Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Abraão de Paula Taveira

Uso de regressão linear múltipla para caracterizar a relação funcional entre o tamanho de cérebro e o nível de inteligência

Campina Grande/PB
2015

Abraão de Paula Taveira

Uso de regressão linear múltipla para caracterizar a relação funcional entre o tamanho de cérebro e o nível de inteligência

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.

Orientadora:

Prof [a]. Dr[a]. Ana Patrícia Bastos Peixoto

Co - orientador:

Prof [o]. Dr. Tiago Almeida de Oliveira

Campina Grande

2015

T232u Taveira, Abraão de Paula.
Uso de regressão linear múltipla para caracterizar a relação funcional entre o tamanho do cérebro e o nível de inteligência. [manuscrito] / Abraão de Paula Taveira. - 2015.
43 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2015.
"Orientação: Profa. Dra. Ana Patricia Bastos Peixoto, Departamento de Estatística".
"Co-Orientação: Prof. Dr. Tiago Almeida de Oliveira, Departamento de Estatística".
1. Análise de regressão. 2. Nível de inteligência. 3. Testes estatísticos. I. Título.

21. ed. CDD 519

Abraão de Paula Taveira

Uso de regressão linear múltipla para caracterizar a relação funcional entre tamanho do cérebro e o nível de inteligência

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: 30 / 06 / 15

Banca Examinadora:

Ana Patrícia Bastos Peixoto

Prof ª. Drª. Ana Patrícia Bastos Peixoto
Orientadora

Ricardo Alves de Olinda

Prof º. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Wanessa Weridiana da Luz Freitas

Prof ª. Msc. Wanessa Weridiana da Luz Freitas
Universidade Estadual da Paraíba

Dedicatória

Dedico este trabalho a Deus, aos meus familiares e amigos por contribuírem grandiosamente na minha formação acadêmica.

Agradecimentos

A realização deste trabalho tornou-se possível devido a colaboração primordial do nosso pai, Deus, que está no comando de tudo, sendo autor e consumador de nossa capacidade intelectual.

Aos meus amados pais José Cardoso Taveira e Jeruza de Paula Taveira, pela transmissão exacerbada de carinho, amor e motivação durante minha formação acadêmica.

As minhas apreciadas irmãs Ana Lídia de Paula Taveira Henriques e Lidiane de Paula Taveira, pela compreensão na minha ausência em momentos especiais da nossa vida.

Aos inesquecíveis orientadores, Profª. Drª. Ana Patrícia Bastos Peixoto e Profº. Dr. Tiago Almeida de Oliveira, pela expressiva competência, disposição e auxílio durante a elaboração deste trabalho.

Aos membros da banca examinadora formada pelo Profº. Dr. Ricardo Alves de Olinda e Profª. Wanessa Weridiana da Luz Freitas por participarem como interlocutores desse trabalho, contribuindo assim para o enriquecimento teórico do mesmo.

Aos meus amigos, do Curso Bacharelado em Estatística, em especial, Rosendo Chagas, Fábio Sandro Dos Santos, Leomir Ferreira Sousa pelo constante apoio e estímulo transmitido durante a construção das atividades teóricas e práticas do trabalho em epígrafe.

A todos o meu sincero agradecimento.

Resumo

Alguns estudos tem mostrado relação positiva e moderada entre a dimensão do crânio e a inteligência medida nos seres humanos. Para tanto, várias variáveis devem ser levadas em consideração na obtenção de tais afirmações. Uma das possibilidades de avaliação das caracteristicas inerentes a estes casos pode ser o uso da análise de regressão múltipla, a qual é utilizada quando se tem o interesse de estudar relações entre uma variável resposta, sobre outras variáveis que influenciam no seu comportamento. Com isso, o objetivo desse trabalho foi verificar se existe relação entre o tamanho do cérebro e o nível de inteligência realizado com dados provenientes de uma pesquisa contendo 40 estudantes do curso de psicologia, com o intuito de ajustar um modelo de regressão linear que melhor se apropriasse aos dados. Realizou-se o teste F de *Snedecor* e o teste t de *Student*, para identificar e selecionar o modelo mais adequado, além do critério de informação de Akaike. Após o ajuste do modelo, utilizou-se o fator de inflação de variância e o método de seleção de variáveis Stepwise para identificar quais variáveis independentes eram significativas para representar a variável dependente nível de inteligência. Em seguida, verificou-se as pressuposições e constatou-se por meio dos testes que o modelo ajustado é confiável. O modelo ajustado conteve as variáveis regressoras gênero, pontuação de coeficiente de inteligência de desempenho com base no quatro Wechsler (PIQ), altura.

Palavras-chave: Análise de regressão. Nível de inteligência. Testes estatísticos.

Abstract

This work discusses some applications the technique of multiple regression analysis, being used when have the interest to study relationships between a variable (dependent variable) on other variables that influence the behavior of the response variable. The study was conducted with data from a survey containing 40 undergraduate students of psychology, whose aim was to establish the relationship between brain size and intelligence level. Whose objective are to verify that exist relationship between brain size and intelligence level, applied the test F of *Snedecor* and the test t of *Student* and Akaike information criterion. After the model fit to the data in issue using the inflation factor of variance and the stepwise variable selection method to identify independent variables were significant to represent the dependent variable level of intelligence. Then, the assumptions are verified and it was found through the tests that the adjusted model adequate. The adjusted model contained the regressive variable, gender, performance intelligence quotient score based on four Wechsler (PIQ), height.

Key-words: Multiple regression analysis. Level of intelligence. Statistical tests.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O que determina o grau de inteligência são as ligações entre os mais de 100 bilhões de neurônios, que são células responsáveis pelo impulso nervoso que transmite informação, com outras células do cérebro. Essas ligações são chamadas de sinapses. Entende-se que, quanto mais ligações, mais inteligente se fica. ou até mesmo que a inteligência melhora ao longo do tempo. Estudos tem mostrado relação positiva e moderada entre a dimensão do crânio e a inteligência medida por escala psicométrica (PIKE, 2013). Uma suposição encontrada em muitos dos trabalhos científicos é que o tamanho da caixa craniana e do cérebro são substituíveis, e se o porte da caixa craniana não for uma poderosa ferramenta da predição da inteligência, o tamanho do cérebro não será uma ferramenta melhor (PETERS, 1995).

Qquando se tem o interesse em estudar à predição da variável dependente, em relação em duas ou mais variáveis independentes, aplica-se a análise de regressão linear múltipla. Com esta análise, também é possível reduzir um número de variáveis para poucas dimensões com o mínimo de perda de informação, identificando os principais padrões de similaridade, associação e correlação entre as variáveis. Por meio da análise de regressão linear múltipla é possível ainda realizar predição, seleção de variáveis, estimação dos parâmetros, realizar inferências tais como testes de hipóteses e intervalos de confiança.

Na regressão múltipla mesmo em algumas situações em que não são encontradas relações casuais entre as variáveis, pode-se utilizar algumas expressões matemáticas, que são úteis para a estimação do valor da variável resposta, quando se tem conhecimento dos valores das variáveis independentes (HOFFMANN; VIEIRA, 1998). O resultado final dessa regressão é uma equação da reta que representa a melhor predição de uma variável dependente a partir de diversas variáveis independentes. Esta equação representa um modelo aditivo, no qual as variáveis preditoras somam-se na explicação da variável critério.

Diante do exposto, este trabalho teve como objetivo, ajustar um modelo de regressão linear múltipla aos dados referentes a um estudo feito por Willerman *et al.* (1991), or meio de critérios de seleção de variáveis *stepwise*, AIC e verificar a relação entre o tamanho do cérebro e o nível de inteligência.

2 Fundamentação Teórica

O conteúdo desta seção relata os principais aspectos da utilização dos modelos de regressão múltipla por meio de artigos práticos, teóricos e livros texto relacionados ao objetivo da pesquisa.

