UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE EDUCAÇÃO E SAÚDE
UNIDADE ACADÊMICA DE FÍSICA E MATEMÁTICA
GRADUAÇÃO EM MATEMÁTICA

JOSÉ JOEDSON LIMA DE SOUZA

VALIDAÇÃO DE UM MODELO LINEAR COM RESÍDUOS NÃO NORMAIS PELO MÉTODO DE BOX-COX

CUITÉ
2023

JOSÉ JOEDSON LIMA DE SOUZA

VALIDAÇÃO DE UM MODELO LINEAR COM RESÍDUOS NÃO NORMAIS PELO MÉTODO DE BOX-COX

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Matemática do Centro de Educação e Saúde da Universidade Federal de Campina Grande, como requisito para obtenção do grau de licenciado em Matemática.

Orientador: Jorge Alves de Sousa, Dr.

Coorientador: Anselmo Ribeiro Lopes, Ms.

CUITÉ
2023

S729v Souza, José Joedson Lima de.

Validação de um modelo linear com resíduos não normais pelo método de Box-Cox. / José Joedson Lima de Souza. - Cuité, 2023.
58 f. : il. color.

Trabalho de Conclusão de Curso (Licenciatura em Matemática) - Universidade Federal de Campina Grande, Centro de Educação e Saúde, 2023.

"Orientação: Prof. Dr. Jorge Alves de Sousa; Ms. Anselmo Ribeiro Lopes".

Referências.

1. Não normalidade. 2. Resíduos. 3. Transformação de dados. 4. Modelo ajustado. I. Sousa, Jorge Alves de. II. Lopes, Anselmo Ribeiro. III. Título.

CDU 519.2

Ficha catalográfica elaborada pelo bibliotecário Ms. Jesiel Ferreira Gomes - CRB-15/256

JOSÉ JOEDSON LIMA DE SOUZA

VALIDAÇÃO DE UM MODELO LINEAR COM RESÍDUOS NÃO NORMAIS PELO MÉTODO DE BOX-COX

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Matemática do Centro de Educação e Saúde da Universidade Federal de Campina Grande, como requisito para obtenção do grau de licenciado em Matemática.

Trabalho aprovado em: 13 de junho de 2023.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
JORGE ALVES DE SOUSA
Data: 22/06/2023 17:22:01-0300
Verifique em https://validar.iti.gov.br

Jorge Alves de Sousa, Dr. (Orientador)
Universidade Federal de Campina Grande (UFCG)

gov.br Documento assinado digitalmente
ANSELMO RIBEIRO LOPES
Data: 22/06/2023 17:07:44-0300
Verifique em https://validar.iti.gov.br

Anselmo Ribeiro Lopes, Ms. (Examinador)
Universidade Federal de Campina Grande (UFCG)

gov.br Documento assinado digitalmente
ALECXANDRO ALVES VIEIRA
Data: 23/06/2023 12:37:19-0300
Verifique em https://validar.iti.gov.br

Alecxandro Alves Vieira, Dr. (Examinador)
Universidade Federal de Campina Grande (UFCG)

Dedico este trabalho primeiramente a Deus, que é minha base e meu porto seguro. Ao meus pais, que batalharam para dar uma melhor educação a seus filhos, e que são meu apoio sempre que os obstáculos aparecem.

AGRADECIMENTOS

A Deus, pela dádiva de tornar meus sonhos realidade, trazendo-me a certeza de que posso todas as coisas naquele que me fortalece.

A minha família, pela força e incentivo, nos momentos difíceis dessa longa jornada.

A todos os professores do curso de matemática da Universidade Federal de Campina Grande, que, com seus conhecimentos e experiências, contribuíram com meu conhecimento.

Enfim, a todos que contribuíram, de forma direta ou indiretamente, para a realização deste trabalho.

“A ciência não pode prever o que vai acontecer. Só pode prever a probabilidade de algo acontecer.” (César Lattes)

RESUMO

Quando nos deparamos com a não normalidade de dados, para que possamos fazer algumas análises estatísticas de forma mais confiáveis, como por exemplo, a análise de regressão, se faz necessário que encontremos alguma forma de transformar esses dados, visando atender em especial, a suposição de normalidade. Para nos auxiliar na transformação de dados reais extraídos do portal (MINISTÉRIO DA SAÚDE, 2022) do Departamento de Informática do Sistema Único de Saúde (DATASUS) e posterior ajuste de um modelo linear realizado no R Core Team (2023), optamos por utilizar a técnica de transformação Box-Cox. Além de análises gráficas visuais, para avaliar o atendimento a esta pressuposição para os resíduos do modelo ajustado, é de extrema importância a realização de testes. Para tanto, nessa pesquisa procedemos com um dos testes mais utilizados para análises de normalidade, o teste de Shapiro-Wilk. O objetivo deste trabalho foi a validação das pressuposições para ajuste de um modelo linear às variáveis idade e tempo de tratamento de pacientes diagnosticados com neoplasia maligna. Nesse cenário, após aplicação do teste de Shapiro-Wilk, verificamos que os resíduos do modelo ajustado ($p\text{-valor} = 1.107e-08$) levavam à rejeição da hipótese de nulidade (H_0), ou seja, os resíduos não seguem a distribuição normal. Dessa maneira, foi aplicado a transformação de Box-Cox nesses resíduos, porém, após feito a transformação foi aplicado novamente o teste nos novos dados, e foi encontrado que os dados permaneciam rejeitando o (H_0), pois o p-valor foi igual a 0.001, sendo menor que o nível de significância sugerido por Fisher, assim, pode-se concluir que essa transformação não a ideal para esses dados em questão. Para próximos trabalhos em que variáveis ajustadas a modelos de regressão linear possuam resíduos que não sigam distribuição normal, sugerimos a aplicação de "Modelos Lineares Generalizados" (MLGs); cuja ideia básica consiste em abrir um leque de opções para variável resposta, permitindo que a mesma pertença à família exponencial uniparamétrica de distribuições.

Palavras-chave: não normalidade; resíduos; transformação de dados; modelo ajustado.

ABSTRACT

When we are faced with the non-normality of data, so that we can perform some statistical analyzes more reliably, such as regression analysis, it is necessary that we find some way to transform these data, aiming to meet, in particular, the assumption of normality. To assist us in transforming real data extracted from the portal (MINISTÉRIO DA SAÚDE, 2022) of the Department of Informatics of the Unified Health System (DATASUS) and later adjust a linear model performed in the R Core Team (2023), we chose to use the Box-Cox transformation technique. In addition to visual graphic analyses, to assess compliance with this assumption for the residuals of the adjusted model, it is extremely important to carry out tests. Therefore, in this research we proceeded with one of the most used tests for analysis of normality, the Shapiro-Wilk test. The objective of this study was to validate the assumptions for adjusting a linear model to the variables age and time of treatment of patients diagnosed with malignant neoplasia. In this scenario, after applying the Shapiro-Wilk test, we verified that the residuals of the adjusted model ($p - valor = 1.107e - 08$) led to the rejection of the null hypothesis (H_0), that is, the residuals did not follow a normal distribution. In this way, the Box-Cox transformation was applied to these residues, however, after the transformation was carried out, the test was applied again to the new data, and it was found that the data remained rejecting (H_0), since the p-value was equal to 0.001, being less than the significance level suggested by Fisher, thus, it can be concluded that this transformation is not ideal for these data in question. For future works in which variables adjusted to linear regression models have residuals that do not follow a normal distribution, we suggest the application of "Generalized Linear Models" (GLMs); whose basic idea is to open a range of options for the response variable, allowing it to belong to the uniparametric exponential family of distributions.

Keywords: non-normality; residuals; data transformation; adjusted model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

DATASUS	Departamento de Informática do Sistema Único de Saúde
MLGs	Modelos lineares generalizados
MS	Ministério da Saúde
Script	Série de instruções para que o PC execute determinadas tarefas segundo programado
TLC	Teorema do Limite Central
V.A	Variável aleatória

LISTA DE SÍMBOLOS

π	A razão entre a circunferência de um círculo e de seu diâmetro
$\ln$	Logaritmo natural
e	Número de Euler, a base do logaritmo natural
$\sum$	Somatório
μ	Letra do alfabeto grego chamada mi
σ	Letra do alfabeto grego chamada sigma
λ	Letra do alfabeto grego chamada lambda
$\mathbb{R}$	Conjunto dos números reais, em notação de intervalo $(-\infty, +\infty)$
$\mathbb{R}^+$	Conjunto dos números reais positivos, em notação de intervalos $(0, +\infty)$
H_0	Hipótese nula (ou da nulidade)
H_1	Hipótese alternativa
$\sin$	Função seno
$\tan$	Função tangente
$\arcsin$	Função inversa do seno ou função de arco seno
$\operatorname{arcsinh}$	Seno hiperbólico inverso ou função arco seno hiperbólico
$\operatorname{arctanh}$	Tangente hiperbólica inversa ou função arco tangente hiperbólico

SUMÁRIO

1 INTRODUÇÃO

Existem várias suposições que devem ser atendidas para que a estatística paramétrica proporcione resultados confiáveis e robustos, por exemplo, quando construímos intervalos de confiança simples, a suposição de que os dados são normalmente distribuídos e não assimétricos para a esquerda ou para a direita, deve ser atendida. Já, para a análise de regressão linear, uma suposição importante é a homoscedasticidade, o que significa que a variância do erro de sua variável dependente é independente de suas variáveis preditoras. Na prática, o que ocorre geralmente é que tais suposições muitas vezes são ignoradas, sendo na maioria das vezes este erro causado por desconhecimento dos pesquisadores. Os autores Box e Cox (1964) enfatizaram a importância de escolher o valor de lambda que otimiza a normalidade e a homogeneidade dos resíduos. Eles propuseram uma abordagem para estimar o valor de lambda a partir dos dados, utilizando a função de verossimilhança. No entanto, vale ressaltar que a escolha do valor de lambda também pode ser feita com base em critérios empíricos ou utilizando métodos gráficos. Nas décadas de 1960 e 1970, antes que os métodos de regressão não paramétricos se tornassem amplamente disponíveis, era comum aplicar uma transformação não linear à variável dependente antes de ajustar um modelo de regressão linear (PRUDENTE, 2009). Isso ainda é feito hoje, sendo a transformação mais comum uma transformação logarítmica da variável dependente, que se ajusta ao modelo linear de mínimos quadrados $\log(Y) = X \cdot \beta + \epsilon$, onde ϵ é um vetor de variáveis independentes normalmente distribuídas. Outras escolhas populares incluem as transformações de potência de Y, como a transformação de raiz quadrada.

