UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

EDNALDO PEREIRA DA SILVA

Ajuste de modelos de regressão em casos notificados de hipertensão em municípios do Estado da Paraíba

Campina Grande - PB

2019

EDNALDO PEREIRA DA SILVA

Ajuste de modelos de regressão em casos notificados de hipertensão em municípios do Estado da Paraíba

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Ricardo Alves Olinda

Campina Grande - PB

2019

S586a Silva, Ednaldo Pereira da.

Ajuste de modelos de regressão em casos notificados de hipertensão em municípios do Estado da Paraíba [manuscrito] / Ednaldo Pereira da Silva. - 2019.

41 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2019.

"Orientação : Prof. Dr. Ricardo Alves de Olinda , Coordenação do Curso de Estatística - CCT."

1. Modelo de regressão linear múltipla. 2. Método de máxima verossimilhança. 3. Variáveis. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

EDNALDO PEREIRA DA SILVA

Ajuste de modelos de regressão em casos notificados de hipertensão em municípios do Estado da Paraíba

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: 16/05/19

Banca Examinadora:

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba
DE/CCT
Orientador

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba
DE/CCT
Examinador

Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba
DE/CCT
Examinador

Agradecimentos

Agradeço a Deus, a ele toda honra, glória e louvor, se não for o SENHOR o construtor da casa, será inútil trabalhar na construção, será inútil levantar cedo e dormir tarde. Todas as minhas conquistas e labuta seria improfícuo se não houvesse objetivo.(Salmos 127)

Externo os meus agradecimentos a minha digníssima mãe Alzira Maria da Silva e a amada esposa Socorro Silva e aos meus queridos filhos; Nathanael, Nathan e Deborah, são eles que me deram forças para chegar até aqui e vencer todos os desafios.

Ao meu orientador Dr. Ricardo Alves Olinda pela competência profissional, em fim a todos que contribuiram para o meu desenvolvimento profissional.

Resumo

O presente estudo teve como objetivo ajustar modelos de regressão linear múltipla utilizando o método de máxima verossimilhança para encontrar os estimadores dos parâmetros, com a finalidade de predizer possíveis casos de hipertensão notificados, do qual, entre os 223 municípios que engloba o Estado da Paraíba, foram selecionados para compor à amostra quarenta municípios, dos quais 12.851 indivíduos foram notificados, no período escolhido de abril de 2010 à abril de 2013, estes dados foram coletados diretamente do *site* do DATASUS, em que o último registro dessas informações foi no ano de 2013. Neste trabalho foi realizado uma seleção de variáveis regressoras segundo o critério *stepwise*, e na sequência foi ajustado um modelo, após o uso do critério AIC, que para este modelo foram incluidas as variáveis sedentarismo e acidente vascular cerebral, que de acordo com os critérios adotados foram consideradas estatísticamente significativas. Os cálculos foram processados por meio da função *lm* do Software R 3.43.

Palavras-chave: Seleção de variáveis, Variável dependente, Hipertensos notificados.

Abstract

The present study aimed to adjust multiple linear regression models using the maximum likelihood method to find the parameter estimators, with the purpose of predicting possible cases of hypertension reported, of which, among the 223 municipalities that comprise the State of Paraíba, were selected to compose the sample forty municipalities, of which 12,851 individuals were notified, in the period chosen from April 2010 to April 2013, these data were collected directly from the DATASUS website, in which the last record of this information was in the year 2013. In this work, a selection of regressor variables according to the stepwise criterion was performed, and then a model was adjusted, after the use of the AIC criterion, that for this model were included the variables sedentarism and stroke, which according to the adopted criteria were considered statistically significant. The calculations were processed using the function lm of software R 3.43.

Key-words: Key words: Variable selection, Dependent variable, Hypertension reported.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O estudo envolvendo modelo de Regressão teve origem no século XIX com Francis Galton. Em um de seus trabalhos ele verificou a relação entre a altura dos pais e dos filhos, (X_i e Y_i), onde através de suas análises ele conclui que de fato a altura dos pais influenciava na altura do filho, constatou que se os pais fossem muito alto ou muito baixos, a estatura do filho tendia a média, por isso, ele chamou de "regressão", ou seja, há uma tendência de os dados regredirem à média (DEMÉTRIO; ZOCCHI, 2006).

A análise de regressão estuda a relação entre uma variável denominada aleatória e dependente, Y , e uma ou mais variáveis independentes, $(X_1, X_2, \ldots, X_k)$. Caso se considere apenas uma variável independente chamamos de análise de regressão linear simples, caso se tenha duas ou mais variáveis, temos a análise de regressão múltipla. A importância do estudo da análise de regressão advém da necessidade do estudo de determinados fenômenos nas Ciências da Natureza (Física, Biologia, Química), nas Ciências Sociais, e nas Ciências da Saúde (RODRIGUES, 2012).

Na área da saúde, quando o interesse é analisar a importância de um conjunto de fatores sobre doenças, eventos ou outras características de interesse, os métodos estatísticos mais usados são os modelos de regressão, os quais ajustam uma equação entre a resposta de interesse (variável dependente) e as causas que se deseja averiguar (variáveis independentes).

O objetivo desse trabalho é ajustar um modelo estatístico de regressão linear múltipla, para predizer a possível relação existente entre a hipertensão e o sedentarismo, sobrepeso, tabagismo, acidente vascular cerebral e doença renal. Foi selecionado dentre as 223 cidades da Paraíba, quarenta municípios e 12.851 indivíduos notificados, no período designado de abril de 2010 à abril de 2013, onde o último registro dessas informações foi no mês de abril de 2013, e a fonte dos dados coletados foi do *site* do DATASUS [1] tecnologia da informação a serviço do SUS, que dispõe de informações que podem ajudar para subsidiar análises objetivas da condição sanitária, tomadas de decisão fundamentada em evidências e construção de programas de saúde para a população.

[1] http://tabnet.datasus.gov.br/cgi/deftohtm.exe?hiperdia/cnv/hdpb.def

2 Fundamentação Teórica

Esta seção descreve em seu conteúdo os conceitos mais importantes da utililização dos modelos de regressão múltiplas, por intermédio de artigos práticos e teóricos pertinentes ao propósito da pesquisa.

2.1 Marco histórico e origem do termo Regressão

A origem do termo regressão se deu por Francis Galton (1822 - 1911) que viveu na Inglaterra durante um período de indiscrição científica que marcou essa época, ele contribuiu para diversas áreas da ciência que incluem, dentre outras, a Antropometria, a Psicologia e a Hereditariedade. Um dos grandes destaques foi a utilização de ferramentas Estatísticas (Regressão e Correlação) para o estudo dos efeitos hereditários e pela aplicação de métodos estatístico ao estudo da evolução (POLIZELLO, 2011).

A Figura 1 representa a relação entre pais e filhos de uma variável métrica. A linha azul representa o valor esperado se os filhos tivessem precisamente o valor da altura média dos pais, enquanto que a linha verde é a reta de regressão ajustada. Os pais que exibem valores elevados da característica têm descendência com um valor médio da característica menor que a média observada daquela medida dos respectivos pais. Por outro lado, os pais que tem valor menor da característica têm os filhos com valores mais elevados que a média entre altura dos respectivos pais. Por essa razão a lei foi chamada de "regressão para a média".

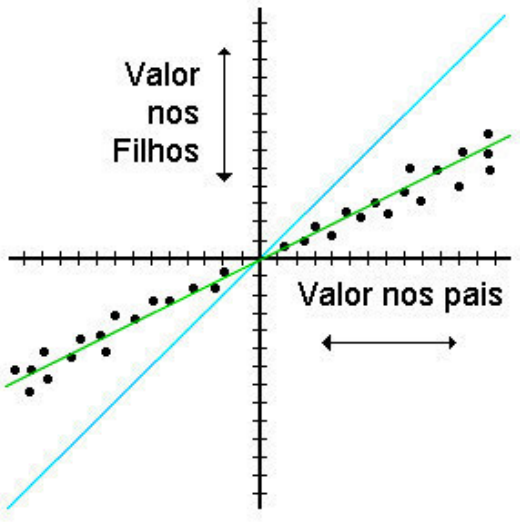

Figura 1 – Gráfico de dispersão da lei de regressão para a mediocridade de Galton

Fonte: https://docs.ufpr.br/ gpj/Apostila-CE066.docx e (RÊGO, 2014)

Galton ainda não tinha a expressão exata da correlação, pois não conhecia seu sinal, foi que Galton pediu para que seu amigo J. D. Hamilton Dickson, professor de matemática na Universidade de Cambridge, para achar a fórmula da superfície encontrada, que corresponde hoje à função normal bidimensional.

Ele expressou-se sobre a co-relação que só depois foi escrita correlação. A fórmula por ele proposta foi modificada por Walter FranK Raphael Weldon (1860 - 1906), professor de Zoologia em Cambridge, que tinha uma ligação muito forte com Galton, onde ele atribuiu com um sinal positivo ou negativo. Mas, a fórmula do coeficiente de correlação, como é conhecida hoje, só foi determinada em 1896, por Karl Person (1857 - 1936).

Segundo Hoffmann et al. (2016), sempre é importante conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista relação causal entre elas, há como relacioná-las por meio de uma expressão matemática, que pode ser útil para se estimar o valor de uma das variáveis, quando se tem conhecimento dos valores das outras. Genericamente, tais relações funcionais podem ser representadas por, $Y = f(X_1, X_2, \ldots, X_k)$, onde Y representa a variável dependente e $X_h (h = 1, 2, \ldots, k)$ representam as variáveis explanatórias.

A utilização de modelos de regressão, tem como objetivos: a predição da variação de Y que é explicada pelas variáveis X, logo pode-se utilizar o modelo para obter valores de Y correspondentes a valores de X que não faziam parte dos dados, o outro é a seleção de variáveis, que não se tem idéia de quais são as variáveis que afetam de forma significativa a variação da variável dependente, a análise de regressão pode efetivamente ajudar no processo de seleção de variáveis, retirando aquelas cuja contribuição não seja relevante para o estudo de interesse, o outro objetivo é estimar os parâmetros, ou ainda, ajustar o modelo aos dados, isto significa obter estimativa para os parâmetros tendo por base o modelo e os dados observados (DEMÉTRIO; ZOCCHI, 2006).

