



UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

Allana Lívia Beserra Paulino

Estudo sobre a interiorização da deficiência física na Paraíba com o uso de regressão linear simples

Campina Grande - PB

Abril de 2016

Allana Livia Beserra Paulino

**Estudo sobre a interiorização da deficiência física na
Paraíba com o uso de regressão linear simples**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Estatística Aplicada 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 especialista em Estatística.

Orientador: Gustavo Henrique Esteves

Campina Grande - PB

Abril de 2016

É expressamente proibida a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano da dissertação.

P328e Paulino, Allana Livia Beserra
Estudo sobre a interiorização da deficiência física na Paraíba
com o uso de regressão linear simples [manuscrito] / Allana Livia
Beserra Paulino. - 2016.
50 p. : il.

Digitado.
Monografia (Estatística Aplicada) - Universidade Estadual da
Paraíba, Centro de Ciências e Tecnologia, 2016.
"Orientação: Prof. Drº Gustavo Henrique Esteves,
Departamento de Estatística".

1. SIAB/DATASUS. 2. Regressão linear simples. 3.
Deficiência física. I. Título.

21. ed. CDD 519.535

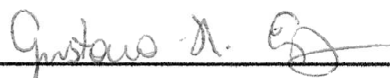
Allana Livia Beserra Paulino

Estudo sobre a interiorização da deficiência física na Paraíba com o uso de regressão linear simples

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Estatística Aplicada 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 especialista em Estatística.

Trabalho aprovado em 22 de abril de 2016.

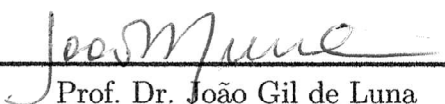
BANCA EXAMINADORA



Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba



Prof. Dr. João Gil de Luna
Universidade Estadual da Paraíba

Aos meus pais Antônio e Lúcia, aos meus irmãos Lúcio, Alisson, Allen e Allan por seu amor e dedicação sempre.

Agradecimentos

Aos coordenadores do curso de Especialização, por seu empenho.

Ao professor Gustavo Esteves pela orientação e paciência.

Ao meu pai Antônio, a minha mãe Lúcia, meus irmãos Lúcio, Alisson, Allen e Allan e a minha tia Fátima pelo carinho, paciência e incentivo.

Aos colegas de classe pelo incentivo, amizade e pelos momentos de sofrimentos juntos.

À todos os professores da UEPB que contribuíram de alguma forma para meu crescimento profissional e também aos funcionários do Departamento de Estatística.

Resumo

O presente trabalho buscou analisar a existência de interiorização da deficiência física para os habitantes dos municípios do estado da Paraíba, através do uso do modelo de regressão linear simples. Os dados utilizados foram obtidos do Sistema de Informação da Atenção Básica (SIAB) referentes ao período de 2004 à 2013, para todos os municípios do estado. A porcentagem de indivíduos com deficiência foi calculada a partir das informações sobre número absoluto de deficientes e a totalidade de habitantes para cada município disponíveis no banco de dados do SIAB. Diversos tipos de bancos de dados de pessoas com deficiência física são disponibilizados nesse sistema, entre eles estão deficiência auditiva, visual e motora, de modo que os indivíduos sejam incapazes de realizar algumas atividades do dia-dia. No mesmo não existia informações sobre indivíduos com deficiência intelectual, logo, não se sabe se esses dados sobre deficiência física usados para tal análise incluem ou não este tipo de problema. A realização das análises foi feita inicialmente por um teste de sequências (*runs test*) para avaliar a aleatoriedade dos dados, após tal teste, foram excluídos 78 municípios que não pareciam apresentar aleatoriedade. Assim, da totalidade de 223 municípios disponíveis no banco de dados, foram analisados 145 deles. Em toda a análise o *software* estatístico R (versão 2.12.0) foi utilizado. Nas análises gráficas, observou-se a dispersão dos resíduos para verificar a validade dos modelos de regressão, e para a suposição de homoscedasticidade (variância constante) foram aplicados os testes de Goldfeld-Quandt e Breusch-Pagan através do pacote *lmtest* do R. Obteve-se normalidade nos resíduos e concluiu-se que quanto mais próximo da capital, João Pessoa, menor a proporção de pessoas com deficiência física, com isso, existe a hipótese de uma interiorização da deficiência no estado da Paraíba, embora os resultados variem ao longo dos anos observados neste estudo.

Palavras-chave: SIAB/DATASUS. Regressão linear. Deficiência física.

Abstract

This study aimed to analyze the existence of internalization of disability for the inhabitants of the municipalities in the state of Paraíba, through the use of simple linear regression model. The data used were obtained from the *Sistema de Informação Básica* (SIAB) for the period from 2004 to 2013 for all cities in the state. The percentage of individuals with disabilities was calculated from the information on the absolute number of disabled and all inhabitants for each municipality available in SIAB database. Several types of databases of people with physical disabilities are available in this system, among them are hearing, visual and motor disabilities, so that individuals are unable to perform some activities of daily life. There was no information about individuals with intellectual disabilities, so it is unknown if these data used for this analysis included this type of problem. The completion of the analysis was initially made by a runs test to evaluate the randomness of the data, and after that, it was excluded 78 municipalities that did not seem to show randomness. Thus, from the total of 223 cities available in the database, 145 of them were analyzed. Throughout the analysis, the R statistical software (version 2.12.0) was used. In the graphical analysis, the residuals dispersion was analysed to verify the validity of regression models, and to evaluate the assumption of homoscedasticity (constant variance), the Goldfeld-Quandt and Breusch-Pagan tests were applied through the R package `lmtest`. We concluded that the residuals were normally distributed and found that the closer the capital, João Pessoa, the lower the proportion of people with disabilities, thus, there is a chance of an inland of disability in the state of Paraíba, although the results vary over the years observed in this study.

Keywords: SIAB/DATASUS. Linear regression. Physical disability.

Lista de ilustrações

Figura 1 – Tabela a análise de variância para um modelo de regressão linear simples	27
Figura 2 – Disposição gráfica do trabalho realizado através do teste de sequências, para se verificar a aleatoriedade dos dados observados.	37
Figura 3 – Gráfico de resíduo da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2010.	39
Figura 4 – Gráfico de resíduo da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2013.	40
Figura 5 – Gráfico da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2010.	40
Figura 6 – Gráfico da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2013.	41

Lista de tabelas

Tabela 1 – Municípios considerados não aleatórios através do teste de aleatoriedade com níveis descritivos inferiores a 0,05 para o total populacional. . . .	36
Tabela 2 – Municípios considerados não aleatórios através do teste de aleatoriedade com níveis descritivos inferiores a 0,05 agora para a proporção populacional.	37
Tabela 3 – O efeito da distância da capital aos municípios sobre a PID para cada 1km de distância da capital a cada município há um evento de $\beta_1\%$ de PID.	39

Sumário

1	INTRODUÇÃO	11
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Teste de sequências - <i>Runs Test</i>	13
2.2	Análise de Regressão Linear	14
2.2.1	Modelo de regressão linear simples	14
2.2.2	Estimação dos parâmetros do modelo	15
2.2.3	Propriedades dos estimadores	16
2.3	Testes e intervalos de confiança para os estimadores	21
2.4	Análise de variância	24
2.4.1	Soma de Quadrados	24
2.4.2	Quadrado Médio	26
2.4.3	Tabela de Análise de Variância	26
2.4.4	Teste F	27
2.5	Seleção de modelos	27
2.5.1	Coeficiente de Determinação	27
2.5.2	Coeficiente de determinação ajustado	28
2.5.3	AIC e BIC	28
2.6	Testes estatísticos	28
2.6.1	Diagnóstico de normalidade	29
2.6.2	Diagnóstico de Homoscedasticidade	29
2.7	Intervalo de confiança para a resposta média e predição	31
2.8	Análise de resíduo na regressão linear simples	32
3	MATERIAL E MÉTODOS	34
4	RESULTADOS OBTIDOS	35
4.1	Resultados dos testes de aleatoriedade	35
4.2	Resultados dos modelos de regressão linear simples	38
5	CONCLUSÃO	42
	REFERÊNCIAS	43
	APÊNDICE A – SCRIPT PARA ANÁLISE NO SOFTWARE R	44

1 Introdução

A prevalência de deficiência física no mundo varia de menos de 1% até mais de 20%. Tal variação acontece principalmente devido a diferentes conceitos de deficiência ou métodos de diagnóstico clínico e a baixa qualidade no planejamento de estudos voltados para o problema. Isso acarreta resultados que não são comparáveis e dificulta o planejamento de políticas públicas voltadas para o problema. Na verdade, a situação é um pouco mais complicada do que simplesmente encontrar uma definição para a deficiência, porque sua etiologia e gravidade também variam significativamente e sua medição depende da finalidade e dos objetivos das pesquisas (AMIRALIAN et al., 2000; Organização Mundial da Saúde, 2003).

A Organização Mundial da Saúde (OMS), em 1976, definiu um padrão de classificação para perdas físicas que as distinguem entre três tipos: deficiência, incapacidade e prejuízo. Posteriormente, após uma atualização deste manual pela OMS em 2001, foi publicado no Brasil a Classificação Internacional de Funcionalidade, Incapacidade e Saúde (CIF), onde deficiências são definidas como problemas nas funções ou nas estruturas do corpo com um desvio importante ou perda (Organização Mundial da Saúde, 2003). De forma geral, a OMS estima que cerca de 10% da população mundial apresente alguma forma de perda (World Health Organization, 2002).

O Brasil conta com um banco de dados que contém, entre outras informações, registros sobre a prevalência de pessoas com deficiência que foi elaborado pelo Ministério da Saúde através do Sistema de Informação da Atenção Básica (SIAB), tendo sido implementado em 1998 para acompanhar as ações e resultados das atividades realizadas pelas equipes do Programa de Saúde da Família (PSF) e do Programa de Agentes Comunitários de Saúde (PACS). Tal sistema contempla informações sobre registros de famílias, habitação, saneamento e estado de saúde. As informações são obtidas pelos agentes através de uma ficha específica e, posteriormente, os dados são consolidados e disponibilizados pelo SIAB/DATASUS (Ministério da Saúde, 2000).

