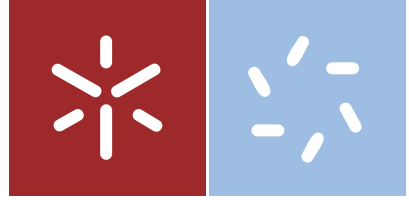
Universidade do Minho
Escola de Ciências

José Arteiro Teixeira Queiroz Neto

Dados Faltantes em Modelos para Dados Longitudinais

julho de 2021

Dados Faltantes em Modelos para Dados Longitudinais
José Arteiro Teixeira Queiroz Neto
UMinho | 2021



Universidade do Minho
Escola de Ciências

José Arteiro Teixeira Queiroz Neto

Dados Faltantes em Modelos para Dados Longitudinais

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Drª Inês Pereira Silva Cunha de Sousa

julho de 2021

Direitos de autor e condições de utilização do trabalho por terceiros

Agradecimentos

Em primeiro lugar, gostaria de agradecer a Deus, pois foi graças a Ele que até aqui cheguei.

Em segundo lugar, quero agradecer a professora Dra. Inês Sousa, minha orientadora, que teve uma contribuição enorme em todo esse trajeto. Em muitos momentos pensei que não seria capaz, porém nunca deixou-me desanimar e sempre mostrou um caminho a seguir e o mais importante, transmitiu-me um conhecimento fulcral para que conseguisse finalizar esse trabalho.

Em terceiro lugar, quero agradecer a minha família que mesmo longe sempre apoiou-me e dava-me força de vontade para concluir, mesmo quando a saudade apertava eles estavam longe fisicamente, entretanto estiveram sempre em meus pensamentos.

Em quarto lugar, quero agradecer aos professores da Universidade do Minho pela atenção e conhecimento transmitido ao longo do mestrado, assim como os meus colegas de mestrado onde muitas vezes passávamos horas nas bibliotecas a estudar e resolver exercícios.

Declaração de integridade

Eu, José Arteiro Teixeira Queiroz Neto, nº PG38829, aluno do Mestrado em Estatística na Escola de Ciências da Universidade do Minho, declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Título: Dados faltantes em modelos para dados longitudinais

Em todas as áreas de conhecimento, quando vamos analisar os dados em busca de informações, nos deparamos com dados faltantes Esse tipo de dados diversas vezes nos fazem tomar decisões erradas devido terem apresentado resultados equivocados em suas análises, e devido isso devemos tomar muito cuido e sermos minuciosos quanto tratamos esse tipo de dados. Para (Little e Rubin, 2019)[13], dados faltantes são valores não observados que seriam significativos para análise se observados; em outras palavras, um valor ausente oculta um valor significativo.

Dados longitudinais são gerados quando os indivíduos são medidos repetidamente ao longo do tempo, e os modelos longitudinais descrevem o processo estocástico dos dados observados. Esses modelos permitem-nos distinguir, sobretudo, a variabilidade dos dados dentro e entre indivíduos (Diggle,2002) [9].

De acordo com (Little e Rubin, 2019)[13], dados faltantes são valores não observados que seriam significativos para análise se observados; em outras palavras, um valor ausente oculta um valor significativo.

O problema de dados faltantes tem sido alvo de muitas pesquisas recentemente devido o fato de que quando se tem uma população com observações faltantes, ao se fazer a inferência estatística, se não usarmos os procedimentos adequados, os resultados podem não ser fiáveis e assim levarão a conclusões equivocadas. O que queremos é muito simples, se F é uma população a ser estimada e $\hat{F}$ é uma estimativa de F, onde a amostra é composta por observações faltantes, então o procedimento que será usado para calcular esses dados deve ser levado em conta o método estatístico que será usado para calcular $\hat{F}$. Logo, se esse procedimento estatístico funcionar corretamente teremos então $\hat{F}$ com valor próximo de F para a média. Estamos a procura de que o viés seja mínimo, assim como o desvio padrão e variância de $\hat{F}$, onde assim poderemos fazer inferência com mais precisão para uma tomada de decisão.

Palavras-chave: Dados faltantes, dados longitudinais, tomada de decisão, valor significativo.

Abstract

Title: Missing data in models for longitudinal data

In all areas of knowledge, when we analyze the data in search of information, we come across missing data. This type of data often makes us make wrong decisions because they have presented wrong results in their analyses, and because of that we must be very careful and be thorough when we handle this type of data. For (Little and Rubin, 2019)[13], missing data are unobserved values that would be significant for analysis if observed; in other words, a missing value hides a significant value.

Longitudinal data are generated when individuals are measured repeatedly over time, and longitudinal models describe the stochastic process of the observed data. These models allow us to distinguish, above all, the variability of data within and between individuals (Diggle, 2002) [9].

According to (Little and Rubin, 2019)[13], missing data are unobserved values that would be significant for analysis if observed; in other words, a missing value hides a significant value.

The problem of missing data has been the subject of much research recently due to the fact that when you have a population with missing observations, when making the statistical inference, if we don't use the proper procedures, the results may not be reliable and thus lead to wrong conclusions. What we want is very simple, if F is a population to be estimated and $\hat{F}$ is an estimate of F, where the sample is made up of missing observations, then the procedure that will be used to calculate these data must take into account the statistical method that will be used to calculate $\hat{F}$. Therefore, if this statistical procedure works correctly, then we will have $\hat{F}$ with a value close to F for the mean. We are looking for the bias to be minimal, as well as the standard deviation and variance of $\hat{F}$, so that we can make inferences more accurately for decision making.

Keywords: Missing data, longitudinal data, decision making, significant value.

Conteúdo

1 Introdução **1**

2 Dados Faltantes **3**

2.1 Padrões de Dados Faltantes 4

2.2 Mecanismos de Dados Faltantes 6

2.3 Abordagens para Tratamento de Dados Faltantes 7

2.4 Análise dos Casos Completos 8

2.5 Análise dos Casos Disponíveis 8

2.6 Métodos Fundamentados em Imputação 9

2.6.1 Imputação Simples 9

2.6.2 Imputação Múltipla 11

2.7 Exemplos de Base de Dados 11

2.7.1 Muscatine Coronary Risk Factor Study (MCRFS) 11

2.7.2 International Breast Cancer Study Group (IBCSG) 12

3 Modelos de Dados Longitudinais **14**

4 Aplicação a uma Base de Dados Real **18**

4.1 Descrição da Base de Dados 18

4.1.1 Descrição das Variáveis 19

4.2 Analise Exploratória de Dados 19

4.2.1 Teste para Normalidade dos Dados (nível de depressão por tipo de tratamento) 21

4.2.2 Teste Para Igualdade de Medianas (nível de depressão por tipo de tratamento) 22

4.2.3 Teste para Normalidade dos Dados (nível de depressão por uso de medicamento antidepressivo) 24

4.2.4 Teste Para Igualdade de Medianas (nível de depressão por uso de medicamento antidepressivo) 24
4.2.5 Teste para Normalidade dos Dados (nível de depressão por tempo de duração do episódio) 26
4.2.6 Teste Para Igualdade de Medianas (nível de depressão por duração de episódio) 28
4.2.7 Teste para Normalidade dos Dados (nível de depressão por tempo de tratamento) 28
4.2.8 Teste Para Igualdade de Medianas (nível de depressão por tempo de tratamento) 30

5 Conclusão e trabalho futuro **32**

Bibliografia **33**

Lista de Figuras

2.1 Exemplos de padrões de dados faltantes; Linhas correspondem a unidades e colunas a variáveis. (a)Padrão univariado, (b)Multivariado com dois padrões, (c)Monótono, (d) Padrão geral, (e)Correspondência de arquivo, (f) Padrão com variáveis latentes. (Little and Rubin,2019)[13] 5

3.1 Exemplos de estruturas de correlação para modelar $\vee_i$. (Pedro, 2019) [1] . . 16

4.1 Nível de Depressão por Tipo de Tratamento 21
4.2 Gráfico Q-Q plot para Normalidade (bdi por treatment) 22
4.3 Nível de Depressão por Uso de Medicamento Antidepressivo 23
4.4 Gráfico Q-Q plot para Normalidade (bdi por drug) 25
4.5 Nível de Depressão por Duração do Episódio 26
4.6 Gráfico Q-Q plot para Normalidade (bdi por length) 27
4.7 Nível de Depressão por Tempo de Tratamento 29
4.8 Gráfico Q-Q plot para Normalidade (bdi por time) 30

Lista de Tabelas

2.1 Freqüências dos padrões de dados faltantes do IBCSG. (Ibrahim e Molenberghs, 2001)[11] 13

4.1 Medidas de Localização das Variáveis Quantitativas.(REFERÊNCIA) 19
4.2 Medidas de Dispersão das Variáveis Quantitativas.(REFERÊNCIA) 20
4.3 Média do Nível de Depressão por Tipo de Tratamento. 23
4.4 Média do Nível de Depressão por Uso de Medicamento Antidepressivo. . . . 25
4.5 Média do Nível de Depressão por Tempo de Duração de Episódio de Depressão. 28
4.6 Média do Nível de Depressão por Tempo de Tratamento. 30

Capítulo 1

Introdução

Em todas as áreas de conhecimento, quando vamos analisar os dados em busca de informações, nos deparamos com dados faltantes Esse tipo de dados diversas vezes nos fazem tomar decisões erradas devido terem apresentado resultados equivocados em suas análises, e devido isso devemos tomar muito cuido e sermos minuciosos quanto tratamos esse tipo de dados. Para (Little e Rubin, 2019)[13], dados faltantes são valores não observados que seriam significativos para análise se observados; em outras palavras, um valor ausente oculta um valor significativo.

Dados longitudinais são gerados quando os indivíduos são medidos repetidamente ao longo do tempo, e os modelos longitudinais descrevem o processo estocástico dos dados observados. Esses modelos permitem-nos distinguir, sobretudo, a variabilidade dos dados dentro e entre indivíduos (Diggle,2002) [9]. Como lidamos com muitas medidas, os dados longitudinais apresentam frequentemente inúmeros dados ausentes e para tratarmos desses dados precisamos identificar os padrões dos dados ausente para que assim possamos utilizar o mecanismo de dados ausente adequado para cada situação.

Este documento é composto por 5 capítulos, onde no Capítulo 1 é feita a introdução do trabalho explicando a estrututa do mesmo. No Capítulo 2 é apresentado um conceito sobre dados faltates e alguns exemplos para melhor entendimento, assim como uma breve explanação sobre os padrões e mecanismos de dados faltantes, todos com exemplos bem leves para boa compreensão. Exploramos também algumas técnicas de abordagens para tratamentos de dados e no fim do capítulo temos dois estudos sobre base de dados contendo dados faltantes para fechar a ideia principal do documento. Já no Capítulo 3 temos uma explanação sobre dados longitudinais mostrando alguns modelos existentes, assim como algumas estruturas de correlação de dados, tudo sendo apenas conceitos fundamentais para melhor compreensão do trabalho. Já no Capítulo 4 temos uma base de dados com dados referente a Depressão apresentando dados faltantes, onde descrevemos todas as variáveis assim como também fize-

mos uma análise exploratória de dados e alguns teste estatísticos. Todas as análises foram efetuadas no $software R$.