2.2 Análise de regressão linear

Na Estatística um dos problemas mais comuns é estudar a relação entre duas variáveis X e Y, isto é, procura-se uma função de X que explique Y. Matematicamente falando, deseja-se encontrar uma função f de forma que

$$Y = f(X_1, X_2, \ldots, X_k).$$

Segundo Demétrio e Zocchi (2006) em geral a relação entre variáveis não é perfeita, ou seja, os pontos não se situam perfeitamente sobre a função que relaciona as duas variáveis, no entanto a natureza das variáveis X e Y pode sofrer variação, isto é, elas podem ser fixas, ou seja, controladas ou podem ser aleatórias, logo, em lugar do modelo

acima, costuma-se descrever a variável Y como a soma de uma quantidade que não é regido por fenômenos aleatóricos e uma quantidade aleatória. A parte aleátoria é denominada erro e pode indicar inúmeros fatores, que em conjunto, podem exercer influência na variável resposta. Admitindo que esse erro seja aditivo, o modelo estatístico fica da seguinte forma

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon, \tag{2.1}$$

em que, $f(X_1,X_2,\ldots,X_k)$ é a parte sistemática, e ε é a componente aleatória do modelo. Logo, pode-se dizer que o erro provoca uma distorção sobre a parte determinística.

Em geral, não existem dados suficientes para estimar f diretamente e por isso, uma alternativa, é admitir que f tem uma forma mais simples. Um caso particularmente importante e bastante usado é o dos modelos lineares, em que considera-se f como sendo uma função linear. Neste caso, tem-se um modelo de regressão linear, que é representado como

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon,$$

sendo $\beta_0, \beta_1, \ldots, \beta_k$ parâmetros desconhecidos, vale lembrar que o modelo é linear nos parâmetros e não impreterivelmente nas variáveis X_i, ou seja, o modelo é linear quando a derivada parcial do modelo em interação aos parâmetros não dependem de nenhum dos parâmetros, caso contrário, o modelo é considerado não linear. Por exemplo os modelos:

$$Y = \beta_0 + \beta_1 ln X_1 + \varepsilon \quad \text{e} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

são lineares, enquanto que,

$$Y = \beta_0\beta_1 + X_2^{\beta_2} + \beta_3 X_3 + \varepsilon \quad \text{e} \quad Y = \frac{\beta_1 X_1}{1 + \beta_0 X_1^{\beta_2}},$$

são não lineares.

Na prática, tem-se n observações sobre Y e $X_1, X_2, \ldots, X_k$, então tem-se,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2.2}$$

2.3 Modelo de Regressão linear simples (MRLS)

Para estabelecer o modelo de regressão linear simples deve-se atentar para os seguintes pressupostos; a relação entre X e Y deve ser linear, os valores de X são fixos, isto é, X não é uma variável aleatória, e a média do erro é nula, logo a $E(\varepsilon_i) = 0$, e para um dado valor de X, a variância do erro ε é sempre σ^2 que é denominada variância residual, portanto a $E(\varepsilon_i^2) = \sigma^2$, então pode-se dizer que o erro é homocedástico ou que

tem-se homocedasticia do erro ou da variável dependente, o erro de uma observação é não-correlacionado com o erro em outra observação, isto é, $E(\varepsilon_i \varepsilon_j) = 0$ para $i \neq j$ e por fim os erros têm distribuição normal, $\varepsilon_i \sim N(0, \sigma^2)$, e ainda é preciso verificar se o número de observações disponíveis é superior ao número de parâmetros da equação de regressão. Para ajustar uma regressão linear simples precisa-se ter, no mínimo, três observações, se só dispor de duas, a determinação da reta é um problema de geometria analítica; neste caso não é possível, fazer nenhuma análise estatística (HOFFMANN et al., 2016).

O modelo estatístico para análise de regresão linear simples pode ser definido pela expressão abaixo:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.3}$$

em que,

i) x_i representa cada observação da variável explicativa X;

ii) β_0 é chamado de intercepto ou coeficiente linear da reta, ou seja, o ponto onde a reta corta o eixo Y, quando $x = 0$;

iii) β_1 representa o coeficiente de regressão ou coeficiente angular da reta, ou seja, o grau que a reta faz com o eixo X, e define também o quanto aumenta, ou diminui, o valor de Y em relação a X;

iv) ε_i é o erro associado a cada observação em relação à reta de regressão linear.

De acordo com a Figura 2 é possível fazer uma interpretação geométrica dos parâmetros do modelo, o β_0 representa o ponto em que a reta regressora corta o eixo dos $y's$, quando $x = 0$, já o β_1 corresponde a inclinação da reta regressora e é denominado coeficiente de regressão ou coeficiente angular.

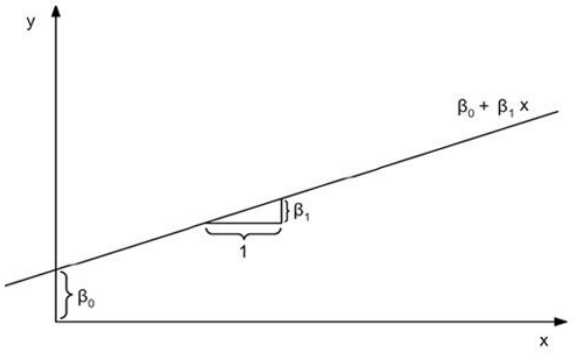

Figura 2 – Reta regressora

2.4 Estimação dos parâmetros do modelo

O foco principal deste trabalho é o uso do modelo de regressão linear múltipla, de modo que não será realizados demonstrações para o modelo de regressão linear simples. Porém, é muito importante apresentar alguns resultados para melhor entendimento de análises realizadas pela regressão linear múltipla.

Conforme Hoffmann et al. (2016) A técnica que consiste em assumir os estimadores que minimizam a soma dos quadrados dos desvios entre valores estimados e valores observados na amostra, é chamado **método dos mínimos quadrados**, este método é uma eficiente estratégia de estimação dos parâmetros da regressão.

Observando a Figura 3 e considerando que a relação linear entre as variáveis Y e X é satisfatória, pode-se, estimar a linha de regressão e solucionar alguns problemas de inferência. O problema de estimar os parâmetros β_0 e β_1 é o mesmo que ajustar a melhor reta em um gráfico de dispersão. Suponha que é traçada uma reta arbitrária $\beta_0 + \beta_1 x$ passando por esses pontos. No valor x_i da variável explicativa, o valor predito por esta reta é $\beta_0 + \beta_1 x_i$, em que o valor observado é Y_i, e os desvios, ou seja os erros associados entre estes dois valores é $\varepsilon_i = Y_i - [\beta_0 + \beta_1 x_i]$, que equivale a distância vertical do ponto à reta arbitrária.

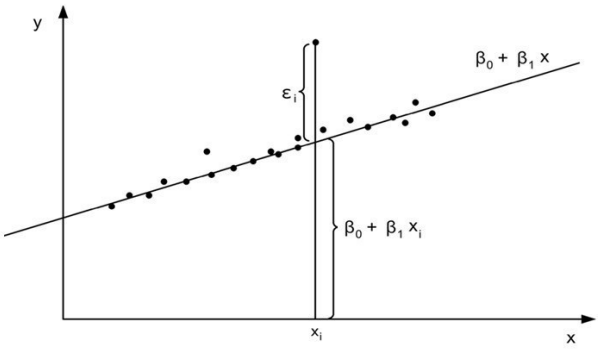

Figura 3 – Reta de Regressão

Fonte: `http://www.portalaction`

Aplicando-se o método de mínimos quadrados obtém-se os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros, que são $\hat{\beta}_0$ e $\hat{\beta}_1$, os estimadores de mínimos quadrados de β_0 e β_1 respectivamente. Assim, obtém-se a reta que melhor explica a relação linear entre as variáveis, que é dada por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, 2, \ldots, n,$$

sendo o intercepto $\hat{\beta}_0$ o valor esperado para a variável dependente y_i quando x_i é igual a zero:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

e o coeficiente angular $\hat{\beta}_1$ é variação esperada na variável resposta, quando a variável independente aumenta uma unidade, e é dado pela seguinte expressão;

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{SS_{xy}}{SS_{xx}}.$$

Outro método para estimação dos parâmetros é o de **máxima verossimilhança**, que de acordo com Hoffmann et al. (2016) esta técnica consiste em adotar como estimadores dos parâmetros, os valores que maximizam a probabilidade no caso da variável aleatória ser discreta ou a densidade de probabilidade no caso da variável ser contínua. Para obter estimadores de máxima verossimilhança é indispensável conhecer ou supor previamente qual é a distribuição da variável em estudo, então determinam-se as estimativas de máxima verossimilhança dos parâmetros β_0 e β_1 do modelo de regressão dado pela equação (2.3). Levando-se em consideração as suposições consideradas para o modelo tem-se que;

$$\varepsilon_i \sim N(0, \sigma^2),$$

por conseguinte, para um certo valor y_i, a função densidade de probabilidade será:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}.$$

Observando-se uma amostra aleatória de n observações (x_i, y_i), sendo os valores de x_i fixos e as observações são independentes, a função de verossimilhança da amostra será, neste caso, definida por

$$L(x_1, \ldots, x_n; \beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\} \tag{2.4}$$

Os estimadores de máxima verossimilhança de β_0,β_1 e σ^2 são aqueles que maximizam a função de verossimilhança $L(\beta_0, \beta_1, \sigma^2 | x_1, \ldots, x_n)$. Uma vez que β_0 e β_1 só aparecem no expoente negativo da equação (2.4), conclui-se que o máximo da função corresponde ao mínimo de:

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Desta forma, as estimativas de máxima verossimilhança dos parâmetros β_0 e β_1 são equivalentes as estimativas de mínimos quadrados, levando em consideração que a distribuição dos erros seja normal. Usando-se o método de máxima verossimilhança, obtém-se também um estimador para a variância dos erros, σ^2, que é

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2.$$

Enfim, estima-se β_0 e β_1 por dois métodos diferentes (mínimos quadrados e de máxima verossimilhança) e em ambos os casos encontra-se os mesmos estimadores para β_0 e para β_1 e ainda encontra-se um possível estimador para $\hat{\sigma}^2$ que é dado pela média dos quadrados residuais (HOFFMANN et al., 2016).