O conceito de deficiência utilizado pelo SIAB é bastante genérico: “A deficiência é um defeito ou condição física ou mental permanente ou de longa duração que de alguma forma dificulta ou impede a pessoa de realizar certas atividades cotidianas de trabalho ou lazer” (WERNER, 1994). Embora o banco de dados tenha informações consolidadas desde 1998, a confiabilidade dos dados tem sido frequentemente questionada para estudos epidemiológicos devido a uma ampla gama de fatores.

O trabalho intitulado Retratos da Deficiência no Brasil mostra que, entre os 50 municípios com maiores índices de pessoas portadoras de deficiência física, boa parte delas

se encontra em municípios localizados no interior dos estados do Nordeste (NÉRI, 2003), de modo que essas cidades apresentam algumas características comuns, como número pequeno de habitantes e localizações geográficas relativamente distantes das capitais litorâneas.

Deste modo, surge naturalmente uma hipótese de interiorização da deficiência física nos estados do Nordeste, de modo que é razoável teorizar que os municípios situados mais no interior apresentam maior proporção de indivíduos com deficiências e/ou incapacidades físicas. Neste sentido, um trabalho de conclusão de curso recente de Melo (2010) do curso de Ciências Biológicas da UEPB, apontou indícios que corroboravam tal hipótese.

Neste trabalho, verificou-se se a distribuição de indivíduos com deficiência era influenciada por fatores geodemográficos, onde foram analisados os dados sobre a ocorrência de pessoas com deficiência física em todos os 223 municípios do estado da Paraíba, que tem aproximadamente 500 km de extensão na direção leste-oeste com elevado número de municípios com baixa densidade populacional, características que permitem testar a hipótese da interiorização da deficiência. Todos os dados de deficiência foram obtidos da plataforma do SIAB/Datasus. Os dados de localização geográfica foram obtidos do IBGE, através da distância (em km) de cada município até a capital do estado, João Pessoa.

A principal ferramenta estatística utilizada no trabalho de Melo (2010) foi a regressão linear simples, que mostrou uma tendência significativa de aumento da proporção de pessoas com deficiência à medida que a distância do município à João Pessoa aumentava. Porém, a principal crítica recebida neste trabalho é que a falta de qualidade dos dados do SIAB não garantia a veracidade dos resultados obtidos.

Assim, o presente trabalho focou atenção na tentativa de reproduzir as análises feitas no trabalho original de Melo (2010), porém, com o uso apenas dos municípios que apresentavam dados aparentemente aleatórios, após a aplicação de um teste de sequências para avaliar a aleatoriedade de uma amostra de valores numéricos. A seguir é apresentada uma breve fundamentação teórica dos métodos utilizados, material e métodos, resultados e discussão.

2 Fundamentação Teórica

O presente trabalho foi desenvolvido em duas etapas principais. A primeira focou atenção em um teste de aleatoriedade, baseado em um teste de sequências, e foi tema principal de um projeto de Iniciação Científica.

Já a segunda etapa do estudo foi desenvolvida no presente trabalho e tratou de utilizar modelos de regressão linear simples nos dados obtidos do SIAB, para todos os municípios que aparentavam apresentar aleatoriedade dos dados após aplicação dos testes.

2.1 Teste de sequências - *Runs Test*

Para os dados de indivíduos com deficiência do SIAB foi feita uma análise quanto a presença ou não de aleatoriedade. Este teste é frequentemente referenciado como teste de sequências (*runs test*, em inglês). Consiste em quantificar o número de sequências em um determinado município, que aumenta ou diminui quando comparado em relação a uma estimativa de tendência local, onde geralmente se usa a média amostral ou a mediana.

Para melhor visualização, temos o cálculo de probabilidade, com intuito de analisar se o dado é considerado aleatório ou se existe ausência de aleatoriedade, na expressão a seguir:

$$f_{R_1, R_2}(r_1, r_2) = \frac{c \binom{n_1-1}{r_1-1} \binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}}, \quad (2.1)$$

sendo, $c = 2$ se o número de sequências do tipo 1 for igual ao número de sequências do tipo 2 ($r_1 = r_2$) e $c = 1$ se o número de sequências do tipo 1 for igual ao número de sequências do tipo 2 mais ou menos 1 ($r_1 = r_2 \pm 1$).

Tanto o número de sequências como seus comprimentos, que são naturalmente inter-relacionados, devem refletir a existência de algum tipo de padrão. Os testes de aleatoriedade podem, portanto, basear-se em qualquer critério ou alguma combinação destes. Poucas ou muitas sequências, uma sequência de comprimento excessivo, ou também muitas sequências de comprimentos excessivos, etc, podem ser usados como critérios estatísticos para a rejeição da hipótese nula de aleatoriedade, uma vez que estas situações devem ocorrer raramente em uma sequência verdadeiramente aleatória, como citado em [Gibbons e Chakraborti \(2003\)](#).

É importante mencionar aqui que o teste de sequências para verificarmos a aleatoriedade dos dados pode ser aplicado tanto com o uso da distribuição exata de probabilidade para a estatística de teste, apresentada na equação (2.1), como com uma distribuição

assintoticamente normal. No nosso caso, dado o tamanho amostral relativamente pequeno, consideramos apenas os resultados obtidos através do teste exato.

2.2 Análise de Regressão Linear

A teoria de regressão teve origem no século *XIX* com Galton. Em um de seus trabalhos ele estudou a relação entre a altura dos pais e dos filhos (X_i e Y_i), procurando saber como a altura do pai influenciava a altura do filho. Notou que se o pai fosse muito alto ou muito baixo, o filho teria uma altura tendendo à média. Por isso, ele chamou de regressão, ou seja, existe uma tendência de os dados regredirem à média (DEMÉTRIO; ZOCHI, 2008).

A análise de regressão é um método utilizado para conhecer os efeitos que algumas variáveis exercem sobre outras. Até mesmo quando não existe uma relação casual entre as variáveis, elas podem se relacionar por meio de algumas expressões matemáticas, que são úteis para a estimação do valor de uma das variáveis, quando se tem conhecimento dos valores das outras variáveis (HOFFMANN, 2006).

2.2.1 Modelo de regressão linear simples

A análise de regressão linear é uma técnica estatística que procura modelar a relação de uma variável dependente (conhecida como resposta) com outras variáveis independentes (conhecidas como explicativas). Assim, pode-se visualizar isso, através da expressão matemática na sua forma mais simples:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (2.2)$$

aonde n é o tamanho da amostra, y_i e x_i representam os valores observados para as variáveis dependente e independente, respectivamente, e ϵ_i é o erro associado ao modelo.

Em casos particulares, existe a forma mais generalizada, em forma matricial

$$Y = X\beta + \epsilon$$

com

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{e} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

em que, Y é um vetor $n \times 1$ cujos componentes corresponde às n respostas; X é uma matriz de dimensão $n \times (p + 1)$ denominada matriz do modelo; ϵ é um vetor de dimensão

$n \times 1$ cujos componentes são os erros e β é um vetor $(p + 1) \times 1$ cujos elementos são os coeficientes de regressão.

Ao estabelecer o modelo de regressão linear simples, pressupomos que:

- A relação entre X e Y é linear.
- Os valores de X são fixos, isto é, X não é uma variável aleatória.
- A média do erro é nula, isto é, $E(y_i) = 0$.
- Para um dado valor de X, a variância do erro y é sempre σ^2 , denominada variância residual, isto é,

$$E(y_i^2) = \sigma^2$$

$$E[Y_i - E(Y_i|X_i)]^2 = \sigma^2$$

Dizemos que o erro é homocedástico ou que temos homocedasticia (do erro ou da variável dependente).

- O erro de uma observação é não-correlacionada com o erro em outra observação, isto é, $E(y_i y_j) = 0$ para $i \neq j$.
- Os erros seguem distribuição normal

$$y \sim N(0, \sigma^2).$$

2.2.2 Estimação dos parâmetros do modelo

Segundo Hoffmann (2006), os estimadores dos parâmetros β_0 e β_1 da regressão linear simples, denotados por b_0 e b_1 , respectivamente, podem ser encontrados pelo método dos mínimos quadrados. Na prática, a partir de uma amostra de n pares de valores (x_i, y_i) com $i = 1, 2, \dots, n$ e após alguns cálculos relativamente simples, b_0 e b_1 são dados por:

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x},$$

onde

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \quad \text{e} \quad S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right).$$

E assim, a reta de regressão pode ser estimada, a partir dos dados, por:

$$\hat{y}_i = b_0 + b_1 x_i \quad i = 1, 2, \dots, n.$$

Ainda de acordo com Hoffmann (2006), assumindo que os erros ϵ_i sejam normalmente distribuídos, é possível mostrar que as distribuições de probabilidades dos estimadores b_1 e b_0 são dadas por:

$$b_1 \sim N \left(\beta_1, \frac{\sigma^2}{S_{xx}} \right) \quad \text{e} \quad b_0 \sim N \left[\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right],$$

onde σ^2 é a variância real dos erros do modelo de regressão. Estes últimos resultados podem ser utilizados para testar hipóteses relacionadas aos parâmetros β_0 e β_1 , bem como para a construção de intervalos de confiança.

2.2.3 Propriedades dos estimadores

De acordo com [Ferreira \(2009\)](#) um dos objetivos da regressão é desenvolver a equação que permitirá ao investigador científico fazer previsões dos valores da variável aleatória Y . Para isso, é necessário ajustar equação, ou seja, os valores dos parâmetros do modelo β_0 e β_1 e da variância residual σ^2 consistentes com os dados disponíveis devem ser determinados. Um dos métodos que pode ser utilizado, para isso, é o dos quadrados mínimos.

1. Valor esperado (média) de $\widehat{\beta}_1$

$$C_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad i = 1, \dots, n,$$

segue que,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n C_i Y_i.$$

Desta forma,

$$E(\widehat{\beta}_1) = E\left(\sum_{i=1}^n C_i Y_i\right) = \sum_{i=1}^n C_i E(Y_i) = \sum_{i=1}^n C_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n C_i + \beta_1 \sum_{i=1}^n C_i x_i.$$

Como

$$\sum_{i=1}^n C_i = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i - n\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i - \sum_{i=1}^n x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

$$\sum_{i=1}^n C_i x_i = \frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1,$$

concluimos que $E(\widehat{\beta}_1) = \beta_1$ (estimador não viciado).