2.1 Marco Histórico

Sendo considerado como um dos personagens mais importantes na evolução da estatística, Sir Francis Galton era o mais novo de nove filhos de um próspero banqueiro, nasceu em uma família socialmente abastada. Apesar de ter um interesse por medicina, teve sua primeira formação acadêmica em Matemática.

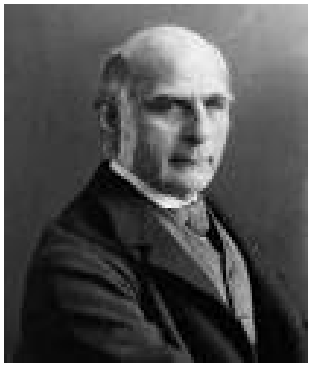

Figura 1: Imagem de Francis Galton. Fonte: Pearson (1930).

Durante sua vida, Galton produziu mais de 340 artigos e livros, além de realizar várias pesquisas, dentre estas, uma que teve suma importância para a evolução da Estatística, criando o conceito estatístico de correlação. Segundo Memória (2004), Galton identificou que a distribuição normal é completamente determinada pela mediana e o desvio semiquartílico.

De acordo com Memória (2004), em um dos seus trabalhos, Galton estudou a relação entre a altura dos pais e dos filhos, procurando saber como a altura do pai influenciava na altura do filho, usando pela primeira vez o termo regressão, pois observando que

os contornos de igual frequência eram constituídos por elipses concêntricas semelhantes dispostas e traçou a linha de regressão à mão.

De acordo com Santos *et al.* (2003), os conhecimentos matemáticos de Galton não eram suficientes, então com o auxílio do professor J.D. Hamilton Dickson, Galton propôs a expressão exata da correlação, sendo esta, modificada pelo professor Walter Frank Raphael Weldon. Por fim, no ano de 1896, a fórmula do coeficiente de correlação foi definida, por Karl Pearson.

Diante dos estudos realizados, a técnica da análise de dados por meio do modelos de regressão tornou-se de comum uso em análises estatísticas. A regressão linear é um método estatístico utilizado para estimar possíveis relações existentes, entre uma variável, a qual é denominada como resposta (Y) em função de outras variáveis, denominadas como variáveis explanatórias ou independentes (X's). Estas análises são realizadas através de modelos estatísticos, onde a partir destes, é possível realizar predição, seleção de variáveis, estimação de parâmetros e inferência. Quando são estudadas as relações entre apenas duas variáveis, a regressão é denominada simples, se determinada variável é analisada em função de duas ou mais variáveis, a regressão é chamada múltipla.

2.2 Modelo de Regressão Linear Simples (MRLS)

Análise de regressão linear simples é um método estatístico que utiliza um modelo estatístico para estudar a relação entre duas variáveis, cujo objetivo é verificar a dependência da variável resposta (Y) em relação a uma variável explicativa (X). Deste modo, o modelo de regressão linear simples é dado por

$$\mathrm{y}_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \; i = 1, 2, ..., n,$$

em que, y_i é a variável resposta para a i-ésima observação; X_i representa cada observação da variável explicativa; β_0 é chamado de intercepto ou coeficiente linear, representando o coeficiente linear da reta, ou seja, o ponto onde a reta corta o eixo Y, quando x=0; β_1 representa o coeficiente angular da reta, isto é, o grau que a reta faz com o eixo X, e define também o quanto aumenta ou diminui, o valor de y_i em relação a X; ε_i é o erro associado a cada observação em relação à reta de regressão linear.

2.2.1 Estimação dos Parâmetros

A estimação dos parâmetros tem por objetivo estimar valores precisos para β_0 e β_1. Esta necessidade é proveniente de uma análise realizada com amostras coletadas em uma população. Com a estimação dos parâmetros é possível reduzir a soma das distâncias entre a função linear e os pontos observados na amostra. Para se obter as estimativas dos parâmetros pode-se utilizado o método dos mínimos quadrados e/ou método da máxima verossimilhança.

O método de mínimos quadrados tem por objetivo encontrar estimativas dos parâmetros para explicar a relação linear entre as variáveis, por minimizar a soma dos quadrados dos resíduos da regressão, de forma a maximizar o grau de ajuste do modelo aos dados observados. Para aplicação do método de mínimos quadrados. É necessário que os erros sejam distribuídos aleatoriamente, onde esta distribuição seja normal e independente.

De acordo com Demétrio e Zocchi (2006), os valores estimados para β_0 e β_1, usando a norma euclideana para avaliar o comprimento de ε, são obtidos da seguinte forma:

$$Z = ||\varepsilon||^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[Y_i - E\left(Y_i\right)\right]^2 = \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1 X_i\right]^2.$$

Portanto, para estimar β_0 e β_1 tais que Z seja mínima. Para isso, obtêm-se as derivadas parciais: (maiores informações ver DEMÉTRIO, ZOCCHI, 2006).

$$\begin{aligned} \frac{\partial Z}{\partial \beta_0} &= 2 \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1 X_i\right](-1) \\ \frac{\partial Z}{\partial \beta_1} &= 2 \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1 X_i\right](-X_i), \end{aligned}$$

e fazendo-se $\frac{\partial Z}{\partial \beta_0} = 0$ e $\frac{\partial Z}{\partial \beta_1} = 0$, obtêm-se as equações normais,

$$\sum_{i=1}^{n} \left[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right] = 0 \Leftrightarrow n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i, \tag{2.1}$$

$$\sum_{i=1}^{n} \left[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right] X_i = 0 \Leftrightarrow \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i, \tag{2.2}$$

De (2.1) tem-se,

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{\beta}_1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{2.3}$$

Substituindo-se (2.3) em (2.2) tem-se,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}.$$

Logo, a reta estimada pelo método de mínimos quadrados é dada por.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, ..., n.$$

Vale ressaltar, caso os erros forem normalmente distribuídos, então as estimativas de mínimos quadrados coincidirá com as estimativas do método da máxima verossimilhança.

2.3 Modelo de Regressão Linear Múltipla (MRLM)

A regressão linear múltipla é um método estatístico que consiste na construção de modelos, a fim de explicar a verdadeira relação entre uma variável em função de outras variáveis, ou como função de polinômios de maior grau de uma única variável. Para tanto, o MRLM exige que alguns pressupostos sejam satisfeitos, tais como:

i) Os erros são independentes;

ii) Os erros tem média igual a zero;

iii) Os erros têm variâncias constantes;

iv) O modelo é aditivo;

v) Existe relação linear entre a variável independente e dependente;

vi) Os erros e as variáveis preditoras são independentes;

vii) Os erros tem distribuição normal multivariada.

O formato geral da equação de regressão linear múltipla é expresso por.

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + \varepsilon$$

em que, y_j é a variável dependente; β_k corresponde aos coeficientes técnicos atrelados as variáveis independentes; X_k corresponde as variáveis independentes.

2.3.1 Métodos de Estimação

Um dos métodos utilizados para estimação dos parâmetros do modelo de regressão linear múltipla é o método da máxima verossimilhança, o qual é definido a seguir.

Sejam $x_1, ..., x_n$ uma amostra aleatória de tamanho n da variável aleatória X com função de densidade (ou probabilidade) $f(x|\theta)$ com $\theta \epsilon \Theta$, onde Θ é o espaço paramétrico (BOLFARINE; SANDOVAL, 2010). A função de verosimilhança de θ correspondente a amostra aleatória observada é expresso por,

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i|\theta).$$

O estimador de máxima verossimilhança de θ é o valor que maximiza L(θ; $x_1, ..., x_n$) e é obtido da seguinte maneira:

i) Encontrar a função de máxima verossimilhança;

ii) Aplicar a função ln;

iii) Derivar em função do parâmetro θ;

iv) Igualar o resultado a zero;

v) Verificar se esse estimador é ponto de máximo.