A modelagem de dados de contagem é amplamente utilizada em várias áreas do conhecimento, como ciências biológicas, educação, saúde pública e agrárias. No entanto, em muitos casos, os dados podem apresentar superdispersão, erros não normalizados e variação heterogênea. Os dados a serem analisados nessa pesquisa serão sobre a neoplasia malígna, um tipo de câncer. O termo ”câncer” engloba um amplo grupo de doenças caracterizadas pelo crescimento e divisão celular descontrolados, resultando na formação de tumores. Os tumores malignos evoluem com o tempo, e podem, eventualmente, invadir tecidos adjacentes ou se espalhar para outros órgãos, dando origem a metástases tumorais o que limita a sobrevida do paciente segundo (BRANDÃO DIAS; SOUZA KUDO; GARCIA, 2020) citados por (BENITES; PEZUK, 2021). À medida que envelhecemos, ocorrem alterações significativas nas funções fisiológicas e imunológicas, bem como nos estímulos que desencadeiam respostas nos tecidos. O envelhecimento está relacionado a mudanças no corpo, incluindo falhas nos processos celulares e biológicos normais, que muitas vezes são acompanhadas por alterações no estado nutricional, comprometendo a saúde do indivíduo (BENITES; PEZUK, 2021). O estado pró-inflamatório reduzido e a capacidade diminuída de resposta imune adaptativa em idosos tornam esses indivíduos mais suscetíveis ao desenvolvimento de patologias, incluindo o câncer (BOUCHLAKA *et al.*, 2013). Atualmente, o câncer corresponde a segunda causa de morte em todo o mundo, representando cerca de 9,6 milhões das mortes em 2018, o que representa uma em cada seis

mortes (WORLD HEALTH ORGANIZATION (WHO), 2018). Para o Brasil, a estimativa para o triênio de 2023 a 2025 aponta que ocorrerão 704 mil casos novos de câncer, 483 mil se excluídos os casos de câncer de pele não melanoma. Este é estimado como o mais incidente, com 220 mil casos novos (31,3%), seguido pelos cânceres de mama, com 74 mil (10,5%); próstata, com 72 mil (10,2%); cólon e reto, com 46 mil (6,5%); pulmão, com 32 mil (4,6%); e estômago, com 21 mil (3,1%) casos novos. Estima-se que os tipos de câncer mais frequentes em homens serão pele não melanoma, com 102 mil (29,9%) casos novos; próstata, com 72 mil (21,0%); cólon e reto, com 22 mil (6,4%); pulmão, com 18 mil (5,3%); estômago, com 13 mil (3,9%); e cavidade oral, com 11 mil (3,2%). Nas mulheres, os cânceres de pele não melanoma, com 118 mil (32,7%); mama, com 74 mil (20,3%); cólon e reto, com 24 mil (6,5%); colo do útero, com 17 mil (4,7%); pulmão, com 15 mil (4,0%); e tireoide, com 14 mil (3,9%) casos novos figurarão entre os principais.

Diante do exposto, o objetivo deste trabalho foi a validação das pressuposições para ajuste de um modelo linear às variáveis idade e tempo de tratamento de pacientes diagnosticados com neoplasia maligna. Para tanto, foram utilizados dados provenientes do portal (MINISTÉRIO DA SAÚDE, 2022) do Departamento de Informática do Sistema Único de Saúde (DATASUS).

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aspectos Históricos e Origem

Em 12 de novembro de 1733 apareceu um panfleto datado e escrito em latim, onde, estava a primeira publicação relacionada a distribuição normal, mas como uma aproximação da distribuição binomial, posteriormente, em 1738, Abraham de Moivre (1667-1754) traduziu esse panfleto para inglês e o publicou em **The Doctrine of Chances** (2ª edição). Contudo, ela era apenas considerada como uma aproximação conveniente pelos primeiros matemáticos, para a distribuição binomial, essa que constitui um modelo probabilístico resultante do binômio de Isaac Newton (1643-1727) físico e matemático inglês. A distribuição normal teve uma maior valorização de sua importância no início do século XIX, quando apareceu em trabalhos de Pierre Simon Laplace (1749-1827) matemático, astrônomo e físico francês e Carl Friedrich Gauss (1777-1855) matemático, astrônomo e físico alemão. Com isso, ela se tornou bastante aceita como base de vários trabalhos estatísticos, principalmente na astronomia (CAIRE, 2013).

Na verdade, esta é uma família de curvas de dois parâmetros que são gráficos da equação

$$\boxed{y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.} \tag{2.1}$$

Essa distribuição é bastante importante em estatística, pois permite modelar uma infinidade de fenômenos naturais, fornecendo, em especial, uma boa aproximação de curvas de frequência para medidas de dimensões e características humanas.

2.1.1 Modelo Normal

Morettin e Bussab (2010) citam que um modelo fundamental em probabilidades e inferência estatística é o modelo de distribuição normal gaussiana. Essa distribuição possui dois parâmetros, a média μ que é onde está centralizada, o desvio padrão σ, e a variância σ^2 que descreve o seu grau de dispersão.

Definição 2.1 (Modelo Normal). Dizemos que a variável aleatória X tem distribuição normal, quando sua densidade é dada por

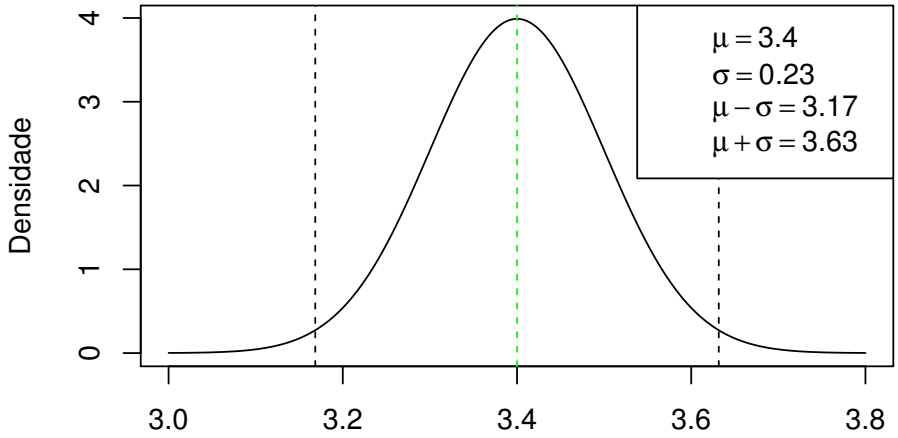
$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ onde } x \in \mathbb{R}, \text{ e } \mu \in \mathbb{R}, \ \sigma^2 \in \mathbb{R}^+. \tag{2.2}$$

Para simplificar a notação, denotaremos a densidade da normal simplesmente por $f(x)$ e escreveremos, simbolicamente, uma variável aleatória x com distribuição normal de média μ e variável σ^2 por,

$$\boxed{X \sim N(\mu, \sigma^2).} \tag{2.3}$$

O gráfico abaixo 1, mostra a curva normal de uma função densidade de probabilidade para uma variável aleatória determinada por valores particulares de μ e σ.

Figura 1 – Curva Normal



Fonte: Elaborada pelo autor.

Podemos demonstrar que (MORETTIN; BUSSAB, 2010):

a) $E(x) = \mu$ e $Var(x) = \sigma^2$;

b) $f(x) \to 0$ quando $x \to \pm\infty$. Ou seja, a curva é assintótica; nunca toca o eixo horizontal, e portanto a função de x jamais se anula;

c) $\int_{-\infty}^{\infty} f(x)dx = 1$. Ou seja, $\forall x \in \mathbb{R}$ a área compreendida pela curva é igual a 1;

d) $\mu - \sigma$ e $\mu + \sigma$ são pontos de inflexão de f, isto é, a curva tem dois pontos de inflexão que são simétricos com relação à média μ;

e) $x = \mu$ é ponto de máximo de f com valor máximo $1/\sigma\sqrt{2\pi}$. Ou seja, a distribuição tem um máximo que corresponde ao seu ponto médio;

f) e $\forall x \in \mathbb{R}$, $f(\mu+x) = f(\mu-x)$. Ou seja, a distribuição f é simétrica em torno da média $x = \mu$.

2.2 Não Normalidade

Assumir a distribuição normal em pesquisa está baseado em dois fundamentos: quando a própria distribuição dos dados é normal, ou quando a distribuição não é normal. De maneira redundante, considera-se que a não normalidade ocorre, quando os erros seguem qualquer distribuição de probabilidade que não seja a normal, por razões intrínsecas ao fenômeno (PINO, 2015).

Existem casos em que a não normalidade é evidente. Por exemplo, quando:

a) há restrições sobre os valores das observações;

b) a distribuição tem caldas pesadas ou deformações em relação à distribuição normal;

c) ou quando a variável aleatória é definida pela razão entre outras duas.

A curva normal é apresentada na forma de erros e é algumas vezes chamada de **Lei dos Erros**. Em qualquer medição, há um número elevado de pequenas fontes de erros; pequenos erros não identificáveis. Nas ciências de observação e experimentais, todos os resultados da observação estão sujeitos a erros (CAIRE, 2013).

Algumas **restrições** que são encontradas em determinadas variáveis que aparecem em estudos com estatísticas agrícolas, particularmente em área plantada e produção, uma delas mais comuns aos valores que as observações podem assumir é que elas sejam estritamente positivas, outras são mais restritivas aos dados de contagem que não devem ser negativos e estritamente inteiros, como por exemplo, números de animais ou números de plantas (PINO, 2015).

Distribuições ditas de **caudas pesadas** atribuem maior probabilidade aos eventos que ocorrem em suas caudas, ou seja, valores distantes das medidas de localização. Por outro lado, as de caudas leves dão maior peso para valores próximos de suas modas, como por exemplo, a distribuição normal, o que sugere maior crédito sobre a informação entregue por elas. Essas caudas pesadas ocorrem em parte quando a variância é muito grande ou até mesmo infinita, como a Distribuição de Cauchy. Em termos de credibilidade entres as fontes, se que forem representadas em pesos de caudas, a de maior crédito seria a de cauda mais leve, portanto, curvas mais ou menos achatadas em relação a uma distribuição normal significam não normalidade (PINO, 2015).

Nas distribuições normais podemos perceber uma simetria no gráfico, ou seja, elas se distribuem da mesma forma tanto acima quanto abaixo do meio da distribuição. Quando isso não ocorre, isto é, quando elas se afastam da normalidade, chamamos essas distribuições de **Distribuições Assimétricas**. No entanto, quando a assimetria é igual a zero ela apresenta uma distribuição simétrica como a normal (PINO, 2015).

Em casos que a não normalidade é evidente, temos a **razão** entre duas distribuições normais, que pela teoria estatística, caso duas dessas variáveis tiverem distribuição normal, a terceira não o terá, mesmo se houver independência entre duas delas (PINO, 2015). Em resumo, considerando $Z = X/Y$, onde X e Y são outras variáveis, deduz-se que:

a) se X e Y forem normais, então, Z terá distribuição de Cauchy (que tem média e variância infinitas);

b) se Z e X forem normais, então, Y terá distribuição de Cauchy;

c) ou se Z e Y forem normais, então, X não terá distribuição normal.