Capítulo 2

Dados Faltantes

Por vezes ao analisarmos um conjunto de dados percebemos que algumas observações não estão registadas, principalmente quando estamos a lidar com um conjunto de dados longitudinais. A falta de registo dessas observações pode ocorrer devido a inúmeros motivos, tais como: em um acompanhamento pós cirúrgico, um enfermeiro/médico interpretou que essa informação não seria importante e resolveu não registar; em uma sondagem de opinião, alguns entrevistados não deixaram claro a sua intenção de voto; em um experimento clínico, em alguns dias não haviam equipamentos suficientes para poder recolher informações de todos os participantes.

No primeiro e no terceiro exemplo podemos tratar a falta de informação como dados faltantes, devido o fato de existirem valores adjacentes reais que deveriam ter sido observados se o comprometimento do profissional hospitalar fosse mais eficaz ou se o experimento clínico tivesse sido melhor planeado. No segundo exemplo é incomum tratarmos como dado faltante em razão da possibilidade do entrevistado estar mesmo em dúvida em quem votar, sendo assim podemos classificar como um estrato "não sei", "sem preferência" ou até mesmo "não vou votar" da população para essa variável.

De acordo com (Little e Rubin, 2019)[13], dados faltantes são valores não observados que seriam significativos para análise se observados; em outras palavras, um valor ausente oculta um valor significativo. Aplicando esta definição nos exemplos de números um e três faz total sentido fazer uma análise estatística que consiga preencher esses valores não observados para que os resultados a partir destas análises sejam mais precisos, mais consistentes e assim possamos tomar decisões melhor fundamentadas. Em detrimento da aplicação da definição temos o exemplo de número dois, onde não faz sentido tentar prever esses valores não observados visto que a criação de estratos da população vem a ser mais apropriado.

O problema de dados faltantes tem sido alvo de muitas pesquisas recentemente devido o fato de que quando se tem uma população com observações faltantes, ao se fazer a inferência

estatística, se não usarmos os procedimentos adequados, os resultados podem não ser fiáveis e assim levarão a conclusões equivocadas. O que queremos é muito simples, se F é uma população a ser estimada e $\hat{F}$ é uma estimativa de F, onde a amostra é composta por observações faltantes, então o procedimento que será usado para calcular esses dados deve ser levado em conta o método estatístico que será usado para calcular $\hat{F}$. Logo, se esse procedimento estatístico funcionar corretamente teremos então $\hat{F}$ com valor próximo de F para a média. Estamos a procura de que o viés seja mínimo, assim como o desvio padrão e variância de $\hat{F}$, onde assim poderemos fazer inferência com mais precisão para uma tomada de decisão.

2.1 Padrões de Dados Faltantes

Existem vários métodos para analisar dados faltantes. O método escolhido deve levar em consideração tanto os padrões de dados faltantes, onde identificamos como os dados estão distribuídos na matriz de dados, como os mecanismos de dados faltantes, que diz respeito à relação entre os dados faltantes e os dados observados.

Para entendermos melhor os padrões de dados faltantes vamos denotar $Y = (y_{ij})$ para um conjunto de dados retangular completo (nxk), ou seja, sem valores faltantes, com a i-ésima linha $y_i = (y_{i1}, ..., y_{ik})$ em que y_{ij} corresponde ao valor da variável Y_j para a unidade i. Agora com dados faltantes iremos definir a matriz de indicador de falta $Z = z_{ij}$, onde $m_{ij} = 1$ para y_{ij} ausente e $m_{ij} = 0$ quando y_{ij} for observado. A matriz Z está definindo um padrão para os dados faltantes. Abaixo temos a descrição de alguns tipos de padrões de dados faltantes.

- **Padrão Univariado:** quando temos um conjunto de dados retangular, onde estamos a analisar uma variável dependente e temos algumas variáveis independentes, porém apenas uma variável independente possuí observações faltantes, em estudos experimentais é muito comum esse tipo de padrão.

- **Multivariado com dois padrões:** quando existem uma única variável com dados faltantes e ela é substituída por um conjunto de variáveis, onde todas são observadas ou ausentes no mesmo conjunto de unidade. Geralmente este tipo de padrão ocorre em pesquisas domiciliares, sondagem de opinião, realizadas através de questionários, onde o respondente tem a opção de recusa de resposta.

- **Padrão monótono:** acontece frequentemente em estudos com dados longitudinais e uma diferenciação altamente importante para se fazer é quando os dados faltantes ocorrerm de forma intermitentemente ou como dropouts (padrão monótono). Vamos ter

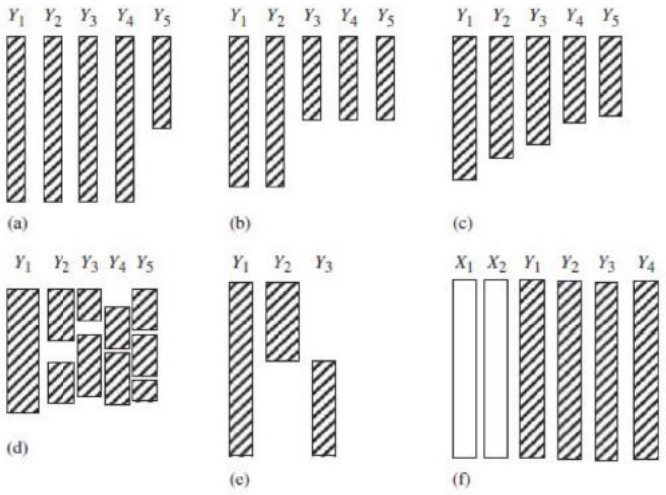


Figura 2.1: Exemplos de padrões de dados faltantes; Linhas correspondem a unidades e colunas a variáveis. (a)Padrão univariado, (b)Multivariado com dois padrões, (c)Monótono, (d) Padrão geral, (e)Correspondência de arquivo, (f) Padrão com variáveis latentes. (Little and Rubin,2019)[13]

em conta uma sequência de medidas $Y_1, Y_2, ..., Y_n$ na i-ésima unidade amostral. Denotamos dados faltantes como dropouts quando ao se observar um Y_j faltante também será observado um Y_k faltante para todo $k \geq j$; Caso contrário os dados faltantes serão chamados de intermitentes (Diggle, Liang e Zegger, 2002)[9].

- **Padrão geral:** existem dados faltantes, porém tudo indica que foi ao acaso. O fato de não ter respondido uma pergunta em determinado questionário, sendo que era uma pergunta básica. Não ir mais a um acompanhamento devido ter mudado de cidade. É o padrão aleatório.

- **Correspondência de arquivo:** a ocorrência de uma grande quantidade de dados faltantes pode ocasionar a não observação de algumas combinações de variáveis. Por exemplo, temos três variáveis onde, Z_1 é um conjunto de variáveis totalmente observadas e comuns as fontes de dados. Z_2 é um conjunto de variáveis observadas e composta a partir da fonte de dados de número 1 e não da fonte de dados número 2. Z_3 é um conjunto de variáveis observadas e composta a partir da fonte de dados de número 2 e não da fonte de dados número 1. Logicamente, não temos informação sobre as associações parciais nem de $Z_2|Z_1$ e nem de $Z_3|Z_1$ nesse tipo de padrão de dados, porém normalmente nesse padrão de dados é assumido zero para essas associações parciais. Na Figura**??**(e) temos a ilustração a situação descrita acima.

- **Padrão com variáveis latentes:** quando temos variáveis explicativas latentes totalmente ausentes e as variáveis de pendentes estão completamente observadas. A figura 1.1(f) nos demonstra como seria essa situação. É um padrão que requer muitas suposições.

2.2 Mecanismos de Dados Faltantes

Esses mecanismos de dados faltantes são de grande importância devido as propriedades dos métodos de dados faltantes terem grande dependência desses mecanismos. Para explicar esses mecanismos vamos considerar um conjunto de dados retangulares onde M é uma variável indicadora que nos fornece o que está observado, $(M = 1)$, e o que está ausente, $(M = 0)$. Vamos nos referir a M como a falta e, como essa falta pode estar relacionada aos dados, vamos classificar as distribuições para M de acordo com a natureza desse relacionamento.(Rubin,1976)[14] e (Little and Rubin,2019)[13] desenvolveram uma tipologia para essas distribuições, onde são citadas por inúmeros autores. Vamos atribuir aos dados completos a notação Y_{com} e iremos particioná-los como $Y_{com} = (Y_{obs}, Y_{mis})$, onde Y_{obs} são os dados observados e Y_{mis} são os dados faltantes. Vamos denotar ϕ para os parâmetros desconhecidos.

Dados faltantes serão chamados de MAR (missing at random) se a distribuição da falta não depender de Y_{mis},

$$P(M = 0|Y_{com}, \phi) = P(M = 0|Y_{obs}, \phi) \tag{2.1}$$

Logo percebe-se que MAR permite que a probabilidade de falta dependa dos dados observados, porém não dos dados faltantes. Uma caso especial do MAR é quando a distribuição de falta também não depende de Y_{obs}, denominado de MCAR (missing completely at random),

$$P(M = 0|Y_{com}, \phi) = P(M = 0|\phi)$$

O mecanismo é chamado de MNAR (missing not at random) quando a equação (2.1) é violada e a distribuição de falta depende de Y_{mis},

$$P(M = 0|Y_{com}, \phi) = P(M = 0|Y_{mis}, \phi)$$

Temos um mecanismo de dados faltantes do tipo MAR quando a probabilidade de falta não depende dos dados faltantes e sim dos observados. Por exemplo, em um ensaio clínico em que o paciente tem de ter um determinado nível de percentagem de uma substancia no sangue, caso contrário será removido do ensaio. Temos um mecanismo de dados ausentes

MCAR quando a probabilidade de falta não depende nem dos dados observados nem dos dados ausentes. Por exemplo, quando o paciente morreu por questões alheias ao tratamento da doença ou simplesmente porque o paciente mudou de localização e não tem mais como o contactar. Temos um mecanismo de dados ausentes MNAR quando a probabilidade de falta depende dos dados ausentes. Por exemplo, quando o paciente deixa de ir em algumas sessões devido sentir alguns efeitos colaterais por causa do tratamento, ou quando o paciente morre devido a doença na qual estava em tratamento.

Os dois primeiros mecanismos são considerados ignoráveis, ou seja, não é necessário especificar um modelo para os dados faltantes. Já o último mecanismo é não ignorável por não ser aleatório, então é necessário especificar um modelo para os dados faltantes.(Rubin and Little,2019)[13] descrevem um exemplo muito simples para explicar os três mecanismos. Vamos considerar duas variáveis, $X = idade$ que está totalmente observada e $Y = renda$ que possui dados faltantes. Se a probabilidade da renda ser observada é a mesma para todos os indivíduos, independente de sua idade ou renda, então os dados são MCAR. Se a probabilidade da renda ser observada varia de acordo com a idade do indivíduo, mas não varia de acordo com sua renda, então os dados são MAR. E se a probabilidade da renda ser observada varia de acordo com a sua própria renda, então os dados são MNAR, pois as observações ausentes estão a depender dos dados ausentes da variável de interesse e, portanto, temos que especificar um modelo para essas observações ausentes, visto que sem um modelo a análise estatística fica enviesada.