2.5 Coeficiente de correlação simples para uma amostra

De acordo com Demétrio e Zocchi (2006), enquanto na análise de regressão é imprescindível identificar a variável dependente, nos problemas de correlação isto não se faz necessário. Aqui, o que se considera relevante é estudar o grau de ligação entre as variáveis X e Y, isto é, uma medida de covariabilidade entre elas. Assim, análise de correlação difere da análise de regressão em dois pontos básicos:

i) Em primeiro lugar não há a idéia de que uma das variáveis é dependente de uma outra ou de um conjunto de outras variáveis. A correlação é considerada como uma medida de influência mútua ou conjunta entre variáveis, ou seja, não se está preocupado em averiguar quem influencia ou quem é influenciado. A análise de regressão, como vem sendo tratada até aqui, tem por objetivo encontrar equações que indiquem o valor médio de Y para valores fixados de X.

ii) Na análise de correlação todas as variáveis são, em geral, aleatórias e a amostra é considerada proveniente de uma distribuição conjunta dessas variáveis. Evidentemente, a distribuição normal bidimensional é, eventualmente, suposta para a variável bidimensional (X, Y).

2.6 Modelo de Regressão Linear Múltipla (MRLM)

Visto que esse trabalho esta envolvendo mais de uma variável, nosso estudo será concentrado no modelo de regressão linear múltipla, pois esse método nós dá condição de trabalhar com várias variáveis explicativas conjuntamente.

Segundo Hoffmann et al. (2016) temos uma regressão linear múltipla quando admitimos que o valor da variável dependente é função linear de duas ou mais variáveis independentes (explanatórias).

De acordo com Demétrio e Zocchi (2006), tem-se uma regressão linear múltipla quando se certifica que a variável resposta Y é função de duas ou mais variáveis regressoras.

A diferença entre a regressão linear múltipla e a regressão linear simples é que na múltipla são consideradas duas ou mais variáveis explicativas (independentes). As variáveis independentes são as ditas variáveis explicativas, uma vez explicam a variação de

y, na regressão linear múltipla assume-se que há uma relação linear entre uma variável dependente e entre as variáveis independentes (preditoras) (RODRIGUES, 2012).

O modelo estatístico de uma regressão linear múltipla com k variáveis regressoras (X_1, X_2, X_k), é representada por

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, 2, \ldots, n, \tag{2.5}$$

sendo, β_0, β_1,..., β_k, parâmetros desconhecidos; e ε_i é o erro amostral, e os erros amostrais são independentes com distribuição $N(0, \sigma^2)$, $i = 1, 2, \ldots, n$. E ainda o modelo de regressão linear múltipla pode ser representado pela seguinte expressão:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i, \tag{2.6}$$

em que, y_i é a variação dos valores das variável resposta x_k, ε_i são os erros associados ao modelo, k é o número de variáveis explicativas para o modelo e n é o tamanho da amostra.

Utilizando a notação matricial, o modelo de regressão linear múltipla fica da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.7}$$

sendo,

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad e \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Segundo Demétrio e Zocchi (2006), $\boldsymbol{Y}$ é um vetor, de dimensão $n \times 1$, da variável aleatória Y; $\boldsymbol{X}$ é a matriz, de dimensões, $n \times p$, conhecida do delineamento, e como sempre ocorre em modelos de regressão, a menos que multicolinea, é de posto completo $p = K + 1$, sendo $p = k + 1$ o número de parâmetros; $\boldsymbol{\beta}$ é o vetor, de dimensão $p \times 1$, de parâmetros desconhecidos, $\boldsymbol{\varepsilon}$ é o vetor, de dimensão $n \times 1$ e de variáveis aleatórias não observáveis.

2.7 Pressuposições para o modelo

De maneira similar ao que foi tratado em regressão linear simples, têm-se as seguintes suposições:

i) A variável dependente Y é função linear das variáveis explicativas X_j, $j = 1, 2, \ldots, k$;

ii) As variáveis explicativas X_j são fixas;

iii) $E(\varepsilon_i) = 0$, ou seja, $E(\varepsilon) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensões $n \times 1$;

iv) Os erros são homocedásticos, isto é, $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$;

v) Os erros são independentes, isto é, $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$;

vi) Os erros têm distribuição normal.

De acordo com Hoffmann et al. (2016), da mesma forma que na regressão linear simples, as pressuposições (i), (ii) e (iii) são fundamentais para demonstrar que os estimadores de mínimos quadrados são não-tendenciosos e as cinco primeiras pressuposições possibilitam demonstrar que tais estimadores são lineares não-tendenciosos de variância mínima. E a pressuposição (vi) é essencial para realizar testes de hipótese e para construir intervalos de confiança para os parâmetros.

2.8 Estimação dos parâmetros do modelo

Conforme Demétrio e Zocchi (2006), o número de parâmetros a serem estimados é $p = k+1$. Se há apenas p observações, a estimação dos parâmetros reduz-se a um problema matématico de resolução de um sistema de p equações a p incógnitas, não sendo possível fazer qualquer análise estatística. deve-se, portanto, ter $n > p$. O método utilizado para a estimar $p(p < n)$ parâmetros é utilizando o método dos quadrados mínimos.

Sejam β o vetor de parâmetros e ε o vetor dos erros.

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

Temos que:

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$$

A soma dos quadrados dos erros é dada por:

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 \cdots \boldsymbol{\varepsilon}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix} = \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2 \boldsymbol{\varepsilon}_2 + \cdots + \boldsymbol{\varepsilon}_n = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^2$$

A soma de quadrados ainda pode ser escrita da seguinte forma:

$$\begin{aligned} Z &= \boldsymbol{\varepsilon'\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= (\boldsymbol{Y}' - \beta'\boldsymbol{X}')(\boldsymbol{Y} - X\boldsymbol{\beta}) \\ &= \boldsymbol{Y'Y} - \boldsymbol{Y'X\beta} - \beta'\boldsymbol{X'Y} + \beta'\boldsymbol{X'X\beta} \end{aligned}$$

em que as matrizes, $\boldsymbol{Y'X\beta} = \beta'\boldsymbol{X'Y}$, são equivalentes, por uma ser a transposta da outra, e possuem um único elemento, logo

$$Z = \boldsymbol{Y'Y} - 2\beta'\boldsymbol{X'Y} + \beta'\boldsymbol{X'X\beta}. \tag{2.8}$$

O ponto de mínimo da função Z para os valores de $\boldsymbol{\beta}$, é aquele que torna a equação distinto de Z

$$\frac{\boldsymbol{\partial Z}}{\boldsymbol{\partial\beta}} = \boldsymbol{\partial(Y'Y - 2\beta'X'Y + X'X\beta) = 0}$$

$$\boldsymbol{\partial(Y'Y) - 2\partial(\beta'X')Y + (\partial\beta'X')X\beta = 0}$$

$$\boldsymbol{-2(\partial\beta')X'Y + (\partial\beta')X'X\beta + \beta'X'X(\partial\beta) = 0}$$

como, $\boldsymbol{(\partial\beta')X'X\beta = \beta'X'X(\partial\beta)}$, são matrizes simétricas com um único elemento, logo pode-se reecrever a equação acima de maneira que $\boldsymbol{-2(\partial\beta')X'Y + 2(\partial\beta')X'X\beta = 0}$, e também:

$$\boldsymbol{(\partial\beta'(X'X\beta - X'Y) = 0)},$$

para que $\boldsymbol{\partial\beta' = 0}$ é necessário que

$$\boldsymbol{X'X\beta = X'Y}, \tag{2.9}$$

de onde se pode concluir que, a partir da matriz inversa de $\boldsymbol{X'X}$, pode-se chegar ao estimador dos parâmetros β, pré multiplicando as duas por, $\boldsymbol{(X'X)^{-1}}$ logo é dado por

$$\boldsymbol{\widehat{\beta} = (X'X)^{-1}X'Y}$$

.