De acordo com Andrews e Mallows (1974) é preciso haver condições necessárias e suficientes para que uma variável aleatória Z seja gerada com a razão X/Y, onde X e Y são independentes e X tem distribuição normal padrão.

2.2.1 Efeitos de Desvios da Normalidade

Um dos procedimentos estatísticos mais comuns é o da suposição de normalidade, onde pode-se mostrar que é possível tolerar um afastamento aceitável da normalidade, com um pequeno

efeito prático no estudo de variância convencional. Os autores Hotelling e Pabst (1936) falam que a não normalidade não costuma dar origem a erros graves na interpretação de médias simples, que na maior parte dos casos são quase normais, diferente do que acontece com a distribuição das estatísticas de segunda ordem pela falta de normalidade, como é evidente em seus erros padrão. O uso da estimação em um modelo de regressão vai depender do grau em que as suposições do modelo são satisfeitas, assim como a heterocedasticidade e normalidade dos erros. Os autores Bernier, Feng e Asakawa (2011) falam que a falta de normalidade não introduz viés na estimação dos parâmetros, mas sim, na dos desvios padrões, afetando a validade dos intervalos de confiança e dos testes de hipótese.

O autor Kronmal (1993) questiona problemas que surgem quando as variáveis do tipo razão surgem como dependentes ou independentes num modelo de regressão, com isso, ele recomenda que no contexto de modelo linear completo onde o intercepto está presente, sejam utilizadas razões, assim ele mostra que seu uso pode levar a inferências enganosas. Pino (2015) destaca alguns casos em que a não normalidade pode ocorrer, como:

a) na estimação de máxima verossimilhança, que pressupõe uma distribuição de probabilidade para poder deduzir as fórmulas de estimação de seus parâmetros;

b) na estimação por intervalo. Na estimação por ponto não é necessário supor uma distribuição, exceto para estimadores de máxima verossimilhança;

c) está associada à assimetria da distribuição, quando as medidas de localização (média, mediana e moda) deixam de coincidir. De modo geral, a não normalidade não conduz a erros muito sérios na interpretação de médias simples, embora deva ser assinalado que a média é mais sensível a outliers[1] do que a mediana;

d) na aplicação de testes de significância baseados na suposição de normalidade, como o teste t de Student ou o teste F, e na análise de variância, esses efeitos podem se mostrar sérios;

e) quando se comparam grupos. O efeito da não normalidade não é sério quando se comparam médias em experimentos com controle interno, como ocorre com a maioria deles, porém, é mais sério quando se comparam variâncias de grupos independentes de observações;

f) ou em casos de heterocedasticidade, ou falta de homogeneidade das variâncias, que costuma ser motivo de preocupação maior. Quando se comparam médias de dois grupos de observações pelo teste t, uma suposição básica é a de que a variância em cada grupo de observações é a mesma, caso contrário, as probabilidades calculadas serão diferentes daquelas dadas nas tabelas de significância.

[1] Segundo Anscombe (1960), um outlier é uma observação com resíduo anormalmente alto.

2.2.2 Predisposições dos Dados

A análise de diagnóstico é fundamental para verificar possíveis desvios das suposições feitas no modelo normal linear, especialmente em relação ao componente aleatório e à parte sistemática desse modelo. É importante destacar observações com interferências desproporcionais e compreender os resultados do ajuste (LOPES, 2012). Ele fala que para implementação de novas técnicas na agricultura é necessário a interpretação correta dos resultados experimentais, mas apenas quando os pressupostos do modelo matemático são satisfeitos, e para uma adequada discriminação dos efeitos do tratamento na análise da variância, é necessário que sejam atendidas as hipóteses de aditividade dos efeitos, bem como a independência conjunta, aleatoriedade, identidade e distribuição normal dos erros, com média zero e variância comum σ^2.

Na estatística, os **dados de contagem** são aqueles em que as medições só podem assumir valores inteiros não negativos 0,1,2,3, ..., e esses números inteiros são obtidos por meio de contagem, não de classificação. Neles encontramos a definição de contagem, que é o número de eventos por unidade de observação. Podemos citar alguns problemas envolvendo contagens (BONAT; ZEVIANI; JÚNIOR, 2017), tais como:

a) o número de acidentes em uma rodovia por semana;

b) o número de automóveis vendidos por dia;

c) ou o número de gols marcados por times de futebol por partida, entre outras.

2.3 Nível de Significância

Também conhecido como nível α, esse é o limite que determina se o resultado de uma pesquisa pode ser considerado estatisticamente significativo, ou não. Isso, após de se realizar os testes estatísticos planeados. Na maioria das vezes, ele é definido como 5% (ou 0,05), sendo que, pode ser utilizados outros níveis dependendo da pesquisa. Essa porcentagem representa a probabilidade de rejeitar a H_0 quando é verdadeira (EUPATI, 2023).

Segundo EUPATI (2023), a probabilidade de um resultado ser devido ao acaso, se a H_0 não existir nenhuma diferença, no caso, for verdadeira. Então, é conhecida como "valor de p". Assim, um resultado é significativo estatisticamente se tiver um valor de $p \leqslant 0{,}05$, ou seja, igual ou inferior ao nível de significância; e com isso, não será considerado uma ocorrência ocasional.

Podemos observar na tabela abaixo de Efron e Gous (1997), ilustrada no livro Morettin e Bussab (2010), a escala de Fisher[2], contra H_0 (ou a favor de H_1).

Tabela 1 – Escala de significância de Fisher

valor-p	0,10	0,05	0,025	0,01	0,005	0,001
Evidência	marginal	moderada	substancial	forte	muito forte	fortíssima

Fonte: Morettin e Bussab (2010)

[2] Fisher utilizou como ponto de referência o valor 0,05, considerando-o como moderado, ou seja, valores do valor-p menores do que 0,05 indicam que devemos rejeitar a H_0.

2.4 Testando a Suposição de Normalidade

Ao fazer várias inferências válidas sobre parâmetros populacionais, é necessário ter como pressuposto a normalidade dos dados de amostra. Diversos métodos de estimativa e testes de hipótese foram desenvolvidos levando em conta a suposição de que a amostra aleatória pertencia a uma população gaussiana. Nesse contexto, para observar se a distribuição é normal, a primeira coisa a se fazer é plotar a distribuição de frequência das observações para ver se há alguma assimetria. Quando há indícios de que a distribuição não é normal, é preciso testar a premissa de que as observações seguem uma distribuição normal, levando em consideração a possibilidade de que isso não seja verdade. Para tanto, existem vários testes categorizados da seguinte maneira (PINO, 2015),

a) testes baseados na função de distribuição empírica;

b) testes baseados em regressão e correlação;

c) e testes baseados em momentos.

2.4.1 Testes Baseados na Função de Distribuição Empírica

Entre os vários testes importantes baseados na distribuição empírica, serão apresentados três principais que consistem em compará-la com a função de distribuição acumulada da normal.

Um teste de uma amostra baseado na diferença entre a função de distribuição cumulativa $F(x)$ e a função de distribuição empírica da amostra $S(x)$ é o Teste de Kolmogorov-Smirnov (KS) dado pela:

Definição 2.2 (Teste de Kolmogorov-Smirnov (KS)). Dado por

$$D_n = \sup_x |F(x) - S(x)|, \text{ onde } \sup \text{ é o supremo do conjunto de distâncias.}$$

A estatística desse experimento é usada para avaliar H_0, ou seja, se a função de distribuição acumulada $F(x)$ é igual a alguma função de distribuição, sob hipótese, $S(x)$. Especificamente, se

$$H_0 : F(x) = S(x), \text{ e } H_1 : F(x) \neq S(x).$$

Já para testar H_0, temos o

Definição 2.3 (Teste de Cramer-von Mises). Dada por

$$W^2 = \sum_{i=1}^{n} [U_i - (2i-1)/2n]^2 + \frac{1}{12n}.$$

Para atribuir pesos maiores às observações nas caudas da distribuição, usamos a:

Definição 2.4 (**Teste de Anderson-Darling**). Dada por

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln U_i + (2n+1-2i)\ln(1-U_i)\right].$$

Pino (2015), explica que alguns autores (como Chen, Lockhart e Stephens (1993)) desenvolveram a teoria para que tanto o teste de Cramer-von Mises (definição 2.3) e o teste de Anderson-Darling (definição 2.4) possam ser aplicados após o uso da transformação de Box-Cox.

2.4.2 Testes Baseados em Regressão e Correlação

Segundo Seier (2002), os testes baseados em regressão e correlação baseiam-se no fato da equação (2.3), em particular poder ser expressa como $y = \mu + \sigma x$, para o caso em que $X \sim N(0,1)$. Em nosso contexto, apresentamos aqui um dos testes mais conhecidos.

Definição 2.5 (**Teste de Shapiro-Wilk** (PINO, 2015)). É dado por

$$w = \frac{1}{(n-k)s^2}\left(\sum_{i=1}^{n} a_i \hat{z}_i\right),$$

com

$$a' = (a_1, ..., a_n) = \frac{c'V^{-1}}{(c'V^{-2}c)^{1/2}},$$

onde $c' = (c_1, ..., c_n)$ e V são, respectivamente, o vetor de valores esperados e a matriz de covariâncias das estatísticas de ordem da normal padrão.

2.4.3 Testes Baseados em Momentos

Nos testes baseados em momentos, a avaliação da assimetria e da curtose da função de distribuição empírica em relação à função de distribuição normal é realizada por meio de testes específicos (CHEN; LOCKHART; STEPHENS, 1993 apud PINO, 2015).

2.4.3.1 Teste de Bartlett

Segundo Dutt-Ross (2020), o **Teste de Bartlett** é utilizado para verificar se as amostras apresentam homogeneidade de variâncias, ou seja, variâncias iguais. A maioria dos procedimentos estatísticos exigem a avaliação dessa hipótese, bem como, vários testes estatísticos pressupõem que as variações são equivalentes entre os grupos. Afim de, definirmos o teste de forma mais concisa, denotamos:

a) o número de repetições do grupo i por n_i;

b) o número de tratamentos por a;

c) a variância do tratamento i por s_i^2;

d) e a variância conjunta por s_c^2, a qual é obtida usando a equação, $s_c^2 = \frac{\sum_{i=1}^{a}(n_i-1)s_i^2}{\sum_{i=1}^{a}(n_i-1)}$.