Os resultados seguintes são válidos para o modelo de regressão linear múltipla, porém, para maior clareza de apresentação, será utilizado o modelo de regressão linear simples (CHARNET *et al.*, 1999). A amostra aleatória sob o modelo de regressão linear é dada por:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad e \quad \varepsilon_i \sim N(0; \sigma^2),$$
$$Cov[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j \quad i, j = 1, ..., n.$$

Consequentemente,

$$y_i \sim N\left(\beta_0 + \beta_1 X_i; \sigma^2\right), \quad i = 1, ..., n, \quad \textit{independentes.}$$

ou seja, a densidade de y_i é dada por:

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

A função de verossimilhança é igual a densidade conjunta, contudo, consideram-se $y_1, y_2, ..., y_n$ fixos e os parâmetros β_0, β_2 e σ^2 como argumentos da função. Notação: $L\left(\beta_0, \beta_1, \sigma^2\right)$,

$$\begin{aligned} L\left(\beta_0, \beta_1, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - \left(\beta_0, \beta_1, \sigma^2\right)\right]^2\right\} \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \left(\beta_0, \beta_1, \sigma^2\right)\right]^2\right\}. \end{aligned}$$

Para definir estimadores de máxima verossimilhança, considera-se β_0, β_1 e σ^2 como variáveis matemáticas (para usar a mesma notação) e deve-se achar os valores que maximizam $L\left(\beta_0, \beta_1, \sigma^2\right)$, em que $\beta_0 \in \Re, \beta_1 \in \Re$ e $\sigma^2 > 0$. Lembrando que a função exponencial é monótona crescente, e notando que $-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \left(\beta_0 + \beta_1 X_i\right)\right]^2 \leq 0$, o máximo de $L\left(\beta_0, \beta_1, \sigma^2\right)$ é obtido para os valores de β_0 e β_1 que minimizam $\sum_{i=1}^{n}\left[y_i - \left(\beta_0 + \beta_1 X_i\right)\right]^2$, seja qual for o valor de σ^2. Portanto, os estimadores de máxima verossimilhança dos parâmetros β_0 e β_1 coincidem com os estimadores de quadrados mínimos $\widehat{\beta_0}$ e $\widehat{\beta_1}$.

O valor de σ^2 que maximiza $L\left(\beta_0, \beta_1, \sigma^2\right)$ é o mesmo valor que maximiza $l\left(\beta_0, \beta_1, \sigma^2\right) = lnL\left(\beta_0, \beta_1, \sigma^2\right)$,

$$l\left(\beta_0, \beta_1, \sigma^2\right) = \ln\left(2\pi\right)^{\frac{-n}{2}} + \ln\left(\sigma^2\right)^{\frac{-n}{2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \left(\beta_0 + \beta_1 X_i\right)\right]^2,$$

assim,

$$\frac{\partial l\left(\hat{\beta}_0, \hat{\beta}_1, \sigma^2\right)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} - \frac{\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2}{2} - \frac{1}{\sigma^4},$$

igualando a zero,

$$-\frac{n}{2\widehat{\sigma^2}}+\frac{\sum_{i=1}^{n}\left[y_i-\left(\widehat{\beta_0}+\widehat{\beta_1}X_i\right)\right]^2}{2\left(\widehat{\sigma^2}\right)^2}=0,$$

$$\widehat{\sigma^2}=\frac{\sum_{i=1}^{n}\left[y_i-\left(\widehat{\beta_0}+\widehat{\beta_1}X_i\right)\right]^2}{n}.$$

o qual difere do estimador de mínimos quadrado apenas no denominador, ou seja, n ao invés de $(n-2)$ o que o torna um estimador viesado.

2.3.2 Regressão Linear Múltipla na Forma Matricial

A regressão linear múltipla também pode ser expressa na forma matricial. Nesse sentido, o vetor $\boldsymbol{\beta}$ tem dimensão maior que 2 e a matriz $\mathbf{X}$ é definida de acordo com as suposições do modelo em questão. Na forma matricial o número de colunas de $\mathbf{X}$ é igual ao número de elementos em $\boldsymbol{\beta}$ e o número de linhas de $\mathbf{X}$ é o tamanho da amostra. A primeira coluna da matriz $\mathbf{X}$, é um vetor de dimensão n formado pelos valores correspondentes as observações da amostra, cujos elementos são todos iguais a 1, correspondendo ao coeficiente β_0 do Modelo de Regressão Linear Múltipla.

Utilizando a notação matricial o MRLM pode ser expresso por

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

em que,

$$\mathbf{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

A estimação dos parâmetros do MRLM, pode ser feita pelo método dos quadrados mínimos ou pelo método da máxima verossimilhança.

No método dos mínimos quadrados, o número de parâmetros a serem estimados é $p=k+1$. Se existirem n observações, a estimação dos parâmetros reduz-se a um problema matemático de resolução de um sistema de n equações e p incógnitas, não sendo possível fazer qualquer análise estatística. Então, deve-se ter $n>p$ (**??**). Sendo assim, o modelo adotado é dado por:

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

Com isto, tem-se que tanto melhor será o modelo quanto menor for o comprimento de $\boldsymbol{\varepsilon}$. Usando a norma Euclideana para o comprimento de $\boldsymbol{\varepsilon}$, tem-se.

$$Z = ||\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta},$$

logo,

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

Maiores detalhes sobre derivadas de matrizes ver Charnet *et al.*,(1999).

Fazendo-se $\frac{\partial Z}{\partial \beta} = 0$, tem-se:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y},$$

que é o sistema de equações normais, em que

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i2} & \cdots & \sum_{i=1}^n X_{ik} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} & \cdots & \sum_{i=1}^n X_{i1}X_{ik} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i2}^2 & \cdots & \sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ik} & \sum_{i=1}^n X_{i1}X_{ik} & \sum_{i=1}^n X_{i2}X_{ik} & \cdots & \sum_{i=1}^n X_{ik}^2 \end{bmatrix},$$

e

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1}Y_i \\ \sum_{i=1}^n X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^n X_{i1}Y_k \end{bmatrix}.$$

Soma de Quadrados

A soma de quadrados é uma forma matemática, com sua aplicação é obtida a soma de quadrados total (SQT), soma de quadrados dos resíduos (SQE) e a soma de quadrados de regressão (SQR_{eg}). Segundo Charnet *et al.* (1999), soma de quadrados total (SQT), mede a variabilidade dos valores observados em torno de sua média, enquanto que a SQE, nos dá uma medida do ajuste do modelo considerado. Sendo assim, quanto menor a SQE, melhor o ajuste aos valores observados.

A obtenção da soma de quadrados é dada por:

i) $SQT = Y'Y - n\bar{y}^2$;

ii) $\hat{\beta} = (X'X)^{-1}X'Y$, com isso, $SQR_{eg} = \hat{\beta}'X'Y - n\bar{y}^2$;

iii) SQE é obtido por subtração, ou seja, SQT - SQR_{eg}.

2.3.3 Tabela Anova e Testes de Hipóteses

Quando é pressuposto um modelo de regressão para relacionar uma variável aleatória, Y, com base em p variáveis preditoras, é aconselhável que se faça a verificação da precisão do modelo. Com isso, uma tabela de análise de variância (ANOVA) deve ser construída, pois a partir da mesma é aplicado o teste F de Snedecor. A apresentação da tabela ANOVA, (CHARNET *et al.*, 1999) é vista na Tabela 1.

Tabela 1: Análise de varância para o modelo de regressão linear múltipla (CHARNET et al. 1999).