Conforme Hoffmann et al. (2016), a primeira etapa dos cálculos para obtenção das estimativas dos parâmetros é a construção das matrizes

$$\boldsymbol{X'X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{1j} & \sum_{i=1}^{n} X_{2j} & \cdots & \sum_{i=1}^{n} X_{kj} \\ \sum_{i=1}^{n} X_{1j} & \sum_{i=1}^{n} X_{1j}^2 & \sum_{i=1}^{n} X_{1j}X_{2j} & \cdots & \sum_{i=1}^{n} X_{1j}X_{kj} \\ \sum_{i=1}^{n} X_{2j} & \sum_{i=1}^{n} X_{1j}X_{2j} & \sum_{i=1}^{n} X_{2j}^2 & \cdots & \sum_{i=1}^{n} X_{2j}X_{kj} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{kj} & \sum_{i=1}^{n} X_{1j}X_{kj} & \sum_{i=1}^{n} X_{2j}X_{kj} & \cdots & \sum_{i=1}^{n} X_{Kj}^2 \end{bmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad e \quad \boldsymbol{X'Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_j \\ \sum_{i=1}^{n} X_{1j}Y_j \\ \sum_{i=1}^{n} X_{2j}Y_j \\ \vdots \\ \sum_{i=1}^{n} X_{kj}Y_j \end{bmatrix}.$$

Na condição em que a matriz $\boldsymbol{X'X}$ for não singular, e deste modo invertível, conclui-se que o estimador para o vetor de parâmetros $\widehat{\beta}$ é definido por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

De acordo com Hoffmann et al. (2016), Essas matrizes são necessárias em várias outras fases da análise de regressão linear múltipla. Do sistema de equações normais pode-se obter outros resultados de interesse conforme pode ser observado:

$$\begin{aligned} \boldsymbol{X'Y} - \boldsymbol{X'X\beta} &= 0 \\ \boldsymbol{X'}(\boldsymbol{Y} - \boldsymbol{X\beta}) &= 0 \\ \boldsymbol{X'\varepsilon} &= 0. \end{aligned}$$

Por essa ligação matricial pode-se dizer que:

$$\sum_{i=1}^{n} \varepsilon_i = 0.$$

Tornando-se nula a soma dos desvios, conclui-se que:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$

2.9 Análise de variância da regressão linear múltipla

A análise de variância, baseia-se na decomposição da soma dos quadrados total, SQT, que equivale à variação da variável resposta, já a soma dos quadrados explicada ou regressão SQR, corresponde à variação da variável resposta que é explicada pelo modelo e a soma dos quadrados dos resíduos, SQE, refere-se à variação da variável resposta que não é explicada pelo modelo (RODRIGUES, 2012).

i) Soma de quadrados total (SQ_{Total})

A soma de quadrado total, mede a variação total das observações em torno da média. Tem-se a expressão

$$SQ_{Total} = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = \boldsymbol{Y'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

ii) A soma de quadrados da regressão (SQ_{Reg})

A soma de quadrado de regressão, mede a quantidade de variação da variável dependente explicada pela equação de regressão linear múltipla. Então a expressão é definida por

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}})^2 = \sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n}$$

$$= (\boldsymbol{X}\hat{\beta})'\boldsymbol{X}\hat{\beta} - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta} - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{Y} - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n}.$$

De acordo com a Equação (2.9), então:

$$SQ_{Reg} = \hat{\beta}'\boldsymbol{X}'\boldsymbol{Y} - \frac{\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2}{n},$$

ou ainda a soma de quadrados da regressão pode ser dada pela expressão:

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

iii) A soma de quadrados da resíduos (SQ_{Res})

Para calcular soma de quadrados dos desvios, ou soma de quadrados residual, é necessário relembrar a Equação (2.8), em que

$$\begin{aligned} \boldsymbol{\varepsilon'\varepsilon} &= \boldsymbol{Y'Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta'}\underbrace{\boldsymbol{X'X\beta}}_{\boldsymbol{X'Y}} \\ &= \boldsymbol{Y'Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta X'Y} \\ SQ_{Res} &= \boldsymbol{Y'Y} - \boldsymbol{\beta'X'Y}. \end{aligned}$$

Sendo assim, após alguns cálculos relativamente simples, pode-se mostrar que a soma de quadrados de resíduos pode ser escrita como:

$$SQ_{Res} = SQ_{tot} - SQ_{Reg}$$

A partir dos resultados obtidos, pode-se obter o esquema do quadro da análise da variância.

Tabela 1 – Análise de variância para o modelo de Regressão linear

Fonte de variação	**GL**	**SQ**	**QM**	**F**
Regressão	$K = p - 1$	$SQReg$	$\frac{SQReg}{p-1}$	$\frac{QMReg}{QMRes}$
Resíduo	$n - p$	$SQRes$	$\frac{SQRes}{n-p}$	-
Total	$n - 1$	$SQTotal$	-	-

2.10 Testes de hipóteses e intervalo de confiança

Na análise de regressão múltipla, o teste F compõe um teste mais geral. Por meio de sua utilização determina-se se qualquer das variáveis independentes no modelo tem poder de explicação. Cada variável pode então ser testada individualmente com o teste t para determinar se é uma das variáveis significativas (RODRIGUES, 2012).

i) Teste F para significância do modelo de regressão linear múltipla

O teste F é utilizado para verificar se as variáveis independentes conjuntamente contribuem de maneira significativa para explicar a variação da variável dependente, ou seja, se realmente existe a regressão. Definindo-se às hipóteses

$$\begin{cases} H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \ \beta_j \neq 0 \text{ para pelo menos um} \quad \text{j = 1,2,...,k} \end{cases}$$

Nesse caso testa-se a existência da regressão linear no modelo, fazendo-se uso da estatística F citada na Tabela 1 que é:

$$F = \frac{QMReg}{QMRes} \sim F_{(k,n-p)},$$

em que, k é o número de variáveis independentes e $p = k + 1$. Então, após encontrar o valor F calculado, o F tabelado, e atribuir o nível α de significância, pode-se decidir que:

Se a estatística $F_{calculado} > F_{tabelado}$, rejeita-se a hipótese H_0 e conclui-se ao nível α de significância que há regressão linear entre as variáveis do modelo.

ii) Teste de significância para os coeficientes de regressão (Teste t-student)

As vezes torna-se necessário testar as hipóteses acerca dos coeficientes de regressão, para determinar o potencial de cada regressor no modelo. Logo a quantidade a ser testada para cada β_i será dada pela fórmula.

$$T = \frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)} \sim t_{n-p} \qquad j = 1, 2, \ldots, k, \tag{2.10}$$

em que $S(\hat{\beta}_j)$ é o estimador da variância de β_j, e T_i segue distribuição t de Student com $n - p$ graus de liberdade e que é usado para verificar as seguintes hipóteses

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0, \quad \text{j = 1,2,...,k.} \end{cases}$$

Portanto, se $|T_{calculado}| > T_{tabelado}$, então rejeita-se H_0 e conclui-se que, ao nível α de significância, $\beta_j \neq 0, j = 1, 2, \ldots, n$ e assim pode-se afirmar que esta variável é importante para explicar a regressão linear. Caso contrário, esta variável não tem influência na regressão linear múltipla.

Intervalo de confiança para os coeficientes de regressão

Outra forma de se avaliar a significância dos parâmetros do modelo é por meio da construção de intervalos de confiança. Podendo-se obter um intervalo de confiança que inclua o verdadeiro valor do parâmetro β_j com $j = 1, 2, \cdots, k$, a um certo nível de significância α, que se queira.

Considerando-se a estatística de teste dada em (2.10), um intervalo com $100(1-\alpha)\%$ de confiança para o coeficiente da regressão $\beta_j, j = 1, 2, \ldots, k$, é definido por

$$IC(\beta_j) = \left[\ \hat{\beta}_j - t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\hat{\beta}_j)}; \hat{\beta}_j + t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\hat{\beta}_j)} \ \right].$$

2.11 Coeficiente de Determinação Múltiplo

Conforme Demétrio e Zocchi (2006), o coeficiente de determinação é utilizado como uma medida descritiva da qualidade do ajuste atingido, e indica a proporção da variação de Y que é explicada pela regressão, por conseguinte esse valor do coeficiente de determinação deve ser usado com cuidado, pois depende do número de observações da amostra, tendendo a crescer quando a amostra diminui, além do que, é sempre possível torná-lo maior, pela adição de um número suficiente de termos.

Quando há valores repetidos, R^2 nunca será igual a 1, pois o modelo não será eficaz para explicar a variabilidade devido ao erro puro. Embora o R^2 aumente se acrescentar uma nova variável ao modelo, isto não significa impreterivelmente que o novo modelo seja superior ao anterior, a menos que a soma de quadrados residual do novo modelo

seja reduzida de uma quantia igual ao quadrado médio residual original, porque o novo modelo terá um quadrado médio residual maior devido à perda de um grau de liberdade, na verdade esse novo modelo poderá ser pior do que o anterior. Para obtenção deste valor do coeficiente de determinação do modelo de regressão linear múltipla utiliza-se a seguinte expressão:

$$R^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQRes}{SQTotal}, \quad com \quad 0 \leq R^2 \leq 1 \tag{2.11}$$

como o (R^2) não obtém um valor esperado de Y confiável, então é preferível que utilize o coeficiente de determinação ajustado (R^2_{ajust}) para graus de liberdade, representado por

$$1 - R^2_{ajust} = 1 - \frac{\frac{SQ_{Res}}{n-p}}{\frac{SQ_{Tot}}{n-1}} = \frac{n-1}{n-p}(1 - R^2),$$

ou melhor

$$R^2_{ajust} = R^2 - \frac{p-1}{n-p}(1 - R^2), \tag{2.12}$$

em que, n é o tamanho da amostra e p é o número de parâmetros. O R^2_{ajust} dá uma idéia da proporção da variação de Y explicada pelo método de regressão uma vez que leva em conta o número de regressores. Ao contrário do que acontecia no coeficiente não ajustado, o ajustado nem sempre aumenta ao passo que é acrescentada uma nova variável ao modelo.

i) Critério de informação de Akaike (AIC)

De acordo com Paula (2013) O método sugerido por Akaike é um processo de minimização que não envolve testes estatísticos. A ideia primordial é selecionar um modelo que seja parcimonioso, ou em outras palavras, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança $L(\beta)$ cresce com o aumento do número de parâmetros do modelo, uma proposta admissível seria encontrar o modelo com menor valor para a função,

$$AIC = -L(\hat{\beta}) + p, \tag{2.13}$$

em que, p indica o número de parâmetros. No caso do modelo normal linear, pode-se mostrar que AIC fica expresso, quando σ^2 é desconhecido, na forma

$$AIC = nlog\{D(\boldsymbol{y}; \hat{\mu})/n\} + 2p,$$

em que $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

ii) Critério de informação de Bayes (BIC)

$$BIC = -2logL + logn$$

sendo L a função de verossimilhança e n o número de parâmetros ajustados. O critério BIC penaliza mais fortemente modelos com um maior número de parâmetros do que o AIC tendendo, desta forma, a selecionar modelos com um menor número de parâmetros (DEMÉTRIO; ZOCCHI, 2006).