Definição 2.6 (**Teste de Bartlett** (UNIVERSIDADE ESTADUAL DE LONDRINA, 2023))**.** É dado por

$$B = \frac{M}{C},$$

onde

$$M = \sum_{i=1}^{a}(2_i-1)\ln s_i^2 - \sum_{i=1}^{a}(2_i-1)\ln s_i^2$$

e

$$C = 1 + \frac{1}{3(a-1)}\left[\sum_{i=1}^{a}\left(\frac{1}{n_i-1}\right) - \frac{1}{\sum_{i=1}^{a}(n_i-1)}\right].$$

2.4.3.2 Teste de Jarque-Bera

Segundo Stephanie Glen (2023), esse teste de normalidade é do tipo de teste multiplicar de Lagrange, que é geralmente usado para grandes conjuntos de dados, porque outros testes de normalidade não são confiáveis quando o n é muito grande, como por exemplo: o teste de Shapiro-Wilk não é confiável com o n maior que 2.000. Com isso, o teste compara a assimetria e a curtose dos dados para verificar se eles seguem uma distribuição normal. Os dados podem assumir várias formas, incluindo: dados da série temporal; erros em um modelo de regressão; ou dados em um vetor.

Definição 2.7 (**Teste de Jarque-Bera**)**.** Esse teste é dado por:

$$JB = n\left[\frac{\left(\sqrt{b_1}\right)^2}{6} + \frac{(b_2-3)^2}{24}\right],$$

onde n é o tamanho da amostra, $\sqrt{b_1}$ é o tamanho de assimetria da amostra e b_2 é o coeficiente de curtose.

2.5 Estratégias para lidar com a não normalidade

De acordo com Pino (2015), quando os dados não apresentam distribuição normal, para lidarmos com esses dados de forma mais estratégica, podemos escolher um dos seguintes métodos:

a) **robusto**: para estimar a centralidade e a dispersão dos dados, a estatística clássica recorre à média aritmética e ao desvio padrão, desde que a distribuição dos dados seja normal. Por outro lado, uma nova metodologia estatística é a robusta, que utiliza como estimativas a mediana e o desvio absoluto mediano, que se destaca pela sua eficiência para tratamento de resultados de ensaio de proficiência e por não serem sensíveis a afastamentos da normalidade (BIASOLI *et al.*, 2007).

b) **assintótico**: Pino (2015) aponta que os métodos comuns de análise podem ser utilizados e que as variáveis possuem distribuição aproximadamente normal, tendo em vista que, à medida que o tamanho da amostra dessa distribuição aumenta, ela se aproxima da normal. Esses métodos analisam o comportamento de estatísticas quando os tamanhos das amostras se aproximam do infinito. É preciso levar em conta que as amostras são finitas na prática e, por isso, é importante determinar o tamanho adequado para obter resultados satisfatórios (COSTA, 2017).

c) **transformação de dados**: quando nos deparamos com a não normalidade, e devido a vários fatores (custo, acesso e outros), não podemos aumentar o tamanho da amostra n. É importante observar, conforme descrito por Meyer (1983), que pelo **Teorema do Limite Central** (TLC)[3], para amostras aleatórias simples retiradas de uma população com qualquer distribuição, a distribuição amostral da média aproxima-se de uma distribuição normal.

Com o advento dos softwares estatísticos livres, dentre esses métodos, destaca-se no meio científico a **transformação de dados**. Também, segundo Pino (2015), uma correlação entre a variância e a média pode implicar em assimetria excessiva.

Neste sentido, apresentamos nas subseções abaixo algumas dessas transformações, em particular a transformação potência e as transformações angulares.

2.5.1 Transformação de Box-Cox

Os autores (BOX; COX, 1964) apresentam a técnica de transformação de Box-Cox, que é uma família paramétrica de transformações que inclui várias transformações comumente usadas, como a transformação logarítmica e a transformação de potência. Eles discutem as propriedades e características dessas transformações e destacam a importância de escolher o parâmetro adequado para obter uma transformação ótima. Conhecida também como, **Transformação Potência**, consiste em encontrar um λ tal que os dados transformados se aproximem de uma distribuição normal. No trabalho deles, são apresentados exemplos práticos de aplicação da técnica de transformação de Box-Cox em diferentes contextos, como análise de regressão e análise de variância. Os autores ilustram como a escolha adequada da transformação pode melhorar a interpretação dos resultados e a adequação dos modelos estatísticos aos dados observados. Dessa maneira, podemos dizer que o trabalho de Box e Cox é uma contribuição importante para a teoria e prática da análise estatística, oferecendo uma abordagem sistemática e rigorosa para a transformação de variáveis dependentes, com ênfase na técnica de transformação de Box-Cox. Ele fornece aos pesquisadores uma ferramenta valiosa para lidar com dados que não atendem às suposições da análise clássica, permitindo uma análise mais robusta e eficiente.

[3] O TLC afirma que, a soma S de n variáveis aleatórias independentes X, com qualquer distribuição e variâncias semelhantes, é uma variável com distribuição que se aproxima da distribuição de Gauss (distribuição normal) quando n aumenta.(TAMPLIN, 2023)

Definição 2.8 (Transformação de Box-Cox). É dada por

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+c)^{\lambda}}{\lambda}, & \text{para} \quad \lambda \neq 0; \\ \\ \log(y), & \text{para} \quad \lambda = 0, \ \text{e} \ y > -c. \end{cases}$$

Onde y representa a observação original, $y^{(\lambda)}$ representa a observação transformada, λ e c são parâmetros desconhecidos e $\log$ representa o logaritmo natural. Em particular, segundo Pino (2015), para alguns valores de λ na definição 2.8, nomeamos as transformações. A saber:

a) **transformação cúbica** para $\lambda = 3$, a qual é dada por

$$y^{(\lambda)} = (y+c)^3/3;$$

b) **transformação quadrática** para $\lambda = 2$, a qual é dada por

$$y^{(\lambda)} = (y+c)^2/2;$$

c) **transformação linear** para $\lambda = 1$, a qual é dada por

$$y^{(\lambda)} = y + c;$$

d) **transformação raiz quadrada** para $\lambda = 1/2$, a qual é dada por

$$y^{(\lambda)} = 2\sqrt{(y+x)};$$

e) **transformação logarítmica** para $\lambda = 0$, a qual é dada por

$$y^{(\lambda)} = \log(y+c);$$

f) **transformação raiz quadrada inversa** para $\lambda = -1/2$, a qual é dada por

$$y^{(\lambda)} = -2/\sqrt{y+c};$$

g) **transformação inversa** para $\lambda = -1$, a qual é dada por

$$y^{(\lambda)} = -1/y + c;$$

h) **transformação quadrática inversa** para $\lambda = -2$, a qual é dada por

$$y^{(\lambda)} = -2/(y+c)^2;$$

i) e **transformação cúbica inversa** para $\lambda = -3$, a qual é dada por

$$y^{(\lambda)} = -3(y+c)^3.$$

2.5.2 Transformações Angulares

São transformações que envolvem as funções inversas do seno, do seno hiperbólico e da tangente hiperbólica, ou seja, que tem como resultado um ângulo. Em outras palavras, estas transformações estão relacionadas diretamente com as funções $\arcsin$, $\operatorname{arcsinh}$ e $\operatorname{arctanh}$ (PINO, 2015).

Definição 2.9 (**Transformação inversa do seno**). Dada por

$$y^{(\lambda)} = \begin{cases} \sqrt{n}\arcsin\left(\sqrt{y + \frac{a}{n}}\right) & , \text{ para } -\frac{a}{n} \leqslant y \leqslant 1 - \frac{a}{n}; \\ 0 & , \text{ caso contrário.} \end{cases}$$

Definição 2.10 (**Transformação angular (ou arco-seno)**). Dada por

$$y^{(\lambda)} = \arcsin\left(\sqrt{y}\right).$$

Definição 2.11 (**Transformação seno hiperbólico inverso**). Dada por

$$y^{(\lambda)} = \operatorname{arcsinh}(y) = \frac{1}{\lambda}\log\left(\lambda y + \sqrt{\lambda^2 y^2 + 1}\right), \text{ para } \lambda > 0.$$

Definição 2.12 (**Transformação tangente hiperbólica inversa**). Dada por

$$y^{(\lambda)} = \operatorname{arctanh}(y) = \frac{1}{2}\log\left(\frac{1+y}{1-y}\right).$$

2.5.3 Efeitos da Transformação

Pino (2015) cita que, após utilizar as transformações sobre os dados para garantir normalidade, os dados transformados são submetidos aos procedimentos de análise estatística. Porém, em alguns casos é preciso fazer a transformação inversa para fornecer resultados sobre a variável original, como os intervalos de confiança por exemplo. Com isso podem surgir alguns problemas ao utilizar um modelo linear para uma variável de resposta transformada, e é preciso reverter a transformação das previsões para que elas possam ser expressas nas unidades originais de observação.

3 MATERIAL E MÉTODOS

Nesse capitulo, iremos falar um pouco sobre os dados coletados em MINISTÉRIO DA SAÚDE (2022); como fizemos uma higienização dessa base de dados coletada e como tratamos esses dados. Por fim, sobre o programa R Core Team (2023), onde foi desenvolvido todo o processo.

3.1 Obtenção dos dados

Nessa pesquisa, vamos extrair e analisar dados do ano de 2022, disponíveis na estrutura do Painel de Oncologia do DATASUS (MINISTÉRIO DA SAÚDE, 2022). Dentre todos os dados que fazem parte desse arquivo, vamos tratar como parâmetros a serem estudados:

a) a **IDADE**, mostrando a Idade que tinham os pacientes;

b) o **TEMPO DE TRATAMENTO**, mostrando o tempo de tratamento que os pacientes tiveram de acordo com o seu diagnóstico;

c) e o **DIAGNÓSTICO**, mostrando os tipos de diagnósticos aos quais os pacientes possuíam, que conforme (MINISTÉRIO DA SAÚDE, 2022) têm entradas denotadas por:

- **1**, para neoplasia maligna[1];
- **2**, para neoplasia *in situ*;
- **3**, para neoplasia com comportamento incerto ou desconhecido;
- e **4**, para neoplasia de pele ou neoplasia na glândula tireoide.

A partir dos dados de casos de neoplasia maligna, foram realizadas análises descritivas e inferências, a respeito das variáveis idade e tempo de tratamento. Na avaliação descritiva foi observada a característica da variável da média do tempo de tratamento em relação a idade dos pacientes. Na avaliação inferencial, primeiramente foi realizada a análise gráfica dos dados dos tempo tratamento, buscando encontrar comportamento estacionário, ou seja, que suas propriedades estatísticas, como média, variância e autocorrelação, não mudem significativamente ao longo do tempo ou ao longo de diferentes partes dos dados. Dessa maneira, quando os dados exibem comportamento estacionário, é mais fácil aplicar métodos estatísticos e modelos preditivos, pois as propriedades dos dados podem ser consideradas constantes. Posteriormente ajustou-se os resíduos ao modelo linear com a finalidade de investigar a presença de tendências estatisticamente significativas na variável. E, assim através do ajuste e pelo teste de normalidade foi decidido se aceita ou rejeita H_0, ou seja, confirmar a hipótese de comportamento estável dos dados ou rejeitá-la a favor da presença de tendência. Com a modelagem de regressão, foi feita a verificação da precisão do ajuste por meio da análise residual, com ajuda do teste de normalidade Shapiro-Wilk. Para realização dos ajustes e aplicação dos testes de hipóteses foi utilizado o

[1] A neoplasia maligna é um tumor maligno, formado por células que se apresentam de forma diferente daquelas presentes do tecido normal. (SANTOS, 2022)

software R Studio Versão 2023.03.1+446, usando os pacotes **MASS** para carregar a função de Box-Cox, **nortest** para carregar a função do teste de Shapiro-Wilk, **ggplot2** para carregar ferramentas de gráficos, **readr** para abrir e ler arquivos, **dbplyr** para realizar transformação de dados e o pacote **lmtest** para carregar o teste de Breusch-Pagan.