F.V. (Fonte de Variação)	*G.l.* (Graus de Liberdade)	*S.Q.* (Soma de Quadrados)	*Q.M.* (Quadrado Médio)	F_0
Regressão	p	SQR_{eg}	$\frac{SQR_{eg}}{p}$	$\frac{QMR_{eg}}{QME}$
Erro	$n-p-1$	SQE	$\frac{SQE}{(n-p-1)}$	
Total	$n-1$	SQT		

Com a construção da tabela ANOVA possibilita-se realizar mais uma etapa para verificação do modelo de regressão múltipla, ou seja, a aplicação do teste de hipóteses. Logo, será testado as hipóteses H_0 *versus* H_1 em que,

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

Quando rejeita-se H_0, conclui-se que há contribuição significante de uma ou mais variáveis regressoras no estudo de Y. A estatística do teste tem, sob H_0, distribuição F com $(n-p-1)$ graus de liberdades. Segundo Charnet *et al.* (1999), a estatística F_0 é obtida a partir do quadrado médio correspondente a partição da regressão extra do quadrado médio do erro.

2.3.4 Coeficiente de Determinação e Coeficiente de Determinação Ajustado

O coeficiente de determinação é uma medida de ajustamento, que tem como sua simbologia R^2. Com o coeficiente determinação calculado é possível identificar o quanto o modelo consegue explicar, isto é, qual grau de explicação o modelo terá em relação aos valores observados.

$$R^2 = \frac{SQR_{eg}}{SQT}.$$

O resultado do R^2 varia entre 0 e 1, com isso, quanto mais próximo de 1, melhor o ajuste do modelo considerado. É relevante ressaltar, que R^2 tem uma forte ligação com o modelo proposto e não somente aos dados, se ajustamos modelos diferentes, teremos distintos valores de R^2 para cada modelo. Entretanto, segundo Demétrio e Zocchi (2006), como o R^2 não obtêm um valor esperado de Y confiável para o modelo de regressão linear múltipla, logo o coeficiente de determinação ajustado (R^2_{ajust}) é recomendado para avaliação da escolha do modelo, pois leva em consideração o número de variáveis em um conjunto de dados. Dessa forma, o mesmo é expresso por

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

em que, p é o número de parâmetros.

2.3.5 Intervalo de Confiança

Através do intervalo de confiança é possível obter informação referente à precisão das estimativas. Com isso, quanto menor a amplitude do intervalo maior será a precisão. Segundo Demétrio e Zocchi (2006), seja $Q = q\left(y_1, y_2, \ldots, y_n, \beta\right)$, uma função da amostra aleatória $y_i, y_2, \ldots, y_n$ e de β, o parâmetro de interesse e tem uma distribuição que independe de β, então Q é uma quantidade pivotal. Assim, para qualquer γ fixo, tal que $0 < 1$, existem q_1 e q_2 dependendo de γ, tais que

$$P\left[q_1 < Q < q_2\right] = 1 - \gamma,$$

e a partir dessa expressão, pode-se obter um intervalo de confiança para β com um coeficiente de confiança $1 - \gamma$

$$\hat{\beta}_j \sim N\left(\beta_j, S\left(\hat{\beta}_j\right)\right).$$

Portanto,

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{S\left(\hat{\beta}_j\right)}} \sim N\left(0,1\right),$$

tem-se que,

$$\frac{1}{\sigma^2} SQ\text{Res} \sim \chi^2_{n-p} \Leftrightarrow W = \left(n-p\right)\frac{QM\text{Res}}{\sigma^2} \sim \chi^2_{n-p}$$

Sendo, Z e QMResíduos independentes. Então,

$$Q = \frac{Z}{\sqrt{\frac{W}{n-p}}} \sim t_{n-p},$$

que é fundamento para a construção dos intervalos de confiança para os parâmetros β_j.

Desse modo,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{S\left(\hat{\beta}\right)}} \sqrt{\frac{\left(n-p\right)\sigma^2}{\left(n-p\right)QM\text{Res}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{S\left(\hat{\beta}\right)}} \sim t_{n-p},$$

é uma quantidade pivotal e um intervalo de confiança para β, com um coeficiente de confiança $1-\gamma$ é,

$$P\left[-t_{\frac{\gamma}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{\sqrt{S\left(\hat{\beta}_j\right)}} \leq t_{\frac{\gamma}{2}}\right] = 1-\gamma,$$

logo,

$$P\left[\hat{\beta} - t_{\frac{\gamma}{2}}\sqrt{S\left(\hat{\beta}_j\right)} \leq \beta \leq \hat{\beta} + t_{\frac{\gamma}{2}}\sqrt{S\left(\hat{\beta}_j\right)}\right] = 1-\gamma.$$

De acordo com a simetria da distribuição t, pode-se escrever

$$IC\left[\beta_j\right]_{1-\gamma} : \beta_j \pm t_{n-p;\frac{\gamma}{2}}\sqrt{S\left(\hat{\beta}_j\right)}.$$

O desenvolvimento de um modelo de regressão, exige uma série de conjecturas. Dessa forma, é preciso realizar uma análise da veracidade e confiabilidade deste modelo. Uma das maneiras é verificar as discrepâncias entre os valores observados e os valores ajustados, ou seja, realizando a análise de resíduos. De acordo com a análise de resíduos, o modelo é apropriado se satisfazer os seguintes pressupostos:

i) ε_i e ε_j são independentes ($i \neq j$);

ii) $\text{Var}(\varepsilon_i) = \sigma^2$ (constante), $0 < \sigma_i^2 < \infty, \quad i = 1, ..., n$;

iii) $\varepsilon_i \sim \text{N}(0; \sigma^2)$ (normalidade);

iv) O modelo é linear;

v) Não existir *outilers* (pontos atípicos) influentes.

2.3.6 Análise de diagnóstico

É importante enfatizar que deve ser realizado uma análise para identificar a existência de colinearidade e multicolinearidade entre as variáveis. O modelo dos resíduos é dado por,

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

no qual é possível reescrever o erro como forma linear de $\boldsymbol{Y}$,

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} = \left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right].$$

A esperança e a variância dos resíduos são,

$$\begin{aligned} E\left(\hat{\boldsymbol{\varepsilon}}\right) &= E\left[\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right] = 0, \\ \widehat{Var}\left(\hat{\boldsymbol{\varepsilon}}\right) &= Var\left[\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right] \\ &= \sigma^2\left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]. \end{aligned}$$

Para verificar cada pressuposto citado acima, são utilizadas algumas técnicas, as quais estão evidenciadas nos tópicos abaixo.

2.3.6.1 Diagnóstico de Normalidade

A análise de normalidade dos resíduos é de suma importância para que o ajuste do modelo de regressão linear possa transmitir resultados confiáveis. Para isto, existem duas formas de realizar esta análise.

i) A análise de resíduo pode ser feita através de um gráfico de probabilidade normal chamado, Q-Q plot - Quantil de probabilidade esperado para distribuição normal, em função dos resíduos.

ii) Um dos testes mais utilizados para se testar a normalidade dos dados é o teste de normalidade de Shapiro-Wilk. O teste de Shapiro-Wilk determina uma estatística (W) calculada sobre os valores ordenados, elevados ao quadrado, buscando aferir se uma amostra aleatória é originada de uma distribuição normal. Devido a seu grande poder de resolução, este método é adotado preferencialmente nos testes de normalidade.

A estatística W é calculada da seguinte forma

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

sendo, x_i os valores ordenados das amostras; a_i constantes grandes a partir de meio, variâncias e covariâncias da ordem estatística de uma amostra de tamanho n e uma distribuição normal.

