Tratamento de Dados Ausentes em Estudos Longitudinais

Stella Maris Lemos Nunes Baracho / UFVJM

1 Introdução

Em muitas situações há interesse em se estudar o comportamento de uma característica (variável) ao longo do tempo. Os dados deste tipo de estudo são denominados longitudinais, e, podem ser definidos como dados resultantes de observações coletadas em tempos distintos em uma mesma unidade amostral.

Devido à natureza dos estudos longitudinais, que podem durar meses ou até anos, e requerem observações nas mesmas unidades amostrais em várias ocasiões, torna-se freqüente a existência de dados ausentes. Embora o ideal em uma situação de dados ausentes seja a recuperação dos próprios dados, isto raramente ocorre na prática. Este texto apresenta algumas formas de tratar os dados ausentes em estudos longitudinais. Um enfoque especial é dado aos procedimentos baseados em imputação.

O presente trabalho foi motivado pelo interesse em se tratar dados reais, originados de um estudo multicêntrico, duplo-cego, de grande importância na determinação da eficácia de um novo medicamento antidepressivo.

2 Dados Ausentes

A ocorrência de não-resposta ou dados ausentes é um problema bastante comum em diversas áreas de pesquisa. Esta situação ocorre com freqüência em experimentos clínicos, estudos epidemiológicos, pesquisas amostrais, etc. As razões que conduzem a uma situação de dados ausentes podem ser as mais diversas possíveis. Dentre elas, pode-se citar:

- Perda do paciente;
- Doenças não relacionadas ao medicamento em estudo;

- Erros de transcrição ou digitação;
- Ineficiência do medicamento;
- Falta de cooperação do paciente;
- Questões mal-formuladas em um questionário.

O problema de lidar com valores ausentes está freqüentemente presente na análise de dados longitudinais. Em um estudo transversal, onde é obtida uma medida instantânea para cada unidade amostral, este tipo de problema pode ser mais facilmente evitado caso não haja falhas no planejamento. Entretanto, em estudos longitudinais, uma mesma unidade amostral é medida várias vezes ao longo do tempo. Neste caso, mesmo que não haja falhas de planejamento, o fato de se obter várias medidas ao longo do tempo em cada unidade amostral já representa um fator complicador na obtenção de dados.

Desde que os níveis dos fatores de um experimento são fixados pelo experimentador, dados ausentes, quando ocorrem, são mais freqüentes na variável resposta (Little e Rubin, 1987). Conseqüentemente, a maior parte da literatura enfoca o problema de lidar com dados ausentes neste tipo de variável. Diz-se que os dados longitudinais são balanceados em relação ao tempo se as observações forem feitas nos mesmos instantes de tempo em todas as unidades amostrais. Se houver observações ausentes, diz-se que a estrutura dos dados é incompleta. Dados incompletos podem resultar no desbalanceamento dos instantes de observação. Devido a este fato, a ocorrência de dados ausentes está sempre associada a um desbalanceamento dos dados.

Uma distinção importante a ser feita é se valores ausentes ocorrem intermitentemente ou como *dropouts*. Para compreender melhor esta distinção, considere a seguinte situação: é tomada uma seqüência de medidas $Y_1, Y_2, ..., Y_n$ na *i-ésima* unidade amostral. Valores ausentes ocorrem como *dropouts* quando ao se observar um Y_j faltoso também será observado Y_k faltoso para todo $k \geq j$; caso contrário os valores ausentes são chamados intermitentes (Diggle, Liang e Zeger, 1994). Alguns autores preferem denominar as situações descritas como padrão de ausência monótono e não-monótono respectivamente. Este texto enfoca os dados ausentes classificados como *dropouts* ou padrão monótono. A justificativa para isto

se deve ao fato de que este tipo é o mais comum em estudos longitudinais (Verbeke e Molenberghs, 2001).

Ao se deparar com um conjunto de dados incompletos deve-se classificar o mecanismo de ausência destes dados a fim de tomar as precauções necessárias na análise. O mecanismo de dados ausentes foi classificado por Little e Rubin (1987) e Davis (2001) como:

1. **Completamente Aleatório (MCAR – Missing Completely At Random):** quando a probabilidade de não-resposta é independente dos valores observados e ausentes (não depende da variável de interesse). Neste caso não é necessário cuidados adicionais na análise;
2. **Aleatório (MAR – Missing At Random):** quando a probabilidade de não-resposta é independente dos valores ausentes;
3. **Informativo (NMAR – Not Missing At Random):** quando a probabilidade de não-resposta depende dos valores ausentes.

O primeiro mecanismo possui uma suposição muito forte e raramente é satisfeito na prática. Quando ele ocorre, os dados não-observados constituem uma subamostra aleatória. O segundo mecanismo também é denominado de Ignorável. O termo Ignorável é usado para indicar que não é necessário especificar um modelo para a não-resposta. É importante ressaltar que é o mecanismo de dados ausentes é que pode ser ignorado e não os dados ou as unidades com dados ausentes (Barroso, 1995). O terceiro mecanismo de ausência é não-ignorável devido à falta de aleatoriedade da não-resposta. Portanto, nesta situação torna-se necessário especificar um modelo para a não-resposta. Little e Rubin (1987) apresentam um exemplo simples em que é feita a distinção destes mecanismos. Uma breve descrição deste exemplo é dada a seguir.