3.2 Análise Estatística

Após a higienização da base de dados, foi realizada uma análise descritiva dos dados relativos às variáveis idade e tempo de tratamento, buscando verificar características de suas distribuições, para tanto, além do cálculo de medidas descritivas, foram construídos o histograma e o boxplot dos dados, considerando somente os casos diagnosticados com Neoplasia Maligna. Além destes, também visando verificar a condição de normalidade foi construído o gráfico q-qnorm da variável resposta tempo de tratamento. Posteriormente, visando estudar a relação existente entre o tempo de tratamento e a idade, foi construído o diagrama de dispersão do tempo em função da idade, e ajustada uma linha de tendência linear. Em seguida, visando obter um modelo que melhor se ajustasse aos dados e evidenciasse melhor a relação entre variáveis resposta e explicativa, respectivamente, tempo de tratamento e idade, foi proposto um modelo de regressão linear simples, e posteriormente, constatada a significância dos parâmetros do modelo foi realizado um diagnostico dos resíduos, seguido da aplicação de testes de normalidade dos mesmos, constatada a não normalidade, foi realizada a transformação Box-Cox e encontrado o valor de λ para que os dados transformados se aproximem de uma distribuição normal. Após a transformação, ajustou-se o novo modelo ajustado, e verificada novamente o comportamento dos resíduos.

3.3 Software R

O R Core Team (2023) é um software livre de linguagem R de programação estatística e gráfica, que vem se especializando na manipulação, análise e visualização de dados, sendo considerada por muitos uma das melhores ferramentas com essa finalidade (DIDÁTICA TECH, 2022).

4 RESULTADOS E DISCUSSÕES

Neste capítulo, utilizando o software R Core Team (2023), mostraremos como aplicar e validar os dados por meio da técnica Box-Cox como ajuste de suas escalas para que se aproximem de uma distribuição normal. O arquivo contém dados sobre a neoplasia maligna e foram obtidos em (MINISTÉRIO DA SAÚDE, 2022).

4.1 Tratamento dos dados

Primeiramente, dentro do ambiente R, vamos precisar carregar alguns pacotes para iniciar nossas análises dos dados, como o **dplyr**, **readr** e **ggplot2**. Para isso, utilizamos o comando **library()**, colocando dentro dos parênteses o nome do pacote que será necessário.

Código-fonte 1 – Carregamento dos pacotes dplyr, readr e ggplot2 no R

```
library(dplyr)
library(readr)
library(ggplot2)
```

Em seguida, importamos os dados coletados para o R usando o comando:

Código-fonte 2 – Importação dos dados coletados do DATASUS/MS para o R

```
data <- read_csv("data.csv", col_types =
cols(DIAGNOSTIC = col_double(), IDADE = col_double()))
```

Como trabalharemos apenas com as variáveis **IDADE**, **TEMPO DE TRATAMENTO** e **DIAGNOSTICO**, criamos uma base de dados que contenha apenas esses dados. Por simplicidade, denotamos essa nova base de dados como **dados**.

Código-fonte 3 – Criação da base de dados **dados**

```
dados <- data[, c("IDADE", "TEMPO_TRAT","DIAGNOSTIC")]
dados
summary(dados)
```

Por meio do comando summary(dados), obtemos um resumo estatístico dos dados. Obtendo valores para os "Máximo e Mínimo", "1º e 3º Quartil", "Mediana", "Média" e "NA's", conforme apresentado na tabela abaixo.

Tabela 2 – Resumo estatístico de **dados**

IDADE	TEMPO DE TRATAMENTO
Min. : 0.00	Min. : -90
1st Qu.: 47.00	1st Qu.: 32
Median : 60.00	Median :99999
Mean : 57.18	Mean :70433
3rd Qu.: 69.00	3rd Qu.:99999
Max. :122.00	Max. :99999
	NA's :13282

Fonte: Elaborada pelo autor.

Analisando essas informações, percebemos que a variável **TEMPO DE TRATAMENTO** possui valores negativos, mas como se tratam de dias vamos colocá-lo em módulo para que possamos avaliar melhor os dados em questão.

Código-fonte 4 – Colocando **dados** em módulo e chamando de **dados_p**

```
dados_p <- rapply(object = dados, how = 'replace', f = abs)
dados_p
summary(dados_p)
```

Assim, a tabela 2 é alterada para uma nova tabela que não possui mais valores negativos para variável **TEMPO DO TRATAMENTO**.

Tabela 3 – Resumo estatístico de **dados_p**

IDADE	TEMPO DE TRATAMENTO
Min. : 0.00	Min. : 0
1st Qu.: 47.00	1st Qu.: 32
Median : 60.00	Median :99999
Mean : 57.18	Mean :70433
3rd Qu.: 69.00	3rd Qu.:99999
Max. :122.00	Max. :99999
	NA's :13282

Fonte: Elaborada pelo autor.

Agora, precisamos especificar que desejamos o retorno de informações apenas para pacientes com valor "**1**" (neoplasias malignas) para variável **DIAGNOSTICO**, bem como eliminar os dados que não tem informação para variável **TEMPO DE TRATAMENTO**, ou seja, eliminar os dados **99999** e **NA's**, que segundo o dicionário do Painel de Oncologia (MINISTÉRIO DA SAÚDE, 2022) significam respectivamente, "Sem informação de tratamento" e "Não ativa".

Código-fonte 5 – Melhorias em **TEMPO DE TRATAMENTO** e **DIAGNOSTICO**

```
attach(dados_p)
names(dados_p) #retorna o nome das colunas
d = which(TEMPO_TRAT!=99999 & DIAGNOSTIC==1)
d1=dados_p[d,]
summary(d1)
```

A tabela 5 abaixo nos mostra que foram retirados os dados que queríamos e renomeamos essa nova base de dados para **d1**.

Tabela 4 – Resumo estatístico de **d1**

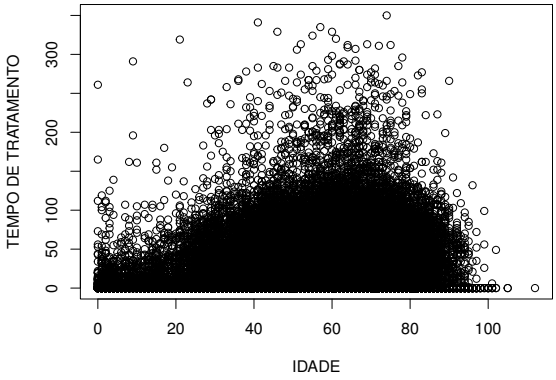
IDADE	TEMPO DE TRATAMENTO
Min. : 0.00	Min. : 0.0
1st Qu.: 49.00	1st Qu.: 0.0
Median : 61.00	Median : 0.0
Mean : 58.05	Mean : 15.4
3rd Qu.: 70.00	3rd Qu.: 20.0
Max. :112.00	Max. : 350.0

Fonte: Elaborada pelo autor.

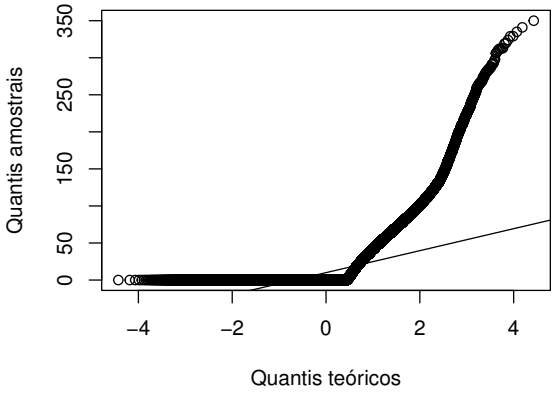
Nesse momento, devemos visualizar as informações da tabela 4, plotando os gráficos desses dados.

Figura 2 – Gráficos para verificação de normalidade de **d1**

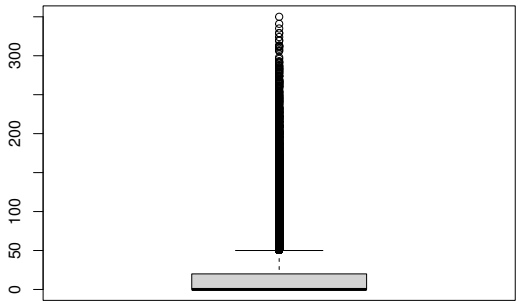
(a) Gráfico: Tempo de Tratamento × IDADE



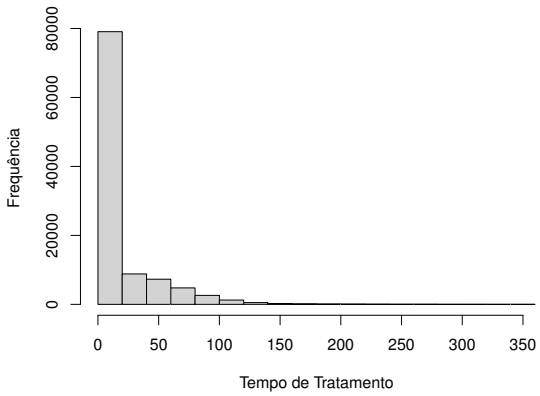
(b) Gráfico Q-Q



(c) BoxPlot: TEMPO DE TRATAMENTO



(d) Histograma: TEMPO DE TRATAMENTO



Fonte: Elaborado pelo autor

Os três últimos nos mostram um pouco sobre a normalidade da variável **TEMPO DE TRATAMENTO**. O código utilizado foi:

Código-fonte 6 – Comandos para gerar gráficos a partir da tabela 4

```
attach(d1)
names(d1)
plot(IDADE,TEMPO_TRAT, main = "")
qqnorm(TEMPO_TRAT, main = "")
qqline(TEMPO_TRAT, main = "")
boxplot(TEMPO_TRAT, main = "")
hist(TEMPO_TRAT, main = "")
```

Observando o gráfico 2b, percebemos que esses dados possuem uma grande quantidade de zeros que não vão ajudar em nossa análise, e portanto devemos retirá-los para uma melhor avaliação, onde renomearemos a nova base de dados para **dados_sem01**[1].