Então, se fórmula as hipóteses:

$$\begin{cases} H_0: & \text{A amostra provém de uma população normal,} \\ H_1: & \text{A amostra não provém de uma população normal.} \end{cases}$$

2.3.6.2 Diagnóstico de Homoscedasticidade

Com a realização do diagnóstico de homoscedasticidade é verificado se as variâncias dos erros são constantes, a fim de satisfazer um dos pressupostos exigidos para que se tenha um modelo de regressão apropriado. Caso a homoscedasticidade não seja satisfeita então, as variâncias não são constantes, ocorrendo assim, uma heteroscedasticidade. Esse diagnóstico pode ser feito das seguintes formas:

i) O gráfico dos resíduos versus valores ajustados (velores preditos) é uma forma de verificar a heteroscedasticidade, pois neste gráfico é possível indicar que não existe uma relação linear entre as variáveis explicativas com a variável resposta por meio de alguma tendência nos pontos.

ii) O teste Goldfeld - Quandt, é indicado para grandes amostras e quando a suposição de normalidade dos erros é assumida. Esse teste, pressupõe que a heteroscedasticidade seja decorrente de relação entre o erro aleatório e uma ou algumas das variáveis regressoras. O teste consiste em ordenar as observações de acordo com a variável explicativa que se acredita ser a responsável pela heteroscedasticidade. Posteriormente, divide-se a amostra ordenada em três partes de tal forma que a parte do meio tenha aproximadamente 20 % dos dados e que as partes 1 e 3 tenham quantidades de dados semelhantes (RODRIGUES; DINIZ, 2006). Logo, as hipóteses testadas são

$$\begin{cases} H_0: & \sigma_1^2{=}\sigma_2^2 = ... = \sigma_k^2, \\ H_1: & \text{Pelo menos um dos } \sigma_i^2 \text{ é diferente } \forall \text{ i=1, ..., k.} \end{cases}$$

A estatística de teste neste caso é dada por:

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

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.

2.3.6.3 Diagnóstico de Independência

Sendo umas das suposições do modelo de regressão linear múltipla, a independência dos resíduos pode ser verificada a partir dos seguintes procedimentos.

i) Gráfico dos resíduos versus a coleta de dados, caso esse gráfico apresente uma tendência dos pontos, logo há indícios de dependência dos resíduos.

ii) Uma outra forma de analisar a independência dos resíduos, é com aplicação do teste Durbin-Watson, que mede a correlação entre cada termo de erro e o termo de erro da observação imediatamente anterior (BRITO *et al.*, 2007). Tendo sua estatística definida como

$$DW = \frac{\sum \left(u_i - u_{t-1}\right)^2}{\sum u_t^2}.$$

2.3.6.4 Diagnóstico de *outliers*

Um *outliers* é uma observação extrema, ou seja, é um ponto com comportamento diferente dos demais. Na regressão, após o ajuste do modelo, é necessário aplicação de várias

medidas de diagnósticos para identificação de possíveis *outliers* no conjunto de dados. Se um *outliers* for influente, ele interfere sobre a função de regressão ajustada. Entretanto, se uma observação ser considerada *outliers* não quer dizer que consequentemente seja um ponto influente.

De acordo com Demétrio e Zocchi (2006), tanto a variável resposta como as covariáveis do modelo podem conter *outliers*. Exitem na literatura várias medidas para identificação de *outliers* (ver Sebert *et al.* (1998) and Bleiberg *et al.* (2004)). A seguir apresentam-se algumas delas.

i) A Medida de Leverage (h_{ii}) mede a importância da i-ésima observação na determinação do ajuste do modelo, em que h_{ii} é o valor do i-ésimo elemento da diagonal principal da matriz H. Valores de $h_{ii} > 2p/n$ devem ser verificados, pois, podem ser pontos de alavanca.

ii) A distância de Cook, representada por D_i mede o afastamento do vetor de estimativas dos coeficientes da regressão provocado pela retirada da i-ésima observação. A estatística é dada por $D_i = \frac{\left(\hat{\beta}_i - \hat{\beta}_{1-i}\right)' X'X \left(\hat{\beta}_i - \hat{\beta}_{1-i}\right)'}{pQMRes}$. Valores de $D_i > 1$ sugerem a avaliação de observações influentes (ALCÂNTARA *et al.*, 2003).

2.3.7 Seleção de Variáveis Regressoras

Conforme Demétrio e Zocchi (2006), em modelos de regressão linear múltipla é necessário determinar um subconjunto de variáveis independentes que melhor explique a variável resposta. Dessa forma, dentre todas as variáveis explicativas disponíveis, é necessário selecionar um subconjunto de variáveis consideradas importantes para a formação do modelo. Para tanto, dentre uma vasta gama de critérios utilizados para a seleção de um subconjunto de variáveis a serem incorporadas ao modelo de regressão linear múltipla (ver Mello *et al.* (2002), Herrera *et al.* (2008)).

2.3.7.1 Critério de informação de Akaike

O critério de informação de Akaike (AIC) foi desenvolvido a partir da distância de Kullback-Leibler (k-L), a qual é uma distância entre o modelo verdadeiro, que geralmente é uma abstração, e o modelo candidato (BELLO, 2010). A expressão do AIC pode ser simplificada, tendo a seguinte forma

$$AIC = n \ln\left(\hat{\sigma}_p^2\right) + 2\left(p+1\right),$$

em que, $\hat{\sigma}_p^2$ é o estimador de máxima verossimilhança da variância do erro.

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{n}.$$

2.3.7.2 Fator de Inflação de Variância

O fator de inflação de variância (VIF), é um técnica que foi desenvolvida para detectar a presença de colinearidade ou multicolinearidade entre as variáveis explanatórias, sendo a multicolinearidade um fenômeno da amostra.

Esse método mostra como a variância é inflada pela multicolinearidade, umas das interpretações quanto a existência da multicolinearidade, é que a média do fator de inflação de variância não pode ser maior que 1 e maior fator de inflação de variância para as variáveis não pode ser maior que 10 (PROTÁSIO T. *et al.*, 2011).

3 Material e Métodos

Os dados utilizados neste trabalho foram provenientes de um estudo realizado por Willerman *et al.* (1991). Os autores selecionaram uma amostra contendo 40 estudantes do curso de psicologia de uma Universidade do Sudoeste dos Estados Unidos, que tem algum histórico de alcoolismo, danos no cérebro, epilepsia e doenças do coração. As sete variáveis obtidas para análises foram: gênero, escores de coeficiente de inteligência na escala completa com base nos quatro Wechsler (FSIQ), escores de coeficiente de inteligência verbais com base nos quatro Wechsler (VIQ), Pontuação de coeficiente de inteligência de desempenho com base nos quatro Wechsler (PIQ), peso corporal, altura e contagem total de pixels a partir dos 18 exames de ressonância magnética (MRI_Count). Os dados encontram-se no Anexo A.

Foi utilizado um modelo de regressão linear múltipla, em que a variável resposta foi a contagem total de pixels a partir dos 18 exames de ressonância magnética (MRI_Count) em função das demais variáveis coletadas. A fim de verificar a relação entre a variável resposta e as covariáveis, foi aplicado o teste t de *Student*, o qual testa a hipótese de associação linear entre as variáveis envolvidas, procedeu-se com o teste F de Snedecor. Após os testes de hipóteses sobre os parâmetros, a análise seguiu com a verificação de algumas pressuposições do modelo de regressão como, a existência de colinearidade das covariáveis, normalidade dos resíduos, homogeneidade de variância dos resíduos, independência dos resíduos. Através da equação $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, é possível calcular o resíduos.

A normalidade dos resíduos foi testada através do teste de *Shapiro-Wilk*, sob a hipótese H_0 de normalidade (SHAPIRO; WILK, 1965). A homoscedasticidade, foi verificada através dos teste de Goldfeld-Quandt, sob a hipótese H_0, em que as variâncias são homogêneas (THURSBY, 1982). Realizou-se os métodos de *stepwise* em conjunto com o critério de seleção de Akaike (AIC) e calculou-se o R^2 ajustado, com o objetivo de selecionar o modelo que melhor explicasse a variável resposta. Todas as análises e gráficos foram obtidos através do *software R 2.15* (*R Development Core Team*, 2015).