Exemplo 2.1 *Considere duas variáveis contínuas, uma sujeita a não-resposta. Suponha que* $X =$ *idade e* $Y =$ *renda.*

Se a probabilidade que a renda seja observada é a mesma para todos os indivíduos, independente de sua idade ou renda, então os dados são MCAR. Se a probabilidade que a renda seja observada varia de acordo com a idade do indivíduo mas não varia de acordo com sua renda, então os dados são MAR. Mas, se a

probabilidade de que a renda seja observada varia de acordo com a própria renda (geralmente a perda é maior entre as pessoas que possuem renda mais alta), então os dados são NMAR. Neste caso, a não-resposta depende dos dados ausentes (Little e Rubin, 1987) e, portanto, a falta de especificação de um modelo para a não-resposta evidentemente vicia a análise estatística.

A maioria dos métodos propostos na literatura para tratar o problema de não-resposta assume que os dados são MAR. Os métodos propostos para tratar dados ausentes neste trabalho também assumirão que o mecanismo de ausência é MAR ou Ignorável, embora alguns deles também sejam adequados sob a suposição MCAR.

3 Tratamento de Dados Ausentes em Estudos Longitudinais

Embora o ideal em uma situação de dados ausentes seja tentar recuperar os próprios dados, isto raramente ocorre na prática. Segundo Verbeke e Molenberghs (2001), três abordagens podem ser dadas para o tratamento dos dados ausentes:

1. Análise dos Casos Completos (ACC);
2. Análise dos Casos Disponíveis (ACD);
3. Procedimentos Baseados em Imputação (PBI).

3.1 Análise dos Casos Completos (ACC)

Ao se deparar com uma situação de dados incompletos, na impossibilidade de corrigi-la, é comum descartar as unidades que possuem observações parciais ou incompletas. Este procedimento é denominado de análise dos casos completos.

Este método pode ser satisfatório desde que o mecanismo de ausência dos dados seja MCAR ou, ainda, quando o mecanismo for MAR e o número de respostas ausentes for pequeno. Entretanto, a estratégia de descartar as unidades incompletas é geralmente inapropriada pois, é de interesse de um pesquisador fazer inferências sobre a população alvo inteira e não apenas sobre a porção da população que fornece respostas completas para todas as variáveis relevantes na análise. Além disso, a retirada das unidades incompletas do estudo evidentemente diminui o tamanho

amostral, aumentando-se assim, a variabilidade dos estimadores e, portanto, diminuindo-se a precisão. Enquanto que usar somente os casos completos tem sua simplicidade e a vantagem de um planejamento completamente balanceado, perde-se a informação dos casos incompletos ao excluí-los da análise.

Em estudos longitudinais, mesmo que a perda seja pequena, a análise dos casos completos pode produzir resultados viciados. Pacientes que possuem observações incompletas (*dropouts)* podem ter um perfil diferente daqueles que permanecem até o final do estudo.

3.2 Análise dos Casos Disponíveis (ACD)

Ao invés de excluir da análise as unidades que possuem observações parciais ou incompletas, uma outra maneira de tratar os dados é através da análise dos casos disponíveis. A maior parte dos modelos implementados em pacotes estatísticos adota esta estratégia. Nesta situação tem-se a vantagem de se usar toda a informação disponível, e, sendo assim, este método torna-se mais eficiente do que a análise dos casos completos. Entretanto, uma desvantagem é que o este método requer que o processo de ausência seja MCAR ou MAR, desde que a taxa de não-resposta seja pequena, o que é uma suposição restrita. Outra desvantagem evidente da análise dos casos disponíveis é o desbalanceamento dos dados, o que geralmente aumenta as dificuldades técnicas além de tornar ineficientes os métodos do tipo ANOVA (Diggle, Liang e Zeger, 1994).

3.3 Procedimentos Baseados em Imputação

Uma forma alternativa de obter um conjunto de dados balanceado em vez de descartar as unidades que possuem observações incompletas é estimá-las através da utilização de procedimentos baseados em imputação. Imputação é uma técnica utilizada em pesquisas estatísticas há mais de 50 anos, que consiste em prever os valores ausentes a fim de completar os dados e então analisar o conjunto de dados obtido (valores observados mais valores imputados), através de métodos estatísticos padrão.

Várias formas de imputação têm sido propostas na literatura, as quais se enquadram em um dos dois tipos de imputação: simples ou múltipla. Através do método de imputação simples cada valor ausente é substituído por um único valor imputado. Há várias formas de se fazer isto, mas, deve-se tomar todo cuidado possível ao utilizar-se desta técnica, pois ela distorce a incerteza dos dados. Em vez de preencher com um mesmo valor cada observação ausente, o processo de imputação múltipla de Little e Rubin (1987) substitui cada valor faltoso por mais de um valor imputado. Os conjuntos de dados completados são analisados e usados para estimar um valor plausível que representa a incerteza sobre o valor a ser imputado.

Grandes esforços têm sido feitos para a obtenção de métodos de imputação aplicados a diversas áreas. Neste texto, o enfoque dado às técnicas de imputação vai de encontro aos estudos longitudinais. O objetivo é imputar dados ausentes do tipo *dropouts* para a variável resposta. Contudo, vale a pena ressaltar que as técnicas aqui apresentadas também se aplicam da mesma forma para os dados intermitentes.