2.3 Abordagens para Tratamento de Dados Faltantes

Muitos pesquisadores, ao se depararem com muitas observações faltantes, acabam por tentar "facilitar"o seu trabalho e apenas excluem essas observações. Se o mecanismo de falta tiver o padrão MCAR ou até mesmo MAR, essa técnica pode ou não prejudicar as análises, porém se o mecanismo for MNAR as análises estarão completamente distorcidas e levarão a conclusões equivocadas. Entretanto, devemos analisar com bastante cuidado tanto o padrão como o mecanismo de dados faltantes para assim podermos aplicar a melhor abordagem possível para podermos fazer análises robustas e obter resultados precisos.

Existem vários métodos para tratamento de dados faltantes, onde o mais apropriado depende fortemente do tipo de mecanismo de dados faltantes, porém sempre que existir a possibilidade de recuperação dos dados, essa será a melhor forma, entretanto na prática isso pouco ocorre devido a natureza do estudo.

Uma importante distinção para se fazer é quando estamos a tratar dados faltantes em variável resposta ou variáveis explicativas. Se estivermos lidando com um conjunto de dados,

onde temos apenas variável resposta com dados faltantes e esses dados possuem o mecanismo de dados faltantes do tipo MAR, então para termos estimativas imparciais podemos fazer uma análise de casos completos, porém se o mecanismo de dados faltantes for do tipo MNAR, então as estimativas serão completamente enviesadas. Se estivermos lidando com um conjunto de dados onde somente as variáveis explicativas apresentam dados faltantes e esses dados apresentem o mecanismo de dados faltantes do tipo MAR ou MNAR, então uma análise baseada em métodos de máxima verossimilhança acaba por apresentar estimativas mais fiáveis do que se usarmos uma análise de casos completos. Em estudos longitudinais é muito comum termos um conjunto de dados onde tanto a variável resposta como as variáveis explicativas apresentem dados faltantes, então uma análise com o método de máxima verossimilhança, assumindo que o mecanismo de dados faltantes é do tipo MAR ou MNAR, é bem mais apropriado do que fazer uma análise dos caso completos, desde que o mecanismo de dados faltantes esteja realmente bem definido. É muito importante salientar que o método de imputação é usado apenas para quando os dados faltantes pertencem as variáveis explicativas e não a variável resposta. Iremos descrever alguns métodos de imputação, tanto simples como múltipla, para fins de conhecimento.

2.4 Análise dos Casos Completos

Trata-se da exclusão das unidades que possuem observações parciais ou faltantes e assim a análise é feita apenas com as unidades que contém todas as observações. Essa estratégia não é apropriada devido ao fazermos uma analise estarmos em busca de fazer inferência sobre toda a população e não apenas na população que contém as variáveis com informações. Essa prática diminui a precisão dos estimadores, pois ao reduzir o tamanho da amostra você está aumentando a variabilidade dos estimadores. Como em estudos longitudinais os indivíduos são medidos ao longo do tempo e os pacientes que apresentam dropout normalmente tem perfis distintos dos paciente que não apresentam, logo para um pesquisador não existe vantagem em obter uma simplicidade através da exclusão dos dados ausentes, sendo que devido isso, ele obterá análises que podem não ser fiáveis e assim causar resultados viciados. Essa técnica é muito usada quando estamos a tratar dados que tenham o mecanismo de dados faltantes do tipo MCAR ou MAR e quando temos poucas observações faltantes.

2.5 Análise dos Casos Disponíveis

Ao usar esta técnica não iremos excluir as unidades que possuem observações parciais ou faltantes, iremos analisar todos as informações disponíveis e trabalhar com elas. Essa

estratégia acaba por ser mais eficaz do que a análise dos casos completos, pois ao utilizarmos todos os dados disponíveis ganhamos um pouco mais de precisão nos estimadores e temos menos variância, isso comparando os dois métodos. Entretanto está técnica precisa que o mecanismo de dados faltantes seja do tipo MCAR ou MAR e também não podem existir muitos valores faltantes.

2.6 Métodos Fundamentados em Imputação

Pensando em uma forma de diminuir os vários problemas em tratar de dados faltantes, (Little e Rubin, 2019)[13] criou um método que não exclui as unidades que possuem observações parciais ou faltantes e também não utiliza apenas os dados disponíveis, ele criou o método de imputação, onde cada valor faltante é substituído por um valor imputado e assim temos um conjunto de dados completos pronto para ser analisado. Dentro do contexto de dados longitudinais, as medidas observadas devem ser usadas para imputar as medidas das observações que não estão observadas com o objetivo de integrar um conjunto de dados completo pronto para análises futuras.

2.6.1 Imputação Simples

Quando para cada observação faltante se dá apenas um valor imputado estamos a lidar com o método de imputação simples. Existem vários métodos de imputação simples, os quais serão descritos brevemente a seguir, porém devemos ser cautelosos ao utilizarmos esta estratégia visto apresenta muitas vantagens devido suas estimativas serem mais precisas do que os métodos tradicionais, assim como o fato de gerar um conjunto de dados completos diminuindo os erros padrão, entretanto as estimativas podem ser distorcidas da realidade e os erros padrão podem ser subestimados.

- **Imputação Através da Última Observação** Essa técnica é muito utilizada e muito simples, para cada dado faltante do paciente será atribuído o valor referente a última observação que foi medida desse mesmo paciente.

- **Imputação Através da Média** Baseia-se na substituição dos dados faltantes, da variável que apresenta dados faltantes, pela valor média desta variável. Nos estudos longitudinais temos a imputação por média dos tempos e dos pacientes, onde na imputação por média dos tempos é calculado a média amostral das observações presentes em todos os tempos e na imputação por média dos pacientes é calculado a média amostral das observações presentes em um único tempo.

- **Imputação Através de Regressão** Um método bem mais promissor é a substituição dos dados faltantes por valores preditos por meio de um modelo de regressão com base nos dados observados. Um modelo para uma amostra multivariada foi desenvolvido por (Buck, 1960)[6]. Todo o método de imputação através da regressão aplicando-se um modelo linear misto pode ser assim descrito:

  1. Modelar os dados através do modelo linear misto (Laird e Ware, 1982) [12], assim:

     $y_i = X_i\beta + Z_i b_i + \epsilon_i, i = 1, ..., n; j = 1, ..., a_i;$
     em que y_i para $j = 1, ..., a_i$ é observado e y_i para $j = a_i + 1, ..., n_i$ é faltante. No caso de dados balanceados, $n_i = m$.

  2. Predizer b_i e Estimar β com os dados que estão observados;

  3. Estimar os dados faltantes através dos coeficientes obtidos em 2 e das matrizes observadas X_i e Z_i, isto é:

     $\hat{y_i} = X_i\hat{\beta} + Z_i\hat{b_i}$, para $j = a_i + 1, ..., n_i$.

  4. Obter o conjunto de dados completos: observados + imputados.

O fato deste método levar em conta o efeito aleatório predito de cada paciente, assim como as estimativas dos efeitos fixos o torna bem eficiente.

Neste trabalho, temos como principal característica nos métodos de imputação simples para dados longitudinais a simplicidade dos modelos. O fato de aumentar a potência dos testes estatísticos trazendo o balanceamento dos dados e de geralmente reproduzir fortes estimativas pontuais para o tipo de mecanismo de dados faltantes escolhido nos da uma enorme vantagem, contudo devemos sempre estar atentos a subestimação da variância quando se utiliza destes métodos. Dependendo do tipo de mecanismo e da quantidade de dados faltantes é que decidimos usar ou não este método, geralmente com MCAR e MAR não é tão usual utilizarmos a imputação simples, devido a facilidade de tratamento destes tipos de mecanismos, porém quando se trata de MNAR devemos utilizar, pois esse tipo de mecanismo é bem mais complexo de se trabalhar e não pode ser ignorável, logo precisamos de um método muito eficaz.

2.6.2 Imputação Múltipla

Pensando em melhorar ainda mais a precisão das estimativas dos métodos de imputação simples [13] criou o método de imputação múltipla, onde esta técnica que consiste em substituir cada dado faltante por dois ou mais dados faltantes imputados, está em destaque na literatura de dados faltantes devido a sua versatilidade que pode ser aplicada em vários contextos. Ao imputarmos inúmeros valores para cada dado faltante a margem de erro é reconhecida claramente. Abaixo está descrito o processo de imputação múltipla:

1. Primeiro temos o processo de imputação, onde para cada dado faltante é gerado M valores ($M \geq 2$);

2. Depois vamos para a análise, onde os M valores gerados serão distribuídos aleatoriamente para que o primeiro valor imputado para cada dado faltante construa o primeiro conjunto de dados completos, o segundo valor imputado para cada dado faltante constrói o segundo conjunto de dados completos e esse critério também ocorre para o restante dos valores imputados. Ao completar os conjuntos de dados, podemos utilizar quaisquer métodos padrão para análise de dados completos.

3. Chegamos assim a parte de combinação, onde os diversos resultados das inúmeras análises serão acompanhamentos em conjunto para que o grau de incerteza na imputação seja considerado.

Todo este processo leva em consideração o tipo de mecanismo de dado faltante. Sem sombra de dúvidas que o passo da imputação é o mais crucial. Uma peculiaridade deste técnica incrível é que o modelo utilizado no passo da imputação não precisa ser o mesmo para a análise dados, visto que as vezes o modelo utilizado na imputação não é conveniente para a análise dos dados.

2.7 Exemplos de Base de Dados

Nesta secção iremos apresentar alguns exemplos de estudos com base de dados contendo dados faltantes para melhor entendimento do leitor.

2.7.1 Muscatine Coronary Risk Factor Study (MCRFS)

Em Muscatine, uma cidade localizada às margens do rio Mississippi nos Estados Unidos, foi realizado o Muscatine Coronary Risk Factor Study (MCRFS), um estudo estudo longitudinal de fatores de risco coronariano em crianças de idade escolar, este estudo foi apresentado

por Woolson and Clarke (1984) e apresenta cinco coortes de crianças, inicialmente com idades entre $5-7, 7-9, 9-11, 11-13, 13-15$ nas quais foram registrados as medidas de altura e peso nos anos de 1977, 1979 e 1981. Obtivemos 4.856 dados coletados, entre crianças do sexo masculino e feminino. Comparando seu peso com as normas específicas de seu género para a idade, as crianças foram classificadas como obesas ou não obesas.

A base de dados contém 14568 observações e 6 variáveis. Entre as variáveis temos o *id* que é um factor com 4856 levels, temos a *gender* que é um factor com dois levels, sendo M para masculino e F para feminino. Temos a variável *age0* que nos apresenta um vetor com a idade inicial do participante. Temos também a variável *age* como sendo um vetor da idade, a variável *year* como sendo um vetor numero que representa o ano da medição e temos a variável *obesity*, na qual é um fator com dois levels, obseso e não obeso.