4 Resultados e Discussão

Procede-se com uma análise gráfica (Figura 2) na qual utilizou-se a variável contagem total de pixels a partir dos 18 exames de ressonância magnética (MRI_Count) como referência e comparou-se a dispersão da mesma em relação as demais. As variáveis analisadas na Figura 2 foram: gênero (x1), escores de coeficiente de inteligência na escala completa com base nos quatro Wechsler (FSIQ), sendo a variável x2, escores de coeficiente de inteligência verbais com base nos quatro Wechsler (VIQ), sendo a variável x3, pontuação de coeficiente de inteligência de desempenho com base nos quatro Wechsler (PIQ), sendo a variável x4, peso corporal sendo a variável x5, altura como variável x6 e contagem total de pixels a partir dos 18 exames de ressonância magnética (MRI_Count) como variável resposta y. É importante ressaltar que neste conjunto de dados houveram algumas observações perdidas para os valores das covariáveis.

Figura 2: Gráfico descritivo das variáveis analisadas MRICount (y), gênero(x1), FSIQ(x2), VIQ(x3), PIQ(x4), peso corporal(x5), altura(x6).

Na Figura 2, é possível visualizar uma relação linear entre as variáveis altura, peso corporal, PIQ, VIQ, FSIQ com a varável resposta (MRI_Count). Uma outra observação

está com relação a multicolinearidade entre as covariáveis. Ainda é possível observar que em média o tamanho do cérebro é maior para gênero masculino. Outra observação adquirida por meio da Figura 2 é a verificação de uma correlação linear moderada entre as variáveis explicativas com a variável resposta, assim como também, altos valores de correlação linear entre as variáveis explicativas exemplo x2 e x3 (FSIQ e VIQ), sendo estas duas escores de inteligência é possível esperar um grau de colinearidade entre elas. Contudo, a partir da análise visual realizada na Figura 2, propõe-se o ajuste do modelo de regressão linear múltipla(modelo completo), envolvendo todas as covariáveis para posteriormente definir o modelo final proposto.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon,$$

em que, MRI_Count é a variável resposta (y), e as variáveis explicativas que compõe o modelo completo são: gênero, FSIQ, VIQ, PIQ, peso corporal e altura, concomitantemente estimados pelos seguintes coeficientes: $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$.

Logo, o modelo completo ajustado,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6.$$

Seguindo-se com as análises, tem-se a tabela de análise de variância para a decomposição das somas de quadrados. Por meio da Tabela 2 é possível verificar a significância do modelo completo, bem como uma importante quantidade que é a estimativa da variância representada pelo quadrado médio dos resíduos.

Tabela 2: Análise de variância para o tamanho do cérebro (MRI_Count) utizando o modelo completo.

Fontes de Variação	*G.L.*	*S.Q.*	*Q.M.*	$F_{cal.}$	Valor P
Regressão	6	27,28	4,55	21,16	<0,0001
Resíduos	33	7,08	0,21	-	
Total	39	34,36	-	-	-

De acordo com a Tabela 3, verifica-se que houve pelo menos uma variável explicativa que foi significativa ao nível nominal de 5% pelo teste F de Snedecor. Conforme Charnet *et al.* (1999) é necessário realizar-se um teste t de student para verificar quais variáveis foram significativas, ou seja, quais apresentam relação linear com a contagem total de pixels medida pela variável MRI_Count. Na Tabela 4, seguem os valores estimados para o modelo completo, referente aos parâmetros com respectivos erros padrões e estatística

t para as variáveis explicativas.

Tabela 3: Estimativas dos parâmetros com respectivos erros padrão e estatística t para as variáveis Gênero (x1), FSIQ (x2), VIQ (x3), PIQ (x4), peso corporal (x5), altura (x6) sob o modelo completo.

Efeitos	Estimativas	Erro Padrão	Valor de t	Valor P
Intercepto	608139,40	62010,70	9,80	<0,0001
x_1	51896,50	20690,10	2,50	0,0172
x_2	-9430,00	4520,60	-2,08	0,0448
x_3	5368,90	2652,70	2,02	0,0511
x_4	6267,10	2462,80	2,54	0,0158
x_5	268,90	1049,80	0,25	0,7994
x_6	3516,80	1826,90	1,92	0,0629

Diante da Tabela 4, pode-se observar que os preditores x1, x2 e x4 obtiveram efeito significativo, ao nível nominal de 5% de probabilidade. Os erros padrão das estimativas não foram superiores a metade dos valores das estimativas. Porém, como o teste t de Student é aplicado ao modelo de regressão múltipla de forma sequencial, a significância dos parâmetros pode ser mudada de acordo com a ordem que os parâmetros entram no modelo, além disso, é preciso identificar o grau de relação entre as variáveis explicativas, pois está relação pode interferir na cálculo da significância dos parâmetros, sendo assim, procedeu-se o cálculo do fator de inflação de variância, com o objetivo de verificar a existência de multicolinearidade entre as covariáveis, tendo seus respectivos resultados apresentados na Tabela 5 e posteriormente o uso do método de seleção Stepwise para cálculo das regressões extras (condicionais).

Tabela 4: Fator de inflação de variância (VIF) para o modelo completo.

x_1	x_2	x_3	x_4	x_5	x_6
1,996238	215,540972	71,375122	55,700608	1,842853	2,425473

Conforme observar-se na Tabela 5, o fator de inflação de variância para as variáveis x_2, x_3 e x_4 foram altos. De acordo com Protásio T. *et al.* (2011), quando o valor do VIF for maior que 10, isso indica que há existência de multicolinearidade entre as variáveis. Como a variável x2 apresentou o maior VIF dentre as demais, optou-se por retirar a mesma e realizar um novo cálculo para o fator de inflação de variância.

Os valores encontrados para o modelo sem considerar a variável x2, apresentaram redução nos valores dos fatores de inflação de variância (todos os valores foram < 3), indicando a ausência de multicolinearidade entre as variáveis. Conforme Gujarati e Porter (2011) é necessário corrigir a multicolinearidade para se evitar os problemas, tais como presença de covariância e variância grande entre as variáveis explicativas que consequentemente levará a intervalos de confiança mais amplos, isto levará a maior aceitação da hipótese nula, gerando-se maior taxa de erro tipo II, o teste t também é afetado devido em sua construção existir a dependência da variância do estimador, isto ocasionará valores insignificantes de sua estatística, porém os valores R^2, medida geral da qualidade do ajustamento, poderá ser alta, entre outras consequências.

Com isso, através do critério de seleção AIC, procedeu-se com o ajuste do modelo aos dados, em que o melhor modelo escolhido foi o que apresentou o menor AIC (864,77) e apontou um ajuste R^2 de 57% (Tabela 4).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_4 x_4 + \hat{\beta}_6 x_6 \tag{4.1}$$

Dando continuidade com as análises, uma nova tabela expõe a análise de variância para o novo modelo proposto 4.1 (Tabela 5).

Tabela 5: Análise de variância para o modelo proposto.

Causas de Variação	G.L.	SQ.	QM.	F_{cal}	Valor P
x_1	1	8,50	8,50	38,14	$< 0{,}0001$
x_4	1	2,79	2,79	12,53	$0,001126^{**}$
x_6	1	1,05	1,05	4,75	$0,035858^{*}$
Resíduo	36	8,02	2,23	-	-
Total	39	20,36	-	-	-

Observa-se na Tabela 5, que todas as variáveis selecionadas para compor o modelo ajustado, foram estatisticamente significativas, ao nível de 1% e 5% de significância de probabilidade, ou seja, os coeficientes das variáveis x1, x4 e x6 são estatisticamente diferente de zero. Os valores desses coeficientes seguem na tabela abaixo.

Tabela 6: Estimativas dos parâmetros com respectivos erros padrão e estatística t para o modelo proposto.

