintes casos:

(1) $X_1$ ser independente de todos os valores de $X_1$. Por exemplo, em uma pesquisa de opinião, a resposta, correspondente à $X_1$, dada (ou não) por cada indivíduo, independe da resposta dada por qualquer outro indivíduo.

(2) $X_1$ depender dos valores de $X_1$. Seguindo o exemplo anterior, quando a resposta correspondente à variável $X_1$, dada por um indivíduo, é influenciada pela resposta do indivíduo anterior.

(3) $X_1$ depender dos valores de $X_2, ..., X_p$, ou seja, quando há dependência entre a variável $X_1$ e todas as outras variáveis explicativas.

(4) $X_1$ depender dos valores de $X_2, ..., X_p$ e $Y$. Isto ocorre quando a variável $X_1$ depende de todas as variáveis explicativas, exceto dela, e da variável resposta.

Afim de formalizar o conceito de mecanismo *missing*, Little (1992) considerou $\mathbf{Z}$ uma matriz $n \times (p+1)$ formada por valores observados e valores *missing*, ou seja, $\mathbf{Z} = (\mathbf{Z}_{obs}, \mathbf{Z}_{mis})$, onde $\mathbf{Z}_{obs}$ é o conjunto dos valores observados e $\mathbf{Z}_{mis}$, o conjunto de valores *missing* de $\mathbf{Z}$. Considerou também uma variável indicadora de valores *missing* $\mathbf{R}$, sendo que $R_{ij} = 1$ quando $X_{ij}$ é observado e $R_{ij} = 0$ quando $X_{ij}$ é *missing*, com $i = 1, ..., n$ e $j = 1, ..., p+1$.

Com as especificações acima, os mecanismos para valores *missing* são dados por meio da distribuição condicional de $\mathbf{R}$ dado $\mathbf{Z}$, indexada por um parâmetro desconhecido $\varphi$, isto é, $Pr(\mathbf{R}|\mathbf{Z}, \varphi)$.

Logo, os mecanismos são definidos por:

1. **MCAR** (*Missing completely at random*) quando

$$Pr(\mathbf{R}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \varphi) = Pr(\mathbf{R}|\varphi), \ \forall \ \mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \tag{2.1}$$

ou seja, quando a distribuição de **R** não depende nem dos valores observados, nem dos valores *missing* em **Z** (Little, 1992).

2. **MAR** (*Missing at random*) quando

$$Pr(\mathbf{R}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \varphi) = Pr(\mathbf{R}|\mathbf{Z}_{obs}, \varphi), \ \forall \ \mathbf{Z}_{mis}, \tag{2.2}$$

isto é, quando a distribuição depende somente dos valores observados de **Z** (Little, 1992).

O mecanismo MAR é o mais utilizado na prática. Alguns autores (como, por exemplo, Didelez 2002) usam MAR-X para especificar que os *missing* dependem apenas dos valores observados nas covariáveis, bem como MAR-Y para representar que os *missing* dependem apenas dos valores observados na variável resposta.

Para exemplificar os dois mecanismos anteriores, considere o Exemplo 2.2. Neste Exemplo, o mecanismo no caso (1) é MCAR, nos casos (3) e (4) os mecanismos são MAR, pois $X_2, ..., X_p$ e $Y$ são completamente observadas e (2) não é MAR, já que $X_1$ não é completamente observado.

O terceiro tipo de mecanismo *missing* é dito *Missing* não ignorável (MNI). Aqui, em contradição ao mecanismo MAR, a distribuição de **R** depende apenas dos valores *missing*.

Para situações onde os casos MAR ou MCAR são válidos, a causa da presença de dados *missing* é considerada ignorável. Já para os casos onde não são válidos, o motivo da ocorrência dos dados *missing* é levado em conta na análise, ou seja, é não ignorável.

Os três mecanismos apresentados acima podem ser exemplificados abaixo:

**Exemplo 2.3.** (Little *et al.*, 1987) Suponha que $X$ = idade e $Y$ = renda, sendo $Y$ completamente observada e $X$ parcialmente observada. Se a distribuição de $X$ é a mesma para todos os indivíduos, sem considerar a idade ou renda, então os dados são MCAR. Agora, se a distribuição de $X$ varia de acordo com a renda e não com a idade, então os dados são MAR. Mas, se a distribuição de $X$ depende da idade e do salário, os dados não são MAR, MCAR nem MNI. Finalmente, se a distribuição da variável $X$ depende apenas da idade, então temos MNI.

A escolha do mecanismo de dados *missing* depende do objetivo da análise. Por exemplo, se o interesse está na distribuição marginal de $Y$, então os dados em $X$ e o mecanismo que conduz os valores *missing* de $X$ são irrelevantes. Se o interesse estiver na distribuição condicional de $X$ dado $Y$ como, por exemplo, quando estamos verificando como a distribuição da idade varia de acordo com a renda, então a análise baseada nas $m$ unidades (número de indivíduos observados na variável $X$) pode ser satisfatória se os dados forem MAR. E, se o interesse for apenas na distribuição marginal de $X$, um mecanismo satisfatório é o mecanismo MCAR.

A literatura em análises de dados incompletos é bem recente. Há vários trabalhos envolvendo modelos normais multivariados com observações incompletas. No entanto, a literatura estatística para dados *missing* em Modelos Lineares Generalizados é bem escassa.

A maioria dos métodos de estimação, presentes em trabalhos científicos, assumem que os dados são MAR. Porém, em muitos problemas práticos, esta suposição é altamente questionável.

# Capítulo 3

# Modelo de Regressão Logística

O modelo de regressão é um componente importante nas análises de dados com interesse na relação entre a variável resposta e as variáveis explicativas (qualitativas ou quantitativas). A regressão logística tem se constituído em um dos métodos mais adequados de modelagem estatística de dados por ter aplicações em diversas áreas, tais como: Saúde, Mercado Financeiro e *Marketing*.

Mesmo quando a resposta de interesse não é do tipo binário (dicotômica, 0 ou 1), podemos dicotomizá-la de modo a obter uma probabilidade de sucesso que seja modelada através de uma regressão logística (Paula, 2004). Em linhas gerais, a regressão logística é um modelo que tem como objetivo predizer a probabilidade de ocorrência de um evento.

Geralmente é chamado "sucesso" o resultado mais importante da variável resposta, representando assim, a presença de uma particular característica de interesse. Nota-se então que a variável resposta é essencialmente qualitativa (ou categórica), visto que a designação "sucesso" é inteiramente arbitrária.

Seguem, três pequenos exemplos de variável resposta dicotômica:

**Exemplo 3.1.** (Pinto, 25/08/2008) Num estudo sobre a participação das esposas no mercado de trabalho, com covariáveis do tipo idade da esposa, número de filhos e renda do marido, a variável resposta $Y$ foi definida do seguinte modo: a mulher participa do mercado de trabalho ("sucesso", $Y = 1$) ou não ("fracasso", $Y = 0$).

**Exemplo 3.2.** (Pinto, 25/08/2008) Em *marketing* podemos desejar saber se alguém comprará ($Y = 1$) ou não ($Y = 0$) um carro na chegada de um novo ano. Aqui as

covariáveis tais como renda anual, número de dependentes na família e valor da prestação do carro são informações importantes para o estudo.

**Exemplo 3.3.** (Cordeiro, 2007) Ensaios do tipo dose-resposta são aqueles em que uma determinada droga é administrada em $k$ diferentes doses, $d_1, ..., d_k$, respectivamente a $m_1, ..., m_k$ indivíduos. Suponha que cada indivíduo responde, ou não, à droga, tal que a resposta é dicotômica, obtendo-se, após um período especificado, $y_1, ..., y_k$ indivíduos que mudam ou não de estado. Por exemplo, quando um inseticida é aplicado a um determinado número de insetos, eles respondem (morrem, $Y = 1$), ou não (sobrevivem, $Y = 0$) à dose aplicada; quando uma droga benéfica é administrada a um grupo de pacientes, eles podem melhorar (sucesso, $Y = 1$), ou não (fracasso, $Y = 0$).

Neste capítulo definimos o modelo de regressão e o mecanismo *missing* utilizado no desenvolvimento do trabalho.

## 3.1 Modelo com uma variável resposta binária

Considere o seguinte modelo:

$$y_i = \mathbf{x_i'}\boldsymbol{\beta} + \epsilon_i \tag{3.1}$$

em que $\mathbf{x_i'} = (1, x_{i1}, x_{i2}, ..., x_{ip})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ e a variável resposta $y_i$ com valores 0 ou 1, onde

$$y_i = \begin{cases} 1, \ \textit{se o elemento possui a caracteristica de interesse} \\ 0, \ \textit{caso contrário} \end{cases}$$

Assume-se que $y_i$ é uma variável aleatória tendo distribuição de probabilidade Bernoulli, ou seja,

| $y_i$ | Probabilidade |
|---|---|
| 1 | $P(y_i = 1) = \pi_i$ |
| 0 | $P(y_i = 0) = 1 - \pi_i$ |

e que $E(\epsilon_i) = 0$. Logo, o valor esperado da variável resposta é dado por:

$$E(y_i) = 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0)$$

$$= 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i$$

o que implica em

$$E(y_i) = \pi_i = \mathbf{x_i'}\boldsymbol{\beta} \tag{3.2}$$

uma vez que

$$E(y_i) = E(\mathbf{x}_i'\beta + \epsilon_i) = E(\mathbf{x}_i'\beta) + E(\epsilon_i) = \mathbf{x}_i'\beta + 0 = \mathbf{x}_i'\beta$$

O resultado anterior nos diz que o valor esperado da variável resposta dado por $\mathbf{x_i'}\boldsymbol{\beta}$ é apenas a probabilidade da variável resposta assumir o valor 1.

Existem algumas características específicas associadas ao modelo de regressão dado pela Equação (3.1). Primeiro, nota-se que se a resposta é binária, certamente os erros $\epsilon_i$ assumem somente dois valores:

$$\epsilon_i = 1 - \mathbf{x_i'}\boldsymbol{\beta}, \ se \ y_i = 1$$

ou

$$\epsilon_i = -\mathbf{x_i'}\boldsymbol{\beta}, \ se \ y_i = 0,$$

conseqüentemente, os erros deste modelo podem não ser normais. Segundo, a variância do erro pode não ser constante, visto que

$$\sigma^2_{y_i} = Var(\epsilon_i) = Var(y_i - \mathbf{x_i'}\boldsymbol{\beta}) = Var(y_i) = E[y_i - E(y_i)]^2$$

$$= (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i)$$

$$= \pi_i(1 - \pi_i) \tag{3.3}$$

Isto indica que a variância das observações (que é a mesma dos erros) é uma função da média e varia, num intervalo fechado, de 0 a 1. Note que existe uma condição na função resposta, já que $0 \leq E(y_i) = \pi_i \leq 1$. Esta restrição pode impedir o uso de uma função de resposta linear, uma vez que os valores preditos podem se estender além do intervalo $[0, 1]$.

Com isso, uma função monótona crescente (ou decrescente), denominada função logística, é freqüentemente usada neste caso e dada por

$$f(z) = \frac{exp(z)}{1 + exp(z)} = \frac{1}{\frac{1}{exp(z)} + 1} = \frac{1}{1 + exp(-z)} \tag{3.4}$$

onde, variando $z$ de $-\infty$ a $\infty$ obtemos valores entre 0 e 1, inclusive. Observe o gráfico abaixo:

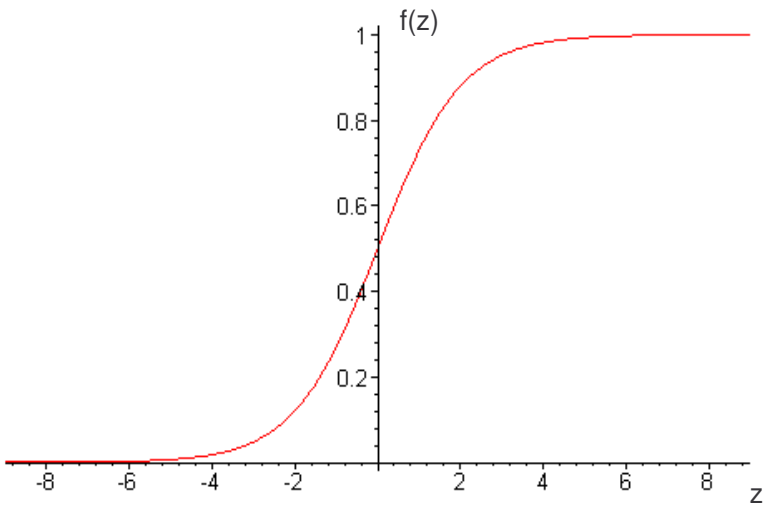


FIGURA 3.1: Função logística

ou seja, quando $z \to \infty$, temos que

$$\lim_{z\to\infty} f(z) = \lim_{z\to\infty} \frac{1}{1+exp(-z)} = \frac{1}{1+exp(-\infty)} = 1$$

e, quando $z \to -\infty$, temos que

$$\lim_{z\to-\infty} f(z) = \lim_{z\to-\infty} \frac{1}{1+exp(-z)} = \frac{1}{1+exp(\infty)} = 0$$

Confirmando assim, um *range* para $f(z)$ de: $0 \leq f(z) \leq 1$.

Para obtermos o Modelo de Regressão Logística a partir da função logística (3.4), consideramos $z$ como a seguinte soma linear

$$z = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

e a seguinte transformação logito

$$z = ln\left(\frac{\pi_i}{1+\pi_i}\right). \tag{3.5}$$

Substituindo a transformação (3.5) na função logística (3.4), obtemos:

$$f(z) = \frac{exp(z)}{1+exp(z)} = \left(\frac{\pi_i}{1-\pi_i}\right) \Big/ \left(1+\frac{\pi_i}{1-\pi_i}\right) = \pi_i = E(y_i). \tag{3.6}$$

Deste modo o Modelo de Regressão Logística Múltipla é dado por:

$$E(y_i) = \pi_i = Pr(y_i = 1|\mathbf{x_i}) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})} \tag{3.7}$$

onde $E(y_i)$ é a probabilidade de sucesso segundo o modelo em questão, $\beta_0$ é o intercepto e $\beta_i$ o coeficiente de regressão de $x_i$, $i = 1, 2, ..., p$.

O modelo de regressão logística é denomidado simples quando $p = 1$, no modelo acima.

Ainda no modelo (3.7), a relação entre a variável resposta $y$ e o vetor de covariáveis $\mathbf{x}$ é descrita por uma curva sigmoidal que tem uma forma que lembra um $S$.

Uma propriedade interessante deste modelo é que pode ser linearizado por meio de funções de ligação, tais como Transformação Logito, Transformação Probito e Transformação Complementar Log-Log.

## 3.2 Modelo Logístico com *Missing*

O seguinte modelo de regressão logística, adotado em Didelez (2002), foi a base do nosso trabalho:

$$Pr(Y = 1|X_1 = x_1, X_2 = x_2; \beta) = \frac{exp\ (\beta'\mathbf{x}^*)}{1 + exp\ (\beta'\mathbf{x}^*)} \tag{3.8}$$

onde

$Y$: variável binária completamente observada;

$X_1$: covariável binária completamente observada;

$X_2$: covariável contínua com alguns valores não observados (*missing*);

$\mathbf{x}^* = (1, x_1, x_2)'$ e

$\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2)$ o vetor de parâmetros a ser estimado.

Adotamos a seguinte notação:

$$Pr(y|x_1, x_2; \boldsymbol{\beta}) = Pr(Y = 1|X_1 = x_1, X_2 = x_2; \boldsymbol{\beta}).$$

O mecanismo para valores *missing* que consideramos, também adotado em Didelez (2002), foi o MAR, visto na Seção 2.1.2. Relembrando:

$$Pr(R|y, x_1, x_2) = Pr(R|y, x_1), \ \ \forall\ y, x_1, x_2 \tag{3.9}$$

onde $R$ é uma variável indicadora, em que

$$R = \begin{cases} 1, \ \textit{se } x_2 \textit{ é observado} \\ 0, \ \textit{caso contrário}; \end{cases}$$

e $y, x_1$ e $x_2$ como definidos anteriormente.

A probabilidade condicional, dada em (3.9), para observações completas foi denotada por Didelez (2002) como sendo $q_{yx_1}, y, x_1 \in \{0, 1\}$.

O modelo de valores *missing* adotado foi o univariado (vide Figura 2.1), pois os valores faltantes estão confinados apenas na variável $x_2$.

# Capítulo 4

# Estimação dos Parâmetros de Interesse e Imputação de Dados

Neste capítulo apresentamos o método de Máxima Verossimilhança para estimação dos parâmetros de interesse e alguns métodos para trabalhar com conjunto de dados quando temos ausência de informação.

## 4.1 Estimador de Máxima Verossimilhança

A estimação por máxima verossimilhança é uma das várias técnicas desenvolvidas para estimar os parâmetros de um modelo de interesse. Outro método bem conhecido é a estimação por mínimos quadrados que é usado para estimar os parâmetros dos modelos de regressão linear, como veremos no Exemplo 4.3.

A partir desta Seção, assumimos o modelo e o mecanismo *missing* apresentados na Seção 3.2, e consideramos que $(y_i, x_{1i}, x_{2i}, r_i), i = 1, ..., N$, é uma amostra independente de $(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{r})$.

Definimos também que o conjunto de dados completos (com dados observados e não observados) é dado por

$$\{(y_i, x_{1i}, x_{2i}, r_i)|\ i \in v\} \cup \{(y_i, x_{1i}, r_i)|\ i \in \bar{v}\}$$

em que $v = \{i|r_i = 1\}$ e $\bar{v} = \{1, ..., N\}\backslash v$, ou seja, em $v$ temos os indivíduos cuja variável

$x_2$ é observada e, em $\bar{v}$, os indivíduos em que a informação referente à variável $x_2$ é *missing* (Didelez, 2002).

Na Seção 3.2, vimos também que Didelez (2002) denotou a probabilidade condicional, dada em (3.9), por $q_{yx_1}, y, x_1 \in \{0, 1\}$. Porém, nas análises feitas consideramos um estimador para essa probabilidade, também proposto pela autora em questão, definido por $\hat{q}_{yx_1}$. Este estimador indica a proporção da unidade amostral dos valores $y, x_1$ e os *missing* $x_2$ sobre todos os valores $y$ e $x_1$. Segue exemplo dessa definição:

**Exemplo 4.1.** Considere o seguinte conjunto de dados

| i | $y$ | $x_1$ | $x_2$ | $r$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 2,5 | 1 |
| 2 | 0 | 1 | 3,4 | 1 |
| 3 | 1 | 0 | - | 0 |
| 4 | 1 | 0 | 4,9 | 1 |
| 5 | 1 | 1 | - | 0 |
| 6 | 0 | 0 | - | 0 |

Com isso, o valor estimado de $q_{y_i x_{1i}} = Pr(r_i = 0|y_i, x_{1i})$ é dado por:

- $\hat{q}_{00} = 1/6$, pois quando $r = 0$, temos um único caso onde $y = 0$ e $x_1 = 0$, em um total de 6 casos;
- $\hat{q}_{01} = 0/6$, pois quando $r = 0$, temos zero casos onde $y = 0$ e $x_1 = 1$, em um total de 6 casos;

E, seguindo o mesmo raciocínio temos que:

- $\hat{q}_{10} = 1/6$;
- $\hat{q}_{11} = 1/6$.

Esta estimação, segundo Didelez (2002), só é possível quando $y$ e $x_1$ são variáveis discretas.

Nas Seções seguintes, mostramos como estimar os parâmetros via método de máxima verossimilhança para conjunto de dados completos e para conjunto de dados incompletos.

### 4.1.1 O estimador para dados completos

Como mencionado anteriormente, um dos métodos mais utilizados para estimar parâmetros de um modelo de regressão linear é o método de mínimos quadrados. Sob suposições usuais (normalidade nos erros com média zero e variância constante) o método em questão fornece estimadores não viciados e não consistentes. Porém, quando o método de mínimos quadrados é aplicado a um modelo que possui respostas binárias, os estimadores não apresentam as mesmas propriedades.

Então, um possível método de estimação para o modelo de regressão logística é o método de máxima verossimilhança, onde os coeficientes são estimados de modo a maximizar a probabilidade de se obter o conjunto de dados observado a partir do modelo proposto. Para o método ser aplicado, constrói-se uma função chamada "função de verossimilhança" que expressa a probabilidade dos dados observados como função dos parâmetros $\beta_0, \beta_1, ..., \beta_p$, sendo os estimadores aqueles que maximizam o valor desta função.

Descrevemos abaixo como ajustar esses valores para o modelo de regressão logística proposto na função (3.8), ou seja, na função

$$Pr(y = 1|X_1 = x_1, X_2 = x_2; \boldsymbol{\beta}) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)},$$

assumindo, exclusivamente nesta subseção, que todas as variáveis são completamente observadas (inclusive $x_2$), $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ e $\mathbf{x}' = (x_1, x_2)$.