Como trata-se de um estudo longitudinal, temos muitas medidas repetidas ao longo do tempo e apresentamos alguns dados faltantes, logo provoca-se uma grande dificuldade estatística que é estimar a taxa de obesidade em função do sexo e da idade. (Baker,1995) [3] descreveu duas principais causas da existência de dados faltantes neste estudo, uma delas é que não foi recebido nenhum formulário de consentimento assinado pelos pais das crianças e o outro é devido a criança não estar na escola no dia da pesquisa. O fato dos pais não assinarem o formulário de consentimento por medo de criar um certo trauma em seu filho pelo fato dele ser considerado obeso ou o fato da criança não ir à escola no dia pesquisa justamente devido o seu peso estar acima da média, o mecanismo de dados faltantes seria provavelmente MAR ou até mesmo MNAR. No caso da criança faltar a escola devido o seu peso estar acima da média, esse exemplo nunca deve ser efetuado com uma analise que ignora o mecanismo de dados faltantes, pois certamente será gerado resultados tendenciosos.

2.7.2 International Breast Cancer Study Group (IBCSG)

Vamos nos basear em um conjunto de dados que se refere a um estudo sobre a qualidade de vida de pacientes que apresentam câncer realizado pelo International Breast Cancer Study Group através de um ensaio clínico onde foi comparado quatro tipos de tratamento com quimioterapia. A principal resposta obtida refere-se ao tempo até a recaída e morte do paciente, entretanto os pacientes foram avaliados através de questionários sobre a qualidade de vida a cada três meses, sendo que alguns não respondiam o questionário, contudo existia uma visita de acompanhamento onde os pacientes tinham que completar uma avaliação e assim não ocorria nenhum dropout, mas exitiam dados faltantes intermitentes. O estudo ocorreu inicialmente com dados dos primeiros 18 meses de tratamento, onde cada paciente foi medido no máximo sete vezes. A avaliação gerou um estudo longitudinal sobre a qualidade de vida, especificamente sobre o humor do paciente, com uma escala de 0 a 100, onde 0 significa melhor

e 100 significa pior. As perguntas eram divididas em uma variável explicativa dicotómica de idioma, com os itens italiano ou sueco; uma covariável contínua para a idade e três variáveis explicativas dicotómicas para o tipo de tratamento da quimioterapia (4 tratamentos).

Existem na base de dados um total de 397 observações, onde foi constatado que o humor não está presente em 71% dos casos, representando 116 casos completos ou 29% das observações. A representatividade dos dados faltantes é bem pequena na baseline, apenas 2% e apresenta uma variação nas demais medições onde não ultrapassa os 31%, porém também não reduz dos 24%. É importante esclarecer que todos os pacientes estavam vivos ao fim do estudo, contudo sabemos que o grau do humor do paciente pode ter influenciado em preencher ou não o formulário de avaliação. Dito isto, o mecanismo de dados ausentes seria do tipo MNAR e qualquer análise que utilizasse outro tipo de mecanismo mostraria estimativas completamente enviesadas. A Tabela**XXX** nos fornece uma frequência dos padrões de dados faltantes.

IBCSG Trial VI patterns of missingness

Number of missing components of y_j	Frequency	Percentage
0	116	29.2
1	116	29.2
2	62	15.6
3	35	8.8
4	30	7.6
5	38	9.6

Tabela 2.1: Freqüências dos padrões de dados faltantes do IBCSG. (Ibrahim e Molenberghs, 2001)[11]

Capítulo 3

Modelos de Dados Longitudinais

Dados longitudinais são gerados quando os indivíduos são medidos repetidamente ao longo do tempo, e os modelos longitudinais descrevem o processo estocástico dos dados observados. Esses modelos permitem-nos distinguir, sobretudo, a variabilidade dos dados dentro e entre indivíduos (Diggle,2002) [9].

Um estudo longitudinal balanceado é aquele que detém todas as unidades de análise observadas e medidas em todos os momentos do tempo, já em um estudo não balanceado as unidade de análise não são todas observadas e os indivíduos são medidos em tempos distintos. Quando queremos fazer um estudo exploratório da base de dados, utilizamos a base no formato largo, onde temos uma linha por indivíduo e assim a análise para este estudo fica mais adequada. Ao se fazer um estudo longitudinal a base de dados tem que estar no formato longo, ou seja, uma linha por observação, para que assim possamos lidar melhor com as informações.

Uma grande vantagem dos estudos longitudinais é o fato de podermos usar cada indivíduo como referência para si mesmo, com isso conseguimos separar o efeito da idade ao longo do tempo para cada indivíduo e ver o que acontece tanto positivo quanto negativamente, e também podemos distinguir o efeito de cohort, que é a diferença entre indivíduos da base de dados, diferentemente dos estudos transversais, que possuem apenas uma única medida em uma variável para cada indivíduo em um certo instante de tempo e por esse motivo não é possível fazer esta distinção. Entretanto como os dados longitudinais possuem uma repetição de medidas no tempo para cada indivíduo, isso levanta a possibilidade de uma correlação entre as observações de um mesmo indivíduo, devido isto, a estrutura da correlação é de fundamental importância na estimação dos parâmetros do modelo. (Diggle, 2002)[9] determinou três categorias para os modelos longitudinais:

- **Modelo de transição:** são os modelos que apresentam em sua modelagem o efeito que o tempo traz para com os indivíduos juntamente com o valor esperado, onde as-

sim tende-se a ter uma resposta que já foi observada anteriormente devido existir um condicionamento na resposta, onde acaba por existir também uma correlação entre as respostas de um mesmo indivíduo, porém explicamos esta correlação devido tratarmos as respostas anteriores como variáveis explicativas, sendo assim Y_{ij} depende das respostas que foram observadas anteriormente.

- **Modelo marginal:** ao utilizarmos um modelo marginal, modelamos a resposta média sobre a subpopulação que tem um valor comum de x como um conjunto de variáveis explicativas, porém separando a correlação intrapessoal para a regressão da resposta nessas variáveis. O modelo se descreve da seguinte forma

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i$$

com $\mathbf{Z}_i \sim MVN(\mathbf{0}, \vee_i)$, $i = 1, ..., m$ onde temos $\mathbf{Y}_i$ como o vetor das variáveis dependentes para o i-ésimo indivíduo, β é um parâmetro desconhecido e é o vetor dos efeitos fixos, $\mathbf{X}_i$ é a matriz desenho dos efeitos fixos. $\vee_i$ é a matriz de covariância, na qual nos fornece a estrutura de correlação entre as medidas do mesmo indivíduo. Existem várias estruturas de correlação que são utilizadas na modelação de $\vee_i$, dentre elas temos:

1. **Auto regressiva de ordem 1** pode ser usada tanto para dados balanceados como não balanceados. A covariância é dependente do tempo, ou seja, a covariância diminui à medida que aumento o intervalo de tempo entre duas observações. O parâmetro auto-regressivo é o ρ e para ser um processo estacionário $|\rho| < 1$.
2. **Não estruturada** não faz sentido usar para dados não balanceados, pois todas as variâncias e covariâncias podem não ser iguais, criando assim uma imensidão de parâmetros e dificultando muito a modelagem.
3. **Simetria composta** pode ser aplicada tanto para dados balanceados como para dados não balanceados. Os erros são independentes, logo as covariâncias e variâncias são constantes entre todas as observações de um mesmo indivíduo.

Na Figura 2.1 temos as estruturas de correlação descritas acima.

- **Modelo de efeitos aleatórios:** dados longitudinais integram três fonte de variabilidade que se podem observar em um variograma empírico, são elas: a variância não explicada, τ^2, que é o erro de medida; a variância dentro das observações de um mesmo indivíduo, δ^2, que é a variabilidade dentro do indivíduo; e a variância entre indivíduos, ν^2, que é a variabilidade das especificações dos indivíduos. Os modelos de feitos aleatórios se baseiam na analise do valor de cada indivíduo, especificando a característica

Estruturas de correlação		
Não estruturada	Simetria composta	Auto-regressiva (ordem 1)
$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_2^2 & \sigma_{21} & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{31} & \sigma_{32} \\ & & & \sigma_4^2 & \sigma_{41} \\ & & & & \sigma_5^2 \end{bmatrix}$	$\sigma^2 \cdot \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ & 1 & \rho & \rho & \rho \\ & & 1 & \rho & \rho \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$	$\sigma^2 \cdot \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$

Figura 3.1: Exemplos de estruturas de correlação para modelar $\vee_i$. (Pedro, 2019) [1]

individual de cada paciente, descrevendo a sua variabilidade em relação ao tempo e as variáveis explicativas. A correlação em série ou a estrutura de correlação no tempo nos permite identificar a variabilidade entre os indivíduos. O ruído branco é a nossa variância não explicada.

Um modelo longitudinal tem a seguinte forma

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + \epsilon_{ij}$$

onde $\mathbf{Y}_{ij}$ é a variável resposta, $\mathbf{X}_{ij}$ é o vetor de p variáveis explicativas associado aos efeitos fixos β e ϵ_{ij} é o nosso erro aleatório. Ao trabalharmos com modelos de efeitos aleatórios iremos reescrever os ϵ_{ij} decompondo a variabilidade presente nesses erros para saber a fonte desta variabilidade. Também modelamos diretamente os $\vee i$ para podermos escolher a estrutura de correlação mais apropriada para o tipo de modelo que iremos utilizar.

1. Modelo de efeitos aleatórios com um efeito aleatório ao nível de especificidade individual

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + U_i + Z_{ij}$$

onde $\mathbf{X}_{ij}$ é o vetor de p variáveis explicativas associado aos efeitos fixos β, $Ui \sim N(0, \nu^2)$ representa o efeito aleatório que contém a especificidade dos indivíduos, onde ν^2 é a variância do efeito aleatório. $Z_{ij} \sim N(0, \tau^2)$ representa a variabilidade não explicada, onde τ^2 é a variância do ruído.

2. Modelo de efeitos aleatórios com dois efeitos aleatórios correlacionados ao nível da especificidade individual

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + U_i^1 U_i^2 tij + Z_{ij}$$

onde $(U_i^1, U_i^2) \sim MVN(0, \sum)$ com $\sum = \begin{bmatrix} \nu_1^2 & \nu_{12} \\ \nu_{12} & \nu_2^2 \end{bmatrix}$ representando a especificidade do intercept e do declive com ν_1^2 sendo a variância do intercept, ν_2^2 sendo a variância da evolução ao longo do tempo e ν_{12} sendo a covariância entre o intercept e o declive. $Z_{ij} \sim N(0, \tau^2)$ representa a variabilidade não explicada, onde τ^2 é a variância do ruído.