4 Imputação Simples para Dados Longitudinais

Dentro do contexto de dados longitudinais, os valores observados podem ser usados de forma simples para imputar valores para as observações ausentes a fim de completar o conjunto de dados para uma análise posterior. Há várias maneiras de usar a informação observada e algumas delas serão brevemente descritas (Little e Rubin, 1987; Verbeke e Molenberghs, 2001).

4.1 Imputação Através da Última Observação

Uma maneira bastante simples de preencher um dado ausente é através da imputação da última observação (IUO). Neste caso, a idéia é substituir os dados ausentes de cada paciente pelo último valor observado neste mesmo paciente. Embora esta técnica possa ser aplicada a padrões de dados monótonos e não-monótonos, geralmente ela é usada em conjunto de dados onde a ausência é caracterizada como *dropouts*. Suposições fortes e freqüentemente não-realistas são feitas para assegurar a validade deste método.

4.2 Imputação Através da Média

Em um estudo longitudinal, duas maneiras distintas de imputação podem ser consideradas através da média:

1. **Média dos Tempos (IMT):** média dos valores observados em tempos distintos para *i – ésima* unidade amostral ($\bar{y}_{i*}$);
2. **Média dos Pacientes (IMP):** média dos valores observados nas diferentes unidades amostrais no *t – ésimo* tempo ($\bar{y}_{*t}$).

No primeiro caso, a idéia é calcular a média das observações presentes para a *i-ésima* unidade amostral nos diferentes tempos, obtendo-se assim, o valor imputado para os dados ausentes da *i-ésima* unidade amostral. Já no segundo caso, o valor imputado é calculado através da média das observações presentes das unidades amostrais em um tempo t_j é substituída pelo mesmo valor.

Ao utilizar as imputações obtidas através da média deve-se levar em consideração que, embora este método não altere a média amostral, ele distorce outros aspectos importantes da distribuição.

4.3 Imputação Através de Regressão

Uma forma mais promissora de imputação é através da substituição dos valores ausentes por valores preditos através de um modelo de regressão dos dados observados. O método foi proposto por Buck (1960) para uma amostra normal multivariada. Para o caso de observações dependentes, situação de interesse neste texto foi feita uma adaptação via modelo linear misto. O processo de imputação através da regressão considerando este modelo pode ser assim descrito:

1. Modelar os dados através do modelo linear misto (Laird e Ware, 1982), isto é:

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n; \qquad j = 1, \ldots, a_i;$$

em que y_i para $j = 1, \ldots, a_i$ é observado e $\boldsymbol{y}_i$ para $j = a_i + 1,..., n_i$ é ausente. No caso de dados balanceados, $n_i = m$.

2. Estimar β e predizer $\mathbf{b_i}$ utilizando os dados observados;

3. Estimar os valores ausentes através dos coeficientes obtidos em 2 e das matrizes observadas $\mathbf{X}_i$ e $\mathbf{Z}_i$, isto é:

$$\hat{y}_i = \mathbf{X_i}\hat{\beta} + \mathbf{Z_i}\hat{b}_i \quad \text{para } j = a_i + 1,...,n_i.$$

4. Obter o conjunto de dados completado: *observados + imputados.*

A predição de dados ausentes através deste método parece bem razoável pois, além de levar em conta as estimativas dos efeitos fixos, também leva em conta o efeito aleatório predito para cada paciente.

Como pode ser observado, os métodos de imputação simples para dados longitudinais propostos neste texto têm como principal característica a simplicidade. Uma vantagem de se utilizar métodos de imputação simples é que estes geralmente fornecem estimativas pontuais consistentes para o mecanismo de ausência considerado. Além disso, recuperam o balanceamento dos dados aumentando o poder dos testes estatísticos. Entretanto, não se deve esquecer que a variância geralmente é subestimada ao se utilizar estes métodos. A vantagem de se obter um planejamento balanceado através de métodos de imputação deve ser avaliada diante da imprecisão das estimativas que serão usadas. Deve-se levar em consideração o mecanismo de ausência e a quantidade de dados ausentes para decidir se é viável ou não usar a imputação simples como uma forma de tratar os dados ausentes.

5 Imputação Múltipla

Alguns métodos de imputação simples podem produzir estimativas pontuais inconsistentes dependendo do mecanismo de ausência de dados. Este problema pode ser superado utilizando-se o método de imputação através de regressão. Entretanto, o problema de subestimar a variabilidade dos estimadores é comum em todos os métodos descritos anteriormente, desde que eles tratam valores imputados como valores observados. Imputação múltipla é uma abordagem analítica que enfoca estes problemas.

Introduzida por Rubin em 1978 a imputação múltipla vem ocupando uma posição de destaque na literatura de dados ausentes devido à sua aplicação em uma variedade de contextos. A idéia é substituir cada valor ausente por dois ou mais valores

imputados. Ao imputar vários valores para cada observação ausente a incerteza é explicitamente reconhecida (Rubin, 1987).

O processo de imputação múltipla consiste basicamente de três passos:

1. **Imputação:** Para cada valor ausente, são gerados *M* valores ($M \geq 2$);
2. **Análise:** Os *M* valores são organizados de forma que o primeiro valor imputado para cada dado ausente produz o primeiro conjunto de dados completado, o segundo valor imputado para cada dado ausente produz o segundo conjunto de dados completado e assim por diante. Cada conjunto de dados completado é analisado usando métodos padrão para dados completos.
3. **Combinação:** Finalmente, os resultados das *M* análises são combinados permitindo que a incerteza da imputação seja considerada.