Efeitos	Estimativas	Erro Padrão	Valor de t	Valor P
Intercepto	702266,9	40837,2	17,197	$< 0{,}0001$
x_1	62435,9	19815,1	3,151	$0,003271^{**}$
x_4	1214,1	336,7	3,606	$0,000935^{**}$
x_6	3459,6	1586,8	2,180	$0,035858^{*}$

A partir da Tabela 6, propõe-se o modelo ajustado, em seguida, procede-se com a verificação das pressuposições para validação do modelo, a qual é realizada através da análise gráfica dos resíduos e com aplicação de testes específicos $\hat{y} = 702266,9 + 62435,9x_1 + 1214,1x_4 + 3459,6x_6$.

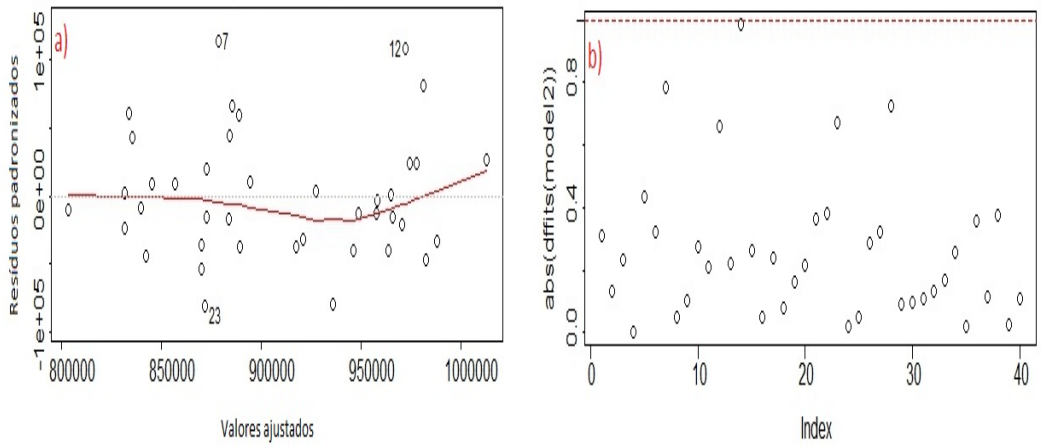

Figura 3: Análise gráfica dos resíduos: resíduos padronizados *versus* valores ajustados (a); resíduos ordenados *versus valores absolutos para o modelo proposto.*

Apresenta-se na Figura 3, a análise gráfica referente a homogeneidade de variância na qual é possível verificar que não há tendência, levando-se a supor que a variância dos resíduos é homoscedástica, corroborando assim com o teste de Goldfeld-Quandt com Valor P de 0,8407 $> \alpha$. Na sequência, apresenta-se os gráficos de probabilidade normal envelopado e o gráfico referente à distância de Cook. Com base em Sousa *et al.* (2008), foram feitas interpretações sobre a normalidade dos resíduos.

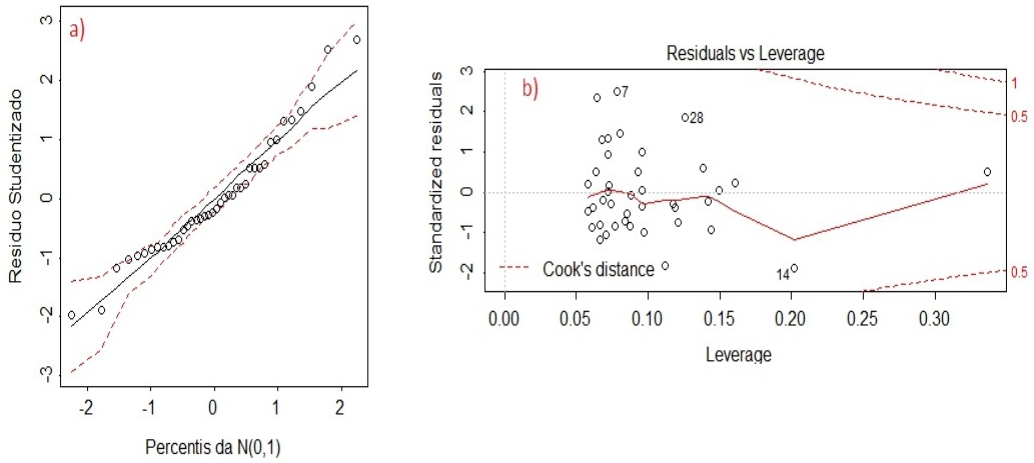

Figura 4: Análise gráfica referente a normalidade dos resíduos: Quantil-Quantil (a) e residuos estudantizados *versus* medida de leverage (b) para o modelo proposto.

Pode-se observar por meio da Figura 4, que não houve desvios de normalidade (a), apesar de ter apresentado uma leve simetria à direita. Em concordância com o gráfico, o teste de *shapiro-wilks* (Valor P = 0,1184 $>$ α), robustece a suposição da normalidade dos resíduos. Observa-se também que não há presença de pontos influentes no gráfico da distância de Cook. O modelo proposto então foi, $\hat{y} = 702266,9 + 62435,9x_1 + 1214,1x_4 + 3459,6x_6$, em que, as variáveis regressoras selecionadas foram: gênero, pontuação de coeficiente de inteligência de desempenho com base nos quatro Wechsler (PIQ), altura. Ou seja, essas variáveis têm influencia na variável resposta concordando com Tramo *et al.* (1998), onde o mesmo também encontrou resultados semelhantes.

5 Considerações Finais

De acordo com as análises realizadas nesse trabalho, ajustou-se um modelo de regressão linear múltipla ao conjunto de dados do pesquisador por Willerman *et al.* (1991). As variáveis regressoras consideradas neste estudo foram: gênero, escores de coeficiente de inteligência na escala completa com base nos quatro Wechsler (FSIQ), escores de coeficiente de inteligência verbais com base nos quatro Wechsler (VIQ), pontuação de coeficiente de inteligência de desempenho com base nos quatro Wechsler (PIQ), peso corporal, altura e contagem total de pixels a partir dos 18 exames de ressonância magnética (MRI_Count).

Após o ajuste com todas variáveis utilizou-se o fator de inflação de variância e o critério de seleção de variáveis, em que, outro modelo foi ajustado. Em seguida, verificou-se as pressuposições e constatou-se por meio dos testes que o modelo ajustado era confiável, resultando no seguinte modelo, $\hat{y} = 702266,9 + 62435,9x_1 + 1214,1x_4 + 3459,6x_6$, em que as variáveis regressoras selecionadas foram: gênero, pontuação de coeficiente de inteligência de desempenho com base nos quatro Wechsler (PIQ), altura. Ou seja, essas variáveis têm influencia na variável resposta, contagem total de pixels a partir dos 18 exames de ressonância magnética.

Referências

ALCÂNTARA, A. A. M.; SANT'ANNA, A. P.; LINS, M. P. E. Restringindo Flexibilidade de Pesos em DEA Utilizando Análise de Regressão MSEA. *Pesquisa Operacional*, SciELO Brasil, v. 23, n. 2, p. 347–357, 2003.

BELLO, L. H. A. *Modelagem em Experimentos Mistura-Processo para Otimização de Processos Industriais*. Tese (Doutorado) — Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro, 152p, 2010.

BLEIBERG, J.; CERNICH, A. N.; CAMERON, K.; SUN, W.; PECK, K.; ECKLUND, L. P. J.; REEVES, C. D.; UHORCHAK, C. J.; SPARLING, M. B.; WARDEN, D. L. Duration of cognitive impairment after sports concussion. *Neurosurgery*, LWW, v. 54, n. 5, p. 1073–1080, 2004.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à Inferência Estatística*. 2. ed. [S.l.]: SBM, 2010.

BRITO, G. A. S.; CORRAR, L. J.; BATISTELLA, F. D. Fatores Determinantes da Estrutura de Capital das Maiores Empresas que Atuam no Brasil. *Revista de Contabilidade e Finanças da USP, São Paulo*, SciELO Brasil, v. 18, n. 43, p. 9–19, 2007.