A função acima fornece-nos a probabilidade condicional de $y$ ser igual a 1 dado $\mathbf{x}$, logo $1 - Pr(y = 1|x_1, x_2)$ fornece a probabilidade de $y$ ser igual a zero dado $\mathbf{x}$. Considerando $P(y = 1|x_1, x_2) = \pi(\mathbf{x})$, temos que $y$ segue uma distribuição Bernoulli com probabilidade $\pi(\mathbf{x})$ (vide Seção 3.1), sendo a função de verossimilhança dada por

$$L(\boldsymbol{\beta}; \mathbf{x}) = f(\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}, \tag{4.1}$$

com $y_i = 0, 1$ independentes, $x_i = \{x_{1i}, x_{2i}\}$ e $i = 1, 2, ..., n$.

Os estimadores pretendidos são os que maximizam a função (4.1). No entanto, o trabalho é matematicamente facilitado se aplicarmos o logaritmo natural, obtendo então

$$l(\boldsymbol{\beta}) = ln[L(\boldsymbol{\beta}; \mathbf{x})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}. \tag{4.2}$$

E, para obter os valores de $\boldsymbol{\beta}$ que maximizam (otimizam) a função (4.2), calculamos a derivada em relação a cada um dos parâmetros. No caso do modelo utilizado, calculamos em relação a $\beta_0, \beta_1, \beta_2$, ou seja,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = 0; \qquad \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = 0; \qquad \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_2} = 0,$$

obtendo as seguintes equações

$$\sum_{i=1}^{n}(y_i - \pi(x_i)) = 0, \ \textit{para encontrarmos } \hat{\beta}_0;$$

$$\sum_{i=1}^{n} x_{1i}(y_i - \pi(x_i)) = 0, \ \textit{para encontrarmos } \hat{\beta}_1;$$

$$\sum_{i=1}^{n} x_{2i}(y_i - \pi(x_i)) = 0, \ \textit{para encontrarmos } \hat{\beta}_2,$$

as quais uma vez solucionadas através de métodos numéricos, como por exemplo *Newton Raphson* ou Quasi-*Newton*, fornecem as estimativas de máxima verossimilhança. No *help* do *Software* SAS 9.0 temos várias técnicas de otimização para os procedimentos NLP e LOGISTIC que têm como objetivo maximizar a função de interesse. Por exemplo:

- No procedimento NLP temos que para um tamanho amostral (n) menor ou igual a 40 e para o método de mínimos quadrados não-linear usamos a técnica (de otimização) de *Newton-Raphson*, por *default*. Para amostras de tamanho $40 < n < 400$, usa-se, por *default*, o método de Quasi-*Newton*. Já, para $n \geq 400$ usa-se a técnica de *Conjugate Gradient*.

- No procedimento LOGISTIC temos as técnicas de *Newton-Raphson* e *Fisher-Scoring*. Ambas resultam nas mesmas estimativas, porém as matrizes de covariância estimadas são diferentes, exceto quando a função de ligação Logito é especificada para dados com resposta binária. A técnica *default* é a de *Fisher-Scoring*.

Os estimadores de máxima verossimilhança são os que têm, normalmente, as propriedades mais adequadas para um estimador.

Uma aplicação para este método é dada no Capítulo 4.

### 4.1.2 O estimador para dados *missing*

Em geral, não há diferença entre a estimação de máxima verossimilhança para dados completos e para dados incompletos.

Considere, segundo Didelez (2002), a seguinte função de verossimilhança gerada por $(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2}, \mathbf{r})$:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} [f(x_{1i}, x_{2i}, y_i, r_i; \beta, \theta)] \\ &= \prod_{i=1}^{n} [f(x_{1i}|\alpha) f(x_{2i}|x_{1i}; \xi) Pr(y_i|x_{1i}, x_{2i}; \beta) f(r_i|y_i, x_{1i}, x_{2i}; \gamma)] \end{aligned} \tag{4.3}$$

onde $\theta = (\alpha, \xi, \gamma)$, $x_2$ é variável aleatória com dados faltantes e $y$ e $x_1$ conforme definidos anteriormente. Os parâmetros $\alpha, \xi, \gamma$ se referem, respectivamente, à distribuição marginal de $x_1$, à distribuição condicional de $x_2$ dado $x_1$ e à distribuição condicional de $r$ dado $y$, $x_1$ e $x_2$ que é Bernoulli com probabilidade $q_{yx_1}$.

Seguindo:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[ f(x_{1i}|\alpha) \underbrace{f(x_{2i}|x_{1i}; \xi)}_{MAR} Pr(y_i|x_{1i}, x_{2i}; \beta) \underbrace{f(r_i|y_i, x_{1i}; \gamma)}_{MAR} \right] \\ &= \prod_{i=1}^{n} \left[ f(x_{1i}|\alpha) f(r_i|y_i, x_{1i}; \gamma) \left\{ \underbrace{Pr(y_i|x_{1i}, x_{2i}; \beta) f(x_{2i}|x_{1i}; \xi)}_{x_{2i}\ observado} \right\}^{r_i} \right. \\ &\qquad \left. \times \left\{ \underbrace{Pr(y_i|x_{1i}; \beta)}_{x_{2i}\ missing} \right\}^{1-r_i} \right] \\ &= \prod_{i=1}^{n} [f(x_{1i}|\alpha) f(r_i|y_i, x_{1i}; \gamma) \{Pr(y_i|x_{1i}, x_{2i}; \beta) f(x_{2i}|x_{1i}; \xi)\}^{r_i} \\ &\qquad \times \left\{ \underbrace{\frac{Pr(y_i, x_{1i}; \beta)}{f(x_{1i})}}_{(*)} \right\}^{1-r_i} \Bigg] \end{aligned}$$

onde $(*)$ é desenvolvido da seguinte forma:

$$\frac{Pr(y_i, x_{1i}; \beta)}{f(x_{1i})} = \int \frac{Pr(y_i, x_{1i}, z; \beta)}{f(x_{1i})} dz = \int \frac{Pr(y_i|x_{1i}, z; \beta) f(x_{1i}, z; \xi)}{f(x_{1i})} dz$$

$$= \int Pr(y_i|x_{1i}, z; \beta) f(z|x_{1i}; \xi) dz.$$

sendo $z$ uma variável de entrada.

Portanto,

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{n} [f(x_{1i}|\alpha) f(r_i|y_i, x_{1i}; \gamma) \{Pr(y_i|x_{1i}, x_{2i}; \beta) f(x_{2i}|x_{1i}; \xi)\}^{r_i} \times \left\{ \int Pr(y|x_{1i}, z; \beta) f(z|x_{1i}; \xi) dz \right\}^{1-r_i}], \quad (4.4)$$

onde $f$ é usado como um símbolo genérico para uma densidade qualquer.

Construída a função de verossimilhança (4.4), o próximo passo seria aplicar o logaritmo natural e maximizar em relação aos parâmetros de interesse, porém não conseguimos fazer isso sem o conhecimento de $f(\cdot|x_1; \xi)$.

Especificando $f(\cdot|x_1; \xi)$ para um parâmetro $\xi$ desconhecido, podemos maximizar (4.4) em relação a $\boldsymbol{\beta}$ e $\xi$, simultaneamente. E, como não temos interesse nos parâmetros $\alpha$ e $\gamma$, a função de verossimilhança pode ser reescrita da seguinte forma:

$$L(\boldsymbol{\beta}, \xi) = \prod_{i \in v} Pr(y_i|x_{1i}, x_{2i}; \boldsymbol{\beta}) f(x_{2i}|x_{1i}; \xi) \prod_{j \in \bar{v}} \int Pr(y_j|x_{1j}, z; \boldsymbol{\beta}) f(z|x_{1j}; \xi) dz, \quad (4.5)$$

com $v$ e $\bar{v}$ definidos no início da presente Seção.

Em geral, a maximização da função acima tem que ser feita numericamente devido à integração no segundo produtório. Isto pode ser, em partes, simplificado pelo uso do algoritmo EM ("Expectation Maximization"), como feito em Didelez(2002), ou ainda pelo uso de algum tipo de Quadratura Gaussiana com N pontos (Legendre ou Laguerre, por exemplo).

O algoritmo EM pode ser caracterizado como um método genérico de estimação de parâmetros por máxima verossimilhança para um conjunto de dados incompletos, sempre buscando uma maneira simples de se obter os estimadores de máxima verossimilhança quando esta, originalmente, é complicada, ou ainda, quando o parâmetro de interesse não é diretamente observável somente com a amostra disponível. Quando existem dados *missing* no conjunto de dados originais, o algoritmo, em uma etapa específica, entra para "completar" este conjunto de dados e assim, permitir a aplicação do método (Little *et al.*, 1987).

Podemos dizer que o algoritmo tem um caráter iterativo e pode claramente ser dividido em duas etapas, uma sendo o cálculo da Esperança (etapa E) e outra a maximização (etapa M) de uma função de verossimilhança em cada uma de suas iterações. Segundo Park (2005), uma vantagem deste método comparado a outras técnicas de otimização é a facilidade de sua construção e a convergência quase certa para o valor real. No entanto, *Burkett (2002)* afirma que o passo $E$ do algoritmo EM para modelos de regressão logística com covariáveis *missing* não é um processo tão simples.

Neste trabalho, fizemos uso da Quadratura Gaussiana como forma de aproximação da integral presente na função de verossimilhança (4.5). Em linhas gerais, a Quadratura é um método que discretiza a integral e retorna dados completos ponderados.

Na Seção abaixo, descrevemos o uso da Quadratura Gaussiana. Para maiores detalhes, ver *Einwoegerer (2006)*.

## 4.2 Uso da Quadratura Gaussiana

Considere uma função contínua em um intervalo $[a, b]$ e sua primitiva $F(z)$ como sendo conhecida. A integral definida desta função no intervalo definido acima é dada por

$$\int_a^b f(z)dz = F(b) - F(a), \tag{4.6}$$

onde $F'(z) = f(z)$.

Porém, em alguns casos, o valor da primitiva $F(z)$ não é conhecido ou não é fácil de obter, dificultando ou impossibilitando assim o cálculo dessa integral. Em situações práticas, a função a ser integrada geralmente não possui uma fórmula analítica, mas sim uma tabela de pontos tornando inviável a utilização da equação (4.6).

Com isso torna-se necessário o uso de técnicas numéricas para se calcular o valor da integral de $f(z)$ nas duas situações citadas anteriormente. Em poucas palavras, a solução numérica de uma integral simples é dita Quadratura.

Os métodos de resolução mais utilizados são:

1. As fórmulas de Newton-Côtes que fornecem valores a $f(z)$, onde os valores de $z$ são igualmente espaçados. Exemplos: Regra do Trapézio e Regra de Simpson.

2. A fórmula de Quadratura Gaussiana que utiliza pontos diferentemente espaçados, sendo este espaçamento determinado por meio de certas propriedades de polígonos ortogonais. Exemplos: Quadratura de Gauss Legendre, Quadratura de Gauss Laguerre.

Dos métodos de resolução mencionados, vamos nos deter na fórmula de Quadratura Gaussiana (ou fórmula de Gauss).

A fórmula de Gauss para o cálculo da integral numérica fornece um resultado bem mais preciso que as fórmulas de Newton Côtes para um número semelhante de pontos. Na aplicação da Quadratura Gaussiana, os pontos não são mais definidos pelo analista que utiliza o método, e sim por um critério definido.

O método de integração aproximada consiste em aproximar uma integral por uma combinação linear de valores da função integranda, ou seja,

$$\int_a^b W(z)f(z)dz \approx \sum_{h=0}^{k-1} w_h f(z_h), \tag{4.7}$$

com $-\infty \leq a < b \leq \infty$ e $a \leq z_h \leq b$. Os pontos $z_h$ (dito abscissas ou raízes), com $h = 0, 1, ..., k$, são usualmente pontos do intervalo de integração, os números $w_h$ os respectivos pesos e $k$ o número de nós.

Para muitas funções, os pesos e as abcissas já encontram-se tabelados e presentes na Literatura, tipo Einwoegerer (2006), Carvalho (2000) e em *Softwares*, como o R 2.7.1.

Os pontos, como dito anteriormente, não são igualmente espaçados, mas sim escolhidos de forma que os $k$ valores apropriadamente ponderados resultem numa integral exata quando $f(z)$ é polinômio de grau $2k + 1$ ou menor (Einwoegerer, 2006).

Quando $f(z)$ não é polinômio, a aproximação dada em (4.7) não é exata, logo deve-se incluir um fator de correção específico para cada tipo de Quadratura Gaussiana. Este fator de correção é descrito no Capítulo seguinte.

A escolha de qual Quadratura usar é definida de acordo com os limites de integração e com a função peso, dada por $W(z)$.

**Exemplo 4.2.** Se $a = -1, b = 1$ e $W(z) = 1$, usamos a Quadratura de Gauss Legendre. Agora, se $a = 0, b = \infty$ e $W(z) = exp(-z)$, usamos a Quadratura de Gauss Laguerre.

No Capítulo 4, apresentamos uma aplicação para a Quadratura de Gauss Laguerre.

## 4.3 Trabalhando com dados *missing*

Nesta Seção apresentamos algumas técnicas para se trabalhar com conjunto de dados onde há indivíduos com falta de informação. Dentre elas: Caso Completo, Caso Completo Corrigido e Imputação Múltipla.

Na Seção anterior, falou-se sobre o método de Quadratura Gaussiana que, em um de seus passos, utiliza os valores pré definidos, no lugar dos dados que eram *missing*.

### 4.3.1 Análise de Caso Completo

Segundo Little (1992), o tratamento padrão usado em pacotes estatísticos quando há *missing* no banco de dados é a Análise de Caso Completo (CC), onde simplesmente descartamos os casos com quaisquer dados faltantes. É também conhecido como *listwise* ou *pairwise deletion*, é de fácil implementação e consiste em aplicar métodos de valores completos a um conjunto reduzido de dados. Porém, ao descartar os casos incompletos podemos perder informações que nem sempre são consideradas desprezíveis; isto depende muito do tamanho da amostra, do número de dados *missing* e do tipo de informação perdida. Parece razoável então explorar caminhos para incorporar os casos incompletos dentro da análise.

Uma preocupação crucial é se a seleção dos dados completos (descartando os dados *missing*) nos leva a estimadores viciados. Sob a suposição MCAR, os casos completos são efetivamente uma amostra aleatória da amostra original, logo o descarte dos dados incompletos não torna os estimadores viciados. Porém, se tivermos um conjunto de dados com presença de *missing* e retirarmos uma amostra aleatória desses dados, a probabilidade da amostra ser constituída apenas de dados completos é mínima. Por esta razão, dizemos que a natureza dos vícios também depende do mecanismo *missing* (MAR, MCAR ou MNI) utilizado na análise.

Segundo Little *et al.* (1987), se os dados completos formam uma amostra aleatória da amostra original, ou seja, se MCAR é uma suposição razoável, as informações descar-

tadas podem ser usadas para estudo. Um procedimento simples é a comparação da distribuição de uma variável particular $X_j$ baseada nos dados completos com a distribuição de $X_j$ baseada nos casos incompletos.

Após a escolha do mecanismo a ser utilizado, o próximo passo é estimar os parâmetros de interesse. Um dos possíveis métodos para estimar os parâmetros é o de mínimos quadrados em que a soma de quadrados dos resíduos deve ser mínima. Note que qualquer conjunto de dados incompletos pode determinar um zero residual por escolhas apropriadas desses valores, ou seja, se escolhermos "a dedo" quais valores serão *missing*, conseguiremos obter um zero residual, conforme dito em Little(1992).

A seguir, o exemplo 4.3 apresenta um caso cujo objetivo é comparar os parâmetros estimados em um conjunto com ausência de *missing* e outro com presença de *missing*.

**Exemplo 4.3.** *Modelo de Regressão Linear*

O conjunto de dados descritos na Tabela 4.1, extraído do censo do IBGE de 2000, presente em Paula (2004), apresenta o número médio de anos de estudo e a renda média mensal (em reais) do chefe ou chefes do domicílio de vários estados do Brasil.

Seja $W$ a variável resposta referente à renda média e $X$ a covariável escolaridade, conforme mostra a tabela abaixo:

TABELA 4.1: Escolaridade e renda média domiciliar no Brasil

| i | X | W | i | X | W | i | X | W |
|---|---|---|---|---|---|---|---|---|
| 1 | 5,7 | 685 | 10 | 5,7 | 722 | 19 | 4,5 | 513 |
| 2 | 4,5 | 526 | 11 | 6,3 | 814 | 20 | 3,5 | 383 |
| 3 | 4,7 | 536 | 12 | 6,0 | 782 | 21 | 4,6 | 517 |
| 4 | 4,5 | 520 | 13 | 5,5 | 689 | 22 | 4,0 | 448 |
| 5 | 3,6 | 343 | 14 | 8,2 | 1499 | 23 | 7,1 | 970 |
| 6 | 4,3 | 462 | 15 | 6,0 | 683 | 24 | 5,4 | 681 |
| 7 | 4,1 | 460 | 16 | 4,9 | 662 | 25 | 6,4 | 800 |
| 8 | 3,7 | 454 | 17 | 5,5 | 627 | 26 | 5,4 | 775 |
| 9 | 6,8 | 1076 | 18 | 3,9 | 423 | 27 | 5,7 | 731 |

Construindo o gráfico de dispersão de $X$ por $Y = log(W)$ (Figura 4.1) observamos uma possível relação linear entre as variáveis, uma vez que os pontos se aproximam de

uma reta. Logo, o modelo pode ser ajustado por um modelo de regressão linear dado por $Y = \beta_0 + \beta_1 X$.

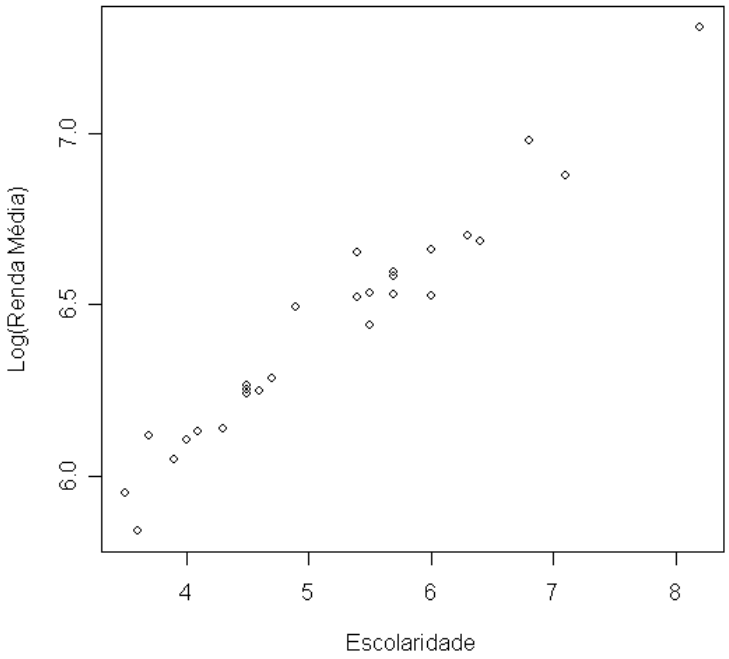


FIGURA 4.1: Gráfico de Dispersão entre as variáveis Escolaridade e Log(Renda Média)

Os parâmetros de regressão $\beta_0$ e $\beta_1$ são constantes desconhecidas e são estimados a partir dos dados amostrais. Neste caso de regressão linear usamos o método de mínimos quadrados que tem como objetivo minimizar os resíduos ($\hat{\epsilon}$), ou seja, minimizar o comprimento do vetor $\hat{\epsilon} = (\hat{\epsilon}_1, \hat{\epsilon}_2, ..., \hat{\epsilon}_{27})$, no presente caso.

Sabendo que os estimadores, via mínimos quadrados, são dados por

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}, \tag{4.8}$$

aplicamos os dados de interesse na equação (4.8) e obtemos os seguintes resultados:

| | Parâmetro Estimado | Erro Padrão | t-valor | p-valor |
|---|---|---|---|---|
| $\beta_0$ | 4,9819 | 0,0672 | 74,11 | $< 2 \times 10^{-16}$ |
| $\beta_1$ | 0,2790 | 0,0126 | 22,11 | $< 2 \times 10^{-16}$ |

A análise acima nos mostra que a covariável é significativa para o modelo e que o modelo ajustado é dado por $\hat{y} = 4,9819 + 0,2790x$.

Considere agora, que os indivíduos $1, 4, 5, 10, 14, 15, 17, 18, 20, 21, 22, 23, 24$ e $27$ apresentam valores *missing* na variável $X$, num total de 51% de dados faltantes.