O passo de imputação certamente é o mais crítico. Neste momento está sendo considerado o mecanismo de ausência de dados. A suposição do mecanismo MAR permite gerar imputações a partir da distribuição dos dados ausentes condicionada nos dados observados. O modelo utilizado no passo de imputação não - necessariamente tem que ser o mesmo modelo utilizado no passo de análise (Rubin, 1987). Esta flexibilidade torna o processo de imputação múltipla ainda mais atrativo, pois nem sempre o modelo utilizado para fazer a imputação é o mais adequado para analisar os dados.

Schafer (1997) mostra que o processo de imputação múltipla pode ser altamente eficiente. Se a fração de informação ausente é γ, a eficiência relativa (na escala da variância) de uma estimativa pontual baseada em *M* imputações para uma baseada em um número infinito de imputações é aproximadamente $(1 + \gamma / M)^{-1}$. Isto pode ser melhor observado através da Tabela 5.1.

Tabela 5.1: Eficiência da Imputação Múltipla (%)

	γ				
M	**0,1**	**0,3**	**0,5**	**0,7**	**0,9**
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

A falta de familiaridade com a imputação múltipla leva a acreditar que além da complexidade inerente ao processo, outra dificuldade seria lidar com muitos conjuntos de dados completados. Entretanto, isso raramente ocorre neste processo. Quando, por exemplo, 20% das informações são ausentes, uma estimativa baseada em $M = 3$ imputações terá um erro-padrão somente $\sqrt{1+0{,}2/3} = 1{,}033$ vezes maior do que a estimativa com $M = \infty$. Considerando $\gamma = 0{,}5$ (50% de dados ausentes) uma estimativa baseada em $M = 5$ imputações terá um erro-padrão somente de $\sqrt{1+0{,}5/5} = 1{,}049$ vezes maior do que a estimativa com $M = \infty$. Portanto, torna-se claro que um número pequeno de imputações é suficiente ao se realizar o processo de imputação múltipla.

5.1 Processo de Imputação Múltipla para Dados Longitudinais

Será dada abaixo uma descrição detalhada dos passos do processo de imputação múltipla para dados longitudinais proposto neste texto. Para tal, levar-se-á em consideração o modelo de regressão linear múltipla para a realização do passo de imputação e o modelo linear misto para o passo de análise dos dados completados. Obviamente a utilização do modelo linear misto no passo de imputação seria mais indicado para o processo de imputação múltipla de dados longitudinais. Entretanto, a dificuldade computacional associada a esta tarefa se tornou empecilho para sua realização. A justificativa para a escolha do modelo de regressão linear múltipla para o passo de imputação se baseia no fato de que a perda de eficiência é mínima ao se considerar o estimador de mínimos quadrados ordinários ao invés de estimador de mínimos quadrados generalizados para os efeitos fixos (Diggle, Liang e Zeger, 1994). Diante disso torna-se claro que as estimativas pontuais dos efeitos fixos são praticamente as mesmas para os dois modelos. Como no passo de imputação o interesse está justamente nas estimativas pontuais dos β´*s*, o modelo de regressão linear múltipla é bem razoável para a tarefa de imputação.

Considere o modelo de regressão linear múltipla

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, N, \ (N = mn).$$

Como suposição associada a este modelo tem-se que $\boldsymbol{\varepsilon_i} \sim N\ (0, \sigma^2)$. Conseqüentemente $\boldsymbol{y_i} \sim N\ (\mathbf{X}_i \beta, \sigma^2)$. Para uma situação de dados ausentes, $\boldsymbol{y_i}$ para $i = 1,\ldots, a$ é observado e $\boldsymbol{y_i}$ para $i = a + 1,\ldots, N$ é ausente. Os dados observados e o vetor de parâmetros a eles associados podem ser representados por:

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_a \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12}\ldots & x_{1p-1} \\ 1 & x_{21} & x_{22}\ldots & x_{2p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{a1} & x_{a2}\ldots & x_{ap-1} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Considere $\beta = (\mathbf{X'X})^{-1}\mathbf{X'y}$ o estimador de mínimos quadrados ordinários dos parâmetros e $\hat{\varepsilon} = \mathrm{y} - \mathbf{X}\hat{\beta}$ o vetor dos resíduos associados ao modelo de regressão para os dados observados. O processo de imputação múltipla se baseia nestas quantidades como ponto de partida.

A justificativa teórica para imputação múltipla é melhor compreendida usando conceitos Bayesianos. Seja $\boldsymbol{\delta}$ o vetor dos coeficientes associados ao modelo de regressão múltipla considerando. A distribuição a posteriori de $\boldsymbol{\delta}$ envolve apenas as unidades de **y** observadas. A justificativa para isto se deve ao fato do mecanismo de ausência dos dados se MAR (ou ignorável). Utilizando uma distribuição imprópria convencional para $\boldsymbol{\delta}$, Pr($\boldsymbol{\delta}$) α constante, a distribuição a posteriori de σ^2 é dada por:

$$(\sigma^2 \mid \mathbf{y}) \sim \hat{\varepsilon}'\hat{\varepsilon} / \chi^2_{a-p},$$

Em que χ^2_{a-p} é uma variável aleatória com distribuição χ^2 com $(a - p)$ graus de liberdade $(a > p)$. A distribuição a posteriori de $\boldsymbol{\beta}$ condicionada a σ^2 é dada por (Rubin, 1987):

$$(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) \sim \mathrm{N}(\hat{\beta}, \sigma^2 (\mathbf{X'X})^{-1}).$$