[1] Vale ressaltar que os dados estão em um processo de tratamento, e que os testes de normalidade serão feitos quando reduzirmos o máximo que pudermos de dados considerados desnecessários, para que façamos uma análise mais interessante.

Código-fonte 7 – Removendo zeros de **TEMPO DE TRATAMENTO**

```
dados_sem_0 = which(TEMPO_TRAT!=0)
dados_sem_0
dados_sem01=dentro_01[dados_sem_0,]
dados_sem01
View(dados_sem01)
```

Novamente com o comando summary(dados_sem01), verificamos conforme apresentado na tabela 5 que os zeros foram retirados da nossa variável **TEMPO DE TRATAMENTO**.

Tabela 5 – Resumo de **dados_sem01**

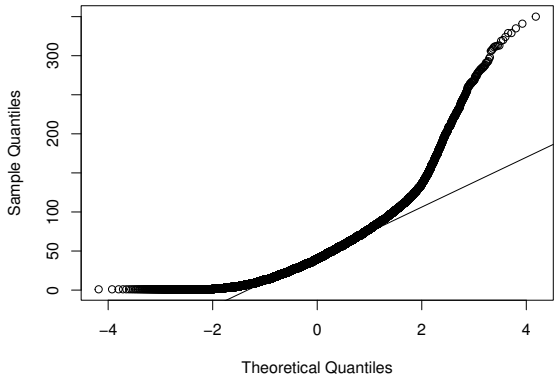
IDADE	TEMPO DE TRATAMENTO
Min. : 0.00	Min. : 1.00
1st Qu.: 49.00	1st Qu.: 21.00
Median : 60.00	Median : 40.00
Mean : 58.06	Mean : 46.77
3rd Qu.: 69.00	3rd Qu.: 64.00
Max. :112.00	Max. : 350.00

Fonte: Elaborada pelo autor.

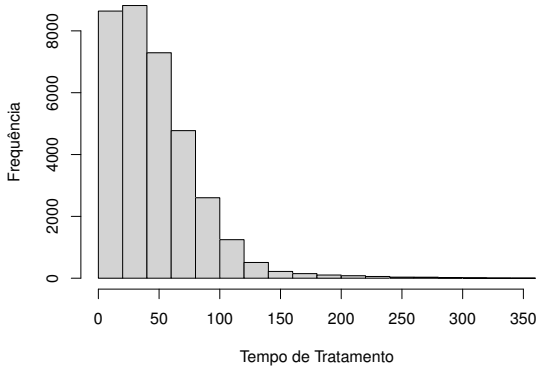
Além disso, observando o gráfico 3a, sabemos que os únicos zeros ainda existentes correspondem às crianças com idade abaixo de um ano.

Figura 3 – Gráficos para verificação de normalidade de **dados_sem01**

(a) Gráfico Q-Q



(b) Histograma



Fonte: Elaborado pelo autor

Visualmente, comparando o gráfico 2b com o gráfico 3a, percebemos que, de fato, os zeros foram eliminados.

Para utilizar o teste de Shapiro-Wilk, precisamos que a quantidade de dados sejam menores que 2000. Porém, ao inserirmos o comando **dim(dados_sem01)** podemos observar que há 34587 linhas e 3 colunas, ou seja, bem maior do que o teste utiliza em sua análise.

Para contornar essa situação, vamos utilizar como nova base de dados a **media** entre as variáveis **TEMPO DE TRATAMENTO** e **IDADE**. Para isso, executamos o comando **tapply** (comando que permite aplicar uma função em duas variáveis ao mesmo tempo) com a função **mean** (função essa que calcula a média dos dados), obtendo uma base de dados com 101 dados.

Podendo assim, de fato, aplicar o teste desejado.

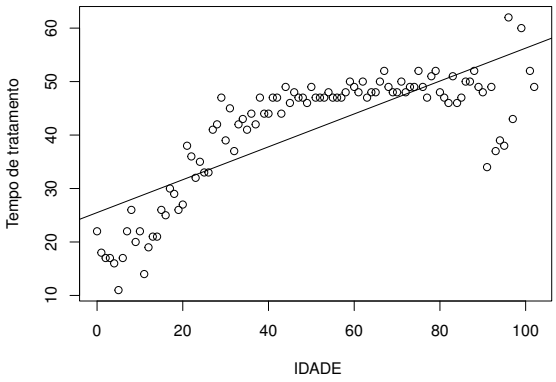
Código-fonte 8 – Média entre as variáveis **IDADE** e **TEMPO DE TRATAMENTO**

```
media <- tapply(TEMPO_TRAT, IDADE, FUN = mean, simplify =
TRUE, drop = TRUE)
View(media)
names(media)
dim(media)

#diagnóstico visual dos dados

regre <- lm(tem_tra ~ ida, data = cap7)
regre
par(mfrow=c(1,1))
plot(tem_tra ~ ida, xlab = " IDADE", ylab = "Tempo de tratamento")
abline(regre)
hist(tem_tra, xlab = "Tempo de tratamento", ylab = "Frequência", main = "")
```

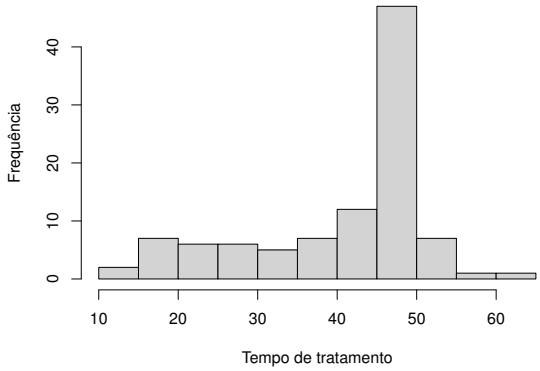
Feito isso, agora podemos fazer um diagnóstico visual dos dados.

Figura 4 – Gráficos para verificação de normalidade da base de dados **media**

(a) Gráfico da média



(b) Histograma da média



Fonte: Elaborado pelo autor

Percebam que em comparação com o gráfico 2a, nossos dados agora no gráfico 4a estão bem mais visuais.

Afim de termos uma melhor análise, criaremos um novo script para a média encontrada. Para isso, será utilizado o comando **write.table()** do pacote **readr()**, que nos permite importar o arquivo "media.csv".

Código-fonte 9 – Uso do comando **write.table()**

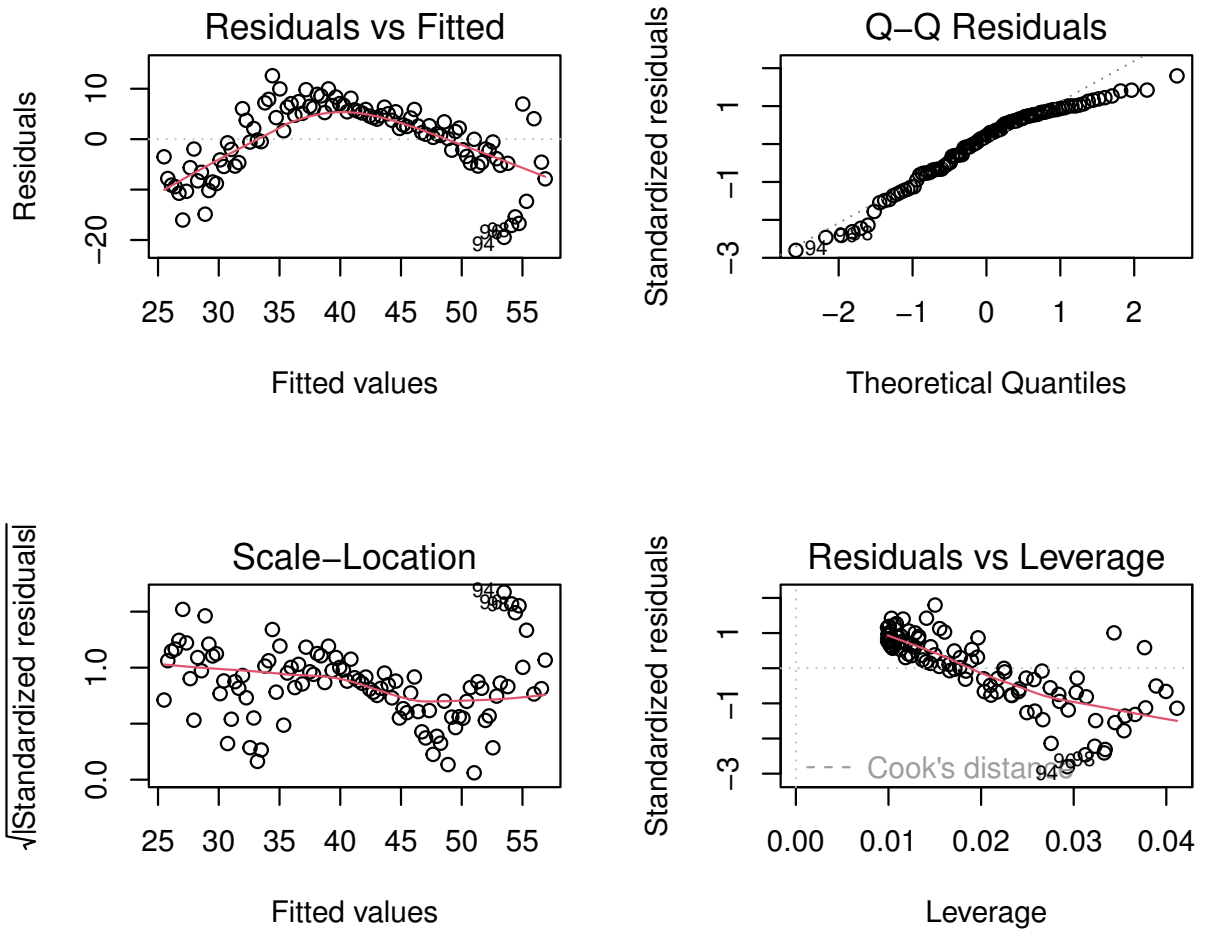
```
write.table(dados, file =
"c:/Users/Joedson/Documents/R/TCC/dados_onc/media.csv",sep=";")
```

A seguir, vamos ajustar o modelo de regressão linear simples e inspecionar os resíduos utilizando a função **lm**.

Código-fonte 10 – Utilizando a função **lm**

```
ajuste <- lm(tem_tra ~ ida)
summary(ajuste)
par(mfrow = c(2,2))
plot(ajuste)
```

Figura 5 – Gráficos de resíduos do modelo ajustado



Fonte: Elaborado pelo autor

Podemos observar nesses gráficos alguns casos de não normalidade como a super dispersão dos dados, caudas pesadas e distribuições assimétricas.

Quando aplicamos o comando summary(ajuste), obtemos algumas informações. Algumas delas são sobre os "Residuals"[2] (da qual esperamos ter "mediana" próxima de 0); e sobre os "Coefficients"[3], bem como, o erro padrão obtido no valor de 7.055.