Pelo método de Caso Completo, desconsideramos os $y's$ correspondentes aos valores *missing* obtendo assim, um novo banco de dados com 13 observações completas. Construindo novamente o gráfico de dispersão desse banco de dados, notamos uma possível relação linear entre as variáveis, indicando, como anteriormente, um modelo de regressão linear. Estimando os parâmetros para esse novo banco, obtemos os seguintes valores:

| | Parâmetro Estimado | Erro Padrão | t-valor | p-valor |
|---|---|---|---|---|
| $\beta_0$ | 5,0577 | 0,1214 | 41,67 | $1,85 \times 10^{-13}$ |
| $\beta_1$ | 0,2703 | 0,0231 | 11,69 | $1,53 \times 10^{-7}$ |

E, tendo novamente a covariável significativa, o modelo ajustado é dado por $\hat{y} = 5,0577 + 0,2703x$.

Agora, a fim de comparar os valores dos parâmetros estimados do conjunto de dados completos com o conjunto de dados *missing*, calculamos a diferença absoluta entre eles para sabermos quão longe estão um do outro. Veja os resultados abaixo:

| | Dados completos | Dados com *mising* | Diferença |
|---|---|---|---|
| $\hat{\beta}_0$ | 4,9819 | 5,05765 | 0,08 |
| $\hat{\beta}_1$ | 0,2790 | 0,27031 | 0,009 |

Note que não existe muita diferença entre as estimativas dos parâmetros. Isso ocorre porque ao retirar pontos (dados *missing*) do banco de dados original continuamos com uma relação linear entre as variáveis muito parecida com a relação linear inicial (dados completos). Isso pode ser observado no seguinte gráfico:

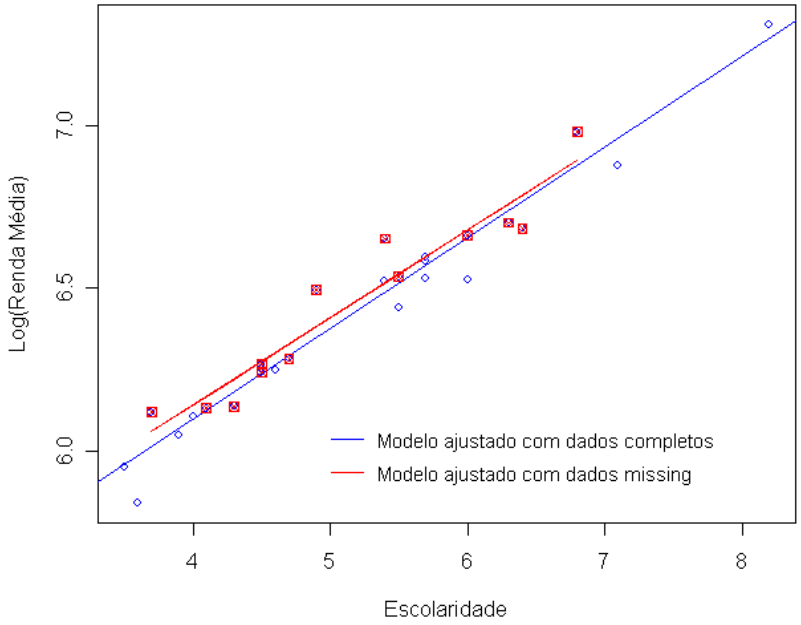


FIGURA 4.2: Gráfico dos Modelos Ajustados com os dados observados nas amostras de tamanhos 27 e 13

Lembrando que $n = 27$ é o tamanho da amostra inicial (sem dados *missing*) e $n = 13$ é o tamanho da amostra após a retirada de 14 casos que tinham *missing* na variável $X$.

Com isso, vemos que este é um caso onde os valores *missing* não provocam muitas modificações nos parâmetros do modelo estimado.

Little (1992), em seu trabalho sobre dados *missing*, mencionou um método de imputação em que os valores *missing* eram tratados como sendo parâmetros, estimando-os como tais. Esta técnica foi considerada uma estratégia pobre pelo próprio autor.

Porém, apenas como ilustração da estratégia de Little, consideremos os dados *missing*, definidos acima, como sendo parâmetros. Para estimá-los fazemos o seguinte

$$\left.\frac{\partial S}{\partial x_j}\right|_{\hat{x_j}\hat{\beta}_0,\hat{\beta}_1} = 0 \Longrightarrow -2(y_j - \hat{\beta}_0 - \hat{\beta}_1\hat{x}_j)\hat{\beta}_1 = 0, \tag{4.9}$$

onde

$$S = \|\hat{\epsilon}\|^2 = \sum_{j\in\Delta}(\hat{\epsilon}_j)^2 = \sum_{j\in\Delta}(y_j - \beta_0 - \beta_1 x_j)^2,$$

sendo $\Delta = \{1, 4, 5, 10, 14, 15, 17, 18, 20, 21, 22, 23, 24, 27\}$.

E, resolvendo as equações normais dadas em (4.9), encontramos os seguintes valores estimados para os $x's$ *missing*:

| $x_1$ | $x_4$ | $x_5$ | $x_{10}$ | $x_{14}$ | $x_{15}$ | $x_{17}$ |
|---|---|---|---|---|---|---|
| 5,4447 | 4,4252 | 2,8859 | 5,6394 | 8,3419 | 5,4339 | 5,118 |

| $x_{18}$ | $x_{20}$ | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{27}$ |
|---|---|---|---|---|---|---|
| 3,6614 | 3,2939 | 4,4038 | 3,8739 | 6,7317 | 5,4231 | 5,6852 |

Após a estimação, imputamos esses valores nos lugares dos dados *missing*, obtendo assim, um conjunto sem presença de dados faltantes, ou seja, com 27 casos. Seguindo a análise, as novas estimativas são dadas por

| | Parâmetro Estimado | Erro Padrão | t-valor | p-valor |
|---|---|---|---|---|
| $\beta_0$ | 5,0577 | 0,0440 | 114,95 | $2 \times 10^{-16}$ |
| $\beta_1$ | 0,2703 | 0,0084 | 32,13 | $2 \times 10^{-16}$ |

Logo a variável $x$ é significativa ao modelo, sendo este dado por $\hat{y} = 5,0577 + 0,2703x$. Note a semelhança entre este modelo, obtido com uma amostra de tamanho 27 e o modelo obtido com a amostra incompleta. Logo o método de imputação utilizado criou um modelo igual ao modelo obtido com a amostra reduzida. Isto vai de encontro com a afirmação de Little (1992) que disse ser este processo de imputação uma estratégia pobre.

**Exemplo 4.4.** *Modelo de Regressão Logística*

Neste exemplo trabalhamos com dados simulados.

Os resultados apresentados aqui são valores médios referentes aos 1.000 conjuntos de dados de tamanho amostral 300 simulados. Para maiores detalhes da geração dos dados, ver Capítulo 5.

Todos os 1.000 conjuntos de dados simulados são da seguinte forma:

TABELA 4.2: Conjunto de dados com 300 indivíduos

| i | y | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 1 | 1 | 2,4297 |
| 2 | 1 | 1 | 0,9057 |
| 3 | 0 | 0 | 0,7900 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 298 | 1 | 0 | 0,1636 |
| 299 | 1 | 0 | 2,0238 |
| 300 | 1 | 1 | 1,6896 |

sendo $x_1$ e $x_2$ variáveis independentes e $y$ a variável dependente.

Como podemos observar, a variável resposta é dicotômica, logo um modelo de regressão logística dado por

$$P(y = 1|\mathbf{x}) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

pode ser utilizado.

Estimando os parâmetros de interesse via método de máxima verossimilhança (apresentado de forma detalhada na Seção 4.1.1), obtemos as seguintes estimativas:

TABELA 4.3: Estimativa dos parâmetros da amostra completa (sem *missing*)

| Amostra | Parâmetro | Estimativa | Variancia | Vício | EQM |
|---|---|---|---|---|---|
| | $\beta_0$ | -0,04906 | 0,09850 | 0,00241 | 0,10090 |
| 300 | $\beta_1$ | 1,08518 | 0,26618 | 0,00726 | 0,27344 |
| | $\beta_2$ | 1,61057 | 0,11811 | 0,01223 | 0,13033 |

onde o erro quadrático médio estimado (EQM) é dado por:

$$EQM = \frac{\sum(\hat{\beta}_i - \bar{\hat{\beta}})^2}{B - 1} + (\bar{\hat{\beta}} - \beta_{real})^2$$

sendo $B = 1.000$ (número de replicações) e $\bar{\hat{\beta}}$ a média dos parâmetros estimados.

Cabe mencionar que os parâmetros reais usados na geração dos dados são iguais a $0,1$ e $1,5$ para, respectivamente, $\beta_0, \beta_1$ e $\beta_2$.

Considere agora os seguintes percentuais de dados *missing* na variável $x_2$, 10% e 50% de 300. Com isso, criamos dois novos conjuntos de dados com 270 e 150 indivíduos com informações faltantes, respectivamente. Vale ressaltar que os casos considerados *missing* foram escolhidos de forma aleatória.

Seguindo o mesmo procedimento do Exemplo anterior, desconsideramos os indivíduos que apresentam dados *missing* na variável $x_2$ e estimamos os parâmetros a partir do conjunto de dados completos (sem os valores *missing*). As seguintes estimativas dos parâmetros, erros quadráticos médios, variâncias e vícios são dadas como seguem:

TABELA 4.4: Estimativa dos parâmetros para 10% de *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,04891 | 0,11229 | 0,00239 | 0,11468 |
| | $\beta_1$ | 1,08398 | 0,28862 | 0,00705 | 0,29567 |
| | $\beta_2$ | 1,61985 | 0,13815 | 0,01436 | 0,15251 |

TABELA 4.5: Estimativa dos parâmetros para 50% de *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,06144 | 0,14105 | 0,00377 | 0,14482 |
| | $\beta_1$ | 1,12775 | 0,78013 | 0,01632 | 0,79645 |
| | $\beta_2$ | 1,65089 | 0,18969 | 0,02277 | 0,21245 |

Observe que os maiores erros são obtidos quando há 50% de casos *missing* na amostra, conforme esperado. Note também que as três métricas (erro quadrático médio, variância e vício) aumentam conforme o número de *missing* aumenta.

Diferentemente dos erros obtidos no Exemplo anterior, onde ajustamos um modelo de regressão linear, os erros apresentados nas Tabelas 4.3, 4.4 e 4.5 possuem uma maior diferença entre si do que a mostrada no exemplo anterior.

Neste caso, os erros nos informam que ao excluir os casos *missing* da análise podemos perder informações importantes.

No Capítulo 5 entramos em maiores detalhes sobre a estimação dos parâmetros em modelos de regressão logística.

### 4.3.2 Caso Completo Corrigido

Segundo Didelez(2002), o estimador obtido pelo método de Caso Completo Corrigido (CCC) pode ser viciado quando consideramos a suposição MAR. Este estimador é composto pela estimativa obtida via estimador de Caso Completo mais um fator de correção que leva em conta a proporção de dados *missing* presente no conjunto de dados.

Na referência citada acima, Didelez define o estimador de caso completo corrigido para o modelo de regressão logística com duas covariáveis, da seguinte forma:

$$\hat{\beta_0}^{CCC} = \hat{\beta_0}^{CC} + log\frac{\hat{q_{00}}}{\hat{q}_{10}} \tag{4.10}$$

$$\hat{\beta_1}^{CCC} = \hat{\beta_1}^{CC} + log\frac{\hat{q}_{10}\hat{q}_{01}}{\hat{q}_{00}\hat{q}_{11}} \tag{4.11}$$

$$\hat{\beta_2}^{CCC} = \hat{\beta_2}^{CC} \tag{4.12}$$

Para maiores detalhes sobre este estimador, ver Vach *et al.* (1997).

Note que estes estimadores utilizam as observações incompletas se a correção dos termos usa $\hat{q}_{yx_1}$, onde $\hat{q}_{yx_1}$ é dado como a proporção da unidade amostral com valores $y, x_1$ e os *missing* $x_2$ sobre todos os valores de $y$ e $x_1$, ver Seção 4.1.

Uma aplicação para este estimador foi dada no Capítulo 4.

## 4.4 Imputação de Dados

Os métodos de imputação de dados são classificados em dois tipos: imputação simples e imputação múltipla. Nesta Seção, apresentamos algumas técnicas presentes em cada um dos métodos de imputação.

### 4.4.1 Imputação simples

Um dos métodos de imputação simples mais conhecido é a imputação pela média. Neste método ocorre a substituição de cada valor *missing* por um único valor. Esse valor

é a média da variável considerando apenas os casos completos. Por exemplo, considere $x_2$ uma variável com n= 20, onde cinco desses indivíduos são *missing*; seguindo a técnica, estes cinco casos são substituídos pela média dos quinze valores observados. Pode-se também substituir os valores *missing* pela média condicional nos valores observados de outras variáveis. E, com os dados já imputados, este método trata o conjunto de dados como se fosse completo (sem *missing*), seguindo com as análises necessárias normalmente.

Porém, a imputação simples não reflete a incerteza sobre as predições de um valor *missing*. E a variância estimada resultante dos parâmetros estimados relacionados à variável que possui *missing* é influenciada, tendendo a zero.

### 4.4.2 Imputação Múltipla

Diferentemente da imputação simples, a imputação múltipla não estima cada valor *missing* através da simulação de valores. A imputação múltipla substitui cada valor *missing* por um conjunto de valores plausíveis que representam a incerteza sobre o valor certo a ser imputado. O conjunto de imputações múltiplas é então analisado utilizando procedimentos padronizados para dados completos e combinações dos resultados dessas análises. Não importa qual análise dos dados completos é usada, o processo de combinação dos resultados de diferentes conjuntos de dados é essencialmente o mesmo.

Segundo Giacon (2007), a inferência na imputação múltipla envolve três fases distintas:

- Os dados *missing* são completados $m$ vezes para gerar $m$ conjuntos de dados completos. Alguns dos possíveis métodos de imputação estão listados na Tabela 4.6.

- Os $m$ conjuntos de dados completos são analisados através do uso de procedimentos padronizados.
- Os resultados dos $m$ conjuntos de dados completos são combinados para inferência.

Nesta Seção listamos alguns métodos para imputação múltipla disponíveis no procedimento MI e MIANALYZE do *Software* SAS 9.0.

PROC MI é um procedimento de imputação múltipla que cria múltiplos conjuntos de dados imputando os dados incompletos. Faz isto utilizando métodos que incorporam apropriadamente a variabilidade através de $m$ imputações. Uma vez que os $m$ conjuntos de dados são analisados usando procedimentos padronizados, outro novo procedimento, PROC MIANALYZE, é usado para gerar inferências estatísticas válidas sobre estes parâmetros através dos resultados combinados dos $m$ conjuntos de dados completos. Ou seja, a partir das $m$ imputações, $m$ diferentes conjuntos de dados são computados e, por meio do PROC MIANALYZE, são combinados, gerando assim, inferências estatísticas válidas sobre os parâmetros.

Existem vários métodos de imputação disponíveis no procedimento MI. O método de escolha depende do modelo de dados *missing* e do tipo de variável a ser imputada. Veja alguns na tabela abaixo:

TABELA 4.6: Métodos de Imputação em PROC MI

| Modelo *Missing* | Tipo de Variável Imputada | Método Recomendado |
|---|---|---|
| Monótono | Contínua | Regressão Simples |
| Monótono | Categórica (Nominal) | Método Função Discriminante |
| Monótono | Categórica (Ordinal) | Regressão Logística |
| Arbitrário | Contínua | MCMC |

Fonte: SAS Institute Inc. (2002)

Sendo modelo de *missing* monótono como definido na Seção 2.1.

Na Tabela 4.6 vemos que para utilizar o método de Regressão Simples além de termos normalidade nos dados, devemos ter modelo *missing* monótono e a variável com ausência de informação ser contínua. Mas, como o modelo adotado neste trabalho não segue algumas destas condições, não podemos utilizar o método de Regressão Simples.

Para imputar dados em uma variável categórica e tendo modelo *missing* monótono, podemos usar o método de regressão logística ou o método de função discriminante, dependendo do tipo da variável imputada (ordinal ou nominal).

Já para variáveis contínuas em um conjunto de dados com modelo *missing* arbitrário, usamos o método da Cadeia de Markov Monte Carlo (MCMC) tanto para a imputação de valores que são *missing* quanto para fazer com que o conjunto de dados

adquira um modelo de valor *missing* monótono. O método MCMC pode ser aplicado em nosso modelo de regressão a fim de transformar o modelo *missing* em monótono, visto que é univariado (arbitrário em uma variável) sendo a variável imputada, contínua.

Com o modelo monótono temos maior flexibilidade na escolha dos métodos de imputação, como visto na Tabela 4.6 acima.

Os métodos de Regressão Logística e MCMC são sugeridos para estudos futuros, como forma de complementação deste trabalho.

# Capítulo 5

# Simulação e Resultados

A simulação presente neste estudo fez uma comparação entre os métodos propostos para amostras de tamanho 300 e 500. Para estes tamanhos amostrais consideramos $5\%, 10\%, 30\%$ e $50\%$ de dados *missing*, em relação às amostras iniciais.

Cada conjunto de dados foi replicado com o uso do *Software SAS* 9.0, 1.000 vezes. Os parâmetros estimados, apresentados nas próximas páginas, são a média dos resultados obtidos em cada conjunto de tamanho amostral diferente. Por exemplo: considere 1.000 conjuntos de dados de tamanho amostral 270 (já com 10% de *missing* sobre a amostra de tamanho 300). Para cada um dos conjuntos de dados estimam-se os parâmetros de interesse ($\beta_0, \beta_1$ e $\beta_2$, na presente análise), obtendo assim, 1.000 parâmetros de cada tipo. Com isso, os valores dos parâmetros apresentados aqui são a média dos 1.000 estimadores, isto é,

$$\overline{\hat{\beta}} = \frac{\sum_{i=1}^{B} \hat{\beta}_i}{B}, \quad B = 1, 2, ..., 1.000$$

O mesmo ocorre para as seguintes métricas apresentadas: desvio padrão, variância, vício, erro quadrático médio, intervalo de confiança assintótico e intervalo de confiança empírico. As métricas variância, vício e erro quadrático médio estimado foram usadas como formas de comparação entre os métodos. Já os intervalos de confiança, apenas como forma de complementação da análise.

A variável resposta do conjunto de dados simulados foi obtida considerando os seguintes Passos:

**Passo 1:** Geramos $x_1$ de uma Bernoulli com probabilidade $0,4$.

**Passo 2:** Geramos $x_2|x_1$ de uma Qui Quadrado com 2 graus de liberdade.

**Passo 3:** Consideramos os valores dos parâmetros $\beta_0 = 0, \beta_1 = 1$ e $\beta_2 = 1,5$.

**Passo 4:** Substituímos as observações e os valores dos parâmetros no modelo abaixo, obtendo a probabilidade de $y|x_1, x_2$,

$$Pr(y|x_1, x_2; \beta) = \frac{exp\ (\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp\ (\beta_0 + \beta_1 x_1 + \beta_2 x_2)}. \tag{5.1}$$

**Passo 5:** Geramos $y$ de uma Bernoulli com a probabilidade encontrada no Passo 4. Obtendo assim a resposta, $y$, é do tipo binária, assumindo valores 0 ou 1.

Veja na Tabela abaixo um exemplo de geração da variável resposta, seguindo os cinco Passos descritos anteriormente:

TABELA 5.1: Exemplo de geração da variável resposta

| | $\beta_0 = 0$ | $\beta_1 = 1$ | $\beta_2 = 1,5$ | |
|---|---|---|---|---|
| i | $x_1$ | $x_2\|x_1$ | $p$ | $y$ |
| 1 | 1 | 1,91011 | 0,97947 | 1 |
| 2 | 0 | 9,77855 | 0,99999 | 1 |
| 3 | 1 | 2,15188 | 0,98562 | 1 |
| 4 | 0 | 0,68117 | 0,73531 | 1 |
| 5 | 1 | 1,03975 | 0,92821 | 1 |
| 6 | 0 | 2,48260 | 0,97642 | 1 |
| 7 | 0 | 0,93053 | 0,80151 | 1 |
| 8 | 1 | 6,05345 | 0,99995 | 1 |
| 9 | 1 | 2,72359 | 0,99285 | 1 |
| 10 | 1 | 2,92716 | 0,99546 | 1 |

Tendo os conjuntos de dados completos, criamos os conjuntos com dados *missing*. Para estes, retiramos, de forma aleatória, a quantidade de dados que queremos ser *missing*. Gerados os conjuntos de dados com e sem dados faltantes, consideramos os parâmetros ($\beta_0, \beta_1$ e $\beta_2$) como sendo desconhecidos e utilizamos algum método para estimá-los. Os estudos foram feitos em relação aos conjuntos de dados sem *missing* e com *missing*.