O processo de imputação múltipla pode ser brevemente descrito por:

1. **Algoritmo de Imputação**

   - Calcular $\hat{\beta} = (X'X)^{-1}X'y$ e $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$ a partir dos dados observados.
   - Gerar uma variável aleatória χ^2_{a-p}, ou seja:

$$k = \chi^2_{a-p}$$

   e então calcular σ^2_* dado por:

$$\sigma^2_* = \hat{\varepsilon}'\hat{\varepsilon}/k$$

   - Gerar uma variável aleatória $N_p(\hat{\beta}, \sigma^2_*(\mathbf{X'X})^{-1})$ e armazená-la em um vetor de dimensão *p,* ou seja:

$$\beta_* = N_p(\hat{\beta}, \sigma^2_*(\mathbf{X'X})^{-1})$$

   - Gerar uma variável aleatória $y_i \sim N(X_i\beta_*, \sigma^2_*)$ para todo $i = a+1, ..., N$.
   - Um novo valor imputado para os dados ausentes é inicializado ao gerar um novo valor do parâmetro σ^2_* (o que consiste em gerar um novo valor para *k*). Assim, para se obter *M* valores imputados para cada dado ausente, deve-se repetir o algoritmo de imputação *M* vezes.

2. **Análise**

   Usando os dados completados, o modelo considerado para a análise é o modelo linear misto dado por:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\lambda} + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i.$$

   Para não haver problemas na notação, foi feita a opção de denotar por $\boldsymbol{\lambda}$ o vetor de efeitos fixos do modelo linear misto no processo de imputação múltipla. Seja $\boldsymbol{\lambda}$ o vetor de parâmetros do modelo a ser estimado. Usando um método de estimação a escolha (máxima verossimilhança ou máxima

verossimilhança restrita), estima-se $\boldsymbol{\lambda}$ e a respectiva variância, **U** (denominada de variância dentro da imputação) M vezes. As análises dos M conjuntos de dados completados resultam em $\hat{\lambda}^{(m)}$ e $\mathbf{U}^{(m)}$ para $m = 1, 2, ..., M$.

3. **Combinação**

Depois de obter as M imputações para os dados ausentes e analisar os M conjuntos de dados completados, os resultados das análises são combinados. Sejam:

$$\bar{\lambda} = \frac{1}{M}\sum_{m=1}^{M}\hat{\lambda}^{(m)},$$

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M}U^{(m)},$$

$$\mathrm{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\lambda}^{(m)}-\bar{\lambda})(\hat{\lambda}^{(m)}-\bar{\lambda})',$$

em que:

- $\bar{\lambda}$ é a média da M estimativas obtidas para $\boldsymbol{\lambda}$;
- $\bar{U}$ é a média das variâncias obtidas dentro do processo de imputação;
- B é a variância entre imputações.

A quantidade

$$\mathrm{T} = \bar{U} + (1 + M^{-1})\,\mathrm{B},$$

é a variância total de $\bar{\lambda}$. Inferências são baseadas na aproximação

$$\mathrm{T}^{-1/2}\,\bar{\lambda} \sim t_{v},$$

em que t_v é a distribuição t com v graus de liberdade dados por:

$$v = (M-1)\left[1+\frac{\bar{U}}{(1+M^{-1})B}\right]^{2}.$$

Assim, uma estimativa intervalar de 100(1- α)% para λ (escalar) é dada por:

$$\bar{\lambda} \pm t_{v,1-\alpha/2} T^{1/2},$$

e o valor p para testar a hipótese nula $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ contra a hipótese alternativa $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$ é dado por:

$$\Pr\left[F_{1,v} > (\lambda_0 - \bar{\lambda})^2 T^{-1} \right]$$

em que $F_{1,v}$ é uma variável aleatória F com 1 e v graus de liberdade (Rubin, 1987).

6 Aplicação

Apresenta-se abaixo uma aplicação dos métodos propostos neste texto em um conjunto de dados reais. Os dados são referentes a um estudo sobre a eficácia e tolerabilidade do medicamento tianeptina no tratamento de episódios depressivos maiores. Trata-se de um estudo multicêntrico, duplo-cego, controlado com placebo. A tianeptina é um antidepressivo cuja eficácia é avaliada através do escore obtido na escala MADRS, sendo que quanto maior o escore obtido, pior o quadro clínico do paciente.

Foram acompanhados 123 pacientes em um estudo de natureza longitudinal. Como principal característica, o conjunto de dados provenientes deste estudo era incompleto. Sendo assim, dados referentes a algumas avaliações foram excluídos ou imputados e os conjuntos de dados obtidos foram analisados usando os métodos de tratamento de dados ausentes.

6.1 Os Dados da Tianeptina

O principal objetivo deste estudo foi comparar a eficácia e tolerabilidade do medicamento tianeptina com as do placebo em episódios de intensidade moderada ou grave. O ensaio foi realizado em três centros: Rio de Janeiro, Campinas e Belo Horizonte, sendo os dados resultantes analisados conjuntamente (Lopes Rocha, 1995). A amostra foi constituída por homens e mulheres com idade entre 18 e 60 anos. A eficácia terapêutica desse antidepressivo foi avaliada através do escore

obtido na escala MADRS (nível de depressão) após 7, 14, 21, 28 e 42 dias de tratamento. Além da covariável grupo, avalia-se neste estudo a importância da contribuição das covariáveis início (medida do nível de depressão no dia base) e tempo (7, 14, 21, 28 e 42 dias após o início do tratamento).