3. Modelo de feitos aleatórios com um efeito aleatório e uma estrutura de correlação temporal contínua no tempo

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + U_i^1 + W_i(tij) + Z_{ij}$$

onde $\mathbf{X}_{ij}$ é o vetor de p variáveis explicativas associado aos efeitos fixos β, $Ui \sim N(0, \nu^2)$ representa o efeito aleatório que contém a especificidade dos indivíduos, onde ν^2 é a variância do efeito aleatório, $W_i(tij)$ é a estrutura de correlação temporal contínua dentro do indivíduo. Pode ser qualquer tipo de estrutura de correlação, simetria composta, não estruturada, auto regressiva de ordem 1, entre outras, o que importa é que a estrutura escolhida torne o modelo o mais adequado possível. $Z_{ij} \sim N(0, \tau^2)$ representa a variabilidade não explicada, onde τ^2 é a variância do ruído.

4. Modelo de efeitos aleatórios com as três fontes de variabilidade

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\beta + \mathbf{d}_{ij}U_i + W_i(tij) + Z_{ij}$$

onde $\mathbf{X}_{ij}$ é o vetor de p variáveis explicativas associado aos efeitos fixos β, $\mathbf{d}_{ij}$ corresponde ao vetor de variáveis explicativas para os efeitos aleatórios, $U_i \sim MVN(0, \sum)$ e representa os efeitos aleatórios, $W_i(tij)$ reproduz a estrutura de correlação temporal e $Z_{ij} \sim N(0, \tau^2)$ representa a variabilidade não explicada, onde τ^2 é a variância do ruído.

Capítulo 4

Aplicação a uma Base de Dados Real

O transtorno antidepressivo vem sendo investigado pela área da saúde há bastante tempo e existem diversos métodos que verificam a intensidade de cada crise e o que levam os pacientes a desenvolverem essa doença. A DBI foi desenvolvida coletando as descrições textuais de cada paciente sobre os seus sintomas e utilizando esses termos foi estrutura uma escala que refletisse a intensidade e a severidade de um dado sintoma (Beck, 2009)[5]. A DBI-II é uma versão atualizada da BDI, onde também é composta por um questionário que contém 21 respostas de múltipla escolha e cada resposta recebe um um valor de 0-3. Existem 4 categorias, onde 0-13 apresentam depressão mínima, 14-19 depressão leve, 20-28 depressão moderada e 29-63 apresentam depressão severa (Beck, 1996) [4].

Iremos iniciar este capítulo com uma descrição da base de dados e em seguida vamos explorar a base de dados BtheB, na qual podemos encontra-lá no package "HSAUR".

Todas as análises aqui efetuadas serão produzidas através do $software R$.

4.1 Descrição da Base de Dados

A base de dados em análise encontra-se no formato largo. A base contém 8 variáveis com 100 observações de 100 pacientes. Os dados não são balanceados, pois foram medidos em tempos distintos. Dos 100 pacientes, 48 foram tratados com o tratamento usual e os outros 52 foram tratados com o novo tratamento, 56 pacientes afirmaram que já tomaram algum medicamento antidepressivo e 44 não fizeram uso de medicamento antidepressivo, existem 49 indivíduos que tiveram uma crise de depressão que durou menos de 6 meses e 51 pacientes tiveram uma crise de depressão que permaneceu por mais de 6 meses. O nível de depressão dos pacientes são medidos em 5 momentos, antes do início do tratamento, após dois meses de tratamento, após quatro meses de tratamento, após seis meses de tratamento e após oito meses de tratamento.

4.1.1 Descrição das Variáveis

- **drug:** O paciente já fez uso de medicamentos antidepressivos. Um factor de dois níveis, não ou sim;
- **length:** A duração do episódio atual de depressão. Um factor dom dois níveis, menos de 6 meses ou mais de 6 meses;
- **treatment:** Grupo de tratamento. Um factor com dois níveis, TAU (tratamento usual) ou BtheB (Beat and Blues);
- **bdi.pre:** Inventário de depressão de Beck antes do tratamento.
- **bdi.2m:** Inventário de depressão de Beck após 2 meses de tratamento.
- **bdi.4m:** Inventário de depressão de Beck após 4 meses de tratamento.
- **bdi.6m:** Inventário de depressão de Beck após 6 meses de tratamento.
- **bdi.8m:** Inventário de depressão de Beck após 8 meses de tratamento.

4.2 Analise Exploratória de Dados

A seguir apresentam-se as medidas de localização das variáveis quantitativas da base de dados:

Estatísticas	bdi.pre	bdi.2m	bdi.4m	bdi.6m	bdi.8m
Mínimo	2	0	0	0	0
1º Quartil	15	8	6	3	3
Mediana	22	15	13	10	10,5
Média	23,33	16,92	14,81	12,76	11,13
3º Quartil	30,25	23	20	20	15,25
Máximo	49	48	53	47	40
NA's	0	3	27	42	48

Tabela 4.1: Medidas de Localização das Variáveis Quantitativas.(REFERÊNCIA)

A Tabela 4.1 ilustra que antes do tratamento 25% dos pacientes já apresentavam depressão leve, sendo que em média os pacientes apresentam depressão moderada, entretanto mais de 25% dos pacientes apresentam sintomas de depressão severa. Podemos observar que após 2 meses de tratamento temos 25% dos pacientes apresentando depressão mínima, onde em média os pacientes apresentam depressão leve e 25% apresentam depressão severa, entretanto temos 3 dados faltantes. Após 4 e 6 meses de tratamento podemos notar que 75% dos pacientes não apresentam depressão severa, entretanto temos 27 e 42 dados faltantes,

respectivamente, nesses períodos. Nos 8 meses após o início do tratamento, temos 48 dados faltes, porém percebemos que a escala diminuiu consideravelmente, pois 75% dos pacientes apresentam sintomas de depressão leve, com uma média na escala dos 11,13 pontos, então, apresentam depressão mínima.

A seguir apresentam-se as medidas de dispersão das variáveis quantitativas da base de dados:

Estatísticas	bdi.pre	bdi.2m	bdi.4m	bdi.6m	bdi.8m
Variância	117,52	116,35	139,71	124,40	86,59
Desvio Padrão	10,84	10,79	11,82	11,15	9,31
Amplitude Amostral	47	48	53	47	40
Amplitude Interquartil	15,25	15	14	17	12,25
Coeficiente de Variação	46,47	63,76	79,82	87,42	83,57

Tabela 4.2: Medidas de Dispersão das Variáveis Quantitativas.(REFERÊNCIA)

Observando a Tabela 4.2 nota-se que do início ao 6 meses de tratamentos temos uma variância superior a 116 e no fim do tratamento apresenta-se uma variância um pouco menor do que as demais, entretanto com um coeficiente de variação superior aos 83%. Começamos o tratamento com uma amplitude amostral de 47, tivemos um leve aumento, chegando aos 53 nos 4 meses de tratamento, porém ao fim do tratamento tivemos uma declínio positivo chegando aos 40 pontos na escala, porém podemos perceber que existem pacientes com sintomas de depressão severos ao fim do tratamento.

Transformamos a base de dados em formato longo e unificamos todos os momentos que foram medidos ao longo do tratamento, gerando assim 380 observações únicas. A seguir temos um gráfico do bdi por tipo de tratamento:

A Figura 4.1 sugere que o inventário de depressão de Beck tanto para quem iniciou o tratamento com a droga TAU quanto para quem utilizou a droga BtheB começaram com os sintomas de depressão mínima até a depressão severa. O gráfico também sugere que o nível de depressão mediano dos pacientes que utilizam a droga TAU é diferente dos pacientes que utilizam a droga BtheB, sendo que com a droga TAU os pacientes apresentam sintomas de depressão leve, quase passando para depressão moderada, já os pacientes que fazem uso da droga BtheB apresentam sintomas de depressão leve, quase regredindo para depressão mínima.

A seguir iremos realizar testes estatísticos a fim de averiguar se há evidência estatística de que a mediana do nível de depressão é igual para cada tipo de tratamento. Para decidir qual o teste adequado precisamos analisar a normalidade dos dados, então vamos fazer um gráfico para visualizar os dados, porém iremos aplicar um teste estatístico para confirmar ou não a normalidade dos dados.

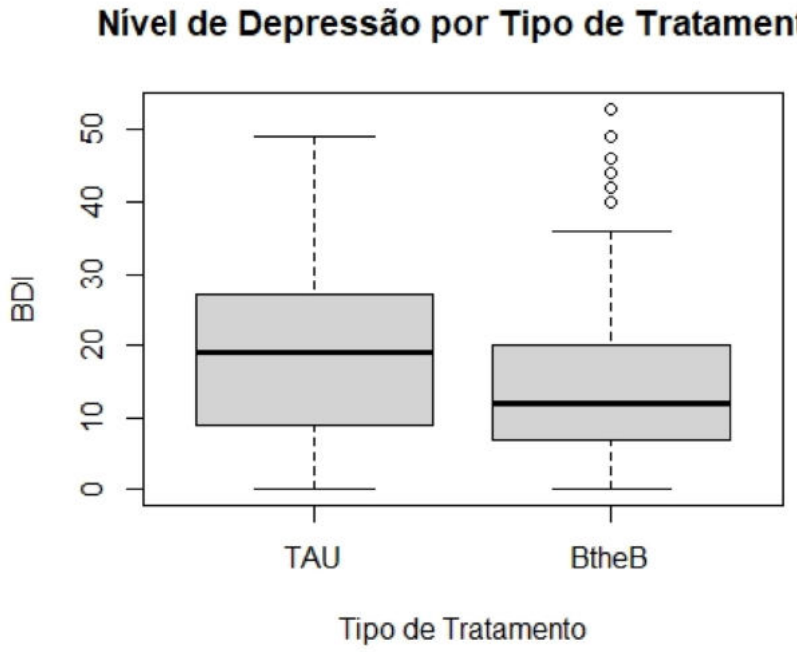


Figura 4.1: Nível de Depressão por Tipo de Tratamento

4.2.1 Teste para Normalidade dos Dados (nível de depressão por tipo de tratamento)

H_0: Os dados seguem uma distribuição Normal
H_1: Os dados não seguem uma distribuição Normal

O tamanho do grupo de pacientes que fizeram o tratamento utilizando a droga TAU possui 183 observações e o grupo de pacientes que fizeram o tratamento utilizando a droga BtheB possui 197 observações.

A análise da Figura 4.2 sugere que os dados não seguem uma distribuição Normal.

Com base nessa análise efetuou-se o teste de shapiro para normalidade dos dados, onde constatamos que os dados realmente não seguem uma distribuição Normal, com valor de prova de 0.001545 para os pacientes que foram sujeitos ao tratamento usual e $8.768e$-09 para os pacientes que foram sujeitos ao novo tipo de tratamento.

Como os dados não seguem uma distribuição Normal, iremos realizar um teste não paramétrico, o teste de Kruskal-Waills para comparar as medianas em cada grupo de tratamento.