Neste Capítulo, apresentamos as estimativas dos parâmetros de interesse de acordo com as técnicas discutidas. Estas estimativas foram comparadas afim de obtermos quais os melhores e os piores métodos de estimação/imputação estudados, de acordo com o tamanho amostral e com o percentual de *missing*.

## 5.1 Conjuntos de dados completamente observados

Os conjuntos de dados de tamanhos 300 (seguindo o mesmo raciocínio para as amostras de tamanhos 500), com todos os indivíduos observados, são do tipo:

TABELA 5.2: Conjunto de dados de tamanho amostral 300

| $n$ | $y$ | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 1 | 1 | 7,43328 |
| 2 | 1 | 0 | 1,95072 |
| 3 | 0 | 0 | 0,10512 |
| 4 | 1 | 1 | 5,09924 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 297 | 1 | 1 | 2,76514 |
| 298 | 0 | 1 | 0,87523 |
| 299 | 0 | 1 | 1,19688 |
| 300 | 1 | 0 | 4,85405 |

sendo $y$ e $x_1$ variáveis categóricas e $x_2$ variável contínua, com as distribuições definidas no início do presente Capítulo.

Nesta Seção, consideramos apenas os conjuntos de dados completos, ou seja, os conjuntos em que todos os indivíduos (dos tamanhos amostrais considerados) possuem as variáveis $y, x_1$ e $x_2$ completamente observadas. A partir destes dados, estimamos os parâmetros do modelo de regressão logística dado em (5.1), via máxima verossimilhança.

A Tabela 5.3 apresenta as estimativas encontradas para as diferentes amostras, juntamente com a variância estimada, o vício estimado e o erro quadrático médio estimado de cada parâmetro.

TABELA 5.3: Parâmetros estimados para os diferentes tamanhos amostrais

| Amostra | Parâmetro | Estimativa | Variancia | Vício | EQM |
|---|---|---|---|---|---|
| 300 | $\beta_0$ | -0,04906 | 0,09850 | 0,00241 | 0,10090 |
| | $\beta_1$ | 1,08518 | 0,26618 | 0,00726 | 0,27344 |
| | $\beta_2$ | 1,61057 | 0,11811 | 0,01223 | 0,13033 |
| 500 | $\beta_0$ | -0,02821 | 0,07976 | 0,00080 | 0,08055 |
| | $\beta_1$ | 1,04709 | 0,11644 | 0,00222 | 0,11866 |
| | $\beta_2$ | 1,52082 | 0,06173 | 0,00043 | 0,06216 |

onde o erro quadrático médio estimado é dado pela soma da variância estimada (primeiro fator) com o vício estimado (segundo fator), isto é:

$$E\hat{Q}M = \frac{\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}})^2}{B-1} + (\bar{\hat{\beta}} - \beta_{real})^2 \tag{5.2}$$

sendo $B = 1.000$ (número de replicações).

Podemos verificar que conforme o tamanho amostral cresce, a variância, o vício e o erro quadrático médio estimado diminuem, ou seja, quanto maior a amostra, melhor as estimativas do conjunto de dados simulados.

Outra medida estatística usualmente calculada nas análises de dados é o intervalo de confiança. Nesta análise foram considerados o intervalo de confiança assintótico, dado pela expressão:

$$ICAssintótico = \hat{\beta} \pm 1,96\sqrt{Var(\hat{\beta})} \tag{5.3}$$

e o intervalo de confiança empírico definido pelos percentis $\alpha$ e $1-\alpha$ da amostra com B estimativas de parâmetros; sendo $\hat{\beta}^{(\alpha)}$ e $\hat{\beta}^{(1-\alpha)}$ os limitantes inferior e superior do intervalo considerado (Pereira, 2.000).

Em nosso trabalho consideramos $\alpha = 2,5\%$. Com isso os limitantes do intervalo são dados pelos $\hat{\beta}'s$ que pertencem às posições $2,5\%$ e $97,5\%$ do percentil.

O intervalo de confiança assintótico é construído com a suposição de normalidade para amostras grandes, já o empírico é construído apenas com base na distribuição da amostra de estimativas. Através desta medida podemos dizer se o parâmetro é ou não significativo para o modelo.

A Tabela 5.4 apresenta estes intervalos para as estimativas dos parâmetros.

TABELA 5.4: Intervalos de Confiança Assintótico e Empírico para amostra sem *missing*

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| 300 | $\beta_0$ | (-0,66419; 0,56607) | 1,23025 | (-0,65456; 0,55930) | 1,21386 |
| | $\beta_1$ | (0,07396; 2,09640) | 2,02245 | (0,17664; 2,41211) | 2,23547 |
| | $\beta_2$ | (0,93698; 2,28416) | 1,34719 | (0,93840; 2,50314) | 1,56474 |
| 500 | $\beta_0$ | (-0,58173; 0,52531) | 1,10705 | (-0,69358; 0,54676) | 1,24034 |
| | $\beta_1$ | (0,37826; 1,71592) | 1,33766 | (0,45631; 1,80521) | 1,34890 |
| | $\beta_2$ | (1,03386; 2,00778) | 0,97392 | (1,04545; 1,97357) | 0,92812 |

Analisando os intervalos acima, podemos concluir que o intercepto não é significativo para o modelo, nos dois tamanhos amostrais, pois o zero pertence ao intervalo (no assintótico e no empírico), o que era esperado, já que o parâmetro $\beta_0$ real é igual a zero. Já os parâmetros $\hat{\beta}_1$ e $\hat{\beta}_2$ são significativos, pois o zero não pertence aos respectivos intervalos. Notamos também que a amplitude dos intervalos de confiança empíricos são, na maioria dos casos, maiores que as amplitudes dos intervalos de confiança assintóticos. E que diminuem conforme o tamanho amostral aumenta.

## 5.2 Conjunto de dados com *missing*

A partir desta Seção, consideramos os conjuntos de dados de tamanhos amostrais 300 e 500 com percentuais de *missing* dados por $5\%, 10\%, 30\%$ e $50\%$.

Os casos com *missing* referem-se apenas à variável $x_2$, sendo $y$ e $x_1$ completamente observadas. Note abaixo o *layout* dos diversos conjuntos de dados:

TABELA 5.5: Conjunto de Dados Incompletos para amostra de tamanho 300

| $n$ | $y$ | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 1 | 1 | 7,43328 |
| 2 | 1 | 0 | - |
| 3 | 0 | 0 | 0,10512 |
| 4 | 1 | 1 | 5,09924 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 297 | 1 | 1 | 2,76514 |
| 298 | 0 | 1 | 0,87523 |
| 299 | 0 | 1 | - |
| 300 | 1 | 0 | 4,85405 |

Os casos apresentados na Tabela 5.5 com ausência de informação são representados por um traço (-).

Novamente usamos o método de máxima verossimilhança para encontrar as estimativas dos parâmetros. Fizemos uso deste estimador em todo o Capítulo 5.

Os métodos estudados para se trabalhar com dados *missing* em conjuntos amostrais foram:

1. Estimador de Caso Completo (CC);
2. Estimador de Caso Completo Corrigido (CCC);
3. Imputação pela Média (IM);
4. Estimador de Máxima Verossimilhança com uso da Quadratura Gaussiana (EMVG) e
5. EMVGM, uma combinação entre os estimadores citados nos itens 1 e 4, acima. Esta combinação foi um método criado pelos autores.

Mostramos com detalhes a estimação EMVG e EMVGM, visto que os outros métodos foram detalhados no Capítulo 4.

A função de verossimilhança utilizada no método EMVG, definida na Seção 4.1.2, foi dada por:

$$L(\boldsymbol{\beta}, \xi) = \prod_{i \in v} Pr(y_i|x_{1i}, x_{2i}; \boldsymbol{\beta}) f(x_{2i}|x_{1i}; \xi) \prod_{j \in \bar{v}} \int Pr(y_j|x_{1j}, z; \boldsymbol{\beta}) f(z|x_{1j}; \xi) dz.$$

E, aplicando o logaritmo natural em ambos lados, obtivemos a seguinte função:

$$l(\boldsymbol{\beta}, \xi) = \sum_{i \in v} \ln[Pr(y_i|x_{1i}, x_{2i}; \boldsymbol{\beta}) f(x_{2i}|x_{1i}; \xi)] + \sum_{j \in \bar{v}} \ln \left[ \int Pr(y_j|x_{1j}, z; \boldsymbol{\beta}) f(z|x_{1j}; \xi) d \right], \tag{5.4}$$

sendo:

$v = \{i|r_i = 1\}$, ou seja, o indivíduo pertence a $v$ quando for observado em $x_2$;

$\bar{v} = \{1, ..., N\} \backslash v$;

$$Pr(y|x_{1i}, x_{2i}; \beta) = \frac{exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}, e$$

$$f(x_{2i}|x_{1i}; \xi) = \frac{1}{2} \times exp\left\{-\frac{x_{2i}}{2}\right\}, x_{2i} > 0.$$

Fazendo as devidas substituições na função (5.4), encontramos:

$$l(\boldsymbol{\beta}, \xi) = \sum_{i \in v} \ln \left[ \frac{exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})} \times \frac{1}{2} \times exp\left\{-\frac{x_{2i}}{2}\right\} \right] + \sum_{j \in \bar{v}} \ln \left[ \int_0^\infty \frac{exp\ (\beta_0 + \beta_1 x_{1j} + \beta_2 z)}{1 + exp\ (\beta_0 + \beta_1 x_{1j} + \beta_2 z)} \times \frac{1}{2} \times exp\left\{-\frac{z}{2}\right\} \right] dz. \tag{5.5}$$

Considerando $z = 2c$, temos:

$$l(\boldsymbol{\beta}, \xi) = \sum_{i \in v} \ln \left[ \frac{exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})} \times \frac{1}{2} \times exp\left\{-\frac{x_{2i}}{2}\right\} \right] + \sum_{j \in \bar{v}} \ln \left[ \int_0^\infty \frac{exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)}{1 + exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)} \times exp\{-c\} dc \right]. \tag{5.6}$$

Observe que o segundo fator da função acima possui o cálculo de uma integral. Esta integral pode ser vista da seguinte forma:

$$\int_0^\infty \frac{exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)}{1 + exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)} \times exp\{-c\} dc = \int_0^\infty F(c) \times exp(-c) dc,$$

onde $F(c) = \frac{exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)}{1 + exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c)}$.

Note que podemos utilizar a Quadratura de Gauss Laguerre (ver Seção 4.2), sendo esta definida da seguinte forma:

$$\int_0^\infty F(c) exp(-c) dc = \sum_{h=0}^{k-1} w_h F(c_h) + E_k,$$

onde

$$E_k = \frac{(k!)^2}{(2k)!} \times \frac{d^{(2k)} F(\zeta)}{d\zeta^{(2k)}}, \quad \zeta \geq 0,$$

com $|E_k| \leq \frac{(k!)^2}{(2k)!} \max |F^{(2k)}(\zeta)|$.

Quando $F(c)$ é um polinômio de grau $2k + 1$ ou menor, o erro $E_k$ é zero, ou seja, a aproximação é exata. Porém, quando $F(c)$ não é uma função polinomial, temos que incluir um fator de correção ($E_k$) no cálculo.

Em nosso trabalho $F(c)$ não é um polinômio, logo o erro é diferente de zero. Com isso temos que:

$$l(\boldsymbol{\beta}, \xi) = \sum_{i \in v} \ln \left[ \frac{exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + exp\ (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})} \cdot \frac{1}{2} \cdot exp \left\{ -\frac{x_{2i}}{2} \right\} \right] +$$

$$\sum_{j \in \bar{v}} \ln \left[ \sum_{h=0}^{k-1} w_h \cdot \frac{exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c_h)}{1 + exp\ (\beta_0 + \beta_1 x_{1j} + 2\beta_2 c_h)} + E_k \right], \tag{5.7}$$

onde $c_h$ são as raízes, $w_h$ os pesos e $k$ o número de nós.

Os valores das raízes e dos pesos são pré-fixados. Abaixo mostramos estes valores para $2, 3, 4$ e $10$ nós.

TABELA 5.6: Raízes e Pesos para Quadratura de Gauss Laguerre

| Nós | Raízes ($c_h$) | Pesos ($w_h$) |
|---|---|---|
| 2 ($k = 2$) | 0,58579 | 0,85355 |
| | 3,41421 | 0,14645 |
| 3 ($k = 3$) | 0,41577 | 0,71109 |
| | 2,29428 | 0,27852 |
| | 6,28995 | 0,01039 |
| 4 ($k = 4$) | 0,32255 | 0,60315 |
| | 1,74576 | 0,35742 |
| | 4,53662 | 0,03889 |
| | 9,39507 | 0,00054 |
| 10 ($k = 10$) | 0,13779 | 0,30844 |
| | 0,72945 | 0,40111 |
| | 1,80834 | 0,21806 |
| | 3,40143 | 0,06208 |
| | 5,55249 | 0,00950 |
| | 8,33015 | 0,00075 |
| | 11,84378 | 0,00002 |
| | 16,27925 | 0,00000 |
| | 21,99658 | 0,00000 |
| | 29,92069 | 0,00000 |

Os valores apresentados na Tabela 5.6 podem ser encontrados na Literatura e em alguns *Softwares*, como R 2.7.1 (pacote statmod, comando *gauss.quad*).

A simulação presente neste trabalho fez uso de dez nós, logo

$$E_{10} = -\frac{(10!)^2}{(20)!} \times \frac{d^{(20)}F(\zeta)}{d\zeta^{(20)}}, \quad \zeta \geq 0,$$

onde

$$F(\zeta) = \frac{exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta)}{1 + exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta)}. \tag{5.8}$$

E, por meio do *Software Maple* 11, calculamos a derivada de ordem 20 de $F(\zeta)$

dada em (5.8), obtendo o seguinte resultado:

$$\frac{d^{(20)}F(\zeta)}{d\zeta^{(20)}} = F(\zeta)^{(20)} = -1.048.576 \times \beta_2^{20} \times exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta) \times (exp\,(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta) - 1) \times ((exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{18} - 1.048.554 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{17} + 3.463.715.961 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{16} - 1.023.045.639.024 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{15} + 71.985.471.942.420 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{14} - 180.772.319.795.407 \times exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{13} + 19.790.873.105.145.828 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{12} - 104.957.308.999.318.032 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{11} + 283.630.951.724.635.278 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{10} - 395.931.266.069.521.660 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{9} + 283.630.951.724.635.278 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{8} - 104.957.308.999.318.032 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{7} + 19.790.873.105.145.828 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{6} - 1.807.723.197.954.072 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{5} + 71.985.471.942.420 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{4} - 1.023.045.639.024 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{3} + 3.463.715.961 \times (exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{2} - 1.048.554 exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta) + 1)/(1 + exp(\beta_0 + \beta_1 x_{1j} + 2\beta_2\zeta))^{21}$$

com $\zeta \geq 0$.

O valor de $\zeta$ a ser usado é o valor que maximiza o módulo da função $F(\zeta)^{(20)}$, em relação a $\zeta$, considerando $\beta_0, \beta_1$ e $\beta_2$ como sendo os valores estimados pelo método de caso completo.

## 5.3 Método EMVGM

Após a estimação dos parâmetros via EMVG, sugerimos um método, denotado EMVGM, dado pela combinação entre os estimadores de Caso Completo e os estimadores obtidos via EMVG. Isto é:

$$EMVGM = p \times EMVG + (1-p) \times CC, \quad para\ \hat{\beta}_1\ e\ \hat{\beta}_2$$

e,

$$EMVGM = (p \times EMVG + (1-p) \times CC) \times m^8, \quad para\ \hat{\beta}_0$$

para $p$ pertencente ao intervalo $(0,01; 0,99)$ e $m$ como sendo o percentual de *missing* dado na amostra.

O critério usado na escolha dos $p$'s foi o estimador EMVGM possuir as três métricas estudadas menores que as métricas obtidas no método de Caso Completo (CC). Este critério foi verificado 100 vezes para os $p's$ iguais a $0,01; 0,02; ...; 0,99$, encontrando os seguintes valores que satisfaziam as condições mencionadas acima:

| | n = 300 | n = 500 |
|---|---|---|
| 5% | $p \in (0,01; 0,21)$ | $p \in (0,01; 0,07)$ |
| 10% | $p \in (0,01; 0,27)$ | $p \in 0,01; 0,09)$ |
| 30% | $p \in (0,01; 0,41)$ | $p \in (0,01; 0,11)$ |
| 50% | $p \in (0,01; 0,54)$ | $p \in (0,01; 0,16)$ |

As análises que seguem utilizaram $p = 0,21$ para todos os parâmetros na amostra de tamanho 300 e $p = 0,07$ para todos os parâmetros na amostra de tamanho 500, ou seja, a escolha do $p$ foi obtida pela intersecção dos intervalos dentro de cada tamanho amostral. Outras escolhas poderiam ser feitas, como, por exemplo, escolher os $p$'s dentro de cada faixa de *missing*, não considerando o tamanho amostral.

Deste ponto em diante, mostramos as diversas estimativas dos parâmetros com suas devidas interpretações. Primeiramente apresentamos alguns gráficos das métricas versus o percentual de *missing*. Logo após, os intervalos de confiança assintóticos e empíricos para cada estimativa. Finalmente, apresentamos os dados, em tabelas, que foram usados na construção dos gráficos.

Esses gráficos foram construídos com os resultados das amostras de tamanhos 300 e 500 com presença de $5\%, 10\%, 30\%$ e $50\%$ de *missing* em relação aos totais de casos.

O critério de decisão sobre qual método forneceu as melhores estimativas foi obter as menores métricas: erro quadrático médio, variância e vício, para todos os $\hat{\beta}$'s, salvo algumas alterações na interpretação dos vícios (para alguns parâmetros).

Primeiramente analisamos o parâmetro estimado $\hat{\beta}_0$ na amostra de tamanho 300. Observe a seguinte Figura e, em seguida, a análise realizada.

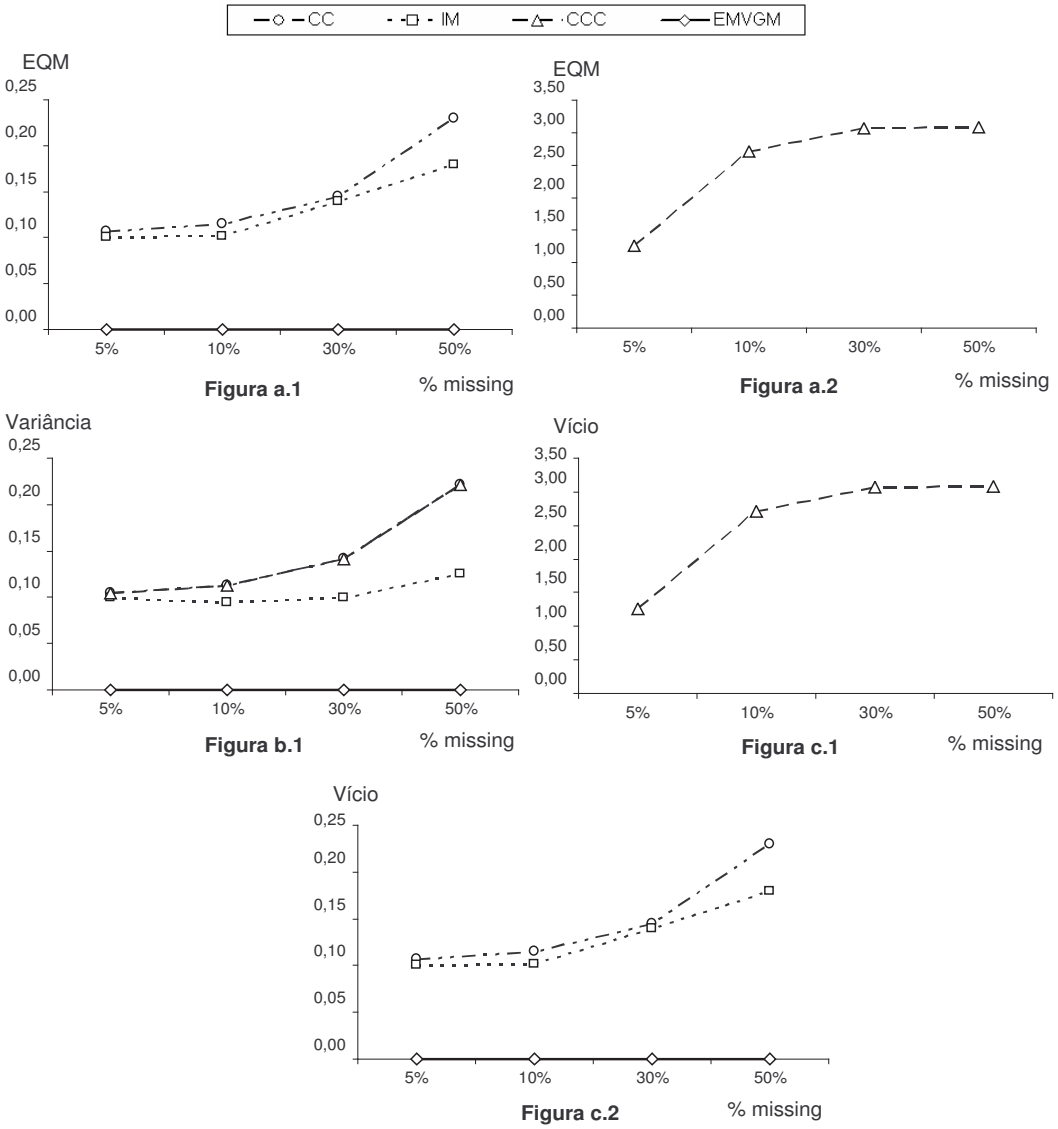


FIGURA 5.1: Erro Quadrático Médio (Figura a.1 e a.2), Variância (Figura b.1) e Vício (Figura c.1 e c.2) para $\hat{\beta}_0$ para amostra de tamanho 300.