O estudo foi realizado com 123 pacientes. Desses, 61 foram submetidos ao tratamento com placebo e 62 foram submetidos ao tratamento com a tianeptina. Um ponto de grande importância a ser destacado neste estudo é que 16 pacientes possuíam dados ausentes em uma ou mais ocasiões. De forma geral, as estatísticas descritivas por ocasião estão apresentadas na Tabela 6.1.1.

Tabela 6.1.1: Descrição do Nível de Depressão (MADRS) por Tempo

Estatísticas	**Tempo**				
Descritivas	**Dia7**	**Dia 14**	**Dia 21**	**Dia 28**	**Dia 42**
N	123	121	116	113	107
N*	0	2	7	10	16
Média	31,42	26,73	22,88	20,40	17,26
Desvio Padrão	8,14	10,20	11,21	11,67	12,17
Mínimo	8,00	2,00	0,00	0,00	0,00
Máximo	56,00	56,00	56,00	54,00	54,00

Observa-se, a partir da Tabela 6.1.1 que, com o passar do tempo existe uma pequena diminuição no nível de depressão dos pacientes. Pode-se observar também que, com o decorrer do tempo aumenta-se o número de observações ausentes (N*) e ao final do experimento este número é aproximadamente de 6% do número total de observações no estudo. As Tabelas 6.1.2 e 6.1.3 apresentam as estatísticas descritivas do nível de depressão por grupo.

A partir das Tabelas 6.1.2 e 6.1.3 pode-se observar que o número de observações ausentes é o mesmo para os dois grupos. Parece existir uma leve diminuição no nível

Tabela 6.1.2: Descrição do Nível de Depressão por Tempo para o Grupo Placebo

Estatísticas	**Tempo**				
Descritivas	**Dia7**	**Dia 14**	**Dia 21**	**Dia 28**	**Dia 42**
N	61	60	58	56	53
N*	0	1	3	5	7
Média	32,16	27,77	24,55	22,07	19,74
Desvio Padrão	9,07	10,80	11,53	12,52	13,16
Mínimo	8,00	2,00	0,00	0,00	0,00
Máximo	56,00	56,00	56,00	54,00	54,00

Tabela 6.1.3: Descrição do Nível de Depressão por Tempo para o Grupo Tianeptina

Estatísticas	**Tempo**				
Descritivas	**Dia7**	**Dia 14**	**Dia 21**	**Dia 28**	**Dia 42**
N	62	61	58	57	54
N*	0	1	4	5	8
Média	30,69	25,70	21,21	18,75	14,83
Desvio Padrão	7,11	9,56	10,71	10,62	10,68
Mínimo	10,00	3,00	0,00	0,00	0,00
Máximo	48,00	43,00	48,00	45,00	40,00

de depressão dos pacientes com o passar do tempo para os dois grupos, existindo portanto um efeito placebo, já que pacientes que não receberam droga apresentam melhora no quadro clínico. Ao comparar os dois grupos parece que a diferença entre suas médias aumenta com o tempo, no entanto, a Figura 6.1.1 não torna esta evidência tão clara. As Figuras 6.1.2, 6.1.3 e 6.1.4 mostram o perfil geral dos pacientes bem como o perfil para cada grupo. Elas também sugerem que o nível de depressão tem uma ligeira diminuição com o tempo.

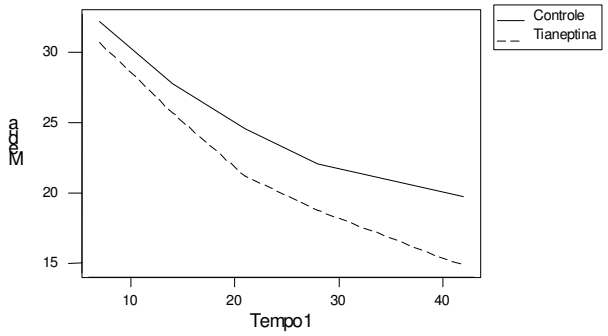

Figura 6.1.1: Gráfico de Evolução das Médias

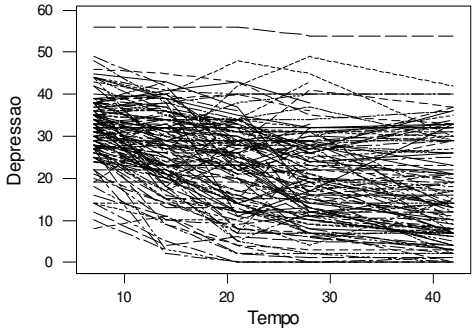

Figura 6.1.2: Gráfico de Perfil Geral dos Pacientes

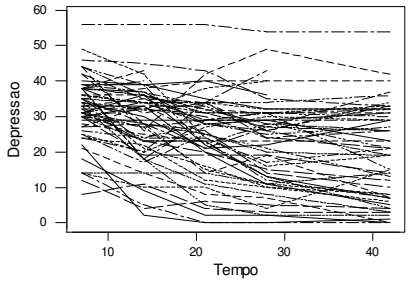

Figura 6.1.3: Gráfico de Perfil para o Grupo Controle

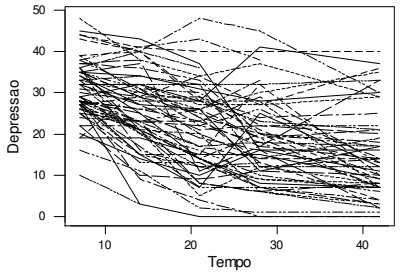

Figura 6.1.4: Gráfico de Perfil para o Grupo Tianeptina

6.2 O Modelo Avaliado

A modelagem dos dados da tianeptina foi realizada através da função *lme* do S-Plus. O método de estimação considerado foi o de máxima verossimilhança restrita. A construção do modelo avaliado foi feita baseando-se nos casos disponíveis que consistem de 580 observações. Considerou-se inicialmente a inclusão de todas as covariáveis e o termo de interação da covariável grupo com o tempo como sugerido pela Figura 6.1.1. Inicialmente, dois modelos foram considerados:

$$\begin{aligned} \textit{Nível de Depressão}_{ij} = {} & (\beta_0 + b_{0i}) + \beta_1 \textit{Grupo}_1 + \beta_2 \textit{Início}_i + \\ & (\beta_3 + b_{1i}) \textit{Tempo}_{ij} + \beta_4 (\textit{Grupo} * \textit{Tempo})_{ij} + \varepsilon_{ij}, \end{aligned} \tag{1}$$

$$\begin{aligned} \textit{Nível de Depressão}_{ij} = {} & (\beta_0 + b_{0i}) + \beta_1 \textit{Grupo}_1 + \beta_2 \textit{Início}_i + \\ & \beta_3 \textit{Tempo}_{ij} + \beta_4 (\textit{Grupo} * \textit{Tempo})_{ij} + \varepsilon_{ij}, \end{aligned} \tag{2}$$

em que:

- $\textit{Nível de Depressão}_{ij}$ é o nível de depressão de *i – ésimo* paciente no *j – ésimo* tempo;
- β_0 é a média geral do nível de depressão dos pacientes;
- b_{0i} é o desvio médio de cada paciente atribuído à média geral do nível de depressão dos pacientes;
- β_1 é o efeito do grupo;
- β_2 é o efeito do início;
- β_3 é o efeito do tempo;
- β_4 é o efeito da interação do grupo como o tempo;
- b_{1i} é o efeito aleatório do tempo;
- ε_{ij} é o erro aleatório;
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I_m)$, $\mathbf{b}_i \sim N(0, B)$,

$$\mathbf{B} = \begin{pmatrix} \sigma_b^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_a^2 \end{pmatrix}, \; Cov(b_i, \varepsilon_i) = 0.$$

Ou seja, os dois modelos diferem pela presença do componente aleatório no termo referente ao tempo. Foi realizado o teste da razão de verossimilhança para comparar os modelos (1) e (2). A Tabela 6.2.1 apresenta o resultado deste teste.

Tabela 6.2.1: Teste da Razão de Verossimilhança

Modelo	**GL**	**logver**	**Valor p**
5.1	9	-1943,84	<0,0001
5.2	7	-1986,30	

Portanto, observando-se o resultado do teste da razão de verossimilhança percebe-se que o efeito aleatório na inclinação é significativo. Para testar os efeitos fixos, foi utilizado o teste de Wald apresentado na Tabela 6.2.2.

Tabela 6.2.2: Resultado do Ajuste do Modelo Considerando ACD

Efeito	**Estimativa(EP)**	**Valor p**
Intercepto	4,28 (3,92)	0,2752
Grupo	-0,23 (1,38)	0,8669
Início	0,80 (0,11)	< 0,0001
Tempo	-0,31 (0,04)	< 0,0001
Grupo*Tempo	-0,13 (0,06)	0,0210

Como pode ser observado, o efeito de grupo não foi significativo. Entretanto, a interação entre o tempo e grupo foi significativa como indicado pelas estatísticas descritivas. Um fato importante a ser ressaltado é que não houve violação das suposições de independência dos erros e da variância constante associadas ao modelo ajustado. Para verificar este fato foi realizada a análise de resíduos. O modelo (1) se

mostrou adequado para ajustar os dados considerando a análise dos casos disponíveis.

6.3 Resultados

Serão apresentados os resultados dos ajustes do modelo (1) considerando análise dos casos completos e procedimentos baseados em imputação. Para o caso de imputação múltipla, o modelo considerado no passo de imputação pode ser representado por:

$$Nível\ de\ Depressão_i = \beta_0 + \beta_1 Grupo_i + \beta_2 Início + \beta_3 Tempo_i + \beta_4 \left(Grupo * Tempo\right)_i + \varepsilon_i$$

em que $i = 1, 2, ..., mn$. Ou seja, para o passo de imputação a dependência entre as observações feitas em um mesmo paciente não foi modelada. Uma vez imputados os dados ausentes, o modelo considerado para analisar os dados foi o (1).

As tabelas 6.3.1 e 6.3.2 apresentam os resultados do ajuste do modelo (1) para as diferentes formas de tratamento de dados ausentes.

Analisando-se os resultados das Tabelas 6.2.2, 6.3.1 e 6.3.2 percebe-se que os efeitos do início e do tempo foram significativos em todas as formas de tratamento de dados ausentes, enquanto que o intercepto e o grupo não foram significativos em nenhuma delas. Um fato interessante ocorreu ao comparar o que aconteceu com o efeito da interação entre o grupo e o tempo. Através da ACD, IMT, IUO e IR esse efeito foi significativo. A ACD considera as observações presentes como se esta fosse uma situação real (ignora completamente as observações ausentes). Ao se desconsiderar as observações ausentes corre-se o risco de estar excluindo da análise observações que estariam traçando um perfil diferente dos que foram observados. Ao se utilizar um método de imputação simples para recuperar os dados ausentes, por mais promissor que seja este método, uma vez imputado, este dado passa a ser considerado como observado. Esta limitação pode distorcer a análise estatística.