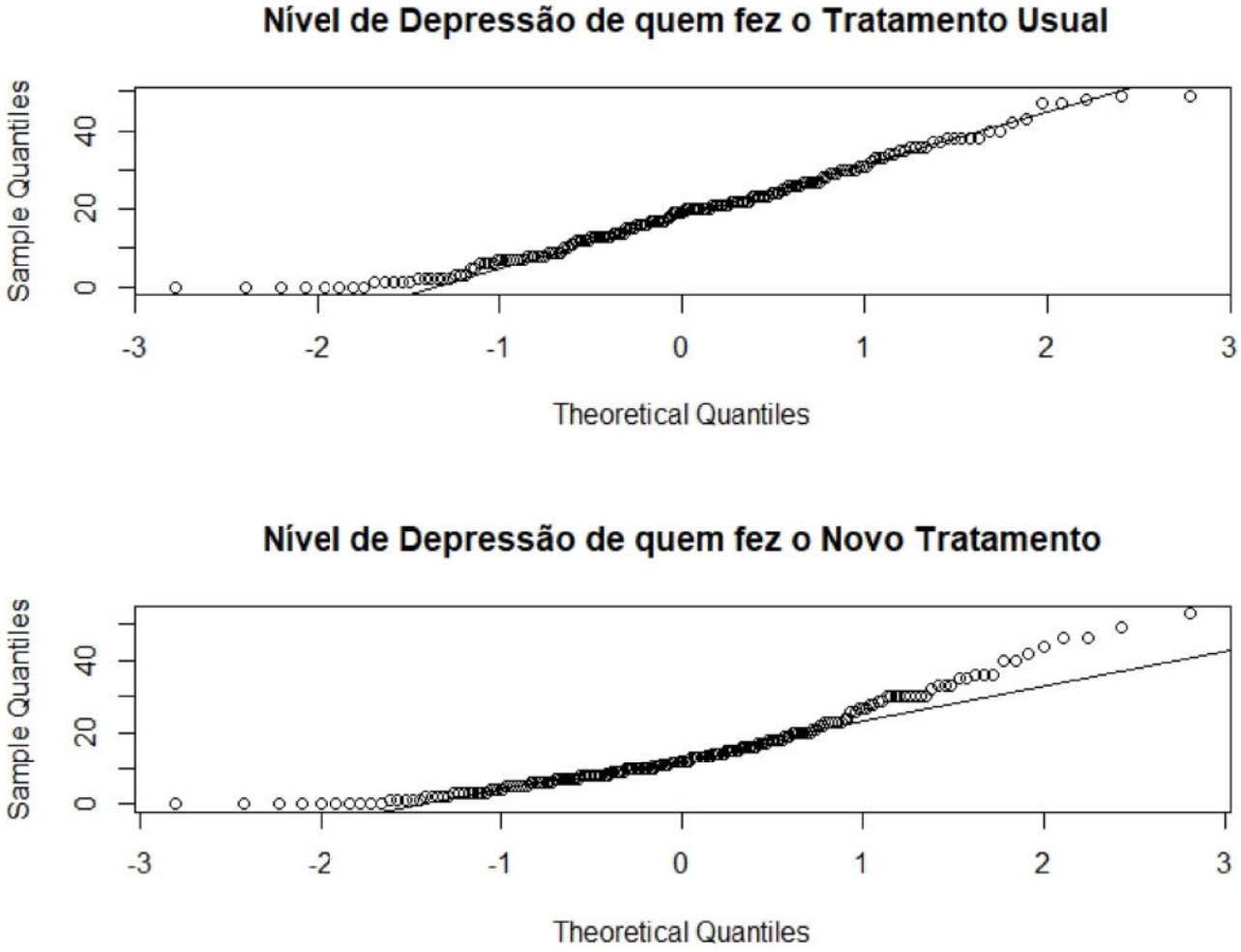


Figura 4.2: Gráfico Q-Q plot para Normalidade (bdi por treatment)

4.2.2 Teste Para Igualdade de Medianas (nível de depressão por tipo de tratamento)

H_0: Não existe diferença nos valores medianos dos 2 grupos
H_1: Existe diferença nos valores medianos dos 2 grupos

Obtivemos um valor de prova de $8.12e$-05. Como o valor de prova é inferior a 0.05, rejeitamos a hipótese nula com um grau de confiança de 95% e concluímos que existem diferenças nos valores medianos dos dois grupos de tratamento.

Tratamento	Média
TAU	19,04
BtheB	14,67

Tabela 4.3: Média do Nível de Depressão por Tipo de Tratamento.

De acordo com a Tabela 4.3 podemos observar que a média do nível de depressão dos pacientes que foram sujeitos ao tratamento BtheB é menor do que a média do nível de depressão dos pacientes que forma sujeitos ao tratamento TAU. Sabendo então que os tratamentos são diferentes e que quanto menor o nível de depressão melhor, então o tratamento mais eficaz foi o BtheB.

A seguir apresentaremos um gráfico do nível de depressão por uso ou não de medicamentos antidepressivos:

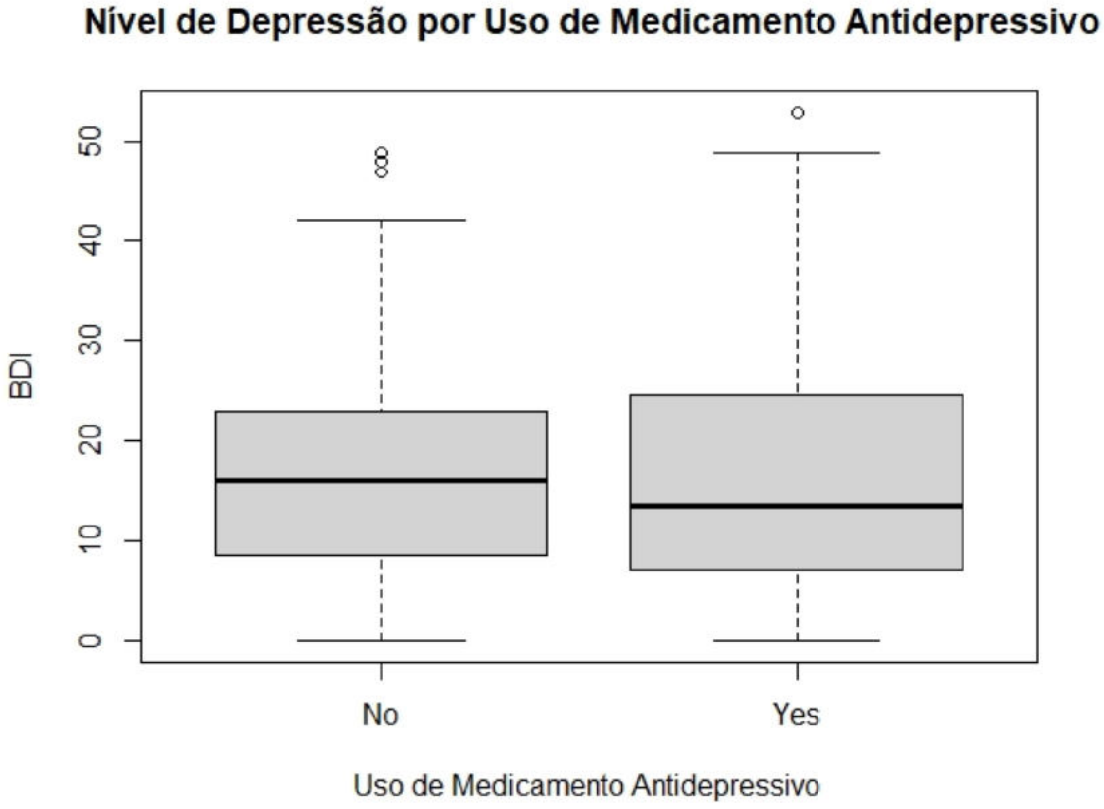


Figura 4.3: Nível de Depressão por Uso de Medicamento Antidepressivo

A Figura 4.3 sugere que a mediana do nível de depressão dos pacientes que utilizaram

algum medicamento antidepressivo está bem próxima da mediana dos pacientes que não utilizaram. Também podemos observar que no grupo que não utilizou medicamento temos alguns outliers que nos mostra que existem pacientes com depressão severa com valores inferiores a 50, sendo que no grupo dos que fizerem uso de medicamentos apresenta-se um outlier, também com sintomas de depressão severa, com valor superior a 50.

Agora iremos realizar testes estatísticos com o intuito de mostrarmos se existe alguma evidência estatística de a mediana do grupo que fez uso de medicamento antidepressivo é diferente do grupo que não utilizou medicamento. Dito isto, iremos fazer um gráfico a fim de visualizar se os dados apresentam normalidade, porém para confirmar iremos aplicar um teste de normalidade dos dados para depois sabermos qual teste utilizar para compararmos a mediana dos dois grupos.

4.2.3 Teste para Normalidade dos Dados (nível de depressão por uso de medicamento antidepressivo)

H_0: Os dados seguem uma distribuição Normal
H_1: Os dados não seguem uma distribuição Normal

O tamanho do grupo de pacientes que não fizeram uso de medicamento antidepressivo possui 212 observações e o grupo de pacientes que fizeram uso possui 168 observações.

A análise da Figura 4.4 sugere que os dados não seguem uma distribuição Normal.

Baseando-se nessa análise efetuou-se o teste de shapiro para normalidade dos dados, onde constatamos que os dados realmente não seguem uma distribuição Normal, com valor de prova de $5.545e$-05 para os pacientes que fizeram uso de medicamento antidepressivo e $6.65e$-07 para os pacientes que utilizaram algum medicamento antidepressivo.

Como os dados não seguem uma distribuição Normal, devemos realizar um teste não paramétrico, o teste de Kruskal-Waills é adequado para comparar as medianas em cada grupo de tratamento.

4.2.4 Teste Para Igualdade de Medianas (nível de depressão por uso de medicamento antidepressivo)

H_0: Não existe diferença nos valores medianos dos 2 grupos
H_1: Existe diferença nos valores medianos dos 2 grupos

Obtivemos um valor de prova de 0.3044. Como o valor de prova é superior a 0.05, não rejeitamos a hipótese nula e concluímos que não existem diferenças nos valores medianos dos

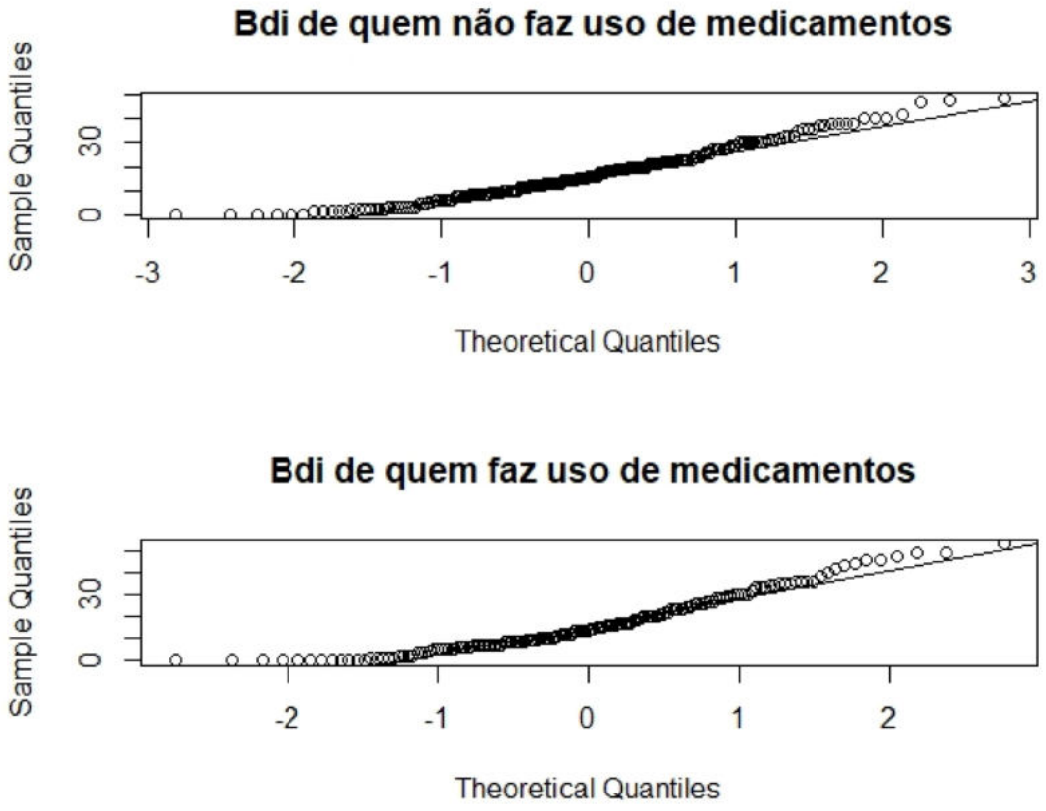


Figura 4.4: Gráfico Q-Q plot para Normalidade (bdi por drug)

dois grupos de pacientes.