2.12 Análise de resíduos no MRLM

Conforme Rodrigues (2012) os resíduos são dados pela diferença entre os valores da variável resposta observada e a variável resposta estimada, isto é, ao realizarmos uma análise de resíduos deseja verificar se o modelo de regressão que está a ser utilizado é satisfatório. No caso da regressão linear múltipla, é indispensável ainda investigar se há colinearidade ou multicolinearidade entre as variáveis explicativas.

De acordo com Demétrio e Zocchi (2006) se as suposições são violadas, têm-se as falhas sistemáticas que é o caso, não linearidade, não-normalidade, heterocedasticidade, não-independência dos erros, e efeito cumulativo de fatores que não foram considerados no modelo e o resultado da análise poder induzir a conclusões duvidosas, outro fator comum é a existência de pontos atípicos, que podem influenciar ou não no ajuste do modelo. E isso pode acontecer devido a:

i) erros grosseiros na variável resposta ou nas variáveis explanatórias;
ii) modelo mal especificado (falta de uma ou mais variáveis, modelo inadequado);
iii) escala errada, talvez os dados sejam melhor descritos após uma transformação do tipo logarítmica ou raiz quadrada;
iv) distribuição da variável resposta errada, por exemplo, tem uma cauda mais longa do que a distribuição normal.

2.13 Diagnóstico de normalidade

A normalidade dos resíduos pode ser conferida quer através de gráficos, quer empregando alguns testes, especificamente através do

i) gráfico Q-Q plot dos resíduos;
ii) histograma dos resíduos padronizados;
iii) teste de Shapiro-Wilk.

i. **Gráfico Q-Q plot dos resíduos**;

Neste gráfico, é possivel visualizar a distribuição de probabilidades dos valores observados com os valores esperados, representada por uma diagonal, segundo uma distribuição normal.

Caso a normalidade se verifique, as observações registadas aproximam-se dessa diagonal, sem nenhum afastamento significativo.

ii. **Histograma dos resíduos padronizados**

Também se pode fazer um histograma dos resíduos no qual se procuram afastamentos evidentes em relação à forma simétrica e unimodal da distribuição normal. Este gráfico apenas deverá ser utilizado em amostras de dimensão elevada, já que quando se trabalha com amostras de dimensão reduzida o histograma não é muito conclusivo.

iii. **Teste de Shapiro-Wilk para normalidade dos resíduos**

De acordo com Demétrio e Zocchi (2006) a partir dos resultados do teste de Shapiro-Wilk é possível concluir que há evidências para dizer que a distribuição residual seja normal, para isso o nível de significância deve ser menor que o valor-p que é uma probalidade do teste, logo é viável testar as hipóteses.

$$\begin{cases} H_0: & \text{A amostra provém de uma população Normal;} \\ H_1: & \text{A amostra não provém de uma população Normal} \end{cases}$$

A estatística do teste de Shapiro-Wilk (1965), é representado pela seguinte expressão

$$W = \frac{\left[\sum_{i=1}^{n} \alpha_i X_{(i)}\right]^2}{(n-1)S^2},$$

em que, α_i são constantes geradas a partir da média, variância e covariância de n ordens.

2.14 Diagnóstico de Homoscedasticidade

Um dos pressupostos do modelo de regressão linear é a de que os erros devem ter variância constante, esta situação é caracterizada por homoscedasticidade. A variância ser constante consiste em pressupor que não existem observações incluídas na variável residual cuja influência seja mais intensa na variável dependente.

Um dos métodos para investigar a hipótese de que os resíduos são homoscedásticos, é a análise do gráfico dos resíduos versus valores ajustados. Este gráfico deve exibir pontos dispostos aleatoriamente sem nenhum padrão exato, como se pode ver, por exemplo na Figura 4.

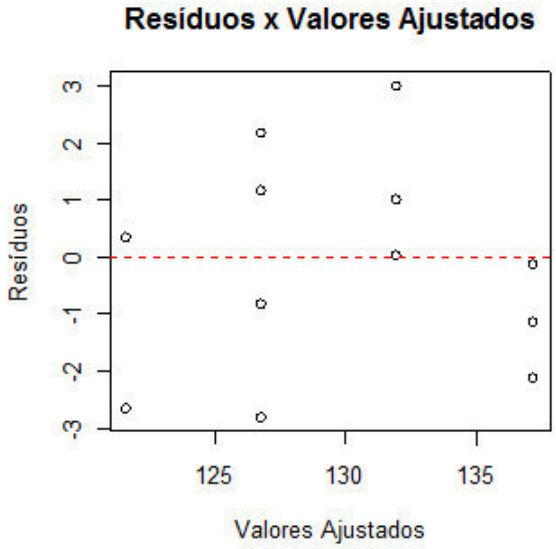

Figura 4 – Confirmação da homoscedasticidade dos resíduos

Fonte: `http://www.portalaction`

Por consequência, se os pontos estão aleatoriamente distribuídos em torno da reta $y = 0$, sem nenhum comportamento ou tendência, temos indicações de que a variância dos resíduos é constante. Já a presença, por exemplo, de "funil"é um indício da existência de heteroscedasticidade (RODRIGUES, 2012).

2.15 Diagnóstico de outliers e observações influentes

Os *outliers* podem ser classificados em severos ou moderados conforme o seu afastamento em referência às restantes observações. Os *outliers* moderados encontram-se fora do intervalo $[Q_1 - 1,5Q; Q_3 + 1,5Q]$ e os *outliers* severos encontram-se fora do intervalo $[Q_1 - 3Q; Q_3 + 3Q]$, em que Q_1 representa o 1º quartil dos dados e Q a amplitude interquartil, isto é, é a diferença entre o 3º e o 1º quartil, $Q = Q_3 - Q_1$.

Se um *outlier* for influente vai interferir sobre a função de regressão ajustada o que significa que a inclusão ou não desse ponto modifica consideravelmente os valores ajustados. Assim, um ponto é influente se a sua exclusão na regressão ajustada acarreta uma alteração considerável nos valores ajustados (RODRIGUES, 2012).

3 Aplicação

Nesta seção encontra-se as principais técnicas que serviram de subsídio para este trabalho, tanto na parte da descrição dos modelos quanto nas inferências.

3.1 Um breve comentário sobre a hipertensão

A hipertensão arterial é considerada um sério problema de saúde pública por sua gravidade e dificuldade no seu controle. É também reconhecida como um dos mais importantes fatores de risco para o desenvolvimento do acidente vascular cerebral e doença renal.

A hipertensão arterial ou pressão alta é uma condição clínica multifatorial caracterizada por níveis elevados e sustentados da pressão arterial, considerando-se valores de pressão arterial maiores ou iguais a 140/90mmHg. A hipertensão normalmente é causada quando há uma resistência e endurecimento maior dos vasos sanguíneos para a passagem do sangue, o que necessita uma força maior do coração para o bombeamento do sangue, isto pode ser um processo natural do corpo, além da questão hereditária há outros fatores de risco controláveis que podem influenciar os níveis de pressão arterial como; o sedentarismo, obesidade, tabagismo.

3.2 Dados utilizados

Os dados utilizados nesse estudo foram fornecidos pelo Ministério da Saúde, no *site* tecnologia da informação a serviço do SUS, onde foi obtido do sistema de cadastramento e acompanhamento de hipertensos no período de abril de 2010 à abril de 2013, que foi o último mês e ano que estas informações foram registradas. Para este estudo foi coletada uma amostra de quarenta municípios do Estado da Paraíba e notificados 12.851 indivíduos com problema de hipertensão. Estas cidades foram selecionadas pelas mesorregiões que por sua vez estão separadas em 23 microrregiões geográficas.

O DATASUS disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, e tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde. A mensuração do estado de saúde da população é uma tradição em saúde pública. Teve seu início com o registro sistemático de dados de mortalidade e de sobrevivência (Estatísticas Vitais). Os dados teste trabalho encontram-se descritos de forma reduzida na tabela 2.

As variáveis utilizadas no presente estudo foram as seguintes:

i) **Hpn**: hipertensos notificados

ii) **Sed**: sedentarismo

iii) **Sobr**: sobrepeso

iv) **Tab**: tabagismo

vi) **Avc**: acidente vascular cerebral

vii) **D.renal**: doença renal

Tabela 2 – Dados de hipertensão em uma amostra de 40 municípios do Estado da Paraíba no período de abril 2010 à abril de 2013

Cidade	**Hpn**	**Sed**	**Sobr**	**Tab**	**Avc**	**D.renal**
Alagoa Grande	624	181	156	133	46	12
Alagoa Nova	642	193	179	140	70	20
Alcantil	180	62	90	15	10	8
Areia	451	169	169	92	32	4
Aroeiras	264	99	57	55	21	6
Barra de Santana	231	60	35	53	26	7
Belém	108	21	27	4	2	1
Boqueirão	48	32	17	17	1	2
Brejo do Cruz	360	116	89	64	18	14
Caaporã	545	149	163	91	20	19
Cabaceiras	259	118	43	31	8	5
Cabedelo	124	35	65	36	17	1
Cajazeiras	514	237	192	142	43	24
Conde	236	164	138	45	11	3
Cruz do Espírito Santo	405	287	214	73	19	1
Esperança	484	148	121	98	28	9
Ingá	273	100	69	34	0	1
Jericó	181	105	72	72	30	20
Juarez Távora	112	30	17	13	1	1
Junco do Seridó	290	107	118	82	8	14
Lagoa de Dentro	118	61	66	18	8	2
Lagoa Seca	517	256	142	92	22	11
Lucena	246	97	78	54	28	2
Mamanguape	981	383	294	265	102	32
Manaíra	126	103	92	23	7	12
Massaranduba	99	42	47	8	3	1
Montadas	140	13	10	3	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Teixeira	88	26	24	11	8	6

3.3 Análise Descritiva das variáveis

Apresenta-se inicialmente uma análise descrita dos dados, na Figura 5 têm-se o boxplot para a variável hipertensos notificados, em quarenta munícipios do estado da Paraíba.