Através da imputação múltipla, mesmo se considerando um número pequeno de imputações para cada observação ausente (2 ou 3), o efeito da interação do grupo com o tempo não mais se mostrou significativo. Ou seja, ao se levar em consideração a incerteza associada ao processo de imputação um efeito deixou de ser significativo.

Isto mostra a importância de se tratar um dado ausente como tal. Sendo assim, acredita-se que o modelo adequado para explicar o nível de depressão dos pacientes seria aquele que considera o início e o tempo como fatores significativos. Portanto, conclui-se que o medicamento tianeptina não diminui significativamente o nível de depressão dos pacientes, e esta diminuição não aparece nem mesmo associada ao medicamento com o tempo. A redução do nível de depressão dos pacientes pode ser explicada pelo nível depressão que este paciente tinha no início de estudo e pela ação do tempo.

Tabela 6.3.1: Resultado dos Métodos

Método	Efeito	Estimativa (EP)	Valor p	N. Obs
ACC	Intercepto	3,59 (3,96)	0,3640	535
	Grupo	0,27 (1,44)	0,8496	
	Início	0,81 (0,11)	< 0,0001	
	Tempo	-0,34 (0,04)	< 0,0001	
	Grupo*Tempo	-0,11 (0,06)	0,0577	
IMP	Intercepto	6,62 (3,95)	0,0945	615
	Grupo	-0,76 (1,44)	0,5973	
	Início	0,74 (0,11)	< 0,0001	
	Tempo	-0,36 (0,04)	< 0,0001	
	Grupo*Tempo	-0,08 (0,05)	0,1406	
IMT	Intercepto	4,81 (3,94)	0,2230	615
	Grupo	-0,57 (1,38)	0,6800	
	Início	0,78 (0,11)	< 0,0001	
	Tempo	-0,29 (0,04)	< 0,0001	
	Grupo*Tempo	-0,11 (0,05)	0,0458	
IOU	Intercepto	4,51 (3,95)	0,2546	615
	Grupo	-0,45 (1,38)	0,7453	
	Início	0,78 (0,11)	< 0,0001	
	Tempo	-0,28 (0,04)	< 0,0001	
	Grupo*Tempo	-0,12 (0,06)	0,0367	
IR	Intercepto	4,12 (3,89)	0,2903	615
	Grupo	-0,24 (1,34)	0,8588	
	Início	0,80 (0,11)	< 0,0001	
	Tempo	-0,31 (0,04)	< 0,0001	
	Grupo*Tempo	-0,13 (0,05)	0,0122	

Tabela 6.3.2: Imputação Múltipla

M	**Efeito**	**Estimativa (EP)**	**Valor p**
2	Intercepto	6,23 (7,96)	0,4539
	Grupo	1,53 (3,71)	0,6809
	Início	0,75 (0,22)	0,0205
	Tempo	-0,34 (0,13)	0,0111
	Grupo*Tempo	-0,20 (0,19)	0,3198
3	Intercepto	7,11 (7,49)	0,3460
	Grupo	0,40 (4,24)	0,9247
	Início	0,71 (0,18)	0,0001
	Tempo	-0,35 (0,17)	0,0415
	Grupo*Tempo	-0,12 (0,23)	0,6063
5	Intercepto	8,42 (7,03)	0,2315
	Grupo	-0,33 (4,26)	0,9379
	Início	0,69 (0,18)	0,0001
	Tempo	-0,35 (0,17)	0,0424
	Grupo*Tempo	-0,07 (0,21)	0,7426
10	Intercepto	7,59 (7,15)	0,2893
	Grupo	0,26 (4,16)	0,9504
	Início	0,71 (0,18)	0,0001
	Tempo	-0,35 (0,16)	0,0255
	Grupo*Tempo	-0,11 (0,21)	0,6102

Referências

[1] Barroso, L.P. (1995). *Imputação de Dados em Painéis para Populações Finitas.* Tese de Doutorado (Instituto de Matemática e Estatística) – USP, São Paulo.

[2] Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Eletronic Computer. *Royal Statistical Society* – B, **22**, 302-306.

[3] Davis, C.S. (2001). *Statistical Methods for the of Repeated Measurementes,* Springer-Verlag, New-York.

[4] Diggle, P.J., Liang, K.-Y and Zeger, S.L. (1994). *Analysis of Longitudinal Data,* OxfordUniversity Press, New York.

[5] Laird, N.M. and Ware, J.H. (1982). Random-Effects Models for longitudinal Data. *Biometrics*, **38**, 963-974.

[6] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data,* J. Wiley & Sons, New york.

[7] Lopes Rocha, Fábio. (1995). *Eficácia e Tolerabilidade da Tianeptina no Tratamento de Epsódios Depressivos Maiores.* Dissertação de Mestrado (Faculdade de Medicina) – UFMG, Belo Horizonte.

[8] Rubin, D.B. (1978). Multiple Impatations in Sample Surveys – a Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

[9] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Date.* Chapman & Hall, London.

[10] Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed models for Longitudinal Data,* Springer-Verlag, New York.