2. Variância de $\widehat{\beta}_1$: De (1) temos que

$$Var(\widehat{\beta}_1) = Var\left(\sum_{i=1}^n C_i Y_i\right).$$

Como Y_i , $i = 1, \dots, n$ são variáveis independentes, segue que

$$Var(\hat{\beta}_1) = \sum_{i=1}^n C_i^2 Var(Y_i) = \sum_{i=1}^n \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Considerando n pares de valores observados $(x_1, y_1), \dots, (x_n, y_n)$, podemos escrever

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

3. Valor esperado (média) de $\hat{\beta}_0$:

$$E(\hat{\beta}_0) = E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E(\bar{Y}) - \bar{x} E(\hat{\beta}_1) = E\left(\sum_{i=1}^n \frac{Y_i}{n}\right) - \bar{x} \beta_1 = \sum_{i=1}^n \frac{E(Y_i)}{n} - \bar{x} \beta_1.$$

Como $E(Y_i) = (\beta_0 + \beta_1 x_i)$, segue que

$$E(\hat{\beta}_0) = \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \bar{x} \beta_1 = \beta_0 + \beta_1 \frac{\sum_{i=1}^n x_i}{n} - \bar{x} \beta_1 = \beta_0.$$

(estimador não viciado).

4. Variância de $\hat{\beta}_0$:

$$Var(\hat{\beta}_0) = Var(\bar{Y} - \hat{\beta}_1 \bar{x}) = Var(\bar{Y}) + Var(\hat{\beta}_1 \bar{x}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{x}).$$

Notemos que

$$\begin{aligned} Cov(\bar{Y}, \hat{\beta}_1 \bar{x}) &= E(\bar{Y} \hat{\beta}_1 \bar{x}) - E(\bar{Y}) E(\hat{\beta}_1 \bar{x}) = \\ &= E(\bar{x} \bar{Y} \hat{\beta}_1) - E\left(\sum_{i=1}^n \frac{Y_i}{n}\right) \bar{x} \beta_1 \\ &= E\left(\bar{x} \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i)}{n} \hat{\beta}_1\right) - \frac{\bar{x} \beta_1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \frac{\bar{x}}{n} \sum_{i=1}^n [\beta_0 \beta_1 + x_i \beta_1^2 + E(\varepsilon_i \hat{\beta}_1)] \\ &\quad - \frac{\bar{x} \beta_1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \\ &= \frac{\bar{x} \beta_1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) + \frac{\bar{x}}{n} \sum_{i=1}^n E(\varepsilon_i \hat{\beta}_1) - \frac{\bar{x} \beta_1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \frac{\bar{x}}{n} \sum_{i=1}^n E(\varepsilon_i \hat{\beta}_1). \end{aligned}$$

Como

$$\begin{aligned} E(\varepsilon_i \hat{\beta}_1) &= \\ E\left[\varepsilon_i \frac{\sum_{j=1}^k (x_j - \bar{x}) Y_j}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] &= \frac{\sum_{j=1}^k (x_j - \bar{x}) E[\varepsilon_i Y_j]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{j=1}^k (x_j - \bar{x}) E[\varepsilon_i (\beta_0 + \beta_1 x_j + \varepsilon_j)]}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

$$= \frac{\sum_{j=1}^k (x_j - \bar{x}) [\beta_0 E(\varepsilon_i) + \beta_1 x_j E(\varepsilon_i) + E(\varepsilon_j \varepsilon_i)]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{j=1}^k (x_j - \bar{x}) E(\varepsilon_j \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0,$$

já que para $i \neq j$,

$$E(\varepsilon_j \varepsilon_i) = 0 \Rightarrow E(\varepsilon_i \hat{\beta}_1) = 0.$$

e para $i = j$,

$$E(\varepsilon_j \varepsilon_i) = \sigma^2 \Rightarrow E(\varepsilon_i \hat{\beta}_1) = \frac{\sum_{j=1}^k (x_j - \bar{x}) \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$\sigma^2 \frac{\sum_{j=1}^k (x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0;$$

segue que

$$Cov(\bar{Y}, \hat{\beta}_1 \bar{x}) = 0.$$

Desta forma,

$$Var(\hat{\beta}_0) = Var(\bar{Y}) + Var(\hat{\beta}_1 \bar{x}) = Var\left(\sum_{i=1}^n \frac{Y_i}{n}\right) + \bar{x}^2 Var(\hat{\beta}_1).$$

Como Y_i , $i = 1, \dots, n$ são independentes, segue que

$$Var(\hat{\beta}_0) = \frac{1}{n^2} \sum_{i=1}^n Var(Y_i) + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n\sigma^2}{n^2} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

Novamente, dados n pares de valores $(x_1, y_1), \dots, (x_n, y_n)$ escrevemos

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

5. Covariância entre $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = E(\hat{\beta}_0 \hat{\beta}_1) - E(\hat{\beta}_0)E(\hat{\beta}_1)$$

$$= E[(\bar{Y} - \hat{\beta}_1 \bar{x}) \hat{\beta}_1] - \beta_0 \beta_1$$

$$= E[\bar{Y} \hat{\beta}_1 - \bar{x} \hat{\beta}_1^2] - \beta_0 \beta_1 = E(\bar{Y} \hat{\beta}_1)$$

$$\begin{aligned}
-\bar{x}E(\hat{\beta}_1^2) - \beta_0\beta_1 &= E\left[\frac{1}{n}\sum_{i=1}^n Y_i\hat{\beta}_1\right] - \bar{x}[Var(\hat{\beta}_1) + (E(\hat{\beta}_1))^2] - \beta_0\beta_1 \\
&= \frac{1}{n}\sum_{i=1}^n E[(\beta_0 + \beta_1 x_i + \varepsilon_i)\hat{\beta}_1] \\
-\bar{x}\left[\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1^2\right] - \beta_0\beta_1 &= \frac{1}{n}\sum_{i=1}^n E[\beta_0\hat{\beta}_1 + \beta_1\hat{\beta}_1 x_i + \varepsilon_i\hat{\beta}_1] - \frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \bar{x}\beta_1^2 - \beta_0\beta_1 \\
&= \frac{1}{n}\sum_{i=1}^n [\beta_0\beta_1 + \beta_1^2 x_i + E(\varepsilon_i\hat{\beta}_1)] - \frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \bar{x}\beta_1^2 - \beta_0\beta_1 \\
&= \beta_0\beta_1 + \beta_1^2\bar{x} + \frac{1}{n}\sum_{i=1}^n E(\varepsilon_i\hat{\beta}_1) - \frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \bar{x}\beta_1^2 - \beta_0\beta_1 = \frac{1}{n}\sum_{i=1}^n E(\varepsilon_i\hat{\beta}_1) - \frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

De (4) temos que

$$E(\varepsilon_i\hat{\beta}_1) = 0$$

e portanto,

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

6. Distribuição amostral para $\hat{\beta}_1$: Em (1), definimos

$$\hat{\beta}_1 = \sum_{i=1}^n C_i Y_i.$$

Como $\hat{\beta}_1$ é combinação linear de normais independentes (combinação linear dos Y_i), segue que $\hat{\beta}_1$ também tem distribuição normal com média e variância dadas respectivamente em (1) e (2) e portanto,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

7. Distribuição amostral para $\hat{\beta}_0$: Como em (6), $\hat{\beta}_0$ também é combinação linear de normais independentes Y_i e portanto, também tem distribuição normal. A média e a variância de $\hat{\beta}_0$ são apresentadas em (3) e (4), respectivamente. Desta forma,

$$\hat{\beta}_0 \sim N\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right].$$

Em relação ao estimador da variância σ^2 ,

8. Valor esperado (média) de QME:

$$QME = \frac{SQE}{n-2}$$

Assim,

$$E(QME) = E\left(\frac{SQE}{n-2}\right) = \frac{E(SQE)}{n-2}.$$

Sabemos que

$$\begin{aligned} SQE &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\ &= \sum_{i=1}^n [Y_i^2 - 2Y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2] \\ &= \sum_{i=1}^n [Y_i^2 - 2Y_i(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) + (\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)^2] \\ &= \sum_{i=1}^n [Y_i^2 - 2Y_i \bar{Y} + 2Y_i \bar{x} \hat{\beta}_1 - 2Y_i x_i \hat{\beta}_1 + [\bar{Y} - \hat{\beta}_1(\bar{x} - x_i)]^2] \\ &= \sum_{i=1}^n [Y_i^2 - 2Y_i \bar{Y} + 2Y_i \bar{x} \hat{\beta}_1 - 2Y_i x_i \hat{\beta}_1 + \bar{Y}^2 - 2\bar{Y} \hat{\beta}_1(\bar{x} - x_i) + \hat{\beta}_1^2(\bar{x} - x_i)^2] \\ &= \sum_{i=1}^n [(Y_i - \bar{Y})^2 - 2\hat{\beta}_1(x_i Y_i - \bar{x} Y_i + \bar{Y} \bar{x} - \bar{Y} x_i) + \hat{\beta}_1^2(\bar{x} - x_i)^2] \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n [(x_i - \bar{x})(\bar{Y} - Y_i)] + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i + \hat{\beta}_1^2 (\sum_{i=1}^n x_i^2 - n\bar{x}^2) = \\ &\quad \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i + \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\hat{\beta}_1} = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i. \end{aligned}$$

Desta forma,

$$\begin{aligned} E(SQE) &= E\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i\right) \\ &= E\left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right) - E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right) \\ &= \sum_{i=1}^n E(Y_i^2) - nE(\bar{Y}^2) - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} E\left[\left(\sum_{i=1}^n (x_i - \bar{x}) Y_i\right)^2\right] \end{aligned}$$

$$= \sum_{i=1}^n [Var(Y_i) + E^2(Y_i)] - n[Var(\bar{Y}) + E^2(\bar{Y})] \\ - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[Var \left(\sum_{i=1}^n (x_i - \bar{x}) Y_i \right) + E^2 \left(\sum_{i=1}^n (x_i - \bar{x}) Y_i \right) \right].$$

Como

$$E \left(\sum_{i=1}^n (x_i - \bar{x}) Y_i \right) = \sum_{i=1}^n (x_i - \bar{x}) E(Y_i) = \sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)$$

e

$$Var \left(\sum_{i=1}^n (x_i - \bar{x}) Y_i \right) = \sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i) = \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2.$$

Segue que

$$E(SQE) = \left(\sum_{i=1}^n \sigma^2 + (\beta_0 + \beta_1 x_i)^2 \right) - n \left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2 \right) \\ - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i) \right]^2 \\ = n\sigma^2 + \sum_{i=1}^n (\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2 \\ - \sigma^2 - \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i \right]^2 \\ = \sigma^2(n-2) + \sum_{i=1}^n (\beta_0 + \beta_1 x_i)^2 - n(\beta_0 + \beta_1 \bar{x})^2 - \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma^2(n-2).$$

Portanto,

$$E(QME) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2,$$

e verificamos que QME é um estimador não viciado para σ^2 .