Drug	Média
No	16,99
Yes	16,51

Tabela 4.4: Média do Nível de Depressão por Uso de Medicamento Antidepressivo.

De acordo com a Tabela 4.4 podemos observar que a média do nível de depressão do pacientes que usaram algum medicamento antidepressivo antes do tratamento não difere muito da média do nível de depressão dos pacientes que não utilizaram. Até mesmo porque o teste de medianas mostrou que há evidências estatísticas que não existem diferenças nos valores medianos dos dois grupos, logo o fato de usar ou não medicamento antidepressivo antes do início do tratamento não influenciou estatisticamente.

A segui iremos apresentar o gráfico do nível de depressão por duração do episódio:

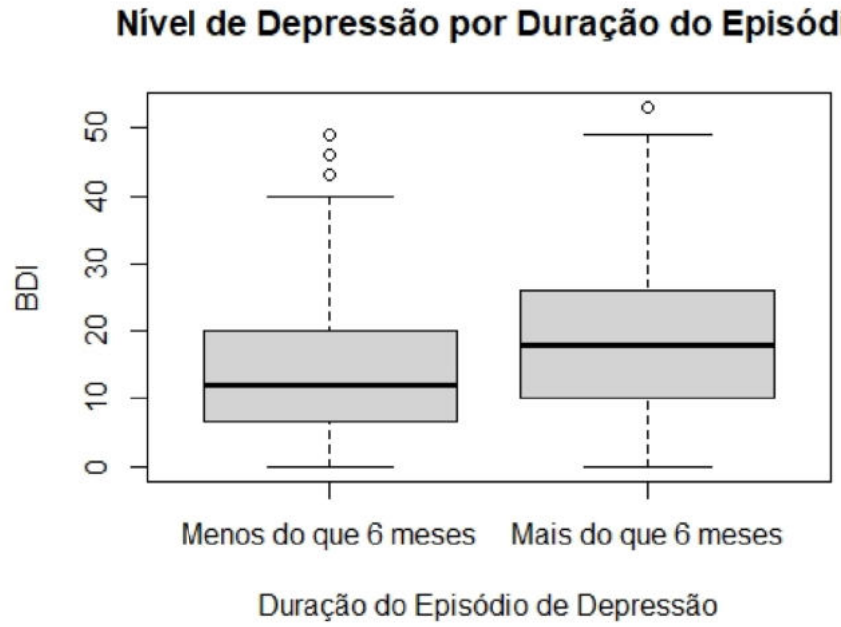


Figura 4.5: Nível de Depressão por Duração do Episódio

Podemos observar na Figura 4.5 que existem pontos outliers nos dois grupos de pacientes. O gráfico sugere que a mediana do nível de depressão dos pacientes que apresentam menos do que seis meses de duração de um episódio é diferente da mediana dos pacientes que apresentam mais do que seis meses de duração. Podemos observar também que os paciente com menos de seis meses de duração de um episódio possuem sintomas de depressão leve, já o outro grupo apresenta sintomas de depressão moderada.

Iremos realizar testes estatísticos a fim de mostrarmos se existe alguma evidência estatística que a mediana dos grupos são diferentes. Sendo assim, iremos fazer um gráfico a fim de visualizar se os dados apresentam normalidade, porém para confirmar iremos aplicar um teste de normalidade dos dados para depois sabermos qual teste utilizar para compararmos a mediana dos dois grupos.

4.2.5 Teste para Normalidade dos Dados (nível de depressão por tempo de duração do episódio)

H_0: Os dados seguem uma distribuição Normal
H_1: Os dados não seguem uma distribuição Normal

O tamanho do grupo de pacientes que apresentam menos do que seis meses de duração de um episódio de depressão possui 171 observações e o grupo de pacientes com mais de seis meses possui 209 observações.

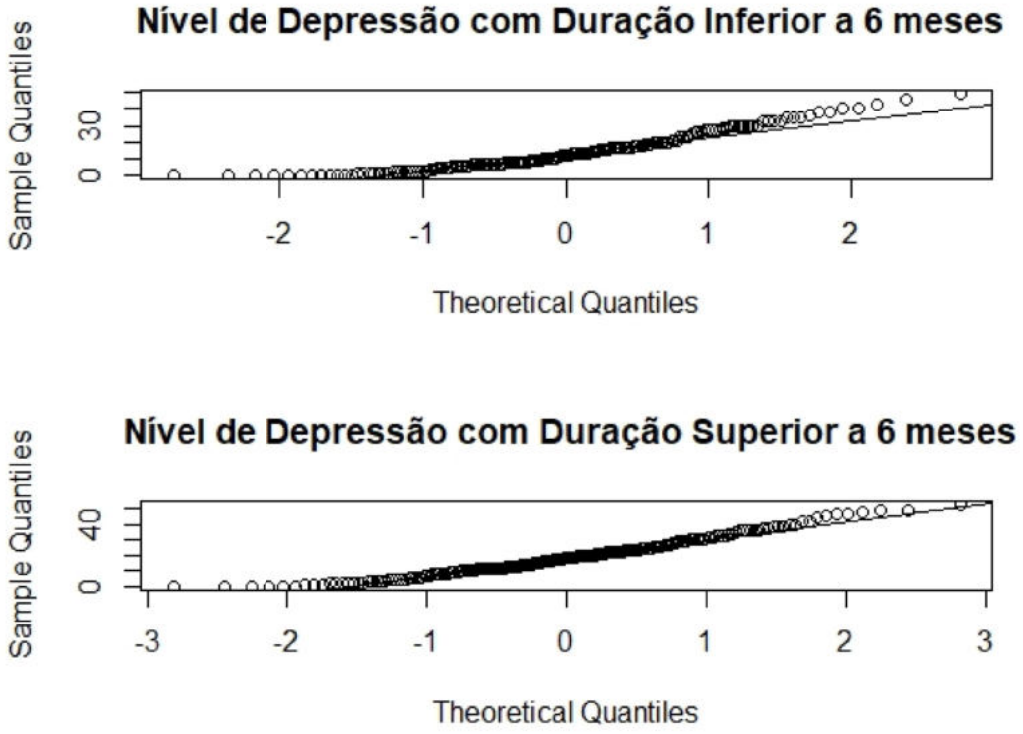


Figura 4.6: Gráfico Q-Q plot para Normalidade (bdi por length)

A análise da Figura 4.6 sugere que os dados não seguem uma distribuição Normal.

Baseando-se nessa análise efetuou-se o teste de shapiro para normalidade dos dados, onde constatamos que os dados realmente não seguem uma distribuição Normal, com valor de prova de $1.62e$-07 para os pacientes que apresentam um episódio de depressão com duração inferior a seis meses e 0.0001417 para os pacientes que apresentam duração superior a seis meses.

Como os dados não seguem uma distribuição Normal, devemos realizar um teste não paramétrico, o teste de Kruskal-Wailles é adequado para comparar as medianas em cada grupo de tratamento.

4.2.6 Teste Para Igualdade de Medianas (nível de depressão por duração de episódio)

H_0: Não existe diferença nos valores medianos dos 2 grupos
H_1: Existe diferença nos valores medianos dos 2 grupos

Obtivemos um valor de prova de $2.883e$-05. Como o valor de prova é superior a 0.05, rejeitamos a hipótese nula com um grau de confiança de 95% e concluímos que existem diferenças nos valores medianos dos dois grupos de pacientes.

Length	Média
<6m	14,19
>6m	18,89

Tabela 4.5: Média do Nível de Depressão por Tempo de Duração de Episódio de Depressão.

De acordo com a Tabela 4.5 podemos observar que a média do nível de depressão dos pacientes que tiveram um episódio de crise inferior a seis menos é menor do que a média do nível de depressão dos pacientes que tiveram um episódio de crise superior a seis meses. Sabendo que os grupos são diferentes, então os pacientes com tempo de duração de crise menor tem vantagens sobre o grupo de pacientes com tempo de crise maior.

A segui iremos apresentar o gráfico do nível de depressão por tempo de tratamento:

A figura 4.6 sugere que a mediana do nível de depressão é diferente para cada período de tratamento. Podemos observar que apenas nos oito meses de tratamento não existe a presença de outliers.

Iremos realizar testes estatísticos a fim de mostrarmos se existe alguma evidência estatística que nos confirme se a mediana dos grupos são diferentes. Sendo assim, iremos fazer um gráfico a fim de visualizar se os dados apresentam normalidade, porém para confirmar iremos aplicar um teste de normalidade dos dados para depois sabermos qual teste utilizar para compararmos a mediana dos dois grupos.