As Figuras a.1 e a.2 mostram que o EQM dos quatro métodos discutidos aumenta conforme o número de informações faltantes aumenta. Nestas Figuras vemos que o EMVGM é superior a IM que é superior a CC, que por sua vez é superior a CCC.

Na Figura b.1 vemos que os métodos CC e CCC possuem as mesmas variâncias, o que era esperado, visto que

$$Var(\hat{\beta}^{CCC}) = Var(\hat{\beta}^{CC} + \log k) = Var(\hat{\beta}^{CC}) + Var(\log k) = Var(\hat{\beta}^{CC})$$

sendo $k$ uma constante e $\hat{\beta}^{CCC}$ o parâmetro estimado via CCC.

As variâncias desses dois métodos são maiores em relação à EMVGM e IM, enquanto que as variâncias obtidas no método EMVGM são menores. Os valores desta métrica aumentam com o aumento do percentual de *missing*.

Finalmente as Figuras c.1 e c.2 nos apresentam os vícios. Como nas métricas acima, o vício cresce juntamente com o crescimento da falta de informação. Os menores vícios são obtidos em EMVGM e os maiores em CCC. Logo, decidimos que o método EMVGM é superior aos outros enquanto que CCC é inferior.

Agora, em relação aos métodos CC e IM, decidimos que o método IM é melhor para estimar $\beta_0$ na presença de 5% e 10% de *missing* enquanto que para 30% e 50% o CC é melhor, salvo algumas considerações de aproximações no vício.

Seguimos analisando a Figura 5.3, que nos mostra os valores relacionados ao parâmetro estimado $\hat{\beta}_1$ na amostra de tamanho 300.

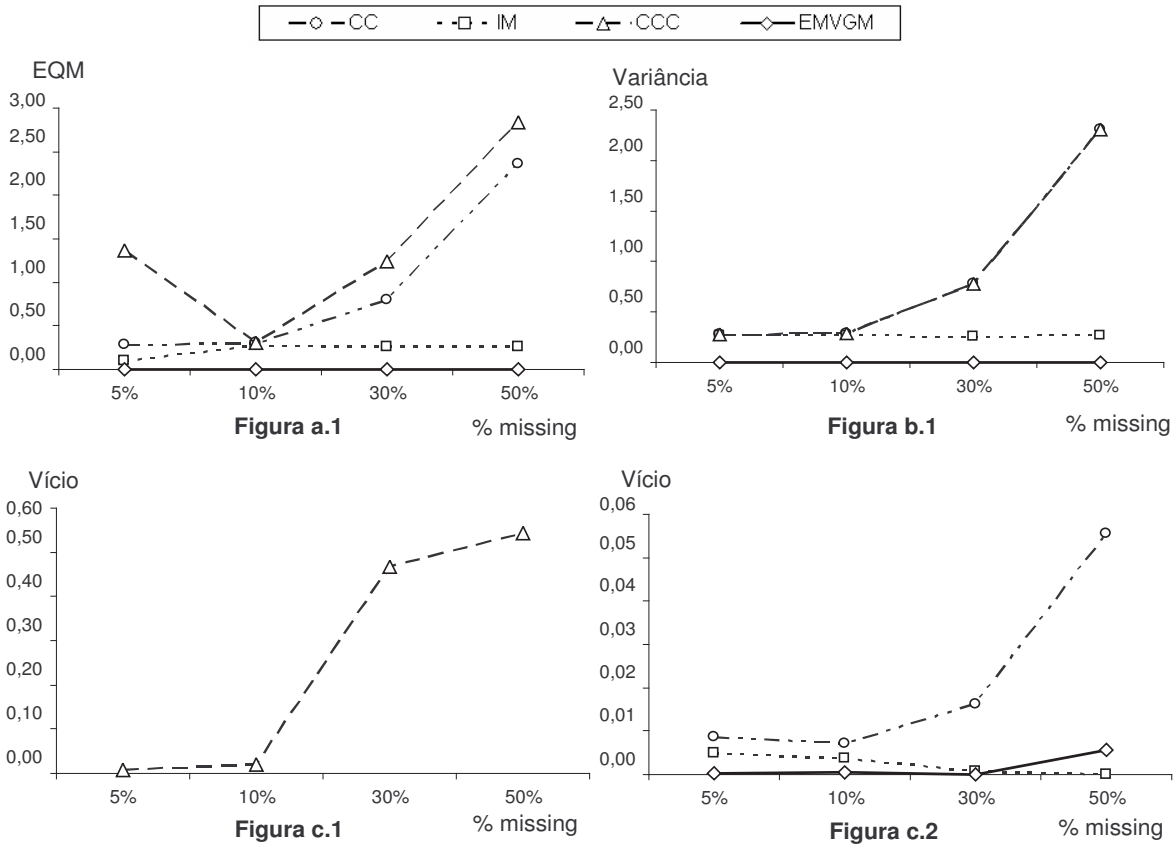


FIGURA 5.2: Erro Quadrático Médio (Figura a.1), Variância (Figura b.1) e Vício (Figura c.1 e c.2) para $\hat{\beta}_1$ para amostra de tamanho 300.

Na Figura anterior, a Figura a.1 mostra a métrica EQM em relação ao percentual de *missing*. Para CC, esta métrica aumenta com o aumento do percentual. Já para os outros três métodos, há uma oscilação, mas sempre com um aumento no percentual 50. O método EMVGM é superior aos outros enquanto CCC é inferior, sendo IM o segundo melhor método e CC o terceiro melhor método para estimar $\beta_1$ na métrica EQM.

As variâncias dadas na Figura *b*.1 aumentam conforme aumenta o número de informações faltantes exceto para os métodos IM e EMVGM que oscilam, mas sempre aumentam no percentual 50. As melhores variâncias são dadas em EMVGM, enquanto que as piores, em CC e CCC. As variâncias de CC e CCC são iguais, conforme mencionado.

Por fim, as Figuras c.1 e c.2 mostram os vícios. Em 5% o vício de CC é maior que em CCC, porém nos outros percentuais a situação se inverte.

Consideramos o método EMVGM superior em relação aos outros para o cálculo do vício enquanto que CCC é o que fornece os piores resultados.

Finalmente a Figura 5.3 apresenta as Figuras para o estimador $\hat{\beta}_2$, no tamanho amostral 300.

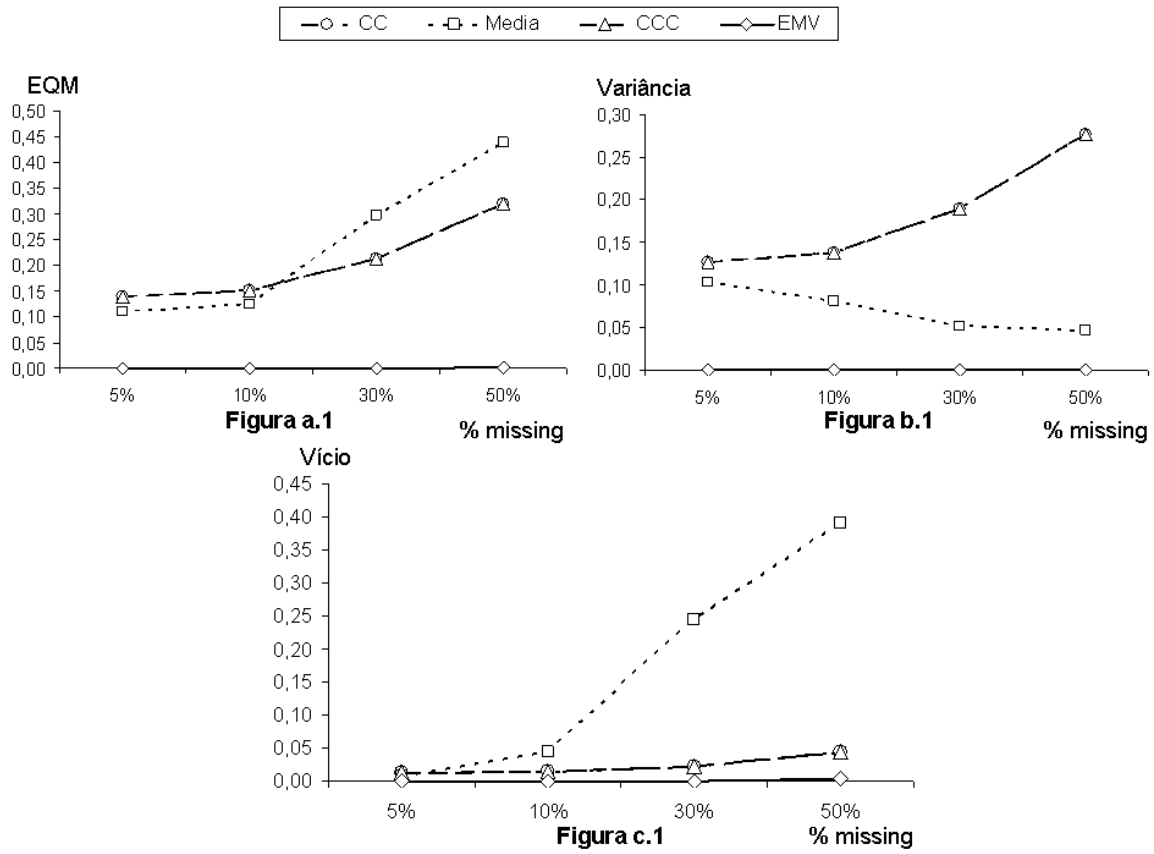


FIGURA 5.3: Erro Quadrático Médio (Figura a.1), Variância (Figura b.1) e Vício (Figura c.1) para $\hat{\beta}_2$ para amostra de tamanho 300.

Abaixo temos a análise dos resultados referente à Figura acima:

A Figura a.1 nos mostra que EMVGM é superior, na métrica EQM, aos demais métodos para todos os percentuais de *missing*, enquanto que para 5% e 10% os métodos CC e CCC são inferiores e para 30% e 50% o método IM passa a ser inferior.

Na análise do $\hat{\beta}_2$, todas as métricas de CC serão iguais a CCC, visto que

$$\hat{\beta}_2^{\,CCC} = \hat{\beta}_2^{\,CC}$$

Na Figura b.1 notamos que EMVGM é sempre superior aos outros métodos enquanto que CC e CCC são inferiores. Vemos também que a variância do parâmetro $\hat{\beta}_2$ (parâmetro relacionado à variável com presença de dados *missing*), no método IM, é influenciada, tendendo a zero.

Finalmente vemos que os vícios, na Figura c.1, aumentam conforme o número de informações observadas diminuem. Os melhores vícios são dados em EMVGM e os piores

em IM para os percentuais de 30% e 50%, sendo que para 5% e 10% os métodos inferiores são CC e CCC, salvo interpretações em alguns vícios.

Após analisados os gráficos das métricas em todos os parâmetros na amostra de tamanho 300, fizemos o mesmo para $n = 500$.

Segue, na Figura 5.4, análise gráfica do parâmetro estimado $\hat{\beta}_0$ para a amostra de tamanho 500.

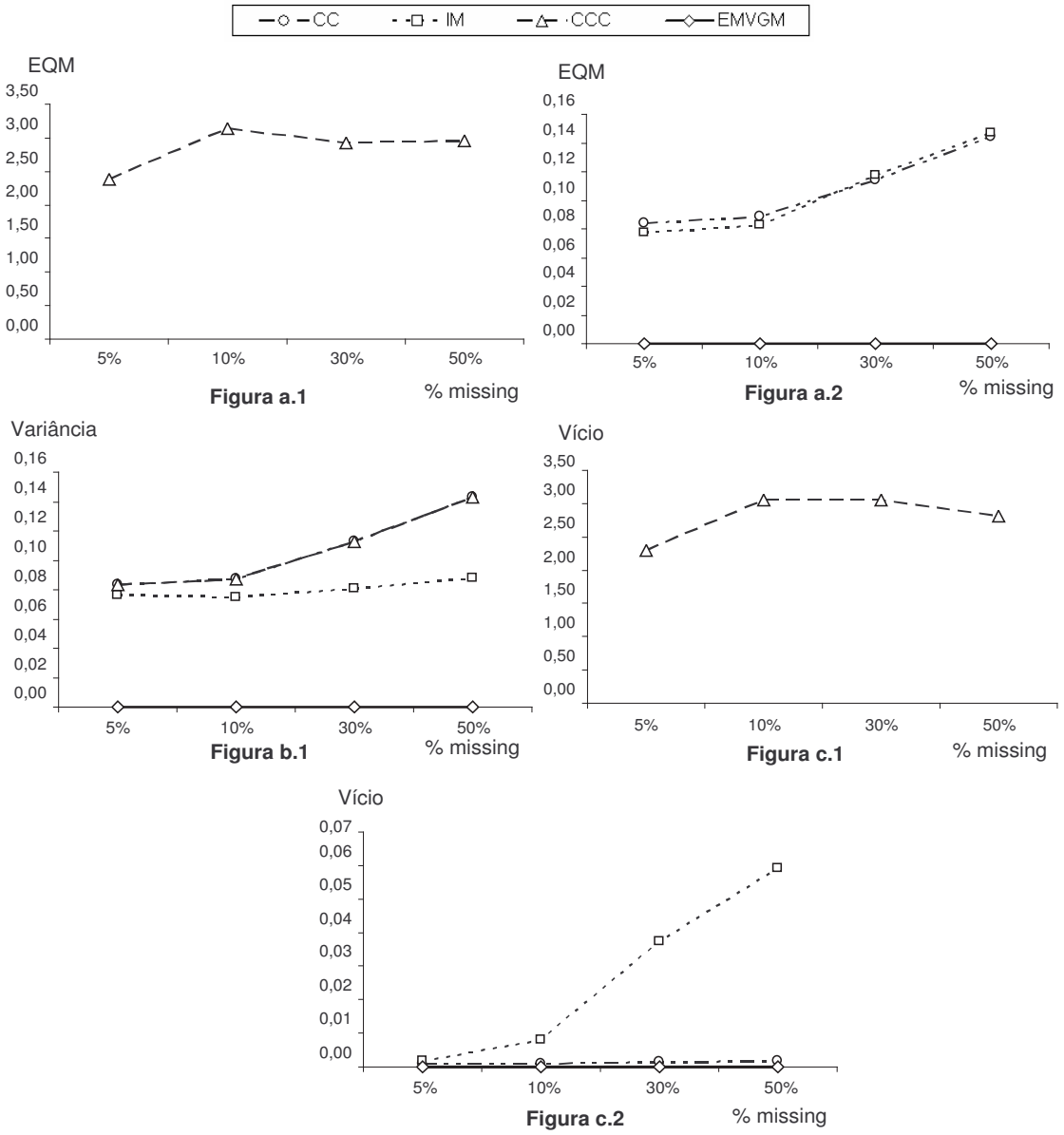


FIGURA 5.4: Erro Quadrático Médio (Figura a.1 e a.2), Variância (Figura b.1) e Vício (Figura c.1 e c.2) para $\hat{\beta}_0$ para amostra de tamanho 500.

As Figuras a.1 e a.2 mostram que o EQM aumenta conforme o número de dados *missing* aumenta, exceto para a curva do método de Caso Completo Corrigido. Nestas Figuras vemos que EMVGM é superior aos três outros métodos e que CCC é inferior. Vemos também que o método IM possui melhor EQM nos percentuais 5% e 10% em relação ao método de CC, invertendo a situação em 30% e 50%.

Na Figura b.1 vemos que as variâncias aumentam conforme a presença de dados *missing* aumenta. Sendo EMVGM o método superior, e CC e CCC os que produzem valores inferiores.

Por fim, as Figuras c.1 e c.2 mostram que os vícios tendem a aumentar com o aumento de dados *missing*, exceto no caso CCC. Os piores vícios são dados no caso CCC e os melhores, no caso EMVGM.

A Figura 5.5 possui as três métricas estudadas para o parâmetro estimado $\hat{\beta}_1$.

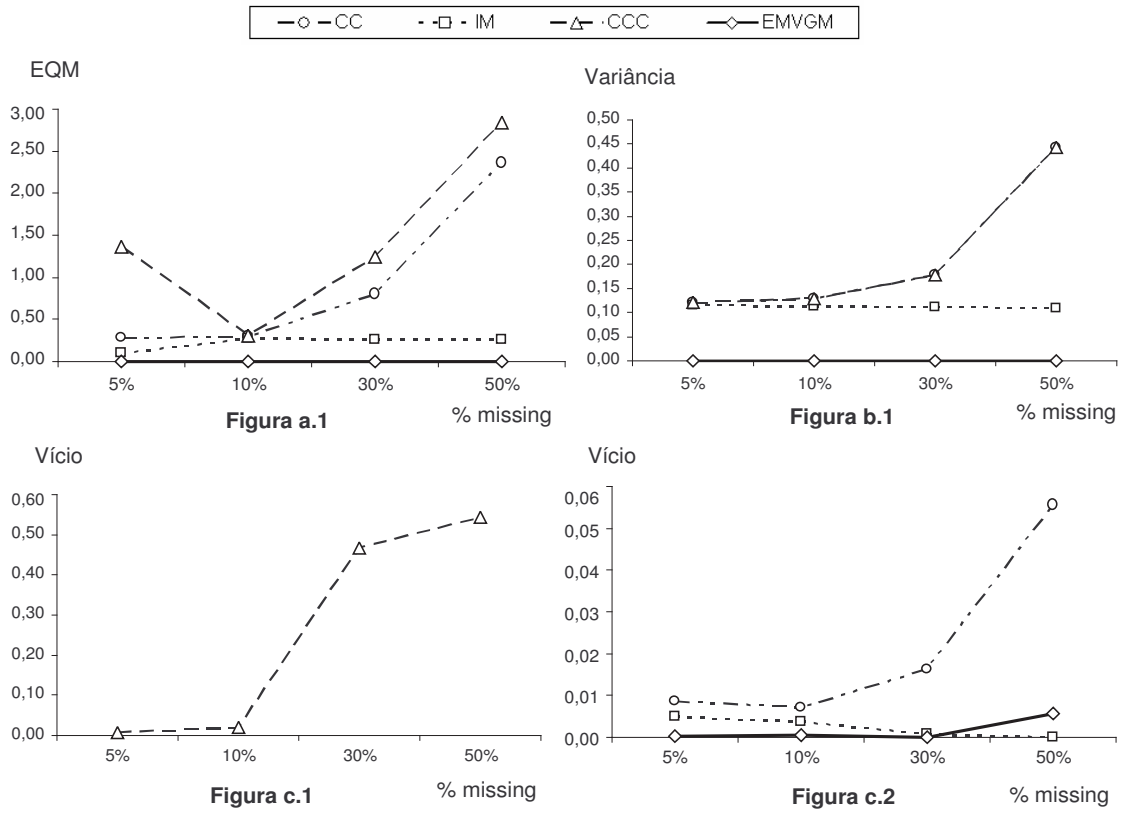


FIGURA 5.5: Erro Quadrático Médio (Figura a.1), Variância (Figura b.1) e Vício (Figura c.1 e c.2) para $\hat{\beta}_1$ para amostra de tamanho 500.

Vemos que, conforme o percentual de *missing* aumenta, o EQM também aumenta

nos casos CC e EMVGM. Para CCC e IM existe uma oscilação, mas sempre com um aumento em 50%. Neste caso, o método CCC é inferior aos outros três, enquanto que EMVGM é superior, ou seja, os valores de EQM são menores neste e maiores naquele. IM pode ser considerado o segundo melhor método e CC o terceiro, nesta parte da análise.