2.3 Testes e intervalos de confiança para os estimadores

Segundo [Bussab e Morettin \(2005\)](#) os procedimentos de construção do intervalo de confiança nos leva aos seguintes intervalos para β_0 e β_1 :

Inferência para β_0

Suponha que desejamos testar a hipótese de que o intercepto é igual a um determinado valor, denotado por β_{00} . Desta forma, sejam as hipóteses

$$\beta_0 \neq \beta_{00}$$

Como visto em "Propriedades dos Estimadores",

$$\hat{\beta}_0 \sim N \left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right).$$

Assim, sob H_0 temos que

$$N_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\text{Var}(\hat{\beta}_0)}} \sim N(0, 1).$$

Além disso, seja

$$\chi = \frac{(n-2)QME}{\sigma^2} \sim \chi^2_{(n-2)}.$$

Como as variáveis aleatórias N_0 e χ são independentes, segue que

$$T = \frac{N_0}{\sqrt{\frac{\chi}{n-2}}} = \frac{\frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\left(\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right)}}}{\sqrt{\frac{(n-2)QME}{\sigma^2}}} = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim t_{(n-2)},$$

ou seja, T tem distribuição t de Student com $n-2$ graus de liberdade. Logo, intervalos de confiança e testes a respeito de β_0 podem ser realizados utilizando a distribuição t . No modelo 1.1.1, queremos testar as hipóteses

$$\beta_0 \neq 0.$$

Assim, a estatística do teste é dada por

$$T_0 = \frac{\hat{\beta}_0}{\sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim t_{(n-2)}.$$

Logo, rejeitamos H_0 com um nível de confiança de $(1 - \alpha)100\%$ se $|T_0| > t_{(1-\alpha/2, n-2)}$. O p -valor associado ao teste é dado por

$$P\text{-valor} = 2 * P \left(t_{(n-2)} > |T_0| \right).$$

Rejeitamos H_0 se o p -valor for menor do que o nível de significância α considerado. Geralmente adotamos $\alpha = 0,05$. Quando não rejeitamos H_0 , podemos utilizar o "Modelo

de Regressão sem Intercepto". O intervalo de confiança para β_0 com $(1 - \alpha)100\%$ é dado por

$$\left[\hat{\beta}_0 \pm t_{\left(1 - \frac{\alpha}{2}; n - 2\right)} \sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} \right]$$

Inferência para β_1 Inferência sobre β_1 é mais frequente já que por meio deste parâmetro temos um indicativo da existência ou não de associação linear entre as variáveis envolvidas. Similarmente ao parâmetro β_0 , consideremos as hipóteses

$$\beta_1 \neq \beta_{10}.$$

De "Propriedades dos Estimadores",

$$\hat{\beta}_1 \sim N \left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Assim, sob H_0 segue que

$$N_1 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0; 1).$$

Novamente, considerando que

$$\chi = \frac{(n - 2)QME}{\sigma^2} \sim \chi_{(n-2)}^2$$

e que N_1 e χ são independentes, obtemos

$$T = \frac{N_1}{\sqrt{\frac{\chi}{n - 2}}} = \frac{\frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\left(\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}}}{\sqrt{\frac{(n - 2)QME}{\sigma^2} \cdot \frac{1}{n - 2}}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{QME}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{(n-2)},$$

ou seja, T tem distribuição t de Student com n-2 graus de liberdade. Logo, intervalos de confiança e testes a respeito de β_1 podem ser realizados utilizando a distribuição t. No modelo em questão, queremos testar as seguintes hipóteses

$$\beta_1 \neq 0.$$

Neste caso, a estatística do teste é

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{QME}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{(n-2)}.$$

Assim, rejeitamos H_0 com um nível de confiança $(1 - \alpha)100\%$ se $|T_0| > t_{(1-\alpha/2, n-2)}$. O p-valor associado ao teste é dado por

$$P - \text{valor} = 2 * P \left(t_{(n-2)} > |T_0| \right).$$

Rejeitamos H_0 se o P-valor for menor do que α . O intervalo de confiança para β_1 com $(1 - \alpha)100\%$ é dado por

$$\left[\hat{\beta}_1 - t_{\left(1 - \frac{\alpha}{2}; n - 2\right)} \sqrt{\frac{QME}{\sum_{i=1}^n (x_i - \bar{x})^2}}; \hat{\beta}_1 + t_{\left(1 - \frac{\alpha}{2}; n - 2\right)} \sqrt{\frac{QME}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

2.4 Análise de variância

Segundo [Ferreira \(2009\)](#) a análise de variância (ANOVA) é uma forma de dividir a variância total em componentes devidos à regressão linear e ao resíduo. Essa partição é obtida para atender a uma série de objetivos. Um desses objetivos é verificar se a parte da variação total explicada pelo modelo é significativamente diferente de zero. Esse teste, todavia, para o modelo de regressão linear simples, é equivalente ao teste da igualdade do coeficiente de regressão a zero. A análise de variância é uma das técnicas de estatística experimental que mais se destacam. Diversos testes de hipótese sobre parâmetro populacionais de modelos lineares são realizados pela análise de variância, que tem por objetivo dividir a variação total em fontes controladas pelo pesquisador e em fontes aleatórias de variação.

A análise de variância é baseada na decomposição da soma de quadrados e nos graus de liberdade associados a variável resposta Y . Em palavras, o desvio de uma observação em relação à média pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão mais o desvio do valor ajustado em relação à média, isto é, podemos escrever $(Y_i - \bar{Y})$ como

$$(Y_i - \bar{Y}) = (Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i). \quad (2.4.1)$$

2.4.1 Soma de Quadrados

Elevando cada componente de (2.4.1) ao quadrado e somando para todo o conjunto de observações, obtemos

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

em que

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = SQT \quad (\text{é a Soma de Quadrados Total});$$

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = SQR \quad (\text{é a Soma de Quadrados da Regressão}) \text{ e}$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = SQE \quad (\text{é a Soma de Quadrados dos Erros (dos Resíduos)}).$$

Desta forma, escrevemos

$$SQT = SQR + SQE,$$

em que decompos a Soma de Quadrados Total em Soma de Quadrados da Regressão e Soma de Quadrados dos Erros.

Prova:

$$\begin{aligned} SQT &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i)^2 = \sum_{i=1}^n ((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \end{aligned}$$

Notemos que

$$\sum_{i=1}^n 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^n 2(Y_i \hat{Y}_i - Y_i \bar{Y} - \hat{Y}_i^2 + \hat{Y}_i \bar{Y}).$$

Como visto em "Algumas propriedades do ajuste de mínimos quadrados",

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = 0 \Rightarrow \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$$

e

$$\sum_{i=1}^n (\hat{Y}_i e_i) = \sum_{i=1}^n \hat{Y}_i (Y_i - \hat{Y}_i) = 0 \Rightarrow \sum_{i=1}^n (\hat{Y}_i Y_i) = \sum_{i=1}^n (\hat{Y}_i^2).$$

Desta forma,

$$\sum_{i=1}^n 2(Y_i \hat{Y}_i - Y_i \bar{Y} - \hat{Y}_i^2 + \hat{Y}_i \bar{Y}) = 2\left(\sum_{i=1}^n \hat{Y}_i^2 - \bar{Y} \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i^2 + \bar{Y} \sum_{i=1}^n \hat{Y}_i\right) =$$

$$= 2(-\bar{Y} \sum_{i=1}^n Y_i + \bar{Y} \sum_{i=1}^n \hat{Y}_i) = 2(-\bar{Y} \sum_{i=1}^n Y_i + \bar{Y} \sum_{i=1}^n Y_i) = 0.$$

e portanto,

$$SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = SQR + SQE.$$

2.4.2 Quadrado Médio

Segundo [Ferreira \(2009\)](#) os quadrados médios referentes a cada fonte de variação são obtidos pela divisão das somas de quadrados correspondentes pelos seus respectivos graus de liberdade.

$$QMR = \frac{SQR}{1} = SQR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (\text{é o Quadrado Médio da Regressão}) \text{ e}$$

$$QME = \frac{SQE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} \quad (\text{é o Quadrado Médio dos Resíduos}).$$

Como visto em "Propriedades dos Estimadores",

$$SQE = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i.$$

Além disso,

$$SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Desta forma,

$$SQR = SQT - SQE = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \left(\sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i \right) = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i,$$

e portanto,

$$QMR = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i \text{ e}$$

$$QME = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i}{n-2}.$$

2.4.3 Tabela de Análise de Variância

De acordo com [Ferreira \(2009\)](#) a tabela de análise de variância para um modelo de regressão linear simples é explicitado por fontes de variação com os respectivos graus de liberdade, somas de quadrados (SQ) e quadrado médio (QM) e a estatística do teste F para a hipótese de nulidade do coeficiente de regressão.