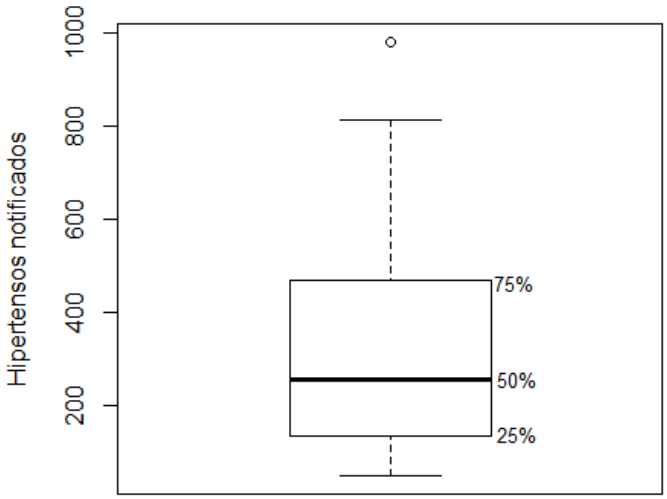

Figura 5 – Análise descritiva da variável hipertensos notificados

Como é possível observar na Figura 5, que o menor elemento da amostra ou o valor mínimo esta abaixo de 200 indivíduos notificados e o valor máximo esta acima de 900 indivíduos, com o primeiro quartil que deixa 25% das observações abaixo e 75% acima, enquanto que o terceiro quartil deixa 75% das observações abaixo e 25% acima e estes valores equivalem a 139,5 e 459,2 respectivamente, já o segundo quartil ou a mediana deixa 50% das observações abaixo e 50% acima e este valor esta em torno de 255,5 conforme a figura.

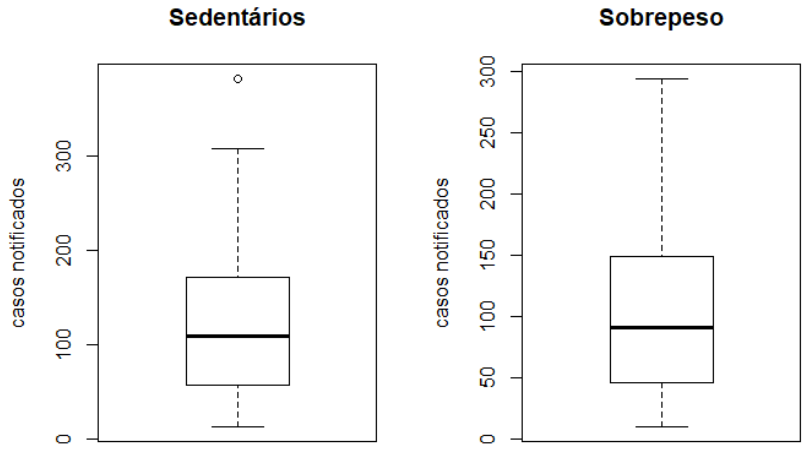

Figura 6 – Análise descritiva das variáveis de casos notificados de sedentarismo e sobrepeso

De acordo com a Figura 6, tem-se que,

Sedentários notificados: Para esta variável em estudo o menor elemento da amostra é 13, e o maior é 383, o primeiro quartil é 58,5 e a mediana equivale a 109,5 e o terceiro quartil é 170,8 e é possível que o valor máximo seja considerado como sendo um provável *outlier*.

Sobrepesos notificados: Com relação a esta variável em estudo o menor elemento da amostra é 10, e o maior é 294, o primeiro quartil é 46,75 e a mediana corresponde a 91 e o terceiro quartil é 145,5 não é observado a presença de pontos discrepantes.

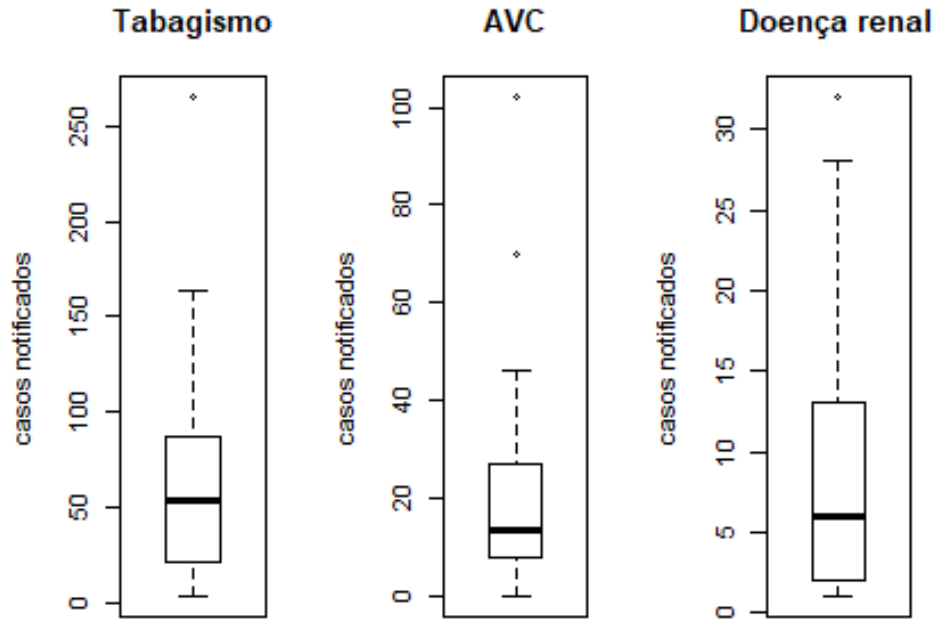

Figura 7 – Análise descritiva das variáveis de casos notificados de tabagismo, acidente vascular cerebral e doença renal

De acordo com a Figura 7, tem-se que

Tabagismo: No tocante a esta variável em estudo o menor elemento da amostra é 3, e o maior é 265, o primeiro quartil é 21,75 e a mediana corresponde a 63,05 e o terceiro quartil é 84,25 é observado a presença de pontos *outlier*.

Acidente vascular cerebral: Em relação a variável em estudo o menor elemento da amostra é 0, e o maior é 102, o primeiro quartil é 8 e a mediana equivale a 13,5 e o terceiro quartil é 26,5 é observado a presença de pontos *outlier*.

Doença renal: Com relação a esta variável em estudo o menor elemento da amostra é 1 e o maior é 32, o primeiro quartil é 2 e a mediana corresponde a 6 e o terceiro quartil é 12,5 é observado a presença de pontos *outlier*.

Com cerca de 12.851 pessoas sofrendo de hipertensão, número preocupante para sistema único de saúde (SUS), conforme nos mostra os dados que são descritos na Tabela 3, que faz referência a proporção de indivíduos hipertensos em quarenta municípios do Estado da Paraíba no período de abril de 2010 à abril de 2013 e como pode-se observar há um grande índice de pessoas sofrendo de hipertensão na cidade de Mamanguape, com cerca de $7,63\%$, em seguida respectivamente vem os seguintes municípios, de Rio Tinto com $6,32\%$, Alagoa Nova com $5,00\%$, Patos com $4,92\%$, Alagoa Grande com $4,86\%$, São Bento

com $4,61\%$, Caaporã com $4,24\%$, Lagoa Seca com $4,02\%$, Cajazeiras com $4,00\%$, em contrapartida foi observado que os menores índices se concentram nos seguintes municípios; Boqueirão com $0,37\%$, Texeira com $0,68\%$, Massaranduba com $0,77\%$, Tavares com $0,78\%$, Belém com $0,84\%$, Lagoa de Dentro com $0,92\%$, Monteiro com $0,92\%$, Manaíra com $0,98\%$, Montadas com $1,09\%$.

Tabela 3 – Proporção de hipertensos em 40 municípios do Estado da Paraíba entre abril 2010 à abril de 2013

Cidade	**Hipertenso**	**%**	**Cidade**	**Hipertenso**	**%**
Alagoa Grande	624	4.86	Lagoa de Dentro	118	0.92
Alagoa Nova	642	5.00	Lagoa Seca	517	4.02
Alcantil	180	1.40	Lucena	246	1.91
Areia	451	3.51	Mamanguape	981	7.63
Aroeiras	264	2.05	Manaira	126	0.98
Barra de Santana	231	1.80	Massaranduba	99	0.77
Belém	108	0.84	Montadas	140	1.09
Boqueirão	48	0.37	Monteiro	118	0.92
Brejo do Cruz	360	2.80	Patos	632	4.92
Caaporã	545	4.24	Pocinhos	147	1.14
Cabaceiras	259	2.02	Queimadas	359	2.79
Cabedelo	124	0.96	Rio Tinto	812	6.32
Cajazeiras	514	4.00	Santa luzia	436	3.39
Conde	236	1.84	São Bento	592	4.61
Cruz do Espírito Santo	405	3.15	Sobrado	175	1.36
Esperança	485	3.77	Soledade	252	1.96
Ingá	273	2.12	Sousa	181	1.41
Jericó	181	1.41	Taperoá	401	3.12
Juarez Távora	112	0.87	Tavares	100	0.78
Junco do Seridó	290	2.26	Teixeira	88	0.68

3.4 Análise de regressão

Dando sequência a análise de regressão linear múltipla entre as variáveis do banco de dados, é necessário definir a variável de interesse ou variável dependente, de acordo com o objetivo do presente trabalho, a variável definida é o números de hipertensos em quarenta municípios selecionados no Estado da Paraíba, deseja-se saber quais variáveis independentes contribui significamente para o aumento da hipertensão nesses municípios.

Propõe-se para o modelo de regressão linear múltipla desenvolvido nesse trabalho cinco variáveis regressoras envolvidas:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 x_i + \beta_4 x_i + \beta_5 x_i + \varepsilon_i, \quad i = 1, \cdots, 40; \tag{3.1}$$

em que, foi considerado números de hipertensos notificados como a variável resposta (y) e sedentários, sobrepeso, tabagismo, acidente vascular cerebral, doença renal, como sendo as variáveis regressoras, representadas por: x_1, x_2, x_3, x_4, x_5, e seus respectivos coeficientes.