Já na Figura b.1, vemos que os métodos CC e CCC possuem um crescimento ao longo do aumento de *missing* na amostra. Neste caso, EMVGM possui variâncias menores que IM, que por sua vez possui variâncias menores que CC e CCC. Ou seja, EMVGM é superior aos outros métodos enquanto que CC e CCC são inferiores.

Finalmente analisamos o vício de $\hat{\beta}_1$. Aqui, diferentemente dos casos vistos até agora, a métrica EMVGM não é superior a todos os outros métodos, já que com 50% de *missing* o vício de IM é menor que o vício obtido em EMVGM.

Exceto no caso IM, os vícios aumentam com o aumento de *missing* na amostra. Vemos que, no geral, EMVGM é superior aos outros enquanto que CCC é inferior.

Seguem as métricas para o estimador $\hat{\beta}_2$.

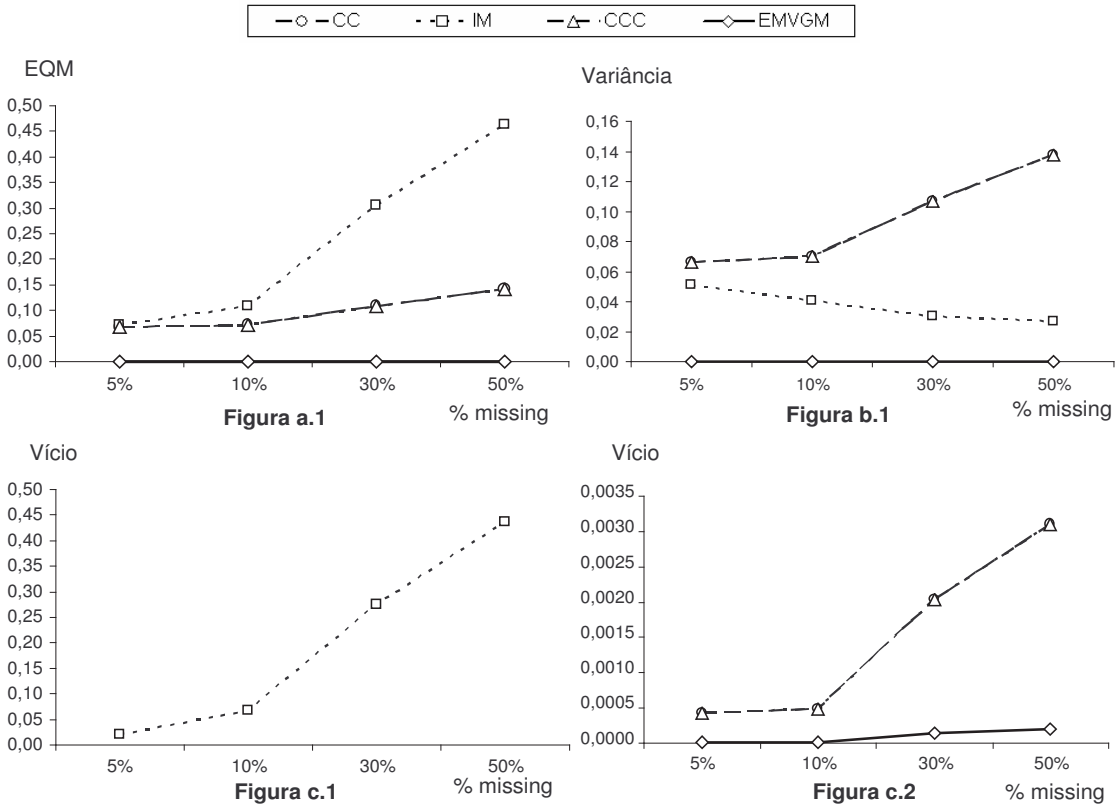


FIGURA 5.6: Erro Quadrático Médio (Figura a.1), Variância (Figura b.1) e Vício (Figura c.1 e c.2) para $\hat{\beta}_2$ para amostra de tamanho 500.

Na Figura 5.6, notamos que há um aumento da métrica EQM conforme o percentual de *missing* aumenta. É fácil notar que EMVGM é superior aos outros três métodos enquanto que IM produz os piores valores, sendo assim, inferior. Na Figura b.1, temos que IM vai tendendo a zero conforme o percentual de informações faltantes aumenta. O contrário ocorre nos outros três métodos.

Observemos agora as Figuras c.1 e c.2, notamos que os vícios aumentam com o aumento das informações faltantes, sendo bem fácil verificar que EMVGM é superior aos três métodos. Sendo inferior o método IM.

Feitas as análises necessárias, observemos agora os intervalos de confiança (IC) assintóticos e empíricos das médias dos 1.000 parâmetros estimados. Estas duas métricas são apresentadas apenas para complementar o estudo, pois a nossa intenção não é ajustar modelos.

TABELA 5.7: Intervalos de Confiança Assintóticos e Empíricos em 5% de *missing*, n=300

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,68074; 0,58460) | 1,26534 | (-0,70342; 0,57415) | 1,27757 |
| | $\beta_1$ | (0,06004; 2,12572) | 2,06568 | (0,19729; 2,38473) | 2,18744 |
| | $\beta_2$ | (0,91919; 2,31435) | 1,39517 | (0,93713; 2,42478) | 1,48765 |
| CCC | $\beta_0$ | (-1,75566; -0,49032) | 1,26534 | (-2,35066; -0,04807) | 2,30259 |
| | $\beta_1$ | (0,05316; 2,11884) | 2,06568 | (0,06326; 2,42649) | 2,36323 |
| | $\beta_2$ | (0,91919; 2,31435) | 1,39517 | (0,93713; 2,42478) | 1,48765 |
| IM | $\beta_0$ | (-0,57804; 0,65496) | 1,23300 | (-0,64632; 0,61301) | 1,25933 |
| | $\beta_1$ | (0,06200; 2,07782) | 2,01582 | (0,16697; 2,35648) | 2,18951 |
| | $\beta_2$ | (0,78544; 2,,4266) | 1,25722 | (0,85968; 2,09219) | 1,23251 |
| EMVGM | $\beta_0$ | $(1,37 \times 10^{-10}; 1,38 \times 10^{-10})$ | $4,9 \times 10^{-13}$ | $(1,2 \times 10^{-10}; 1,6 \times 10^{-10})$ | $3,99 \times 10^{-11}$ |
| | $\beta_1$ | (0,98005; 0,99085) | 0,01080 | (0,28269; 2,01077) | 1,72808 |
| | $\beta_2$ | (1,51694; 1,52508) | 0,00814 | (0,98005; 2,29638) | 1,31634 |

Nos métodos CC e IM os intervalos de confiança, para o parâmetro estimado $\hat{\beta}_0$, contêm o zero, mostrando que $\hat{\beta}_0$ não é significativo para o modelo, conforme esperado. Já em CCC e EMVG $\hat{\beta}_0$ é significativo para o modelo.

Na maioria dos casos as amplitudes dos intervalos de confiança empíricos são maiores que nos intervalos de confiança assintóticos.

Observe agora, na Tabela 5.8, os intervalos de confiança para as médias dos parâmetros nas amostras com 10% de *missing* em relação ao total de casos (300).

TABELA 5.8: Intervalos de Confiança Assintóticos e Empíricos em 10% de *missing*, n=300

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,70571; 0,60789) | 1,31359 | (-0,77715; 0,60275) | 1,37990 |
| | $\beta_1$ | (0,03101; 2,13695) | 2,10594 | (0,14640; 2,53057) | 2,38417 |
| | $\beta_2$ | (0,89136; 2,34834) | 1,45699 | (0,94181; 2,48955) | 1,54774 |
| CCC | $\beta_0$ | (-2,30214; -0,98854) | 1,31359 | (-2,99250; -0,04807) | 2,94443 |
| | $\beta_1$ | (-0,19230; 1,91364) | 2,10594 | (-0,57322; 2,12343) | 2,69665 |
| | $\beta_2$ | (0,89136; 2,34834) | 1,45699 | (0,94181; 2,48955) | 1,54774 |
| IM | $\beta_0$ | (-0,51425; 0,68923) | 1,20348 | (-0,52091; 0,68907) | 1,20998 |
| | $\beta_1$ | (0,05738; 2,06478) | 2,00739 | (0,17815; 2,35476) | 2,17661 |
| | $\beta_2$ | (0,73335; 1,84823) | 1,11489 | (0,82374; 1,93077) | 1,10703 |
| EMVGM | $\beta_0$ | $(3,53 \times 10^{-8}; 3,54 \times 10^{-8})$ | $1,2 \times 10^{-10}$ | $(2,95 \times 10^{-8}; 4,05 \times 10^{-8})$ | $1,10 \times 10^{-8}$ |
| | $\beta_1$ | (0,97088; 0,98223) | 0,01134 | (0,24123; 2,12472) | 1,88349 |
| | $\beta_2$ | (1,51597; 1,52416) | 0,00819 | (0,98061; 2,21163) | 1,23102 |

Nos métodos CC e IM temos que $\hat{\beta}_0$ não é significativo para o modelo. Porém, isto não é verdade em CCC e EMVGM, ou seja, neste método $\hat{\beta}_0$ é significativo. No método CC, o zero está contido no intervalo de $\hat{\beta}_1$, sendo este não relevante na explicação da resposta. Nos outros métodos, $\beta_1$ e $\beta_2$ são significativos.

Na maioria dos casos, as amplitudes dos intervalos de confiança assintóticos são menores que as amplitudes dos intervalos de confiança empíricos.

Acompanhe agora, na Tabela 5.9, os intervalos de confiança para as médias dos parâmetros nas amostras com 30% de *missing* em relação ao total de casos (300).

TABELA 5.9: Intervalos de Confiança Assintóticos e Empíricos em 30% de *missing*, n=300

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,79754; 0,67466) | 1,47220 | (-0,82829; 0,67377) | 1,50206 |
| | $\beta_1$ | (-0,60342; 2,85892) | 3,46234 | (0,07439; 2,64451) | 2,57012 |
| | $\beta_2$ | (0,79725; 2,50453) | 1,70728 | (0,90949; 2,61614) | 1,70665 |
| CCC | $\beta_0$ | (-2,48656; -1,01436) | 1,47220 | (-2,63591; -1,08000) | 1,55591 |
| | $\beta_1$ | (-1,41448; 2,04786) | 3,46234 | (-1,07259; 1,56938) | 2,64197 |
| | $\beta_2$ | (0,79725; 2,50453) | 1,70728 | (0,90949; 2,61614) | 1,70665 |
| IM | $\beta_0$ | (-0,41604; 0,81852) | 1,23456 | (-0,39406; 0,87304) | 1,26710 |
| | $\beta_1$ | (0,02998; 2,02522) | 1,99524 | (0,13717; 2,32300) | 2,18583 |
| | $\beta_2$ | (0,55914; 1,44934) | 0,89019 | (0,59695; 1,48988) | 0,89293 |
| EMVGM | $\beta_0$ | (0,00022; 0,00024) | 0,00002 | (0,00019; 0,00027) | 0,00008 |
| | $\beta_1$ | (0,99533; 1,00961) | 0,01427 | (0,18434; 2,21473) | 2,03039 |
| | $\beta_2$ | (1,52376; 1,53351) | 0,00975 | (0,95508; 2,31164) | 1,35656 |

Ao analisarmos os intervalos acima, vemos que nos métodos CC e IM, $\hat{\beta}_0$ não é significativo para o modelo. Isto não ocorre em CCC e EMVGM, visto que o zero não pertence ao intervalo. Com relação à $\beta_1$ vemos que não é siginificativo nos métodos CC e CCC, uma vez que o zero pertence ao intervalo de $\hat{\beta}_1$. Nos outros métodos, $\beta_1$ e $\beta_2$ são significativos.

Na maioria dos casos, as amplitudes dos intervalos de confiança assintóticos são menores que as amplitudes dos intervalos de confiança empíricos.

Finalmente, a Tabela 5.10 apresenta os intervalos de confiança para as médias dos parâmetros nas amostras com 50% de *missing* em relação ao total de casos (300).

TABELA 5.10: Intervalos de Confiança Assintóticos e Empíricos em 50% de *missing*, n=300

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-1,01278; 0,83162) | 1,84440 | (-1,01389; 0,86253) | 1,87642 |
| | $\beta_1$ | (-1,73748; 4,20900) | 5,94648 | (-0,10869; 2,93183) | 3,04052 |
| | $\beta_2$ | (0,67896; 2,74002) | 2,06106 | (0,88036; 2,98028) | 2,09992 |
| CCC | $\beta_0$ | (-2,67727; -0,83287) | 1,84440 | (-2,35002; -1,23364) | 1,11638 |
| | $\beta_1$ | (-2,70993; 3,23655) | 5,94648 | (-1,36038; 1,36092) | 2,72130 |
| | $\beta_2$ | (0,67896; 2,74002) | 2,06106 | (0,88036; 2,98028) | 2,09992 |
| IM | $\beta_0$ | (-0,46013; 0,92547) | 1,38560 | (-0,49454; 0,93998) | 1,43452 |
| | $\beta_1$ | (0,00380; 2,01516) | 2,01135 | (0,11154; 2,33106) | 2,21952 |
| | $\beta_2$ | (0,45043; 1,29765) | 0,84723 | (0,47911; 1,37426) | 0,89515 |
| EMVGM | $\beta_0$ | (0,01175; 0,01450) | 0,00276 | (0,01018; 0,01681) | 0,00662 |
| | $\beta_1$ | (1,06473; 1,08542) | 0,02069 | (0,02315; 2,42517) | 2,40201 |
| | $\beta_2$ | (1,54558; 1,55973) | 0,01415 | (0,89023; 2,56348) | 1,67325 |

Nos métodos CC e IM os intervalos de confiança para o parâmetro $\beta_0$ contêm o zero, mostrando que $\beta_0$ não é significativo para o modelo, conforme esperado. Já em CCC e EMVGM $\beta_0$ é significativo para o modelo.

Na maioria dos casos, as amplitudes dos intervalos de confiança empíricos são maiores que nos intervalos de confiança assintóticos.

Após apresentarmos os intervalos de confiança empírico e assintótico para a amostra de tamanho 300, fazemos o mesmo para o tamanho amostral 500. Porém não apresentamos a análise neste trabalho, apenas os resultados numéricos. Seguem Tabelas.

Observe a Tabela 5.11 que apresenta os intervalos de confiança para as amostras com 5% de *missing* em relação ao total de casos (500).

TABELA 5.11: Intervalos de Confiança Assintóticos e Empíricos em 5% de *missing*, n=500

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,59409, 0,53829) | 1,13237 | (-0,66658, 0,51224) | 1,17882 |
| | $\beta_1$ | (0,36271, 1,73071) | 1,36800 | (0,46991, 1,79124) | 1,32133 |
| | $\beta_2$ | (1,01602, 2,02518) | 1,00916 | (1,02897, 2,03234) | 1,00337 |
| CCC | $\beta_0$ | (-2,08233, -0,94995) | 1,13237 | (-2,80049, -0,02790) | 2,77259 |
| | $\beta_1$ | (0,25104, 1,61904) | 1,36800 | (-0,33958, 2,08816) | 2,42774 |
| | $\beta_2$ | (1,01602, 2,02518) | 1,00916 | (1,02897, 2,03234) | 1,00337 |
| IM | $\beta_0$ | (-0,49883, 0,58105) | 1,07988 | (-0,57018, 0,56491) | 1,13509 |
| | $\beta_1$ | (0,37177, 1,69673) | 1,32496 | (0,43700, 1,73311) | 1,29611 |
| | $\beta_2$ | (0,91475, 1,80149) | 0,88674 | (0,95558, 1,81736) | 0,86178 |
| EMVGM | $\beta_0$ | $(4{,}54\times10^{-11}; 4{,}55\times10^{-11})$ | $1{,}0\times10^{-13}$ | $(2{,}0\times10^{-11}; 7{,}0\times10^{-11})$ | $4{,}0\times10^{-11}$ |
| | $\beta_1$ | (1,01284; 1,01543) | 0,00259 | (0,47629; 1,70773) | 1,23144 |
| | $\beta_2$ | (1,49429; 1,49655) | 0,00227 | (1,03659; 1,97240) | 0,93581 |

A Tabela 5.12 apresenta os intervalos de confiança para as médias dos parâmetros nas amostras com 10% de *missing* em relação ao total de casos (500).

TABELA 5.12: Intervalos de Confiança Assintóticos e Empíricos em 10% de *missing*, n=500

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,60850, 0,55040) | 1,15891 | (-0,68167, 0,54764) | 1,22931 |
| | $\beta_1$ | (0,34159, 1,75401) | 1,41242 | (0,42108, 1,82725) | 1,40617 |
| | $\beta_2$ | (1,00212, 2,04194) | 1,03982 | (1,03860, 2,05373) | 1,01513 |
| CCC | $\beta_0$ | (-2,32845, -1,16955) | 1,15891 | (-3,20710, -0,87635) | 2,33075 |
| | $\beta_1$ | (-0,17383, 1,23859) | 1,41242 | (-0,91374, 2,03963) | 2,95337 |
| | $\beta_2$ | (1,00212, 2,04194) | 1,03982 | (1,03860, 2,05373) | 1,01513 |
| IM | $\beta_0$ | (-0,44521, 0,62491) | 1,07012 | (-0,50540, 0,58267) | 1,08807 |
| | $\beta_1$ | (0,36415, 1,68077) | 1,31661 | (0,45499, 1,71327) | 1,25828 |
| | $\beta_2$ | (0,84649, 1,63523) | 0,78874 | (0,89672, 1,69346) | 0,79674 |
| EMVGM | $\beta_0$ | ($1{,}2\times10^{-8}$; $1{,}21\times^{-8}$) | $3{,}0\times10^{-11}$ | ($1{,}1\times10^{-8}$; $1{,}55\times10^{-8}$) | $4{,}5\times10^{-9}$ |
| | $\beta_1$ | (1,01316; 1,01598) | 0,00282 | (0,43103; 1,73993) | 1,30890 |
| | $\beta_2$ | (1,49445; 1,49681) | 0,00235 | (1,04306; 1,99018) | 0,94712 |

A Tabela 5.13 apresenta os intervalos de confiança para as médias dos parâmetros nas amostras com 30% de *missing* em relação ao total de casos (500).

TABELA 5.13: Intervalos de Confiança Assintóticos e Empíricos em 30% de *missing*, n=500

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,69708, 0,62036) | 1,31743 | (-0,77924, 0,59270) | 1,37194 |
| | $\beta_1$ | (0,23166, 1,88550) | 1,65385 | (0,37096, 1,94978) | 1,57882 |
| | $\beta_2$ | (0,90382, 2,18656) | 1,28274 | (0,98845, 2,24164) | 1,25319 |
| CCC | $\beta_0$ | (-2,33719, -1,01975) | 1,31743 | (-2,30950, -1,13697) | 1,17253 |
| | $\beta_1$ | (-0,72439, 0,92945) | 1,65385 | (-1,35956, 1,13818) | 2,49774 |
| | $\beta_2$ | (0,90382, 2,18656) | 1,28274 | (0,98845, 2,24164) | 1,25319 |
| IM | $\beta_0$ | (-0,36315, 0,74865) | 1,11179 | (-0,42397, 0,72900) | 1,15297 |
| | $\beta_1$ | (0,34360, 1,64826) | 1,30465 | (0,43630, 1,66564) | 1,22934 |
| | $\beta_2$ | (0,63463, 1,31725) | 0,68263 | (0,65660, 1,34071) | 0,68411 |
| EMVGM | $\beta_0$ | (0,00008; 0,000081) | 0,00000 | (0,00003; 0,00012) | 0,00009 |
| | $\beta_1$ | (1,01984; 1,02344) | 0,00360 | (0,38054; 1,85074) | 1,47020 |
| | $\beta_2$ | (1,51037; 1,51325) | 0,00288 | (0,99294; 2,16011) | 1,16717 |

Finalmente, a Tabela 5.14 apresenta os intervalos de confiança para as médias dos parâmetros nas amostras com 50% de *missing* em relação ao total de casos (500).