Figura 1 – Tabela a análise de variância para um modelo de regressão linear simples

Fonte de variação	GL	Soma de Quadrados	Quadrado Médio
Regressão	1	$SQR = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i$	$QMR = SQR$
Resíduo	$n - 2$	$SQE = \sum_{i=1}^n (Y_i - \hat{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})Y_i$	$QME = \frac{SQE}{(n - 2)}$
Total	$n - 1$	$SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2$	

2.4.4 Teste F

Segundo [Ferreira \(2009\)](#) a distribuição F de probabilidade é a mais importante na estatística, tendo, ainda, um maior destaque na estatística experimental. Essa distribuição é definida pela variável resultante da razão de duas variáveis aleatórias independentes com distribuição qui-quadrado.

Considerando o Modelo de Regressão Linear Simples, a análise de regressão estabelece um teste para avaliar o parâmetro β_1 , isto é, testar as hipóteses

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Seja

$$\frac{SQT}{\sigma^2} = \frac{SQR}{\sigma^2} + \frac{SQE}{\sigma^2}$$

2.5 Seleção de modelos

O critério de seleção do modelo é necessário para comparar e selecionar o modelo que melhor descreve os dados. Com isso, é comum o ajuste de vários modelos que descreve um fenômeno. Dentre esses critérios, fez-se uso do coeficiente de determinação, coeficiente de determinação ajustado, critério de informação Akaike (AIC) e (BIC).

2.5.1 Coeficiente de Determinação

Segundo [Hoffmann \(2006\)](#) seja R_p^2 a notação do coeficiente de determinação de um modelo com p variáveis explicativas, isto é, p coeficientes e o intercepto β_0 .

$$R_p^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT},$$

em que SQR, SQE e SQT são a soma dos quadrados do modelo, soma dos quadrados dos resíduos (erros) e soma dos quadrados total, respectivamente.

Esse critério utilizando o método é que se acrescentar uma variável não significativa teremos um pequeno aumento de R_p^2 . Logo, ele é mais usado para julgar quando parar de adicionar variáveis do que para encontrar o melhor modelo já que R_p^2 nunca diminui quando p aumenta.

2.5.2 Coeficiente de determinação ajustado

Segundo Draper e Smith (1998) e Ratkowsky (1983) o coeficiente de determinação (R^2) é utilizado para escolher o “melhor” modelo, inconvenientemente, ele não considera no seu cálculo o número de parâmetros presentes no modelo. Logo, uma alternativa é o uso do coeficiente de determinação ajustado (R_α^2) que usa uma ponderação em relação ao número de parâmetros do modelo, sendo assim, estimado por:

$$R_\alpha^2 = 1 - \left[\frac{(1 - R^2)(n - i)}{n - p} \right]$$

em que R^2 é o coeficiente de determinação, n é o número de observações, p é o número de parâmetros do modelo, i está ligado ao ajuste de intercepto na curva, sendo igual a 1 se houver intercepto e 0 para caso contrario.

2.5.3 AIC e BIC

O Critério de Informação de Akaike (AIC) é definido como

$$AIC_p = -2\log(L_p) + 2[(p + 1) + 1],$$

em que L_p é a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas consideradas no modelo.

O Critério de Informação Bayesiano (BIC) é definido como

$$BIC_p = -2\log(L_p) + [(p + 1) + 1]\log(n).$$

Tanto o AIC quanto o BIC aumentam conforme SQE aumenta. Além disso, ambos os critérios penalizam modelos com muitas variáveis sendo que valores menores de AIC e BIC são preferíveis.

Como modelos com mais variáveis tendem a produzir menor SQE mas usam mais parâmetros, a melhor escolha é balancear o ajuste com a quantidade de variáveis.

2.6 Testes estatísticos

Logo após a seleção do modelo, para verificar os pressupostos da análise residual, fez-se alguns testes estatísticos. Entre esses pressupostos tem-se a normalidade, homogeneidade

e independência.

2.6.1 Diagnóstico de normalidade

- **Teste de Shapiro-Wilk** - o teste Shapiro-Wilk (SHAPIRO-WILK, 1965) é utilizado para verificação dos pressupostos de normalidade dos resíduos. Testando-se as hipóteses:

$$\begin{cases} H_0 : & \text{Os resíduos provém de uma distribuição normal} \\ H_1 : & \text{Os resíduos não provem de uma distribuição normal} \end{cases}$$

A estatística teste é dada por

$$W = \frac{b^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

em que x_i são os valores das amostras ordenados e b é uma constante da seguinte forma em que $a_{n+i} - 1$ são constantes geradas por meio das médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição normal e seus valores são estabelecidos. Na tomada de decisão rejeita-se H_0 ao nível de significância α se $W < W_\alpha$, esse W_α é valor crítico da estatística W de Shapiro-Wilk.

2.6.2 Diagnóstico de Homoscedasticidade

A heteroscedasticidade é a falta de homoscedasticidade, ou seja, quando existe heteroscedasticidade as variâncias não são constantes em diferentes observações, fazendo com o que o modelo sofra alguns efeitos em seu ajuste. Sua suposição é testada pelas técnicas abaixo:

- **Gráficos dos resíduos versus valores ajudados** - é uma das principais técnicas utilizadas para verificar as suposições de resíduos. Por meio de alguma tendência nos pontos pode-se identificar se existe heteroscedasticidade da variância dos erros. Com isso, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhuma tendência, temos indícios de que a variância dos resíduos é homoscedástica.
- **Teste de Goldfeld-Quandt** - é utilizado para testar a homoscedasticidade dos resíduos. Porém, neste teste, há exigência de amostra relativamente grande. Segundo [Rodrigues e Diniz \(2006\)](#) as n observações são ordenadas de acordo com os valores da variável regressora, dividi-se a amostra ordenada em 3 partes, em que, a parte do meio deve ter 25% dos dados, 1º contendo os menores valores da variável explicativa e a 3º parte contendo os maiores valores da variável explicativa, em que deve-se apresentar praticamente a mesma quantidade de dados. Assim, ajusta-se dois modelos

de regressão, um com os dados da primeira parte e o outro com os dados da terceira parte. Logo, utiliza-se o teste F, com as seguintes hipóteses: em que, σ_i^2 com $i = 1, 2, 3$ é a variância dos resíduos dos três modelos de regressão. A estatística de teste é dado por

$$F_{GQ} = \frac{SQE^b / (n_3 - (p + 1))}{SQE^a / (n_1 - (p + 1))},$$

em que SQE^a e SQE^b são as somas de quadrados dos resíduos da regressão para o grupo inferior (parte 1) e para o grupo superior (parte 3), respectivamente, n_1 é o número de observações da parte 1 e n_3 é o número de observações da parte 3. Chamamos de d o número de observações omitidas (parte 2). Essa estatística tem distribuição $F_{(n_3 - (p + 1), n_1 - (p + 1))}$. Desta forma, considerando um nível de significância $\alpha = 0,05$, rejeitamos a hipótese nula, ou seja, a hipótese de que as variâncias são iguais se $F_{GQ} > F_{(\alpha)}$.

- **Breusch-Pagan** - é um teste multiplicador de Lagrange, o teste de Breusch-Pagan é muito utilizado para testar a hipótese nula de que as variâncias dos erros são iguais (homoscedasticidade) versus a hipótese alternativa de que as variâncias dos erros são uma função multiplicativa de uma ou mais variáveis, sendo que tais variável(eis) pode(m) pertencer ou não ao modelo estudado. É indicado para grandes amostras e quando a suposição de normalidade nos erros é assumida. Inicialmente, ajustamos o modelo de regressão linear (simples ou múltiplo) e encontramos os resíduos $e = (e_1, \dots, e_n)$ e os valores ajustados $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$. Em seguida, consideramos os resíduos ao quadrado e os padronizamos de modo que a média do vetor de resíduos padronizados, que denotaremos por u , seja 1. Esta padronização é feita dividindo cada resíduo ao quadrado pela SQE/n em que SQE é a Soma de Quadrados dos Resíduos do modelo ajustado e n é o número de observações. Desta forma, temos que cada resíduo padronizado é dado por

$$u_i = \frac{e_i^2}{SQE/n}, \quad i = 1, \dots, n,$$

em que

$$SQE = \sum_{i=1}^n e_i^2.$$

Por fim, fazemos a regressão entre $u=(u_1, \dots, u_n)$ (variável resposta) e o vetor \hat{y} (variável explicativa) e obtemos a estatística do teste χ_{BP}^2 calculando a Soma de Quadrados da Regressão de u sobre \hat{y} e dividindo o valor encontrado por 2. Sob a hipótese nula, esta estatística tem distribuição qui-quadrada com 1 grau de liberdade. Resumidamente, se não existe heteroscedasticidade, é de se esperar que os resíduos ao quadrado não aumentem ou diminuam com o aumento do valor predito, \hat{y} e assim, a estatística de teste deveria ser insignificante.

2.7 Intervalo de confiança para a resposta média e predição

A estimativa de um intervalo de confiança para $E(Y | X = x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$ é de grande interesse.

Um estimador pontual de $\mu_{Y|x_0}$ pode ser obtido a partir do modelo ajustado, isto é,

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{Y}(x_0).$$

Notemos que $\hat{\mu}_{Y|x_0}$ é uma variável aleatória normalmente distribuída já que é uma combinação linear das observações Y_i . Além disso, temos que

$$E(\hat{\mu}_{Y|x_0}) = \beta_0 + \beta_1 x_0 = \mu_{Y|x_0} \text{ e}$$

$$\begin{aligned} \text{Var}(\hat{\mu}_{Y|x_0}) &= \text{Var}[\bar{Y} + \hat{\beta}_1(x_0 - \bar{x})] = \text{Var}[\bar{Y}] + \text{Var}[\hat{\beta}_1(x_0 - \bar{x})] = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \end{aligned}$$

ou seja, $\hat{\mu}_{Y|x_0}$ é um estimador não viciado para $E(Y | X = x_0)$. Assim, temos que

$$\frac{\hat{Y}(x_0) - \mu_{Y|x_0}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim N(0, 1).$$

Temos também que

$$\frac{(n-2)QME}{\sigma^2} \sim \chi^2_{(n-2)}.$$

Logo,

$$t = \frac{N(0, 1)}{\sqrt{\frac{\chi^2_{(n-2)}}{(n-2)}}} = \frac{\frac{\hat{Y}(x_0) - \mu_{Y|x_0}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}}}{\sqrt{\frac{(n-2)QME}{\sigma^2 (n-2)}}} = \frac{[\hat{Y}(x_0) - \mu_{Y|x_0}]}{\sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \sim t_{(n-2)},$$

Portanto, o intervalo de confiança para $\mu_{Y|x_0} = E[Y | X = x_0]$ é dado por

$$\left[\hat{Y}(x_0) - t_{\left(1 - \frac{\alpha}{2}; n - 2\right)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} ; \right. \\ \left. \hat{Y}(x_0) + t_{\left(1 - \frac{\alpha}{2}; n - 2\right)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} \right],$$

em que $\hat{Y}(x_0)$ é a resposta média estimada para o nível $x = x_0$. Considerando vários valores para x_0 dentro do intervalo de realização dos dados, encontraremos vários valores para $\hat{Y}(x_0)$. Com isso, ao calcularmos o intervalo de confiança para cada um dos $\hat{Y}(x_0)$, temos um conjunto de intervalos de confiança que representam as bandas de confiança para a reta de regressão.