De acordo com Demétrio e Zocchi (2006), antes de se iniciar qualquer análise de regressão de um conjunto de dados, é muito importante se fazer os gráficos de dispersão para que se tenha noção a respeito do tipo de relação existente entre as variáveis, e da variabilidade e associação entre elas e da presença de pontos atípicos. No entanto esses gráficos devem ser olhados com certo cuidado se há duas ou mais variáveis explicativas, pois essas variáveis não levam em consideração a correlação existente entre elas.

Para isto será exibido nessa seção os diagramas de dispersão da variável dependente já definida em relação as variáveis independentes proposta nesse estudo.

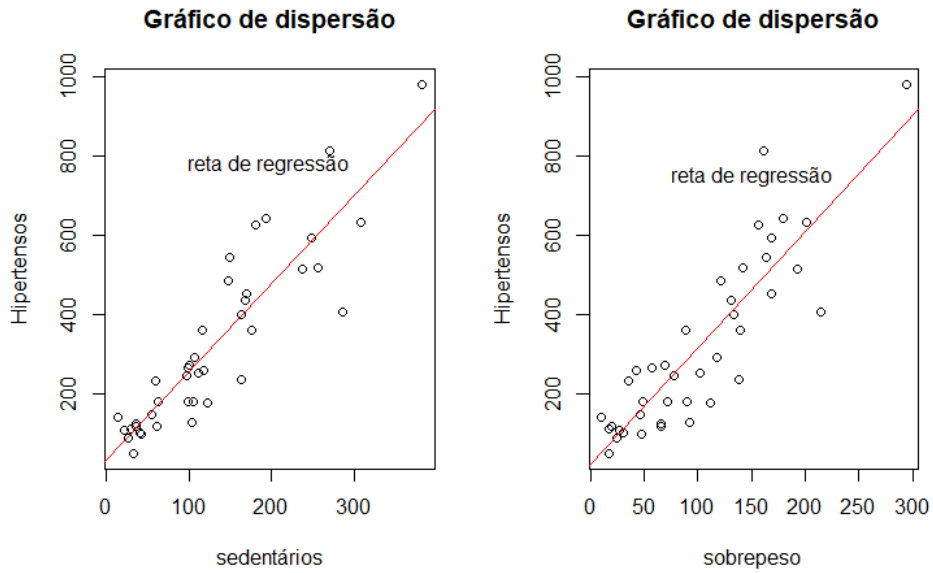

Figura 8 – Diagramas de dispersão da variável hipertensos vs sedentários vs sobrepeso

A Figura 8 mostra que há uma relação linear entre as variáveis Y e X, foi feito uma análise de regressão simples para cada caso e constatado que a relação entre as variáveis hipertensos e sedentários é significativa, tendo em vista que o p-valor do teste t do modelo é inferior ao nível de 5% de probabilidade, com o coeficiente de determinação ajustado $R^2 = 0,81$. E para verificar o comportamento dos resíduos foi aplicado o teste de Shapiro-Wilk e obteve uma estatística $W = 0{,}96$ com valor $p = 0,18 > \alpha = 0,05$ de significância ou seja os resíduos seguem uma distribuição normal. Para a relação das variáveis hipertensos e sobrepeso o p-valor do teste t do modelo é menor que o nível de 5% de probabilidade, com o coeficiente de determinação ajustado $R^2 = 0,77$. E com a utilização do teste de Shapiro-Wilk obteve-se uma estatística de $W = 0{,}97$ com valor $p = 0,31 > \alpha = 0,05$ de significância logo este modelo, foi estatisticamente significativo.

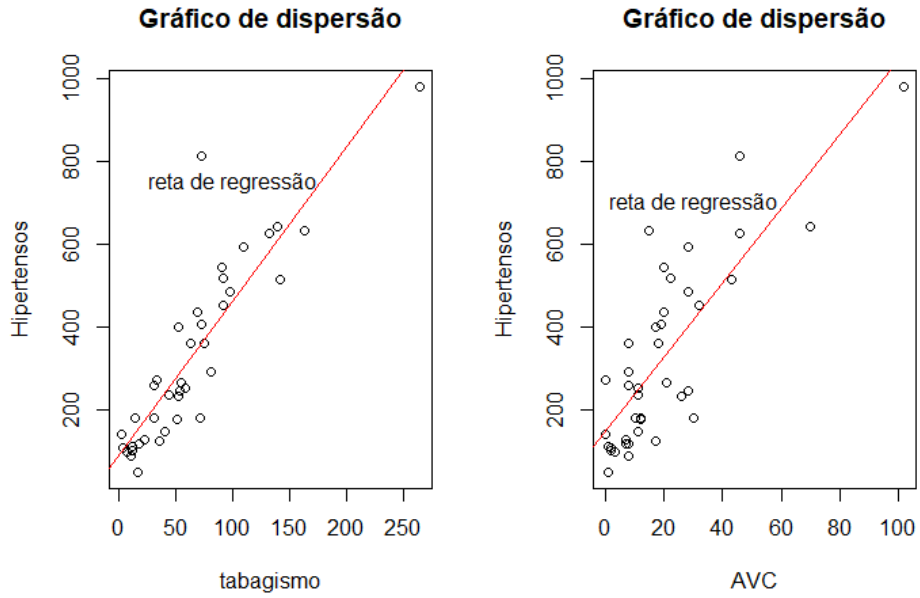

Figura 9 – Diagramas de dispersão da variável hipertensos vs tabagismo vs AVC

A Figura 9 mostra que há uma relação linear entre as variáveis Y e X, foi feito uma análise de regressão simples para verificar se a relação entre as variáveis hipertensos e tabagismo são significativas, logo constatou que o p-valor do teste t do modelo é inferior ao nível de 5% de probabilidade, com o coeficiente de determinação ajustado $R^2 = 0,79$, e para verificar o comportamento dos resíduos foi aplicado o teste de Shapiro-Wilk e obteve uma estatística $W = 0{,}82$ com valor $p = 0,0000 < \alpha = 0,05$ de significância, ou seja os resíduos não seguem uma distribuição normal, para tanto a relação destas variáveis acima não é estatisticamente significativa pela quebra da pressuposição de normalidade dos resíduos. Já para a relação das variáveis hipertensos e acidente vascular cerebral o p-valor do teste t do modelo é menor que o nível de 5% de probabilidade, com o coeficiente de determinação ajustado $R^2 = 0,65$. E utilizando o teste de Shapiro-Wilk para constatação da normalidade dos resíduos, obteve-se uma estatística de $W = 0{,}96$ com valor $p = 0,15 > \alpha = 0,05$, logo este modelo é estatisticamente significativo.

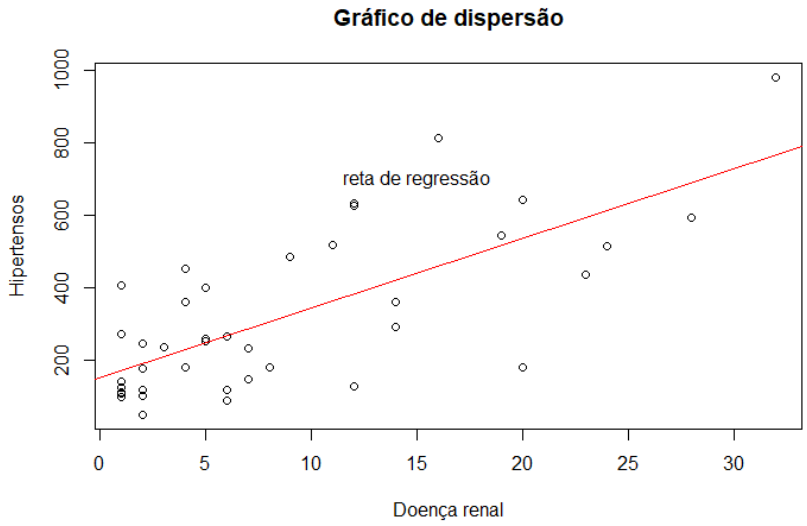

Figura 10 – Diagrama de dispersão da variável hipertensos vs doença renal

A Figura 10 mostra que há uma relação linear entre as variáveis Y e X, foi feito uma análise de regressão simples para verificar se a relação entre as variáveis hipertensos e doença renal é estatisticamente significativo, então constatou que o p-valor do teste t do modelo é inferior ao nível de 5% de probabilidade, com o coeficiente de determinação ajustado $R^2 = 0,51$. E para verificar a normalidade dos resíduos foi aplicado o teste de Shapiro-Wilk obteve uma estatística $W = 0{,}97$ com valor $p = 0,38 > \alpha = 0,05$ de significância, logo este modelo é estatisticamente significativo.

Utilizando-se a função *lm* do software R. Sejam as hipóteses:

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \text{ para pelo menos um,} \quad \text{j} = 0{,}1{,}\cdots, 5. \end{cases}$$

Tem-se o modelo ajustado com o valor p dos coeficientes e a estatística t para as variáveis, apresentado na Tabela 4.