Código-fonte 11 – Resultado dos testes de regressão

```
> summary(ajuste)
Call:
```

[2] Mostra o mínimo/máximo e os quantis para a distribuição dos resíduos.
[3] São os coeficientes da função ajustada e mostra o quão significativos eles são.

```
lm(formula = tem_tra ~ ida)

Residuals:
    Min      1Q  Median      3Q     Max
-19.487  -4.648   1.512   5.433  12.584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.49641    1.39172   18.32   <2e-16 ***
ida          0.30759    0.02401   12.81   <2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.055 on 99 degrees of freedom
Multiple R-squared:  0.6237,    Adjusted R-squared:  0.6199
F-statistic: 164.1 on 1 and 99 DF,  p-value: < 2.2e-16
```

De posse destas informações, vamos testar a normalidade desses resíduos aplicando o teste de Shapiro-Wilk (definição 2.5), carregando o pacote **nortest** no R.

Código-fonte 12 – 1° Resultado do teste de Shapiro-Wilk

```
> testeShapiro
    Shapiro-Wilk normality test
data:  tem_tra
W = 0.85065, p-value = 1.107e-08
```

Além deste, também foi aplicado o teste de normalidade Kolmogorov-Smirnov, pela função **lillie.test** do mesmo pacote, tendo como resultado:

Código-fonte 13 – Resultado do teste de Kolmogorov-Smirnov

```
> lillie.test(tem_tra)
    Lilliefors (Kolmogorov-Smirnov) normality test
data:  tem_tra
D = 0.2268, p-value = 5.54e-14
```

Sendo um p-valor bem menor que o nível de significância de 0,05 da tabela 1. Para verificar a homogeneidade das variâncias foi aplicado o teste de Breusch-Pagan pelo pacote **lmtest**, com a função **bptest(ajuste)**, onde o **ajuste** é a base de dados ajustada.

Código-fonte 14 – Resultado do teste de verificação de homogeneidade de variâncias

```
> bptest(ajuste)
  studentized Breusch-Pagan test
data:  ajuste
BP = 0.021418, df = 1, p-value = 0.8836
```

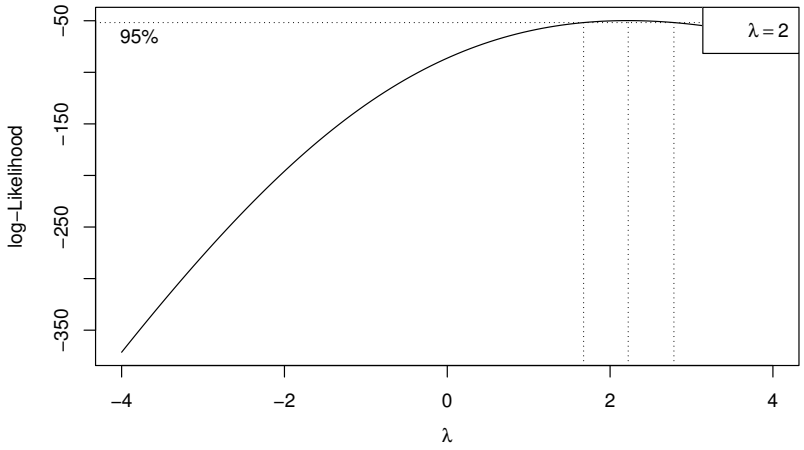
Como o p-valor foi de 0.8 não rejeitamos H_0, o que significa que as variâncias para todas as observações são as mesmas, não possuindo heterocedasticidade (BREUSCH; PAGAN, 1979). Para os testes de normalidade, foi apresentado rejeição de H_0, ou seja, a hipótese de nulidade, já

que seria necessário obter o p-valor maior do que 0.05 (tabela 1), e nesse caso o valor obtido foi $p = 1.107e-08$ para o teste de Shapiro-Wilk, e $p = 5.54e-14$ para o teste de Kolmogorov--Smirnov . Observemos como esses dados se comportam no gráfico de Box-Cox, e para tal carregamos o pacote "library(MASS)" no R.

Código-fonte 15 – Utilizando o pacote **library(MASS)**

```
library(MASS)
dados_boxcox <- boxcox(tem_tra ~ ida, data = cap7, plotit =
T,lamb =seq(-4,4,1/10))

#verificando o valor exato de lambda

dados_boxcox$x[which.max(dados_boxcox$y)]
legend(x = "topright", expression( lambda == 2))
dados_boxcox
```

Figura 6 – Gráfico de Box-Cox



Fonte: Elaborada pelo autor.

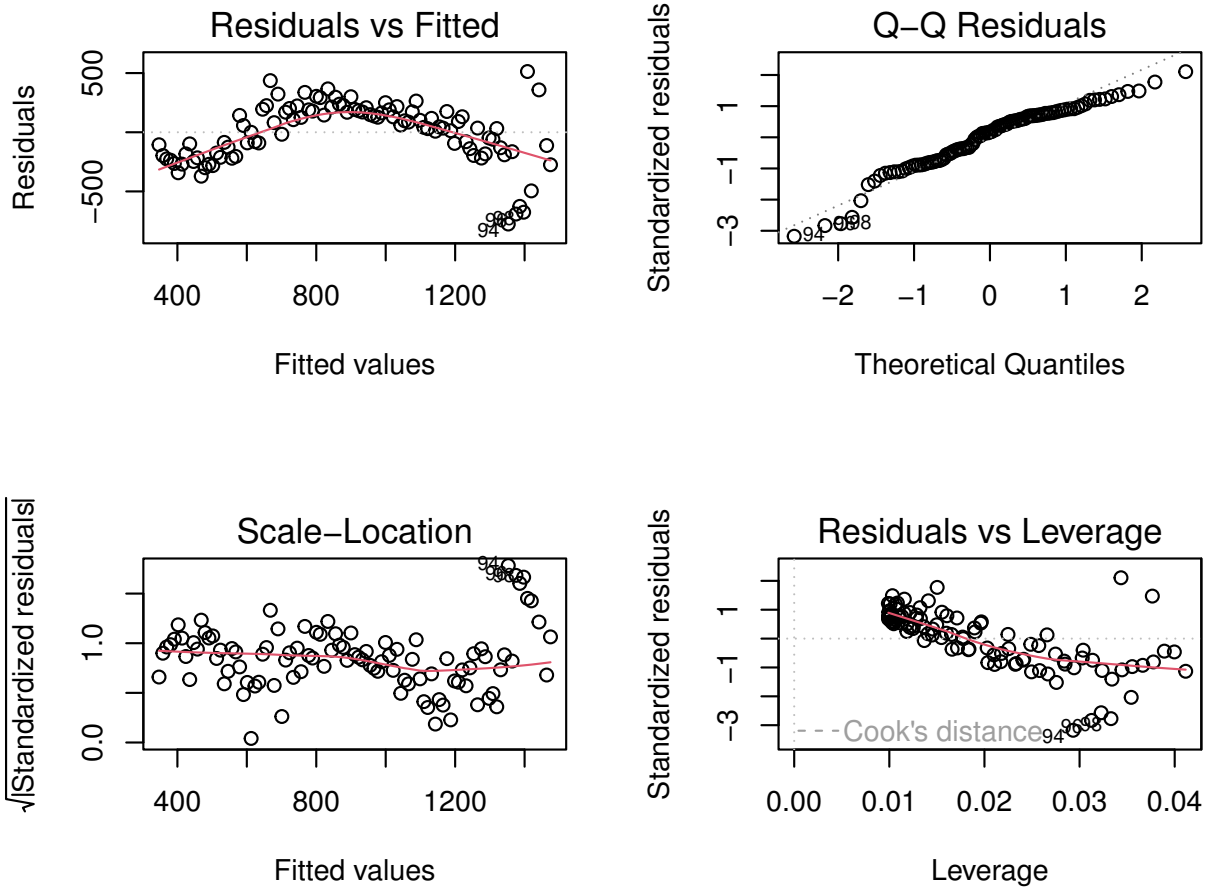
Graficamente, podemos verificar que $\lambda \approx 2$. Sabendo disso, podemos aplicar a Transformação de Box-Cox (definição 2.8 com $\lambda = 2$).

No R, utilizaremos novamente a função **lm** para realizar o ajuste de obtenção da Transformação de Box-Cox com $\lambda = 2$, além de inspecionar graficamente os resíduos. Para tal, fizemos o seguinte procedimento:

Código-fonte 16 – Aplicando a Transformação de Box-Cox

```
#modelo ajustado conforme a transformação de boxcox
ajuste1 <- lm((((tem_tra)^2)-1)/2 ~ ida)
par(mfrow = c(2,2))
plot(ajuste1)
names(ajuste1)
par(mfrow = c(1,1))
boxplot(ajuste1$residuals, main = "")
hist(ajuste1$residuals, main = "", xlab = "Residuos")
```

Figura 7 – Gráfico de regressão após a Transformação de Box-Cox



Fonte: Elaborado pelo autor

Vamos aplicar novamente o teste de Shapiro-Wilk para verificar o p-valor após a transformação.

Código-fonte 17 – 2º Resultado do teste de Shapiro-Wilk

```
> testeShapiro1

    Shapiro-Wilk normality test

data:  ajuste1$residuals
W = 0.95359, p-value = 0.001349
```

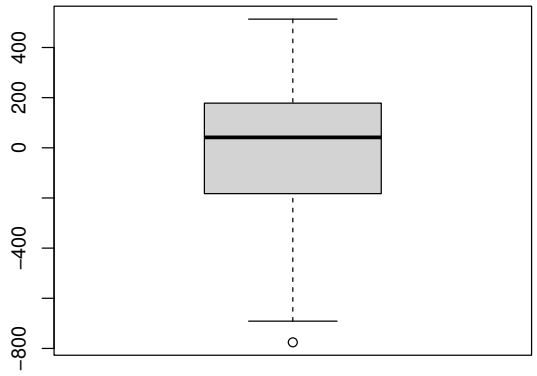
Podemos perceber que mesmo após a transformação, o p-valor foi de 0.001. Assim, mesmo após a transformação, esse valor continua sendo menor que 0.05. Isso ocorreu devido a grande quantidade de dados que haviam em suas caldas, dificultando a transformação. Portanto, observando a tabela 1 de Fisher, devemos rejeitar H_0, ou seja, os dados continuam sendo **não normais**.