2.8 Análise de resíduo na regressão linear simples

Para que uma análise de regressão seja confiável, é importante que as suposições do modelo sejam validas, se não, pode ocorrer não normalidade, presença de pontos atípicos, heterocedasticidade, não independência dos erros, fazendo com que a análise venha a ter conclusões duvidosas. Assim, a análise de resíduo oferece técnicas que nos ajudam a verificar esses imprevistos. Logo, o vetor de resíduo é definido por

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta$$

Assim temos alguns resultados importantes:

$$y \sim N(\mathbf{X}\beta, \sigma^2),$$

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

$$\hat{y}_i = x_i' \hat{\beta} \sim N(x_i' \hat{\beta}, \sigma^2 x_i' (\mathbf{X}'\mathbf{X})^{-1} x_i)$$

Com isso, temos a esperança e a variância dos resíduos respectivamente definidos por

$$E(\epsilon) = E(\mathbf{Y} - \mathbf{X}\beta) = 0$$

$$\text{Var}(\epsilon) = \text{Var}(\mathbf{Y} - \mathbf{X}\beta) = \beta^2(\mathbf{I} - \mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

que pode ser reescrito da forma seguinte

$$\epsilon \sim N(0, \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']).$$

De acordo com Hoffmann (2006), a matriz $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é considerada matriz de projeção \mathbf{H} , em que, é simétrica e idempotente e os valores da diagonal principal da matriz \mathbf{H} são h_{ii} , com $0 < h_{ii} < 1$ e $i = 1, 2, \dots, n$. Em que h_{ii} é o valor observado da influência de x_i a \bar{x} .

Para a verificação das suposições do modelo utiliza-se algumas técnicas, informais (como gráficos) ou formais (como testes), que são mais indicados para se tomar uma decisão. Tendo essas combinações disponíveis, pode-se ter o diagnóstico de problemas para as suposições dos modelos.

3 Material e Métodos

Os dados usados na aplicação deste trabalho foram retirados do Sistema de Informação da Atenção Básica - SIAB (<http://siab.datasus.gov.br/SIAB/index.php>), referente ao período de 2004 à 2013, sobre indivíduos com alguma deficiência de 223 municípios do estado da Paraíba. A porcentagem de indivíduos com deficiência do SIAB foi calculada a partir das informações sobre número absoluto de deficientes e a totalidade de habitantes do município disponíveis no banco de dados do SIAB.

Diversos tipos de bancos de dados de pessoas com deficiência física são disponibilizados nesse sistema, entre eles estão deficiência auditiva, visual e motora, com isso fazendo com que indivíduos sejam incapazes de realizar algumas atividades do dia-dia. No mesmo não existia informações sobre indivíduos com deficiência intelectual, logo, não se sabe se esses dados sobre deficiência física usados para tal análise incluem ou não a deficiência mental.

Para a realização das análises, inicialmente, fez-se um teste de sequências (*runs test*), que por sua vez encontra-se disponível no pacote `randtests`, com isso, após tal teste, fez-se a retirada de 78 municípios que apresentavam não aleatoriedade, assim, dos 223 municípios disponíveis no bando de dados, apenas 145 municípios foram analisados neste trabalho. Foi utilizado o *software* estatístico R (versão 2.12.0). Uma das técnicas da estatística usada nesse banco de dados foi a regressão linear simples para analisar a porcentagem de indivíduos com deficiência nos diversos municípios, como variável dependente, considerando como variável independente o número de habitantes, e distância do município à capital João Pessoa. Fez-se uma análise gráfica dos resíduos para verificar a validade dos modelos de regressão, e para a suposição de homoscedasticidade (variância constante), foram aplicados os testes de Goldfeld-Quandt e Breusch-Pagan através do pacote `lmtest` do *software* de programação estatística R (<http://www.r-project.org>).

4 Resultados Obtidos

Como foi mencionado anteriormente, o trabalho apresentado aqui consistiu de duas partes. A primeira tratava da avaliação da aleatoriedade dos dados por município do estado da Paraíba, que foi desenvolvido como parte do trabalho de iniciação científica intitulado por Testes de Aleatoriedade para os dados do SIAB e análise sobre a interiorização da deficiência física no Brasil no ano de 2015 e as análises de regressão linear simples com os dados que parecem apresentar aleatoriedade na primeira fase do trabalho.

Desta forma, os resultados estão apresentados aqui em duas seções, a primeira que trata dos resultados dos testes de aleatoriedade e a segunda que trata das análises de regressão.

4.1 Resultados dos testes de aleatoriedade

Inicialmente fizemos uma análise preliminar, tanto usando a média como a mediana das observações para definir o ponto de corte para se avaliar como o teste de sequências se comportava. Nesta etapa percebemos que para alguns municípios os níveis descritivos dos testes, tanto com a distribuição exata como na versão assintótica, se apresentavam maiores que o valor um, o que claramente é um erro teórico, dado que o nível descritivo do teste é uma probabilidade e obrigatoriamente deve estar contido no intervalo $(0, 1)$. Isso acontecia devido a um pequeno problema na implementação da função que estávamos usando, que foi corrigido com uma pequena adaptação feita no código.

Nesta mesma análise preliminar, também avaliamos o efeito do uso da mediana e da média na definição do limiar para classificar as observações como positivas ou negativas na realização do teste de sequências, onde percebemos que o uso da mediana era um pouco menos rigorosa, classificando menos municípios como não aleatórios. Isso pode ser explicado pelo fato de a mediana ser menos susceptível a presença de pontos discrepantes entre as observações, o que acontecia em alguns casos. Desta forma, optamos pelo uso da mediana na realização dos testes, cujos resultados são apresentados a seguir.

Na primeira análise realizada com o número total de habitantes de cada município coletado do SIAB, observamos 52 desses municípios considerados como não aleatórios através do teste de sequências, de modo que obtiveram níveis descritivos do teste abaixo de 0,05. A lista destas cidades está relacionada na Tabela 1. Os 171 municípios restantes, que podem ser considerados como tendo um padrão aleatório para o número total de habitantes ao longo do período observado, foram utilizados para análise posterior considerando a proporção populacional de deficientes físicos na análise.

Após a primeira filtragem realizada com o número total de habitantes de cada município, partimos para segunda etapa, que consistiu em aplicar o mesmo teste de aleatoriedade agora para a proporção populacional de deficientes físicos dos municípios restantes. Como já mencionado anteriormente, ainda utilizando a mediana no teste e cálculo dos níveis descritivos pela distribuição exata, temos descrito na Tabela 2, a lista dos 26 municípios, do total de 171 que ficaram para esta etapa da análise, que foram considerados como não aleatórios, novamente considerando aqueles que apresentaram níveis descritivos menores que 0,05.

No teste para avaliar a aleatoriedade dos dados, sempre que um valor da amostra coincide com o valor do ponto de corte para classificar as observações como positivas ou negativas, tal valor é removido da análise resultando em uma redução do número de observações usadas no respectivo cálculo. No nosso estudo em particular, uma ressalva importante deve ser feita para o município de Caiçara, onde o valor da mediana da proporção de deficientes físicos foi de 0,0094, que coincidia exatamente com os valores observados consecutivamente para sete anos dentre os dez observados (de 2006 até 2012) de modo que estas sete observações foram excluídas da análise, sendo analisados apenas as três observações restantes (para os anos de 2004, 2005 e 2013), o que resultou em um nível descritivo inexistente para o teste.

Assim, apenas para a cidade de Caiçara, o teste de sequências acabou sendo

Tabela 1 – Municípios considerados não aleatórios através do teste de aleatoriedade com níveis descritivos inferiores a 0,05 para o total populacional.

Municípios do Estado da Paraíba		
Aparecida	Araçagi	Baía da Traição
Bananeiras	Baraúna	Boa Vista
Brejo do Cruz	Brejo dos Santos	Caaporã
Cachoeira dos Índios	Caldas Brandão	Campina Grande
Capim	Carrapateira	Catingueira
Caturité	Coxixola	Cuitegi
Curral Velho	Diamante	Emas
Esperança	Guarabira	Imaculada
Itabaiana	Itapororoca	Jacaraú
João Pessoa	Logradouro	Malta
Mogeiro	Montadas	Nova Floresta
Olivedos	Pocinhos	Pombal
Prata	Princesa Isabel	Puxinanã
Riachão	Riachão do Poço	Rio Tinto
Salgado de São Félix	São Domingos do Cariri	São João do Cariri
São José de Caiana	São José de Princesa	São José dos Ramos
Soledade	Sossêgo	Várzea
Vista Serrana		

Tabela 2 – Municípios considerados não aleatórios através do teste de aleatoriedade com níveis descritivos inferiores a 0,05 agora para a proporção populacional.

Municípios do Estado da Paraíba		
Alagoa Nova	Alagoinha	Arara
Boa Ventura	Caçara	Cajazeirinhas
Conceição	Curral de Cima	Duas Estradas
Ingá	Jericó	Livramento
Lucena	Mataraca	Paulista
Pedra Branca	Pedra Lavrada	Pilões
Pilõezinhos	Pitimbu	São Domingos
São José do Bonfim	Serra Grande	Serra Redonda
Teixeira	Uiraúna	

inconclusivo e a fim de tentar entender melhor o problema resolvemos reaplicar o teste para este município definindo o limiar manualmente para 0,0095, de modo que as sete observações problemáticas não fossem removidas da análise. Isso resultou em um nível descritivo de aproximadamente 0,07; e dado que este valor está muito próximo do limite de 0,05 e a repetição da proporção de deficientes físicos ao longo de sete anos consecutivos, optamos por manter o município de Caçara como não aleatório, e assim constando da Tabela 2.