TABELA 5.14: Intervalos de Confiança Assintóticos e Empíricos em 50% de *missing*, n=500

| Amostra | Est. | IC Assintótico | Amplitude.A | IC Empírico | Amplitude.E |
|---|---|---|---|---|---|
| CC | $\beta_0$ | (-0,78344; 0,69738) | 1,48082 | (-0,82293; 0,67351) | 1,49644 |
| | $\beta_1$ | (-0,20849; 2,39705) | 2,60555 | (0,42089; 0,51914) | 0,09825 |
| | $\beta_2$ | (0,82852; 2,28272) | 1,45420 | (0,96054; 2,36765) | 1,40711 |
| CCC | $\beta_0$ | (-2,41796; -0,93714) | 1,48082 | (-2,13002; -1,21512) | 0,91490 |
| | $\beta_1$ | (-1,13449; 1,47105) | 2,60555 | (-0,84134; 0,97319) | 1,81453 |
| | $\beta_2$ | (0,82852; 2,28272) | 1,45420 | (0,96054; 2,36765) | 1,40711 |
| IM | $\beta_0$ | (-0,33866; 0,82472) | 1,16338 | (-0,38738; 0,80686) | 1,19424 |
| | $\beta_1$ | (0,31777; 1,60855) | 1,29078 | (0,42258; 1,62640) | 1,20382 |
| | $\beta_2$ | (0,51913; 1,15965) | 0,64053 | (0,54544; 1,17218) | 0,62674 |
| EMVGM | $\beta_0$ | (0,00457; 0,00457) | 0,00000 | (0,00539; 0,00720) | 0,00180 |
| | $\beta_1$ | (1,04814; 1,05334) | 0,00520 | (0,42431; 0,51609) | 0,09178 |
| | $\beta_2$ | (1,51212; 1,51622) | 0,00409 | (0,96029; 2,26950) | 1,30921 |

Segue Tabelas com valores utilizados na construções dos Gráficos apresentados anteriormente.

TABELA 5.15: Parâmetros estimados para n=300 com 5% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,04807 | 0,10419 | 0,00231 | 0,10650 |
| | $\beta_1$ | 1,09288 | 0,27769 | 0,00863 | 0,28631 |
| | $\beta_2$ | 1,61677 | 0,12667 | 0,01364 | 0,14031 |
| CCC | $\beta_0$ | -1,12299 | 0,10419 | 1,26111 | 1,36530 |
| | $\beta_1$ | 1,08600 | 0,27769 | 0,00740 | 0,28508 |
| | $\beta_2$ | 1,61677 | 0,12667 | 0,01364 | 0,14031 |
| IM | $\beta_0$ | 0,03846 | 0,09894 | 0,00148 | 0,10041 |
| | $\beta_1$ | 1,06991 | 0,26444 | 0,00489 | 0,26933 |
| | $\beta_2$ | 1,41405 | 0,19286 | 0,00739 | 0,11025 |
| EMVG' | $\beta_0$ | $1,37 \times 10^{-10}$ | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 0,98545 | 0,00001 | 0,00021 | 0,00022 |
| | $\beta_2$ | 1,52101 | 0,00000 | 0,00044 | 0,00045 |

TABELA 5.16: Parâmetros estimados para n=300 com 10% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,04891 | 0,11229 | 0,00239 | 0,11468 |
| | $\beta_1$ | 1,08398 | 0,28862 | 0,00705 | 0,29567 |
| | $\beta_2$ | 1,61985 | 0,13815 | 0,01436 | 0,15251 |
| CCC | $\beta_0$ | -1,64534 | 0,11229 | 2,70714 | 2,81944 |
| | $\beta_1$ | 0,86067 | 0,28862 | 0,01941 | 0,30803 |
| | $\beta_2$ | 1,61985 | 0,13815 | 0,01436 | 0,15251 |
| IM | $\beta_0$ | 0,08749 | 0,09426 | 0,00765 | 0,10191 |
| | $\beta_1$ | 1,06108 | 0,26224 | 0,00373 | 0,26597 |
| | $\beta_2$ | 1,29079 | 0,08089 | 0,04377 | 0,12466 |
| EMVGM | $\beta_0$ | $3,5 \times 10^{-8}$ | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 0,97655 | 0,00001 | 0,00055 | 0,00056 |
| | $\beta_2$ | 1,52006 | 0,00000 | 0,00040 | 0,00041 |

TABELA 5.17: Parâmetros estimados para n=300 com 30% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,06144 | 0,14105 | 0,00377 | 0,14482 |
| | $\beta_1$ | 1,12775 | 0,78013 | 0,01632 | 0,79645 |
| | $\beta_2$ | 1,65089 | 0,18969 | 0,02277 | 0,21245 |
| CCC | $\beta_0$ | -1,75046 | 0,14105 | 3,06411 | 3,20516 |
| | $\beta_1$ | 0,31669 | 0,78013 | 0,46691 | 1,24704 |
| | $\beta_2$ | 1,65089 | 0,18969 | 0,02277 | 0,21245 |
| IM | $\beta_0$ | 0,20124 | 0,09919 | 0,04050 | 0,13968 |
| | $\beta_1$ | 1,0276 | 0,25907 | 0,00076 | 0,25983 |
| | $\beta_2$ | 1,00424 | 0,05157 | 0,24578 | 0,29735 |
| EMVGM | $\beta_0$ | 0,00023 | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 1,00247 | 0,00001 | 0,00001 | 0,00002 |
| | $\beta_2$ | 1,52863 | 0,00001 | 0,00082 | 0,00083 |

TABELA 5.18: Parâmetros estimados para n=300 com 50% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,09058 | 0,22138 | 0,00820 | 0,22958 |
| | $\beta_1$ | 1,23576 | 2,30117 | 0,05558 | 2,35675 |
| | $\beta_2$ | 1,70949 | 0,27644 | 0,04389 | 0,32033 |
| CCC | $\beta_0$ | -1,75507 | 0,22138 | 3,08027 | 3,30165 |
| | $\beta_1$ | 0,26331 | 2,30117 | 0,54271 | 2,84388 |
| | $\beta_2$ | 1,70949 | 0,27644 | 0,04389 | 0,32033 |
| IM | $\beta_0$ | 0,23267 | 0,12494 | 0,05414 | 0,17908 |
| | $\beta_1$ | 1,00948 | 0,26327 | 0,00009 | 0,26336 |
| | $\beta_2$ | 0,87404 | 0,04671 | 0,39183 | 0,43854 |
| EMVGM | $\beta_0$ | 0,01312 | 0,00000 | 0,00017 | 0,00017 |
| | $\beta_1$ | 1,07508 | 0,00003 | 0,00564 | 0,00566 |
| | $\beta_2$ | 1,55265 | 0,00001 | 0,00277 | 0,00279 |

TABELA 5.19: Parâmetros estimados para n=500 com 5% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,0279 | 0,08345 | 0,00078 | 0,08422 |
| | $\beta_1$ | 1,04671 | 0,12179 | 0,00218 | 0,12397 |
| | $\beta_2$ | 1,5206 | 0,06628 | 0,00042 | 0,06670 |
| CCC | $\beta_0$ | -1,51614 | 0,08345 | 2,29868 | 2,38213 |
| | $\beta_1$ | 0,93504 | 0,12179 | 0,00422 | 0,12601 |
| | $\beta_2$ | 1,52060 | 0,06628 | 0,00042 | 0,06670 |
| IM | $\beta_0$ | 0,04111 | 0,07589 | 0,00169 | 0,07758 |
| | $\beta_1$ | 1,03425 | 0,11424 | 0,00117 | 0,11542 |
| | $\beta_2$ | 1,35812 | 0,05117 | 0,02013 | 0,07130 |
| EMVGM | $\beta_0$ | 0,00000 | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 1,01414 | 0,00000 | 0,00020 | 0,00020 |
| | $\beta_2$ | 1,49542 | 0,00000 | 0,00002 | 0,00002 |

TABELA 5.20: Parâmetros estimados para n=500 com 10% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,02905 | 0,08740 | 0,00084 | 0,08825 |
| | $\beta_1$ | 1,0478 | 0,12982 | 0,00228 | 0,13211 |
| | $\beta_2$ | 1,52203 | 0,07036 | 0,00049 | 0,07085 |
| CCC | $\beta_0$ | -1,74900 | 0,08740 | 3,05900 3,14640 | |
| | $\beta_1$ | 0,53238 | 0,12982 | 0,21867 | 0,34849 |
| | $\beta_2$ | 1,52203 | 0,07036 | 0,00049 | 0,07085 |
| IM | $\beta_0$ | 0,08985 | 0,07452 | 0,00807 | 0,08260 |
| | $\beta_1$ | 1,02246 | 0,11281 | 0,00050 | 0,11331 |
| | $\beta_2$ | 1,24086 | 0,04049 | 0,06715 | 0,10764 |
| EMVGM | $\beta_0$ | 0,00000 | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 1,01457 | 0,00000 | 0,00021 | 0,00021 |
| | $\beta_2$ | 1,49563 | 0,00000 | 0,00002 | 0,00002 |

TABELA 5.21: Parâmetros estimados para n=500 com 30% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,03836 | 0,11295 | 0,00147 | 0,11442 |
| | $\beta_1$ | 1,05858 | 0,17800 | 0,00343 | 0,18143 |
| | $\beta_2$ | 1,54519 | 0,10708 | 0,00204 | 0,10912 |
| CCC | $\beta_0$ | -1,67847 | 0,11295 | 2,81726 | 2,93021 |
| | $\beta_1$ | 0,10253 | 0,17800 | 0,80545 | 0,98345 |
| | $\beta_2$ | 1,54519 | 0,10708 | 0,00204 | 0,10912 |
| IM | $\beta_0$ | 0,19275 | 0,08044 | 0,03715 | 0,11759 |
| | $\beta_1$ | 0,99593 | 0,11077 | 0,00002 | 0,11079 |
| | $\beta_2$ | 0,97594 | 0,03032 | 0,27464 | 0,30496 |
| EMVGM | $\beta_0$ | 0,00000 | 0,00000 | 0,00000 | 0,00000 |
| | $\beta_1$ | 1,02164 | 0,00000 | 0,00047 | 0,00047 |
| | $\beta_2$ | 1,51181 | 0,00000 | 0,00014 | 0,00014 |

TABELA 5.22: Parâmetros estimados para n=500 com 50% de dados *missing*

| Método | Parâmetro | Estimativa | Variância | Vício | EQM |
|---|---|---|---|---|---|
| CC | $\beta_0$ | -0,04303 | 0,14270 | 0,00185 | 0,14455 |
| | $\beta_1$ | 1,09428 | 0,44180 | 0,00889 | 0,45069 |
| | $\beta_2$ | 1,55562 | 0,13762 | 0,00309 | 0,14071 |
| CCC | $\beta_0$ | -1,67755 | 0,14270 | 2,81417 | 2,95688 |
| | $\beta_1$ | 0,16828 | 0,44180 | 0,69176 | 1,13356 |
| | $\beta_2$ | 1,55562 | 0,13762 | 0,00309 | 0,14071 |
| IM | $\beta_0$ | 0,24303 | 0,08808 | 0,05906 | 0,14714 |
| | $\beta_1$ | 0,96316 | 0,10843 | 0,00136 | 0,10978 |
| | $\beta_2$ | 0,83939 | 0,02670 | 0,43641 | 0,46311 |
| EMVGM | $\beta_0$ | 0,00457 | 0,00000 | 0,00002 | 0,00002 |
| | $\beta_1$ | 1,05074 | 0,00000 | 0,00257 | 0,00258 |
| | $\beta_2$ | 1,51417 | 0,00000 | 0,00020 | 0,00020 |

# Capítulo 6

# Conclusão

Após analisar os resultados obtidos na Simulação, podemos afirmar que, dentre os métodos: Análise de Caso Completo (CC), Estimador de Caso Completo Corrigido (CCC), Imputação pela Média (IM) e Estimador de Máxima Verossimilhança com uso da Quadratura Gaussiana Modificado (EMVGM), o método que obteve as melhores estimativas para os parâmetros foi EMVG em todos os percentuais de *missing*.

No caso da amostra de tamanho 300, os percentuais 5% e 10% fez do IM o segundo melhor método. Mesmo assim devemos ter cuidado ao utilizar esta forma de imputação, pois a variância estimada resultante dos parâmetros estimados é influenciada, tendendo a zero. Por fim, o terceiro e quarto melhores métodos são dados por CC e CCC, respectivamente.

Analisando os percentuais de *missing* 30% e 50%, ainda na amostra de tamanho 300, não podemos afirmar muita coisa, apenas que com as métricas discutidas, o EMVGM estima melhor os parâmetros de interesse. Isto ocorre porque as métricas não seguem um padrão de crescimento como em 5% e 10% de *missing* nos outros três métodos.

Para a amostra de tamanho 500, podemos apenas afirmar que o método EMVGM é superior a todos os outros para os quatro percentuais de *missing* considerados. Não conseguimos afirmativas para os outros métodos, pois os resultados não seguem um padrão.

Portanto, percebemos que as técnicas de dados *missing* são melhores aplicadas em amostras de tamanho pequeno. Ou seja, conforme o aumento do tamanho amostral, as

técnicas vão perdendo sua eficiência tornando-se cada vez mais prático trabalhar com o método de Caso Completo.

Finalmente, ao analisarmos os intervalos de confiança assintóticos e empíricos notamos que a amplitude aumenta com o aumento de informações faltantes.

# Propostas Futuras

Para continuidade deste trabalho, propomos estudar os métodos de imputação presentes no *Software SAS* 9.0, como, por exemplo, Método da Regressão Logística e Método de Monte Carlo via Cadeia de Markov.

Bem como utilizar o algoritmo EM na função de verossimilhança como forma de imputação de dados.

# Referências Bibliográficas


[1] Burkett, K. (2002). *Logistic Regression with Missing Haplotypes*. Department of Statistics and Actuarial Science, Simon Fraser University.

[2] Carvalho, J.(2000). *Integração de Funções*. Departamento de Física da F.C.T.U.C. e LIP - Coimbra.

[3] Cordeiro, G.M. e Demétrio, C.G.B. (2007). *Modelos Lineares Generalizados*. Minicurso para o 12 SEAGRO e a 52 Reunião Anual da RBRAS. UFSM, Santa Maria.

[4] Didelez, V. (2002). *ML-and semiparametric estimation In logistic models with incomplete covariate data*. Statistica Neerlandica, Vol.56, No.3.

[5] Einwoegerer, W. (2006). *Quadratura Gaussiana*. Seminário de Dinâmica Orbital I, Instituto Nacional de Pesquisa Espacial (INPE).

[6] Giacon, F.O. (2007) *Imputação Múltipla para Missing Data em Pesquisa Antropométrica na Ergonomia Industrial*. Relatório Técnico, Universidade Federal de São Carlos.

[7] Ibrahim, J.G. (1990). *Incomplete Data in Generalized Linear Model*. Journal of the American Statistical Association, Vol.85, No.411.

[8] Kleinbaum, D.G. (1998). *Logistic Regression - A Self-Learning Text*. Series Editors.

[9] Little, R.J.A. (1992). *Regression with Missing X's: A Review*. Journal of the American Statistical Association, Vol. 87, No. 420.

[10] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with missing data*. John Wiley e Sons, Inc., New York.

[11] Park, C. (2005). *Parameter Estimation of Incomplete Data in Competing Risks Using the EM Algorithm*. IEEE Transactions on Reliability, Vol. 54, No. 2.

[12] Paula, G.A. (2004). *Modelos de Regressão com apoio computacional*. Instituto de Matemática e Estatística - USP.

[13] Pereira, J.E. & Silva, J.F.V. & Dias, W.P. & Souza, G.S. (2000). *Intervalo de Confiança "Bootstrap"Como Ferramenta Para Classificar Raças Do Nematóide De Cisto de Soja*. Pesq. Agropec. Bras., Brasília.

[14] Pinto, R.M.C. *Regressão Logística*. At http://www.famat.ufu.br/espe/eea3/aulas_arquivos/rogerio/regressao%20logistica.pdf. Acessado em 25/08/2008.

[15] Rubin, D.B. (1976). *Inference and missing data*. Biometrika 63, New Jersey.

[16] SAS Institute Inc.(2002), *The LOGISTIC Procedure*, Version 9.0, Cary, NC, USA.

[17] SAS Institute Inc.(2002), *The NLP Procedure*, Version 9.0, Cary, NC, USA.

[18] SAS Institute Inc.(2002), *The MI Procedure*, Version 9.0, Cary, NC, USA.

[19] SAS Institute Inc.(2002), *The MIANALYZE Procedure*, Version 9.0, Cary, NC, USA.

[20] Vach, W. & Illi, S. (1997). *Biased Estimation of Adjusted Odds Ratios From Incomplete Covariate Data Due to Violation of the Missing at Random Assumption*. Biometrical Journal 39, Germany.

[21] Yuan, Y. *Multiple Imputation for Missing Data: Concepts and New Development*. SAS Institute Inc.

# Apêndice A

# Programas no *Software* R 2.7.1

```
/* Exemplo Prático */
/* Dados completos */
x<-c(5.7,4.5,4.7,4.5,3.6,4.3,4.1,3.7,6.8,5.7,6.3,6.0,5.5,8.2,6.0,4.9,5.5,
3.9,4.5,3.5,4.6,4.0,7.1,5.4,6.4,5.4,5.7)
w<-c(685,526,536,520,343,462,460,454,1076,722,814,782,689,1499,683,662,627,
423,513,383,517,448,970,681,800,775,731)
y<-log(w)
plot(x,y,xlab="Escolaridade",ylab="Log(Renda Média)",lwd=2,)
modelo1<-lm(y ~ x)
summary(modelo1)

/* Dados incompletos descartados */
x_i<-c(4.5,4.7,4.3,4.1,3.7,6.8,6.3,6.0,5.5,4.9,4.5,6.4,5.4)
w_i<-c(526,536,462,460,454,1076,814,782,689,662,513,800,775)
y_i<-log(w_i)
plot(x_i,y_i,xlab="Escolaridade (dados incompletos)",ylab="Log(Renda Média
(dados incompletos))",lwd=2)
modelo2<-lm(y_i~ x_i)
summary(modelo2)

/* Gráficos dos modelos ajustados */
plot(x,y,xlab="Escolaridade",ylab="Log(Renda Média)",col="blue")
```

```
abline(modelo1,col="blue")
legend(4.2,6.0,lty=c(1),c("Modelo ajustado em amostra sem dados missing"),
lwd=1,bty="n",cex=0.9, col="blue")
points(x_i,y_i,col="red",pch=7)
lines(x_i,fitted.values(modelo2),col="red")
legend(4.2,5.9,lty=c(1),c("Modelo ajustado em amostra com dados missing"),
lwd=1,bty="n",cex=0.9, col="red")

/* Estimando os valores missing como sendo parâmetros */
/* dados referentes aos x missing */
w_comp<-c(685,520,343,722,1499,683,627,423,383,517,448,970,681,731)
y_comp<-log(w_comp)
/* imputação dos valores de x missing */
input_x<-(y_comp-beta0)/beta1
input_x

/* Estimando os valores com os dados imputados */
x_input<-c(5.4447,4.5,4.7,4.4252,2.8859,4.3,4.1,3.7,6.8,5.6394,6.3,6.0,5.5,
8.3419,5.4339,4.9,5.1174,3.6614,4.5,3.2939,4.4038,3.8739,6.7317,5.4231,6.4,
5.4,5.6852)
w<-c(685,526,536,520,343,462,460,454,1076,722,814,782,689,1499,683,662,627,
423,513,383,517,448,970,681,800,775,731)
y<-log(w)
plot(x_input,y,xlab="Escolaridade",ylab="Log(Renda Média)",lwd=2,)
modelo1<-lm(y ~ x_input)
summary(modelo1)
```

# Apêndice B

# Programas no *Software* Maple 11

```
restart;
f:=2*exp(beta0+beta1*x+2*beta2*n)/(1+exp(beta0+beta1*x+2*beta2*n));
R0 := diff(f,n);
R1 := diff(R0,n);
R2 := diff(R1,n);
R3 := diff(R2,n);
R4 := diff(R3,n);
R5 := diff(R4, n);
R6 := diff(R5, n);
R7 := diff(R6, n);
R8 := diff(R7, n);
R9 := diff(R8, n);
R10 := diff(R9, n);
R11:= diff(R10, n);
R12 := diff(R11, n);
R13 := diff(R12, n);
R14 := diff(R13, n);
R15 := diff(R14, n);
R16 := diff(R15, n);
R17 := diff(R16, n);
R18 := diff(R17, n);
```

```
R19 := diff(R18, n);

R20 := simplify(R19,'size');
```