Por outro lado, comparando os boxplots 2c e 8a, percebemos que os gráficos melhoraram muito. Vale ressaltar que, não foi aplicado novamente o teste de homogeneidade, pois anteriormente os dados já não possuíam heterocedasticidade. Além disso, no gráfico 8b observamos

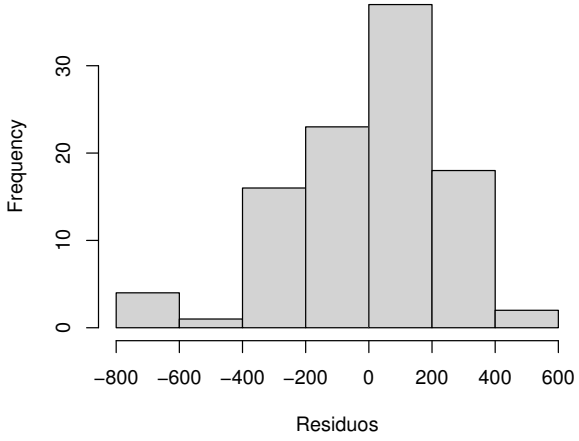
que há uma assimetria a esquerda, causando uma não normalidade.

Figura 8 – Gráficos para verificação de normalidade após transformação dos resíduos

(a) BoxPlot dos Resíduos



(b) Histograma dos Resíduos



Fonte: Elaborado pelo autor

5 CONCLUSÕES

A transformação de Box-Cox é uma técnica estatística utilizada para normalizar a distribuição de dados, cujo resíduos não seguem uma distribuição normal. Isso é importante porque muitos modelos estatísticos e de aprendizado de máquina assumem que os resíduos seguem uma distribuição normal. No transcorrer da nossa análise, sobre uma base de dados extraídas do portal (MINISTÉRIO DA SAÚDE, 2022) e tratada no software livre R, detectamos inicialmente por meio de visualizações gráficas, boxplot, qqplot e histograma, que os resíduos do modelo ajustado, de tempo de tratamento em função da idade não seguiam uma distribuição normal. Entretanto, induções visuais não podem gerar conclusões robustas sobre o comportamento de ajuste de modelos.

Uma forma segura e cientificamente comprovada de se tirar conclusões sobre ajuste de modelos são através de testes de normalidade, sendo, o teste de Shapiro-Wilk considerado o mais recomendado para o mesmo, pois o número de observações passou a ser de 101, sendo menor que 2000, onde o teste não obtinha resultados confiáveis. Na aplicação desse teste aos dados trabalhados junto ao teste de Kolmogorov-Smirnov, verificou que os resíduos do modelo ajustado eram menores que o nível de significância de 0,05, ou seja, levavam a rejeição da hipótese de nulidade (H_0), e dessa forma os resíduos não seguiam distribuição normal. Aplicando o teste de Breusch-Pagan, encontramos que os dados possuíam homocedasticidade, mas não normalidade. Atentando a este fato, foi aplicada a transformação de Box-Cox, considerando o λ sugerido de 2, o que, proporcionou a transformação da variável $(\textbf{TEMPO DE TRATAMENTO})^2$. Resultados observados com a aplicação do teste de Shapiro-Wilk após a transformação, pontuou um p-valor igual a 0.001, que embora tenha reduzido significantemente a probabilidade de rejeição da H_0, mostra que a transformação de Box-Cox não pode ser considerada eficiente para este tipo de dados. Além disso, visualmente podemos observar uma assimetria a esquerda no histograma dos resíduos após a transformação (gráfico 8b).

Para próximos trabalhos com variáveis ajustadas a modelos de regressão linear, onde os resíduos não sigam distribuição normal, sugerimos a aplicação de Modelos Lineares Generalizados (MLGs), onde a ideia básica consiste em abrir um leque de opções para variável resposta, permitindo que a mesma pertença a família exponencial uniparamétrica de distribuições.

REFERÊNCIAS

ANDREWS, D. F.; MALLOWS, C. L. Scale Mixtures of Normal Distributions. **Journal of the Royal Statistical Society. Series B (Methodological)**, Royal Statistical Society, Wiley, v. 36, n. 1, p. 99–102, 1974. ISSN 00359246. Disponível em: http://www.jstor.org/stable/2984774. Acesso em: 1 mai. 2023.

BENITES, Katia Pires; PEZUK, Julia Alejandra. O Tratamento de Câncer de Mama em Idosas, uma Revisão Sobre as Limitações eDificuldades. **Ensaios e Ciência**, v. 25, n. 1, p. 102–109, 2021. DOI: https://doi.org/10.17921/1415-6938.2021v25n1p102-109. Disponível em: https://doi.org/10.17921/1415-6938.2021v25n1p102-109.

BERNIER, Julie; FENG, Yan; ASAKAWA, Keiko. Strategies for handling normality assumptions in multi-level modeling: A case study estimating trajectories of Health Utilities Index Mark 3 scores. **Health reports / Statistics Canada, Canadian Centre for Health Information**, v. 22, p. 45–51, dez. 2011.

BIASOLI, Vinicius *et al.* Aplicação de Estatística Robusta em Ensaios de Proficiência. **ControlLab**, Rio de Janeiro, 2007.

BONAT, Wagner Hugo; ZEVIANI, Walmes M.; JÚNIOR, Eduardo E. Ribeiro. **Modelos de regressão para dados de contagem**: além do modelo Poisson. [*S. l.*], mar. 2017.

BOUCHLAKA, Myriam N. *et al.* Aging predisposes to acute inflammatoryinduced pathology after tumor immunotherapy. **The Journal of Experimental Medicine**, v. 210, n. 11, 2013. Disponível em: www.jem.org/cgi/doi/10.1084/jem.20131219.

BOX, G. E. P.; COX, D. R. An Analysis of Transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, Wiley, v. 26, n. 2, p. 211–252, 1964.

BRANDÃO DIAS, Debora Queila; SOUZA KUDO, Carina Rocha; GARCIA, Daniel Moreno. Impacto de medicamentos biossimilares utilizados na imunoterapia contra o câncer de mama no Brasil. **Brazilian Journal of Natural Sciences**, v. 3, n. 1, p. 274, mar. 2020. DOI: 10.31415/bjns.v3i1.80. Disponível em: https://www.bjns.com.br/index.php/BJNS/article/view/80.

BREUSCH, T. S.; PAGAN, A. R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. **Econometrica**, [Wiley, Econometric Society], v. 47, n. 5, p. 1287–1294, 1979. ISSN 00129682, 14680262. Disponível em: http://www.jstor.org/stable/1911963. Acesso em: 21 jun. 2023.

CAIRE, Elaine. **A história da origem da curva normal**. 2013. Dissertação (Mestrado em Educação Matemática) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista, Rio Claro, 2013. Disponível em: http://hdl.handle.net/11449/91024.

CHEN, Gemai; LOCKHART, Richard; STEPHENS, Michael A. **EDF tests for Normality in linear models after a BOX-COX transformation**. Stanford, California, jul. 1993. Disponível em: https://purl.stanford.edu/vy587cd8259.

COSTA, Felipe Matheus Gonçalves. **Métodos assintóticos em Estatística**. 2017. Trabalho de conclusão de curso (Graduação em Estatística) – Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, 2017. Disponível em: http://dspace.bc.uepb.edu.br/jspui/handle/123456789/16099.

DIDÁTICA TECH. **A linguagem R**. Didática Tech. 2022. Disponível em: https://didatica.tech/a-linguagem-r/. Acesso em: 20 mai. 2023.

DUTT-ROSS, Steven. **Manual de Análise de Dados**. UNIRIO. 2020. Disponível em: https://livro.metodosquantitativos.com/docs. Acesso em: 2 mai. 2023.

EUPATI. **Nível de significância**. 2023. Disponível em: https://toolbox.eupati.eu/glossary/nivel-de-significancia/?lang=pt-pt. Acesso em: 1 jun. 2023.

HOTELLING, Harold; PABST, Margaret Richards. Rank Correlation and Tests of Significance Involving No Assumption of Normality. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 7, n. 1, p. 29–43, 1936. DOI: 10.1214/aoms/1177732543. Disponível em: https://doi.org/10.1214/aoms/1177732543.

KRONMAL, Richard A. Spurious Correlation and the Fallacy of the Ratio Standard Revisited. **Journal of the Royal Statistical Society. Series A (Statistics in Society)**, Wiley, Royal Statistical Society, v. 156, n. 3, p. 379–392, 1993. ISSN 09641998, 1467985X. Disponível em: http://www.jstor.org/stable/2983064. Acesso em: 1 mai. 2023.

LOPES, Alessandro D Lúcio; Diogo V Schwertner; Fernando M Haesbaert; Daniel Santos; Rélia R Brunes; Ana LP Ribeiro; Sidinei J. Violação dos pressupostos do modelo matemático e transformação dedados. **Scielo Brasil**, 2012. DOI: https://doi.org/10.1590/S0102-05362012000300010.

MEYER, Paul L. **Probabilidade**: Aplicações à Estatística. Edição: Ruy de C. B. Lourenço Filho. 2. ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A, 1983.

MINISTÉRIO DA SAÚDE. **Transferência de Arquivos**. 2022. Disponível em: https://datasus.saude.gov.br/. Acesso em: 21 mai. 2023.

MORETTIN, Pedro A.; BUSSAB, Wilton O. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2010. ISBN 978-85-02-08177-2.

PINO, Francisco Alberto. A QUESTÃO DA NÃO NORMALIDADE:uma revisão. **CCTC, Rev. de Economia Agrícola-16**, 2015.

PRUDENTE, ANDREA ANDRADE. **MODELOS NÃO-LINEARES DE REGRESSÃO: ALGUNS ASPECTOS DE TEORIA ASSINTÓTICA**. 2009. Diss. (Mestrado) – UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: https://www.R-project.org/.

SANTOS, Vanessa Sardinha dos. **Neoplasia**. Mundo Educação. 2022. Disponível em: https://mundoeducacao.uol.com.br/doencas/neoplasia.htm/. Acesso em: 20 mai. 2023.

SEIER, Edith. Comparison of tests of univariate normality. **Interstat**, v. 1, jan. 2002.

STEPHANIE GLEN. **Jarque-Bera Test**. StatisticsHowTo.com: Elementary Statistics for the rest of us! 2023. Disponível em: https://www.statisticshowto.com/jarque-bera-test/.

TAMPLIN, True. **Teorema do Limite Central (CLT)**. Strategists Finance. Disponível em: https://www.financestrategists.com/wealth-management/fundamental-vs-technical-analysis/central-limit-theorem/?gclid=CjwKCAjw___ihBhADEiwAXEazJqEmcNeZ-kcdyXreAkX0NLY7F_lWFZdK0FrRVmbkFyJYOUJ-ITdGchoCZ7oQAvD_BwE. Acesso em: 29 abr. 2023.

UNIVERSIDADE ESTADUAL DE LONDRINA. **Teste de Bartlett**. 2023. Disponível em: http://www.uel.br/projetos/experimental/pages/arquivos/Bartlett.html. Acesso em: 25 mar. 2023.

WORLD HEALTH ORGANIZATION (WHO). **Cancer**. 2018. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cancer. Acesso em: 1 jun. 2023.