A seguir apresentamos a Figura 2 que traz uma representação esquemática dos testes realizados e descreve o número e porcentagens de municípios considerados aleatórios por meio do teste de sequências para o total populacional e para a proporção de deficientes físicos nas duas etapas do trabalho utilizando como fonte os dados obtidos do SIAB.

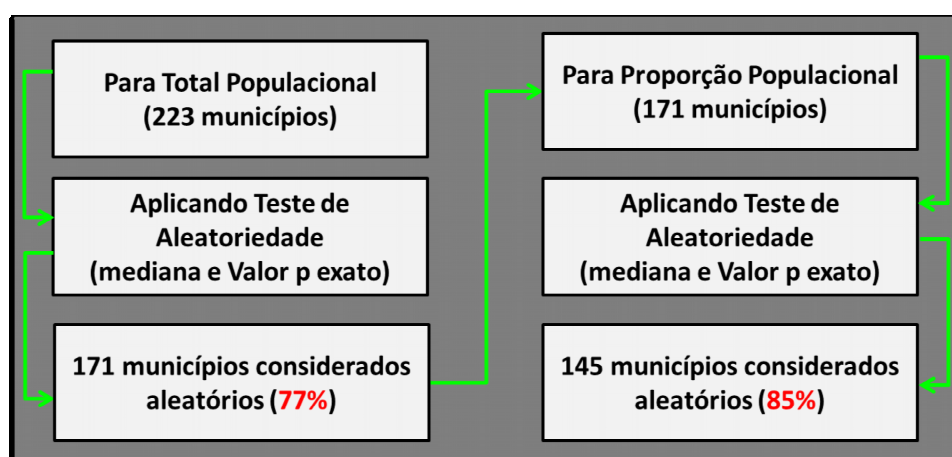


Figura 2 – Disposição gráfica do trabalho realizado através do teste de sequências, para se verificar a aleatoriedade dos dados observados.

4.2 Resultados dos modelos de regressão linear simples

Inicialmente, fez-se o teste de seleção do modelo para verificar qual modelo melhor descreve os dados, comparando os modelos dos anos de 2010 e 2013, respectivamente, através do coeficiente de determinação ajustado, obteve-se um $R_{\alpha}^2 = 0,0128$ e $R_{\alpha}^2 = 0,0103$, nota-se que por esse teste de seleção o maior R_{α}^2 é o do ano de 2010, porém, pelo critério de informação Akaike (AIC) e o de Bayesiano (BIC), obteve-se um $AIC = 221.834$ e $BIC = 230.7642$, e $AIC = 257.4314$ e $BIC = 266.3616$, podemos observar que comparando os modelos através desses critérios, o melhor modelo foi o de 2013, pois obteve um AIC maior, sendo assim, o critério de AIC mais forte, verifica-se que o modelo de 2013 foi o melhor.

No diagnóstico de normalidade para a verificação dos pressupostos de normalidade, para os anos de 2010 e 2013, respectivamente, fez-se o uso dos gráficos de resíduos Figuras 3 e 4. Para verificar se realmente a variância dos resíduos é homocedástica, fez-se o uso do teste de Goldfeld-Quandt e Breusch-Pagan, que para o ano de 2010 os resultados foram $GQ = 1,8349$, o valor $P = 0,005 < \alpha = 0,05$ e $BP = 0,8123$, o valor $P = 0,3674 > \alpha = 0,05$, e para o ano de 2013 $GQ = 1,4455$, o valor $P = 0,0622 > \alpha = 0,05$ e $BP = 0,4011$, o valor $P = 0,5265 > \alpha = 0,05$, onde pode-se dizer que há indícios que a variância dos resíduos é homocedástica. Com isso, é preciso fazer um teste que identifique a normalidade dos resíduos com mais precisão, assim, fez-se o uso do teste de Shapiro-Wilk, obtendo-se $W = 0,9816$ com valor $P = 0,0488 < \alpha = 0,05$ e $W = 0,9786$ com valor $P = 0,0227 < \alpha = 0,05$. Logo, nota-se que há indícios para se rejeitar a hipótese nula, e dizemos que os resíduos não seguem uma distribuição normal.

Através da estatística descritiva dos dados para o ano de 2010, têm-se uma média da porcentagem de indivíduos com deficiência (PID) no estado da Paraíba de 1,48% com amplitude de 0% (município de Santa Cruz) e 3,18% (município de Santana dos Garrotes). Na análise de regressão mostra-se que há uma dependência linear entre a (PID) e a distância do município à capital João Pessoa (aumento 0,03% por km; $p=0,175$; Figura 5). Logo, isso significa que há um aumento médio de 0,03% de indivíduos com deficiência física a cada 1 km no sentido do interior do estado, quando os municípios considerados não aleatórios foram retirados, com um p-valor não significativo, porém tendo uma regressão positiva e linear.

As mesmas análises foram feitas para o ano de 2013 para uma melhor visualização do que ocorre à medida que os anos passam, com isso, a média do PID é de 1,45% com amplitude 0% (município de Araruna) e 2,98% (município de Santana de Mangueira). Na análise de regressão os resultados foram estatisticamente os mesmos, com uma diferença no p-valor que foi de ($p=0,224$; Figura 6), também não é significativo, porém não negativo e linear.

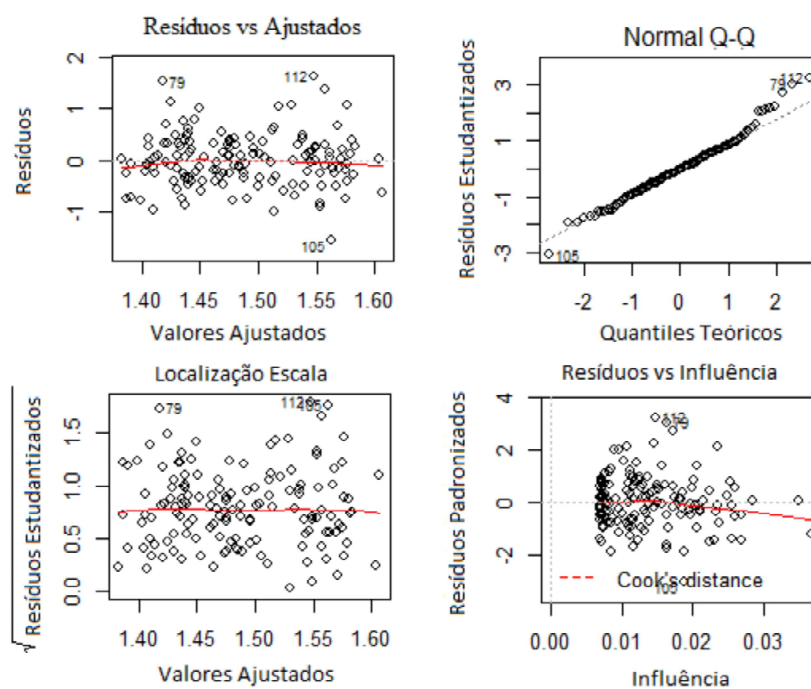


Figura 3 – Gráfico de resíduo da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2010.

Analisando-se a proporção dos anos seguintes anos, vê-se que de 2004 à 2009 existe estatisticamente uma significância que quanto mais próximo à capital João Pessoa, menor é a proporção de deficientes, indicando um processo de interiorização da deficiência, com isso, os anos de 2010 à 2013 obteve-se uma não significância estatisticamente observando-se na Tabela 3.

Tabela 3 – O efeito da distância da capital aos municípios sobre a PID para cada 1km de distância da capital a cada município há um evento de $\beta_1\%$ de PID.

Ano	Signif (p-valor)
2004	Sim (0,005)
2005	Sim (0,001)
2006	Sim (0,002)
2007	Sim (0,002)
2008	Sim (0,002)
2009	Sim (0,010)
2010	Não (0,175)
2011	Não (0,275)
2012	Não (0,063)
2013	Não (0,224)

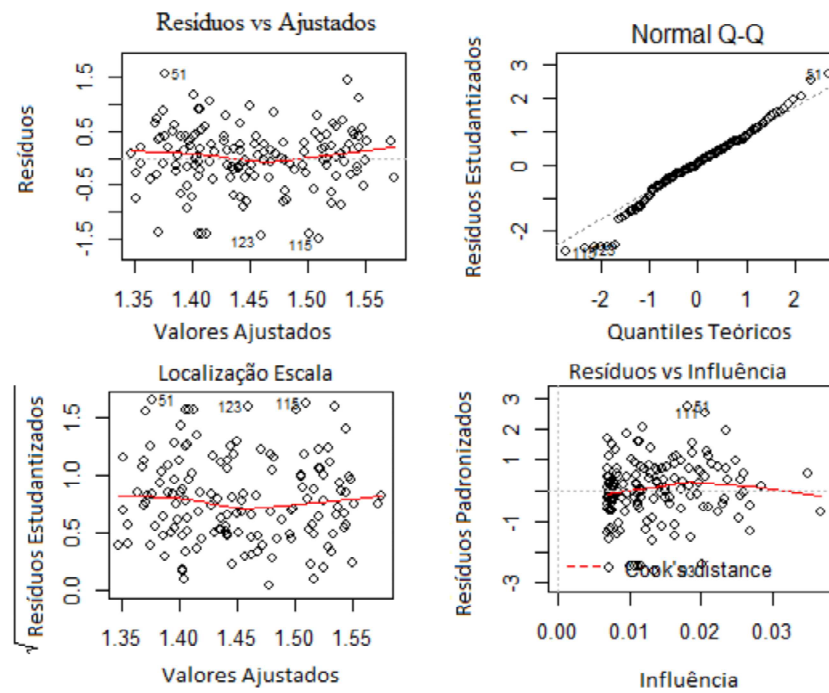


Figura 4 – Gráfico de resíduo da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2013.