# Apêndice C

# Programas no *Software* SAS 9.0

O programa abaixo foi utilizado para os diferentes tamanhos amostrais considerados neste trabalho. Ou seja, para $n = 300$ e $n = 500$. Bem como para os diferentes porcentuais de *missing* adotados, $5\%, 10\%, 20\%$ e $50\%$.

```
/* Geração das Variáveis de Interesse */
%let B=1000;
%macro macro1(mac);
%do i = 1 %to &B;
proc iml;
beta0 = 0; beta1=1; beta2=1.5;
n=300;
x1_&i.  = J(&B,n,0); x2_&i.  = J(&B,n,0);
y_&i.  = J(&B,n,0); p_&i.  = J(&B,n,0);
s_&i.  = J(&B,n,0);
semente = J(&B,1,0); sem1 = J(&B,1,0);
call ranuni((30*&mac),semente);
sem1 = int(100*semente);
do j = 1 to n;
x1_&i.[&i,j] = ranbin(sem1[&i],1,0.4);
x2_&i.[&i,j] = 2*rangam(sem1[&i],1);
end;
```

```
s_&i.[&i,] = beta0+beta1*x1_&i.[&i,]+beta2*x2_&i.[&i,];
p_&i.[&i,] = exp(s_&i.[&i,])/(1+exp(s_&i.[&i,]));
do k = 1 to n;
y_&i.[&i,k] = ranbin(0,1,p_&i.[&i,k]);
end;
C_&i.=t(y_&i.[&i,])||t(x1_&i.[&i,])||t(x2_&i.[&i,])||t(p_&i.[&i,]);
cname = {"_y_&i._x1_&i._x2_&i._p_&i._"};
create d300.dados_macro&mac._&i.  from C_&i.  [ colname=cname ];
append from C_&i.;
run; quit;
%end; %mend;

%macro roda(mac);
%do t = 1 %to &mac;
%macro1(&t.);
%end;%mend;
%roda(1);

/* Estimação do Parâmetros */
%macro macro2(mac);
%do i = 1 %to &B;
proc nlp data=d300.dados_macro&mac._&i.  cov=2 vardef=n outest=est_macro&
mac._&i.  noprint;
max l_&i.;
parms beta0=0, beta1=1, beta2=1.5;
l_&i.=_y_&i._*log(exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._)/(1+exp(beta0+
beta1*_x1_&i._+beta2*_x2_&i._)))+(1-_y_&i._)*log(1-(exp(beta0+beta1*_x1_&i.
_+beta2*_x2_&i._)/(1+exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._))));
run; quit;

data parametros_macro&mac._&i.;
set est_macro&mac._&i.;
keep _TYPE_ beta0 beta1 beta2;
if _TYPE_ ne 'PARMS' then delete;
```

```
run;

data parametros_macro&mac._&i.;
set parametros_macro&mac._&i.;
keep beta0 beta1 beta2;
run;
%end; %mend;

%macro macro3(mac);
data macro&mac.allparametros;
set parametros_macro&mac._1;
run;
%mend;

%macro macro4(mac);
%do i = 2 %to &B;
data macro&mac.allparametros;
set macro&mac.allparametros parametros_macro&mac._&i.;
run;
%end; %mend;

%macro macro5(mac);
proc means data = macro&mac.allparametros maxdec=5;
ods output summary = macro&mac._media_dadosiniciais;
run;
%mend;

%macro roda(mac);
%do t = 1 %to &mac;
%macro2(&t.);%macro3(&t.);%macro4(&t.);%macro5(&t.);
%end;%mend;
%roda(1);

/* Intervalo de Confiança Empirico */
proc univariate data= macro1allparametros;
output out=aicemp0 pctlpre=P_ pctlpts = 0 to 100 by 2.5;
```

```
var beta0;
run;

proc univariate data= macro1allparametros;
output out=aicemp1 pctlpre=P_ pctlpts = 0 to 100 by 2.5;
var beta1;
run;

proc univariate data= macro1allparametros;
output out=aicemp2 pctlpre=P_ pctlpts = 0 to 100 by 2.5;
var beta2;
run;
```

/* Gerando os dados *missing* */

/* Gerando 5% de dados *missing* para n = 300 e estimando os parâmetros */

```
%let B=1000;
proc iml;
r300 = 1:300;
r300 = t(r300);
create r300 from r300;
append from r300;
quit; run;

%macro macro6(mac);
%do i=1 %to &B;
%let semente = %eval(17094*&i.*&mac.);
proc surveyselect data=r300 method=srs samprate=0.95 seed=&semente outall
out=r300_05_macro&mac._&i.;
run;
%end;%mend;

%macro macro7(mac);
%do i=1 %to &B;
data dadosmissing300_05_macro&mac._&i.;
merge d300.dados_macro&mac._&i.  r300_05_macro&mac._&i.;
```

```
drop _p_&i._ col1;
label Selected = r_&i.;
run;
%end;%mend;

%macro macro8(mac);
%do i=1 %to &B;
data dadosmissing300_05_macro&mac._&i.;
set dadosmissing300_05_macro&mac._&i.;
if selected eq '0' then _x2_&i._ = .;
run;
%end;%mend;

%macro macro9(mac);
%do i = 1 %to &B;
proc nlp data=dadosmissing300_05_macro&mac._&i.  cov=2 vardef=n outest=
obsest_05_macro&mac._&i.  nomiss noprint;
max l_&i.;
parms beta0=0, beta1=1, beta2=1.5;
l_&i.
=_y_&i._*log(exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._)/(1+exp(beta0+beta1*
_x1_&i._+beta2*_x2_&i._)))+(1-_y_&i._)*log(1-(exp(beta0+beta1*_x1_&i._+
beta2*_x2_&i._)/(1+exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._)))); run; quit;
data obsparam_05_macro&mac._&i.;
set obsest_05_macro&mac._&i.;
keep _TYPE_ beta0 beta1 beta2;
if _TYPE_ ne 'PARMS' then delete;
run;

data obsparam_05_macro&mac._&i.;
set obsparam_05_macro&mac._&i.;
keep beta0 beta1 beta2; run; %end; %mend;

%macro macro10(mac);
data macro&mac.allparobs_05;
```

```
set obsparam_05_macro&mac._1;
run;
%mend;

%macro macro11(mac);
%do i = 2 %to &B;
data macro&mac.allparobs_05;
set macro&mac.allparobs_05 obsparam_05_macro&mac._&i.;
run;
%end; %mend;

%macro macro12(mac);
proc means data = macro&mac.allparobs_05 maxdec=5;
ods output summary = macro&mac._media_CC_05;
run;
%mend;

%macro roda1(mac);
%do t = 1 %to &mac;
%macro6(&t.);%macro7(&t.);%macro8(&t.);%macro9(&t.);%macro10(&t.);
%macro11(&t.);%macro12(&t.);
%end;%mend;
%roda1(1);

/* Caso Completo Corrigido */
%macro eccc1;
%do i = 1 %to &B;
data macro1_yzero_&i.;
set dadosmissing300_05_macro1_&i.;
if _y_&i._ = 1 then delete;
if Selected = 1 then delete;
run;

data macro1_yzero_x1zero_&i.;
set macro1_yzero_&i.;
```

```
if _x1_&i._ = 1 then delete;
run;

data macro1_yum_&i.;
set dadosmissing300_05_macro1_&i.;
if _y_&i._ = 0 then delete;
if Selected = 1 then delete;
run;

data macro1_yum_x1zero_&i.;
set macro1_yum_&i.;
if _x1_&i._ = 1 then delete;
run;

data macro1_yzeroo_&i.;
set dadosmissing300_05_macro1_&i.;
if _y_&i._ = 1 then delete;
if Selected = 1 then delete;
run;

data macro1_yzero_x1um_&i.;
set macro1_yzeroo_&i.;
if _x1_&i._ = 0 then delete;
run;

data macro1_yumm_&i.;
set dadosmissing300_05_macro1_&i.;
if _y_&i._ = 0 then delete;
if Selected = 1 then delete;
run;

data macro1_yum_x1um_&i.;
set macro1_yumm_&i.;
if _x1_&i._ = 0 then delete;
run;
%end; %mend;
```

```
%eccc1;

%macro eccc2;
%do i = 1 %to &B;
proc iml;
use macro1_yzero_x1zero_&i.;
read all into macro1_yzero_x1zero_&i.;
k_&i.  = nrow(macro1_yzero_x1zero_&i.);
use macro1_yum_x1zero_&i.;
read all into macro1_yum_x1zero_&i.;
l_&i.  = nrow(macro1_yum_x1zero_&i.);
use macro1_yzero_x1um_&i.;
read all into macro1_yzero_x1um_&i.;
m_&i.= nrow(macro1_yzero_x1um_&i.);
use macro1_yum_x1um_&i.;
read all into macro1_yum_x1um_&i.;
n_&i.  = nrow(macro1_yum_x1um_&i.);
if k_&i.  = 0 | l_&i.  = 0 then a_&i.  = 0;
else a_&i.  = log(k_&i./l_&i.);
b_&i.  = l_&i.  * m_&i.;
c_&i.  = k_&i.  * n_&i.;
if b_&i.  = 0 | c_&i.  = 0 then d_&i.  = 0;
else d_&i.  = log(b_&i./c_&i.);
beta0 = -0.04807; beta1 = 1.09288; beta2 = 1.61677;
beta0_CC_&i.  = beta0 + a_&i.;
beta1_CC_&i.  = beta1 + d_&i.;
beta2_CC = beta2;
aa_&i.  =beta0_CC_&i.  || beta1_CC_&i.  || beta2_CC;
create ecc_&i.  from aa_&i.;
append from aa_&i.;
run; quit;
%end; %mend;
%eccc2;
```

```
data ecc_05;
set ecc_1;
run;
%macro merge;
%do i = 2 %to &B;

data ecc_05;
set ecc_05 ecc_&i.;
run;
%end; %mend;
%merge;

proc means data=ecc_05 maxdec=5;
run;

/* Imputação pela Média */
%let B=1000;
%macro media1;
%do i=1 %to &B;
data media_macro1_05_&i.;
set dadosmissing300_05_macro1_&i.;
if _x2_&i._ = '.'  then delete;
%end;%mend;
%media1;

%macro media2;
%do i=1 %to &B;
proc means data = media_macro1_05_&i.  mean maxdec=5;
ods output summary = media1_macro1_05_&i.;
var _x2_&i._;
run;
%end;%mend;
%media2;

%macro media3;
```

```
%do i=1 %to &B;
data media2_macro1_&i.;
set media1_macro1_05_&i.;
array _x2_&i.__Mean_{300} _x2_&i.__Mean_1-_x2_&i.__Mean_300;
do k = 1 to 300;
_x2_&i.__Mean_[k] = _x2_&i.__Mean;
end;
drop _x2_&i.__Mean k;
run;

proc transpose data=media2_macro1_&i.  out = media2_macro1_&i.;
run;

data media2_macro1_&i.;
set media2_macro1_&i.;
rename col1 = media_x2;
run;

data media05_macro1_&i.;
merge dadosmissing300_05_macro1_&i.  media2_macro1_&i.;
drop _name_ Selected;
run;
%end;%mend;
%media3;

%macro media4;
%do i=1 %to &B;
data media05_macro1_&i.;
set media05_macro1_&i.;
if _x2_&i._ = '.'  then _x2_&i._ = media_x2;
drop media_x2;
run;
%end;%mend;
%media4;

%macro media5;
```

```
%do i = 1 %to &B;
proc nlp data=media05_macro1_&i.  cov=2 vardef=n outest=est_media_macro1_&
i.  noprint;
max l_&i.;
parms beta0=0, beta1=1, beta2=1.5;
l_&i.=_y_&i._*log(exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._)/(1+exp(beta0+
beta1*_x1_&i._+beta2*_x2_&i._)))+(1-_y_&i._)*log(1-(exp(beta0+beta1*_x1_&i.
_+beta2*_x2_&i._)/(1+exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._))));
run; quit;

data parametros_macro1_&i.;
set est_media_macro1_&i.;
keep _TYPE_ beta0 beta1 beta2;
if _TYPE_ ne 'PARMS' then delete;
run;

data parametros_macro1_&i.;
set parametros_macro1_&i.;
keep beta0 beta1 beta2;
run;
%end; %mend;
%media5;

%macro media6;
data macro1_mediaallpar;
set parametros_macro1_1;
run;
%mend;
%media6;

%macro media7;
%do i = 2 %to &B;
data macro1_mediaallpar;
set macro1_mediaallpar parametros_macro1_&i.;
run;
```

```
%end; %mend;
%media7;

proc means data = macro1_mediaallpar maxdec=5;
ods output summary = macromedia_CC_05;
run;
%mend;

/* Etimando o valor para o erro da integral, para 5% */
%let B=1000;
%macro macro34(mac);
%do i=1 %to &B;
data CCmis300_macro&mac._05_&i.;
if _N_=1 then set obsparam_05_macro&mac._&i.(keep=beta0 beta1 beta2);
set dadosmissing300_05_macro&mac._&i.;
if selected eq '0' then _x2_&i._ = 0;
run;
%end;
%mend;

%macro macro35(mac);
%do i = 1 %to &B;
proc nlp data=CCmis300_macro&mac._05_&i.  cov=2 vardef=n outest=vm_macro
&mac._05_&i.;
max l_&i.;
parms k=2;
bounds k >= 0;
l_&i.=abs(-1048576*beta2**20*exp(beta0+beta1*_x1_&i._+2*beta2*k)*(exp(be
ta0+beta1*_x1_&i._+2*beta2*k)-1)*((exp(beta0+beta1*_x1_&i._+2*beta2*k))**
18-1048554*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**17+3463715961*(exp(beta
0+beta1*_x1_&i._+2*beta2*k))**16-1023045639024*(exp(beta0+beta1*_x1_&i._+
2*beta2*k))**15+71985471942420*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**14-
180772319795407*exp(beta0+beta1*_x1_&i._+2*beta2*k))**13+197908731051458
28*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**12-104957308999318032*(exp(beta
```

```
0+beta1*_x1_&i._+2*beta2*k))**11+283630951724635278*(exp(beta0+beta1*_x1_
&i._+2*beta2*k))**10-395931266069521660*(exp(beta0+beta1*_x1_&i._+2*beta2
*k))**9+283630951724635278*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**8-10495
7308999318032*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**7+19790873105145828*
(exp(beta0+beta1*_x1_&i._+2*beta2*k))**6-1807723197954072*(exp(beta0+beta
1*_x1_&i._+2*beta2*k))**5+71985471942420*(exp(beta0+beta1*_x1_&i._+2*beta
2*k))**4-1023045639024*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**3+346371596
1*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**2-1048554*exp(beta0+beta1*_x1_&i
._+2*beta2*k)+1)/(1+exp(beta0+beta1*_x1_&i._+2*beta2*k))**21;
run; quit;

data vm_macro&mac._05_&i.;
set vm_macro&mac._05_&i.;
keep _TYPE_ k;
if _TYPE_ ne 'PARMS' then delete;
run;

data vm_macro&mac._05_&i.;
set vm_macro&mac._05_&i.;
keep k;
run;
%end; %mend;

%macro macro36(ma);
%do i=1 %to &B;
data CCmis300_macro&mac._05_&i.;
if _N_=1 then set vm_macro&mac._05_&i.(keep=k);
set CCmis300_macro&mac._05_&i.;
run;
%end; %mend;

%macro macro37(mac);
%do i=1 %to &B;
data erro_macro&mac._&i.;
set CCmis300_macro&mac._05_&i.;
```

```
if selected = 0 then
erro_&i.=-0.000005412*(-1048576*beta2**20*exp(beta0+beta1*_x1_&i._+2*beta
2*k)*(exp(beta0+beta1*_x1_&i._+2*beta2*k)-1)*((exp(beta0+beta1*_x1_&i._+2
*beta2*k))**18-1048554*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**17+34637159
61*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**16-1023045639024*(exp(beta0+bet
a1*_x1_&i._+2*beta2*k))**15+71985471942420*(exp(beta0+beta1*_x1_&i._+2*be
ta2*k))**14-180772319795407*exp(beta0+beta1*_x1_&i._+2*beta2*k))**13+1979
0873105145828*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**12-10495730899931803
2*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**11+283630951724635278*(exp(beta0
+beta1*_x1_&ı._+2*beta2*k))**10-395931266069521660*(exp(beta0+beta1*_x1_&i
._+2*beta2*k))**9+283630951724635278*(exp(beta0+beta1*_x1_&i._+2*beta2*k))
**8-104957308999318032*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**7+197908731
05145828*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**6-1807723197954072*(exp(b
eta0+beta1*_x1_&i._+2*beta2*k))**5+71985471942420*(exp(beta0+beta1*_x1_&i
._+2*beta2*k))**4-1023045639024*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**3+3463
715961*(exp(beta0+beta1*_x1_&i._+2*beta2*k))**2-1048554*exp(beta0+beta1*_
x1_&i._+2*beta2*k)+1)/(1+exp(beta0+beta1*_x1_&i._+2*beta2*k))**21;
else erro_&i.=0;
run;

data dados_erro_macro&mac._&i.;
merge erro_macro&mac._&i.  CCmis300_macro&mac._05_&i.;
run;
%end;%mend;

%macro macro38(mac);
%do i=1 %to &B;
data dados_erro_macro&mac._&i.;
set dados_erro_macro&mac._&i.;
if erro_&i.  >= e-20 then erroo_i.  = 0;
else erroo_&i.  = erro_&i.;
drop beta0 beta1 beta2;
run;
%end; %mend;
```

```
%macro roda40(mac);
%do t = 1 %to &mac;
%macro34(&t.);
%macro35(&t.);
%macro36(&t.);
%macro37(&t.);
%macro38(&t.);
%end;%mend;
%roda40(1);

%macro macro38(mac);
%do i = 1 %to &B;
proc nlp data=dados_erro_macro&mac._&i.  cov=2 vardef=n outest=CCmisest_m
acro&mac._05_&i.  noprint; max l_&i.;
parms beta0=0, beta1=1, beta2=1.5;
l_&i.=(Selected*(log(exp(beta0+beta1*_x1_&i._+beta2*_x2_&i._)/(1+exp(beta
0+beta1*_x1_&i._+beta2*_x2_&i._)*exp(-(_x2_&i._)/2)))))+((1-Selected)*(log
((2*3.084411e-01*exp(beta0+beta1*_x1_&i._+2*0.1377935*beta2)/(1+exp(beta0
+beta1*_x1_&i._+2*0.1377935*beta2)))+(2*4.011199e-01*exp(beta0+beta1*_x1_
&i._+2*0.7294545*beta2)/(1+exp(beta0 + beta1*_x1_&i._+2*0.7294545*beta2))
)+(2*2.180683e-01*exp(beta0+beta1*_x1_&i._+2*1.8083429*beta2)/(1+exp(beta
0+beta1*_x1_&i._+2*1.8083429*beta2)))+(2*6.208746e-02*exp(beta0+beta1*_x1
_&i._+2*3.4014337*beta2)/(1+exp(beta0+beta1*_x1_&i._+2*3.4014337*beta2)))
+(2*9.501517e-03*exp(beta0+beta1*_x1_&i._+2*5.5524961*beta2)/(1+exp(beta
0+beta1*_x1_&i._+2*5.5524961*beta2)))+(2*7.530084e-04*exp(beta0+beta1*_x1
_&i._+2*8.3301527*beta2)/(1+exp(beta0+beta1*_x1_&i._+2*8.3301527*beta2)))
+(2*2.825923e-05*exp(beta0+beta1*_x1_&i._+2*11.8437858*beta2)/(1+exp(beta
0+beta1*_x1_&i._+2*11.8437858*beta2)))+(2*4.249314e-07*exp(beta0+beta1*_x
1_&i._+2*16.2792578*beta2)/(1+exp(beta0+beta1*_x1_&i._+2*16.2792578*beta2
)))+(2*1.839565e-09*exp(beta0+beta1*_x1_&i._+2*21.9965858*beta2)/(1+exp(
beta0+beta1*_x1_&i._+2*21.9965858*beta2)))+(2*9.911827e-13*exp(beta0+beta
1*_x1_&i._+2*29.9206970*beta2)/(1+exp(beta0+beta1*_x1_&i._+2*29.9206970*b
eta2)))+erroo_&i.)));
```

```
run; quit;

data obsparam_macro&mac._05_&i.;
set CCmisest_macro&mac._05_&i.;
keep _TYPE_ beta0 beta1 beta2;
if _TYPE_ ne 'PARMS' then delete;
run;

data obsparam_macro&mac._05_&i.;
set obsparam_macro&mac._05_&i.;
keep beta0 beta1 beta2;
run;
%end; %mend;

%macro macro39(mac);
data macro&mac.allparobs_Cmis_05;
set obsparam_macro&mac._05_1;
run;
%mend;

%macro macro40(mac);
%do i = 2 %to &B;
data macro&mac.allparobs_Cmis_05;
set macro&mac.allparobs_Cmis_05 obsparam_macro&mac._05_&i.;
run;
%end; %mend;

%macro macro41(mac);
proc means data = macro&mac.allparobs_Cmis_05 maxdec=5;
ods output summary = d100.macro&mac._media_Cmis_05;
run;
%mend;

%macro roda41(mac);
%do t = 1 %to &mac;
%macro38(&t.);
```

```
%macro39(&t.);%macro40(&t.);%macro41(&t.);
%end;%mend;
%roda41(1);
```