Tabela 4 – Estimativas dos parâmetros com respectivos erros padrão e estatística t para as variáveis sedentarismo, sobrepeso, tabagismo, acidente vascular cerebral, doença renal

Efeitos	Estimativas	Erro Padrão	Valor de t	Pr($>\|t\|$)
$\hat{\beta}_0$	31,94533	23,50799	1,359	0,18312
$\hat{\beta}_1$	1,40597	0,41433	3,393	0,00177
$\hat{\beta}_2$	0,00055	0,59998	0,001	0,99928
$\hat{\beta}_3$	0,45719	0,70039	0,653	0,51829
$\hat{\beta}_4$	2,82544	1,23653	2,285	0,02868
$\hat{\beta}_5$	2,60599	2,31560	1,125	0,26830

Desta forma obtem-se o seguinte modelo ajustado, conforme a expressão abaixo:

$$\hat{y} = 31,95 + 1,41x_{_1} + 0,0005x_{_2} + 0,46x_{_3} + 2,83x_{_4} + 2,61x_{_5} \tag{3.2}$$

Como é possível observar na Tabela 4, que há variáveis cujo o p valor do teste é superior ao nível de significância, para isto utiliza-se o método de seleção de variável regressora passo a passo (*stepwise*) sobre o critério de informação de Akaike (AIC), no qual através deste procedimento é admissível selecionar quais variáveis vão continuar no modelo, para tanto, ajustou-se inicialmente o modelo reduzido, aquele apenas com o intercepto e vão sendo incluidas as variáveis regressoras, até que o menor valor de AIC fosse obtido, deste modo partiu-se de um $AIC = 469,44$ e obteve-se um $AIC = 466,0853$ com o modelo (3.3):

Tem-se o modelo final definido na Tabela 5:

Tabela 5 – Estimativas dos parâmetros com respectivos erros padrão e estatística t para as variáveis sedentarismo, acidente vascular cerebral

Efeitos	Estimativas	Erro Padrão	Valor de t	Pr(>|t|)
$\hat{\beta}_0$	31,9964	21,6811	1,614	0,1150
$\hat{\beta}_1$	1,6070	0,1906	8,431	0,0000
$\hat{\beta}_4$	3,9945	0,8532	4,682	0,0000

Desta forma obtem-se o seguinte modelo ajustado, conforme a expressão abaixo:

$$\hat{y} = 31,9964 + 1,6070x_{_1} + 3,9945x_{_4} \tag{3.3}$$

Nota-se que o modelo (3.3) da Tabela 5, é um modelo inicialmente satisfatório, aonde pode ser observado que a variável sedentarismo e acidente vascular cerebral apresenta o valor p do teste inferior ao nível de 5% e até 1% de probabilidade sendo assim há fortes evidências para rejeitar a hipótese H_0. Logo assegurar-se que os coeficientes dessas variáveis são estatísticamente diferentes de zero.

Por meio de uma análise gráfica dos resíduos é possível observar se os resultados acima são confiáveis ou não:

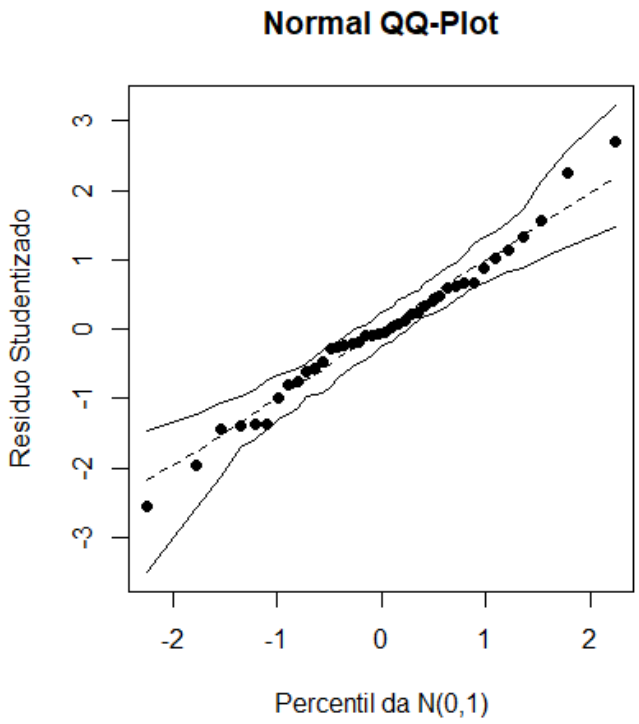

Figura 11 – Gráfico de envelope

Visualizando o gráfico de envelope, nota-se maior acomodação dos pontos entre as bandas de confiança, isto leva-nos a opinar que não apresenta evidência contra a suposição de normalidade nos resíduos. mesmo com esta certeza, aplicou-se um teste rigoroso para identificação da normalidade dos resíduo, o de Shapiro-Wilk e obteve o valor da estatística $W = 0{,}98771$ com valor $p = 0,9356 > \alpha = 0,05$, então conclui-se ao nível de 5% de probabilidade que há indicações para dizer que a distribuição dos resíduos seja normal.

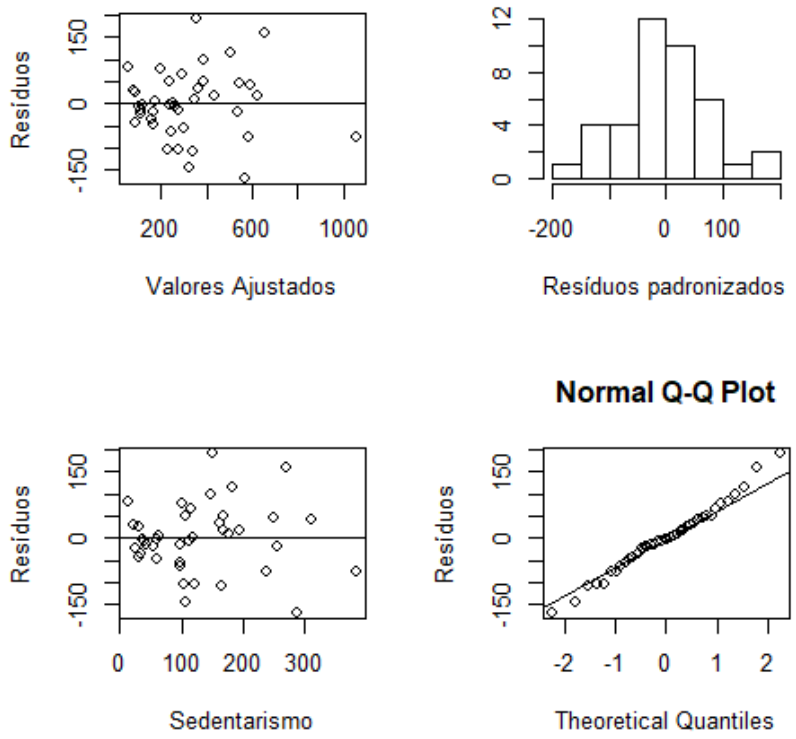

Figura 12 – Gráficos para Análise dos Resíduos

Observe, nestes gráficos, na Figura 12 que há indicativos de que a variância dos erros seja constante, e não há evidências de que os erros não sigam a distribuição Normal e percebe-se de que os erros não são correlacionados entre si. Como é possível observar no histograma dos resíduos padronizados. que a distribuição é relativamente normal.

4 Conclusão

Nesse trabalho ajustou-se modelo de regressão linear múltipla utilizando o método de máxima verossimilhança para encontrar os estimadores dos parâmetros, com a finalidade de averiguar quais das variáveis contribuem significativamente para o crescimento de hipertensos em quarenta municípios do estado da Paraíba no período de abril de 2010 à abril de 2013, para este fim, foi escolhida a variável resposta como sendo os casos de hipertensos notificados, e as variáveis independentes escolhidas foram; sedentarismo, sobrepeso, tabagismo, acidente vascular cerebral e doença renal. Inicialmente foi retirado a variável cuja probabilidade é superior ao nível de significância, para tanto foi necessário utilizar o critério de seleção de variável regressora *stepwise*, daí foi escolhido um modelo reduzido onde apenas a variável sedentarismo e acidente vascular cerebral foram estatisticamente significativas. Comprovando a gravidade do sedentarismo e mostrando que indivíduos sedentários têm mais chance de sofrerem de hipertensão, e ainda revelando a relação entre a hipertensão e indivíduos que sofreram de acidente vascular cerebral, de uso desta informação e ainda para confirmação do estudo verificou-se as suposições para a validação do modelo.

Para comprovar as análises iniciais foi feito uma análise gráfica para modelo ajustado e percebido maior acomodação dos pontos entre as bandas de confiança, isto leva-nos a supor que existem evidência para dizer que os resíduos seguem distribuição normal, mesmo com esta certeza, foi indispensável a aplicação do teste de Shapiro-Wilk, com o intuito de comprovar ou não se os resíduos seguem normalidade, logo que foi aplicado este teste foi confirmado a normalidade dos resíduos, com $W = 0,98771$ e com valor $p = 0,9356$.

Desse forma foi constatado para este modelo que indivíduos que levam uma vida sedentária estão mais propícios a sofrerem com problema de hipertensão nesses municípios em estudo, tendo em vista que, o sedentarismo é um problema vital de saúde pública no mundo e favorece a epidemia crescente de obsedidade e aumento da prevalência de doenças como hipertensão, de acordo com Hoffmann et al. (2016), sempre é importante conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista uma relação causal entre elas, há como relacioná-las por meio de uma expressão matemática, mesmo que não haja uma relação de causa e efeito, por esta razão é intuitivo percebe a relação entre indivíduos que sofreram de acidente vascular cerebral causado por outro fator de risco, e que eventualmente sobreviveram e que passam a integrar o quadro de hipertensos nesses munícipios.

Referências

DEMÉTRIO, C. G. B.; ZOCCHI, S. S. Modelos de regressão. *Piracicaba: ESALQ*, 2006. Citado 9 vezes nas páginas 9, 11, 16, 17, 18, 23, 25, 26 e 34.

HOFFMANN, R. et al. *Análise de regressão: uma introdução à econometria*. [S.l.]: O autor, 2016. Citado 9 vezes nas páginas 11, 13, 14, 15, 16, 18, 19, 20 e 40.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013. Citado na página 24.

POLIZELLO, A. O desenvolvimento das ideias de herança de francis galton: 1865-1897. *Filosofia e historia da biologia*, Associação Brasileira de Filosofia e História da Biologia-ABFHiB, v. 6, n. 1, p. 1–17, 2011. Citado na página 10.

RÊGO, N. L. d. Modelo de regressão linear múltipla com variável dummy: um estudo de caso. 2014. Citado na página 10.

RODRIGUES, S. C. A. *Modelo de regressão linear e suas aplicações*. Tese (Doutorado) — Universidade da Beira Interior, 2012. Citado 6 vezes nas páginas 9, 17, 20, 22, 25 e 27.