4.2.7 Teste para Normalidade dos Dados (nível de depressão por tempo de tratamento)

H_0: Os dados seguem uma distribuição Normal
H_1: Os dados não seguem uma distribuição Normal

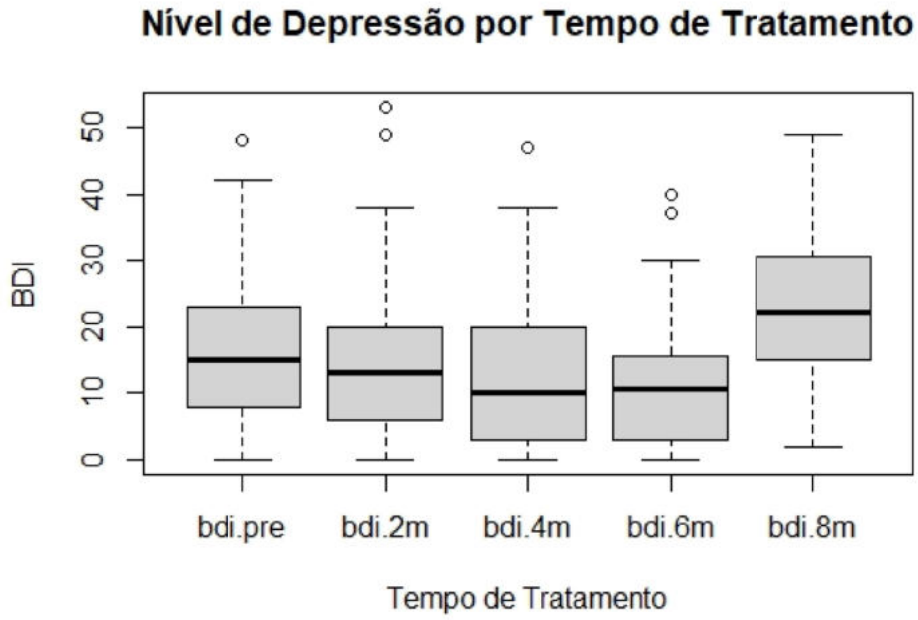


Figura 4.7: Nível de Depressão por Tempo de Tratamento

O tamanho do grupo de pacientes antes de iniciar o tratamento é composto por 100 observações, o grupo que tem dois meses de tratamento possui 97 observações, o grupo de quatro meses possui 73 observação, o grupo de seis meses possui 58 observações e o último grupo possui apenas 52 observações.

A análise da Figura 4.8 sugere que os dados não seguem uma distribuição Normal.

Baseando-se nessa análise efetuou-se o teste de shapiro para normalidade dos dados, onde constatamos que os dados realmente não seguem uma distribuição Normal, apenas os pacientes antes de iniciarem o tratamento que possuem os dados normais, pois apresentam o valor de prova de 0.06105, nos demais períodos de tratamento é apresentado os seguintes valores de prova de 0.002234, $1.654e$-05, 0.0007644, 0.001167.

Como os dados não seguem uma distribuição Normal, devemos realizar um teste não paramétrico, o teste de Kruskal-Waills é adequado para comparar as medianas em cada grupo de tratamento.

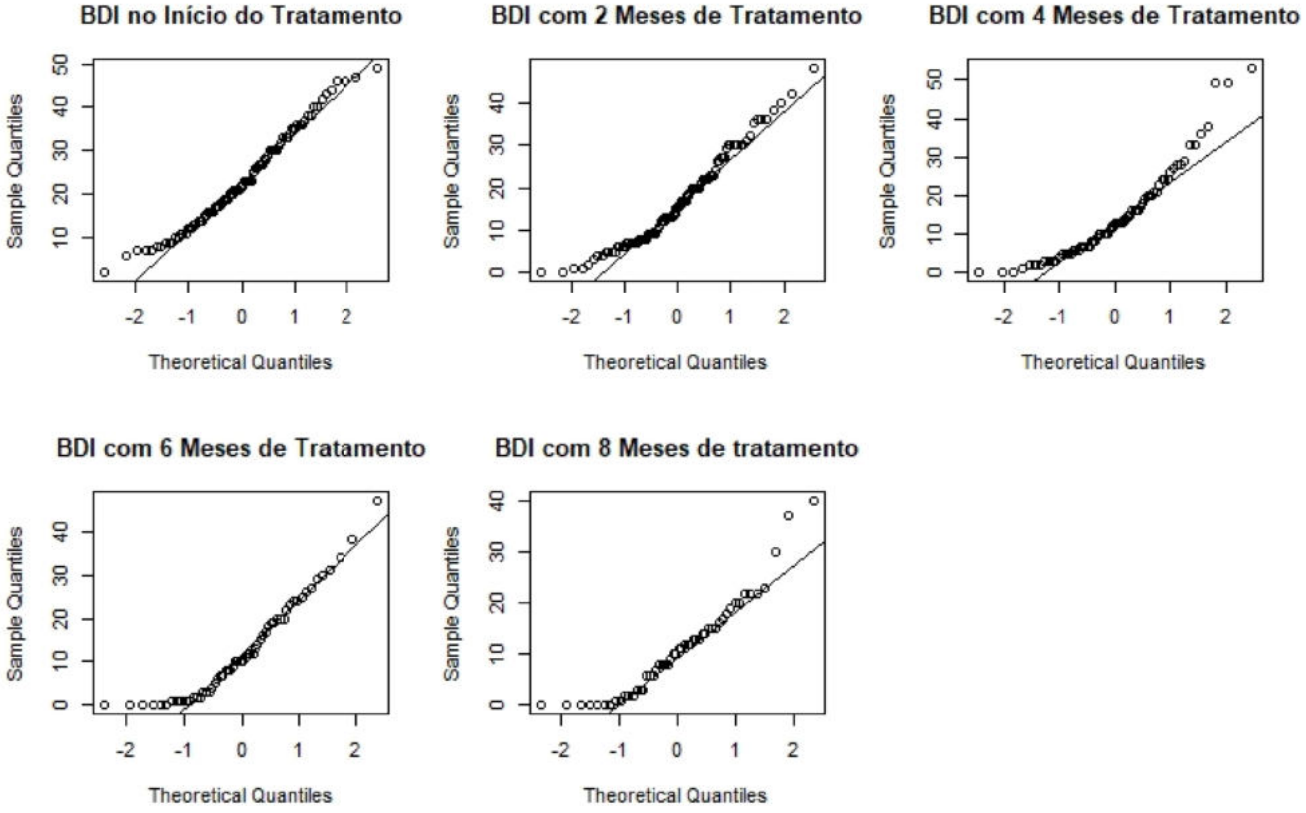


Figura 4.8: Gráfico Q-Q plot para Normalidade (bdi por time)

4.2.8 Teste Para Igualdade de Medianas (nível de depressão por tempo de tratamento)

H_0: Não existe diferença nos valores medianos dos 2 grupos
H_1: Existe diferença nos valores medianos dos 2 grupos

Obtivemos um valor de prova de $1.028e$-11. Como o valor de prova é superior a 0.05, rejeitamos a hipótese nula com um grau de confiança de 95% e concluímos que existem diferenças nos valores medianos dos períodos de tratamento.

Média	bdi.pre	bdi.2m	dbi.4m	bdi.6m	bdi.8m
	23,33	16,92	14,81	12,76	11,13

Tabela 4.6: Média do Nível de Depressão por Tempo de Tratamento.

De acordo com a Tabela 4.6 podemos observar que a média do nível de depressão dos pacientes antes do início do tratamento era classificada com depressão moderada, entretanto ao passar do tempo foi diminuindo e após oito meses de tratamento a média do nível de depressão dos pacientes passou a ser de depressão mínima.

Capítulo 5

Conclusão e trabalho futuro

OS dados faltantes apresentam dificuldades em suas análises, principalmente quando tratam-se de dados longitudinais. Devemos analisar os padrões dos dados faltantes para podermos identificar qual o mecanismo mais adequado para aplicar nesses dados, pois assim nossas análises serão de suma importância para uma boa tomada de decisão.

Na base de dados BtheB, após uma análise exploratória de dados constatou-se que o novo tratamento, Beat the Blue, é mais eficaz do que o tratamento usual, TAU, pois apresentou uma média de 14,67 e o tratamento usual apresentou uma média de 19,04. Constatou-se também que não existem diferenças estatísticas entre o grupo de paciente que já utilizaram algum medicamento antidepressivo e o grupo que não utilizaram. Já para o grupo de pacientes que apresentam um episódio de depressão com duração inferior aos seis meses apresentaram uma média de nível de depressão de 14,19 e o grupo pacientes que apresentam um episódio de depressão com duração superior aos seis meses apresentaram uma média de nível de depressão de 18,89, então quanto menor o tempo do episódio melhor para o paciente. Sobre o tempo de tratamento, concluímos que os pacientes iniciaram o tratamento apresentando em média sintomas de depressão severa com níve lde depressão no valor de 23,33, após dois meses de tratamento já obtiveram uma diminuição dos sintomas e passaram a apresentar em média sintomas de depressão leve, ao fim do quarto mês de tratamento apresentam a mesma média, porem no fim do sexta mês já apresentaram como média de nível de depressão o valor de de 12,76, já com sintomas de depressão mínima. Ao fim do tratamento os pacientes apresentaram em média sintomas de depressão mínima no valor de 11,13, ou seja, menos da metade da média do nível de depressão quando iniciaram o tratamento. Resultado muito satisfatório.

Para trabalhos futuros, seria interessante acompanhar esses pacientes para saber se o medicamento utilizado no tratamento Beat the Blue não os deixou dependentes químicos, pois os medicamentos antidepressivos possui compostos agressivos e muitas vezes o paciente melhora de um problema e entra em outro.

Bibliografia

[1] Afonso, P.M.M. (2019). *Uma abordagem multivariada para modelos conjuntos de dados longitudinais e de sobrevivência.* Dissertação de Mestrado - Universidade do Minho, Braga - Portugal.

[2] Ekholm, A. and Skinner C. (1998). The musctine children's obesity data reanalysed using pattern mixture models. *Appl. Statist.*, Part 2, **47**, 251-263

[3] Baker, S. G. (1995) Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics*, **51**, 1042-1052.

[4] Beck A. T., Steer R. A., Ball R, Ranieri W (Dez de 1996). Comparison of Beck Depression Inventories -IA and - II in psychiatric outpatients. *Journal of personality assessment.* **67 (3)**, 588–97.

[5] Beck, A. T., Alford, B. A. (2009). *Depression: Causes and treatment.* University of Pennsylvania Press.

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

[7] Conaway, M. R. (1992) The analysis of repeated categorical measurements subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **87**, 817-824.

[8] Conaway, M. R. (1994) Causal nonresponse models for repeated categorical measurements. *Biometrics*, **50**, 1102-1116.

[9] Diglle, P.J.; Heagerty P.; Liang, K.J and Zeger S. and Others. (2002). *Analysis of Longitudinal Data*, London, Oxford University Press

[10] ENDERS, C.K. Applied Missing Data Analysis. Guilford Press, Inc.72 Spring Street, New York, 2010.

[11] Ibrahim, J.G., Molenberghs (2001), G. Missing data methods in longitudinal studies: a review. *TEST*, **18**, 1–43.

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

[13] Little, R.J.A and Rubin, D.B. (2019) *Statistical Analysis with Missing Data [Third Edition]*, Wiley Series in Probability and Statistics, book 793.

[14] Rubin, R.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

[15] Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. J. Am. *Statist. Ass.*, textbf88, 125-134.

[16] Little, R. J. A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112-1121.

[17] Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1994) Weighted least squares analysis of repeated categorical measurements with outcomes subject to nonresponse. *Biometrics*, **50**, 11-24.

[18] Paes, A.T. e Poleto, F.Z. (2013). O problema de dados omissos (missing data), *Educação Continuada em Saúde: Einstei*, **11**, 1, 5-7.

[19] Woolson, R.F. and Clarke, W.R. (1984). Analysis of categorical incompletel longitudinal data. *Journal of the Royal Statistical Society*, Series A, **147**, 87-99.