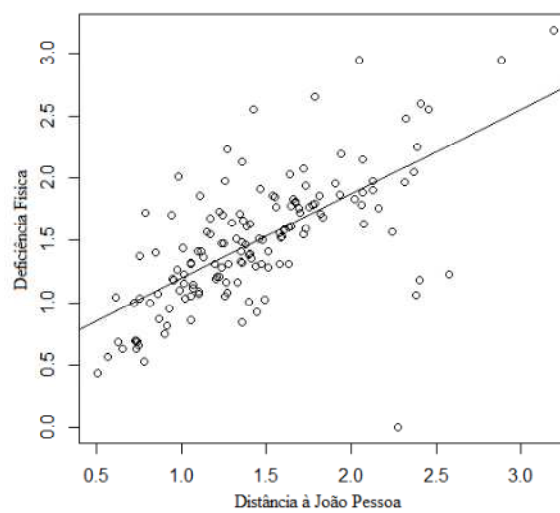


Figura 5 – Gráfico da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2010.

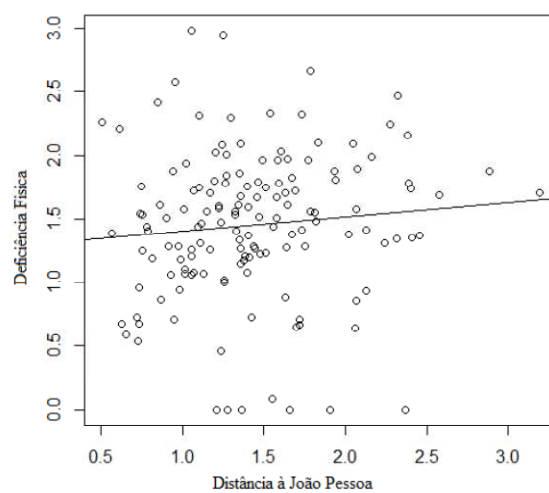


Figura 6 – Gráfico da porcentagem de deficientes físicos contra quanto a distância de João Pessoa para o ano de 2013.

5 Conclusão

Neste trabalho visou-se aplicar um modelo de regressão linear simples aos dados do Sistema de Informação da Atenção Básica - SIAB referente ao período de 2004 à 2013, referente a proporção de indivíduos com alguma deficiência de 223 municípios do estado da Paraíba, para saber se há alguma interiorização da deficiência no estado. No teste de seleção de modelo através do AIC, verificou-se que o modelo que melhor descreve os dados foi o de 2013. Na verificação dos pressupostos de normalidade, para os anos de 2010 e 2013, o ano de 2013 tem variância homocedástica. Concluiu-se que quanto mais próximo da capital, João Pessoa, menor a proporção de pessoas com deficiência física, com isso, existe a hipótese de uma interiorização da deficiência no estado da Paraíba.

Referências

- AMIRALIAN, M. et al. Conceituando deficiência. *Rev. Saúde Pública*, v. 34, n. 1, p. 97–103, 2000. Citado na página 11.
- BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 5. ed. São Paulo: Editora Saraiva, 2005. Citado na página 21.
- DEMÉTRIO, C. G. B.; ZOCHI, S. S. *Modelo de Regressão*. Piracicaba, 2008. Citado na página 14.
- DRAPER, N. R.; SMITH, H. *Applied regression analysis*. 3. ed. New York: John Wiley, 1998. Citado na página 28.
- FERREIRA, D. F. *Estatística Básica*. 2. ed. Lavras: Ed. UFLA, 2009. Citado 4 vezes nas páginas 16, 24, 26 e 27.
- GIBBONS, J. D.; CHAKRABORTI, S. *Nonparametric Statistical Inference*. 4. ed. New York: Marcel Dekker, Inc, 2003. Citado na página 13.
- HOFFMANN, R. *Análise de Regressão : uma introdução à econometria*. 4. ed. São Paulo: Hucitec, 2006. Citado 4 vezes nas páginas 14, 15, 27 e 33.
- MELO, U. S. *A interiorização da deficiência: análise dados do SIAB na Paraíba*. Dissertação (Trabalho de Conclusão de Curso) — Universidade Estadual da Paraíba, Campina Grande, 2010. Citado na página 12.
- Ministério da Saúde. *Manual do Sistema de Informação da Atenção Básica*. Brasília, 2000. Citado na página 11.
- NÉRI, M. *Retratos da deficiência no Brasil (PPD)*. Rio de Janeiro: FGV/IBRE/CPS, 2003. Citado na página 12.
- Organização Mundial da Saúde. *CIF: classificação internacional de funcionalidade, incapacidade e saúde*. São Paulo: Edusp, 2003. Citado na página 11.
- RATKOWSKY, D. A. *Nonlinear regression modeling*. New York: Marcel Dekker, 1983. Citado na página 28.
- RODRIGUES, S. A.; DINIZ, C. A. R. Modelos de regressão heterocedásticos. *Revista de Matemática e Estatística*, v. 24, n. 2, p. 133–146, 2006. Citado na página 29.
- WERNER, D. *Guia de deficiências e reabilitação simplificada*. [S.l.]: CORDE, 1994. Citado na página 11.
- World Health Organization. *Disability and Rehabilitation: Future, Trends and Challenges in Rehabilitation*. Geneva, 2002. Citado na página 11.

APÊNDICE A – *Script* para análise no *software* R

Este apêndice apresenta a sequência de comandos utilizada no *software* R para as análises de regressão desenvolvidas neste trabalho.

```
### Limpar memoria ###
```

```
rm(list=ls(all=TRUE))
```

```
### Chamando banco de dados ###
```

```
Dados<-read.table("C:\\Users\\Allana Livia\\  
Desktop\\Allana\\dados_analise_2.csv",  
header=T, dec=",", sep=";")
```

```
### Atribuindo nomes a cada coluna do banco de dados ###
```

```
prop2004<-(Dados[,2])  
prop2004
```

```
prop2005<-(Dados[,3])  
prop2005
```

```
prop2006<-(Dados[,4])  
prop2006
```

```
prop2007<-(Dados[,5])  
prop2007
```

```
prop2008<-(Dados[,6])  
prop2008
```

```
prop2009<-(Dados[,7])  
prop2009
```

```
prop2010<-(Dados[,8])  
prop2010
```

```
prop2011<-(Dados[,9])  
prop2011
```

```
prop2012<-(Dados[,10])  
prop2012
```

```
prop2013<-(Dados[,11])  
prop2013
```

```
dist<-(Dados[,13])  
dist
```

```
### Organizando os dados para a regressão ###
```

```
### Onde y é a proporção de indivíduos com deficiência ###
```

```
### x é a distância do município em relação à João Pessoa ###
```

```
x1=dist  
x1
```

```
## Ajustando os modelos ##
```

```
y1=prop2004  
y1
```

```
model2004=lm(y1~x1)
summary(model2004)
#####
```

```
y2=prop2005
y2
```

```
model2005=lm(y2~x1)
summary(model2005)
#####
```

```
y3=prop2006
y3
```

```
model2006=lm(y3~x1)
summary(model2006)
#####
```

```
y4=prop2007
y4
```

```
model2007=lm(y4~x1)
summary(model2007)
#####
```

```
y5=prop2008
y5
```

```
model2008=lm(y5~x1)
summary(model2008)
#####
```

```
y6=prop2009
y6
```

```
model2009=lm(y6~x1)
summary(model2009)
#####
```

```
y7=prop2010
```

```
y7
```

```
model2010=lm(y7~x1)
```

```
summary(model2010)
```

```
#####
```

```
y8=prop2011
```

```
y8
```

```
model2011=lm(y8~x1)
```

```
summary(model2011)
```

```
#####
```

```
y9=prop2012
```

```
y9
```

```
model2012=lm(y9~x1)
```

```
summary(model2012)
```

```
#####
```

```
y0=prop2013
```

```
y0
```

```
model2013=lm(y0~x1)
```

```
summary(model2013)
```

```
#####
```

```
### Para os dados de 2010 ###
```

```
### Coeficiente de determinação ###
```

```
cor(y7,x1,use="complete.obs")
```

```
#####
```

```
### Checando as pressuposições do modelo(Multicolinearidade)###
```



```
require(car)
library(car)
vif(model2010)

#####

### Criterio de seleção do modelo ##

library(MASS)
stepAIC(model2010)
AIC(model2010)
require(bbmle)
require(stats4)
BIC(model2010)

#####

##Teste de má especificação do modelo###

library(lmtest)
resettest(model2010,power=2:3,type="fitted")
#####

###Teste de Heterocedaticidade###

#Goldfeld-Quandt#
gqtest(model2010)

#Koenker#
bptest(model2010,studentize=TRUE)

#Breusch-Pagau#
bptest(model2010,studentize=FALSE)

####Normalidade###
shapiro.test(residuals(model2010))

#####
```

```
###Fim de 2010##
```

```
#####
```

```
##Para os dados de 2013##
```

```
##Coeficiente de determinação###
```

```
cor(y0,x1,use="complete.obs")
```

```
#####
```

```
###Checando as pressuposições do modelo(Multicolinearidade)###
```

```
require(car)
```

```
library(car)
```

```
vif(model2013)
```

```
#####
```

```
###Critério de seleção do modelo##
```

```
library(MASS)
```

```
stepAIC(model2013)
```

```
AIC(model2013)
```

```
require(bbmle)
```

```
require(stats4)
```

```
BIC(model2013)
```

```
#####
```

```
##Teste de má especificação do modelo###
```

```
library(lmtest)
```

```
resettest(model2013,power=2:3,type="fitted")
```

```
#####
```

```
###Teste de Heterocedaticidade###
```

```
#Goldfeld-Quandt#
```

```
gqtest(model2013)

#Koenker#
bptest(model2013,studentize=TRUE)

#Breusch-Pagau#
bptest(model2013,studentize=FALSE)

####Normalidade###
shapiro.test(residuals(model2013))

#####

##Para 2010##
Dados

str(Dados)
which.max(Dados[,8])
Dados[112,]
which.min(Dados[,8])
Dados[105,]

##Para 2013##

which.max(Dados[,11])
Dados[111,]
which.min(Dados[,11])
Dados[8,]
```