CHARNET, R.; FREIRE, C. L.; CHARNET, E. M.; BONVINO, H. Análise de Modelos de Regressão Linear com Aplicações. *Campinas, São Paulo, Unicamp, 356p*, 1999.

DEMÉTRIO, C. G. B.; ZOCCHI, S. S. Modelos de Regressão. *Piracicaba: ESALQ*, p. 183, 2006.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica-5*. [S.l.]: McGraw Hill Brasil, 2011.

HERRERA, L. G. G.; EL FARO, L.; ALBUQUERQUE, L. d.; TONHATI, H.; MACHADO, C. H. C. Estimativas de parâmetros genéticos para a produção de leite e persistência da lactação em vacas gir, aplicando modelos de regressão aleatória. *Revista Brasileira de Zootecnia*, SciELO Brasil, v. 37, n. 9, p. 1584–1594, 2008.

HOFFMANN, R.; VIEIRA, S. Análise de regressão: uma introdução à econometria. *São Paulo*, p. 379, 1998.

MELLO, J. Soares de; GOMES, E. G.; MELLO, M. H. C. Soares de; LINS, M. E. Método multicritério para seleção de variáveis em modelos dea. *Pesquisa Naval*, v. 15, p. 55–66, 2002.

MEMÓRIA, J. M. P. Breve História da Estatística. *Área de Informação da Sede-Texto para Discussão (ALICE)*, Brasília, DF: Embrapa Informação Tecnológica: Embrapa-Secretaria de Gestão e Estratégia, p. 111, 2004.

PEARSON, K. *The life, letters and labours of Francis Galton*. [S.l.]: University press, 1930.

PETERS, M. Does Brain size matter? A Reply to Rushton and Ankney. Canadian Psychological Association, v. 49, n. 4, p. 570 – 576, 1995.

PIKE, A. A. The effect of art therapy on cognitive performance among ethnically diverse older adults. *Art Therapy*, Taylor e Francis, v. 30, n. 4, p. 159–168, 2013.

PROTÁSIO T., P.; BUFALINO, L.; TONOLI, G. H. D.; COUTO, A. M.; TRUGILHO, P. F.; JÚNIOR, M. G. Relação Entre o Poder Calorífico Superior e os Componentes Elementares e Minerais da Biomassa Vegetal. *Pesquisa Florestal Brasileira*, v. 31, n. 66, p. 113, 2011.

RODRIGUES, S. A.; DINIZ, C. A. R. Modelo de Regressão Heteroscedástico. *Revista de Matemática e Estatística*, v. 24, n. 2, p. 133–146, 2006.

SANTOS, J. S.; SCARPELINI, S.; BRASILEIRO, S. L. L.; FERRAZ, C. A.; DALLORA, M. E. L.; SÁ, M. F. S. Avaliação do Modelo de Organização da Unidade de Emergência do HCFMRP-USP, Adotando, como Referência, as Políticas Nacionais de Atenção. *Medicina (Ribeirão Preto)*, v. 36, n. 2/4, p. 498 – 515, 2003.

SEBERT, D. M.; MONTGOMERY, D. C.; ROLLIER, D. A. A clustering algorithm for identifying multiple outliers in linear regression. *Computational statistics & data analysis*, Elsevier, v. 27, n. 4, p. 461–484, 1998.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, p. 591–611, 1965.

SOUSA, J. E. R. d.; SILVA, M. d. A.; SARMENTO, J. L. R.; SOUSA, W. H. d.; SOUZA, M. d. S. M. d.; FRIDRICH, A. B. Homogeneidade e Heterogeneidade de Variância Residual em Modelos de Regressão Aleatória Sobre o Crescimento de Caprinos Anglo-Nubianos. *Pesquisa Agropecuária Brasileira*, SciELO Brasil, v. 43, n. 12, p. 1725–1732, 2008.

THURSBY, J. G. Misspecification, heteroscedasticity, and the chow and goldfeld-quandt tests. *The Review of Economics and Statistics*, JSTOR, p. 314–321, 1982.

TRAMO, M. J.; LOFTUS, W.; STUKEL, T.; GREEN, R.; WEAVER, J.; GAZZANIGA, M. Brain size, head size, and intelligence quotient in monozygotic twins. *Neurology*, AAN Enterprises, v. 50, n. 5, p. 1246–1252, 1998.

WILLERMAN, L.; SCHULTZ, R.; RUTLEDGE, J. N.; BIGLER, E. D. In vivo brain size and intelligence. *Intelligence*, Elsevier, v. 15, n. 2, p. 223–228, 1991.

APÊNDICE

```
### Organizando os dados ####

attach(dados)
y=as.numeric(MRI_Count)
x1=Gender
x2=as.numeric(FSIQ)
x3=as.numeric(VIQ)
x4=as.numeric(PIQ)
x5=as.numeric(Weight)
x6=as.numeric(Height)
detach(dados)
pairs(dados, upper.panel = panel.smooth)
summary(dados)

### Ajustando o modelo ###

model=lm(y x1+x2+x3+x4+x5+x6)
anova(model)
require(car)
summary(model)

## Multicolinearidade
library(car)
vif(model)
model1=lm(y x1+x3+x4+x5+x6)
anova(model1)
vif(model1)
summary(model1)
```

```
    ### Critérios de seleção de modelos ###
library(MASS)
stepAIC(model1)
require(bbmle)
require(stats4)


model2 = lm(y x1+x4+x6)
model2


    ## Teste de Heterocedasticidade
gqtest(model2)


    ## Normalidade
require(stats)
qqnorm(rstandard(model2))
qqline(rstandard(model2))
plot(model2)


    ## Envelope ##
win.graph()
envelope.normal¡-function(form=form,k=k,alfa=alfa){
alfa1 = ceiling(k*alfa)
alfa2 = ceiling(k*(1-alfa))
glm1 = lm(formula=form)
X = model.matrix(glm1)
n = nrow(X)
p = ncol(X)
H = X
h = diag(H)
lmi = lm.influence(glm1)
si = lmi$sigma
rp = residuals(glm1)
ts = rp/(si*(1-h)0̂.5)
```

```
ident = diag(n)
epsilon = matrix(0,n,k)
e = matrix(0,n,k)
e1 = numeric(n)
e2 = numeric(n)
for(i in 1:k) epsilon[,i] = rnorm(n,0,1)
e[,i] = (ident-H)
u = diag(ident-H)
e[,i] = e[,i]/(u0̂.5)
e[,i] = sort(e[,i])

for(i in 1:n)
eo = sort(e[i,])
e1[i] = eo[alfa1]
e2[i] = eo[alfa2]

xb = apply(e,1,mean)
faixa = range(ts,e1,e2)
par(pty="s")
qqnorm(e1,axes=F,xlab=,ylab=,type="l",ylim=faixa,lty=2,col="red")
par(new=TRUE)
qqnorm(e2,axes=F,xlab=,ylab=,type="l",ylim=faixa,lty=2,col="red")
par(new=TRUE)
qqnorm(xb,axes=F,xlab=,ylab=,type="l",ylim=faixa,lty=1)
par(new=TRUE)
qqnorm(ts,xlab="Percentis da N(0,1)",ylab="Residuo
Studentizado",ylim=faixa,main=)
}
```

```
envelope.normal
envelope.normal(model2,k=40,alfa=0.05)
require(stats)
qqnorm(model2)
require(MASS)
shapiro.test(residuals(model2))
```

ANEXO - A

Tabela 7: Dados coletados referentes aos 40 estudantes de que participaram do estudo.

Gender	FSIQ	VIQ	PIQ	Weight	Height	MRI_Count
Female	133	132	124	118	64,5	816932
Male	140	150	124	.	72,5	1001121
Male	139	123	150	143	73,3	1038437
Male	133	129	128	172	68,8	965353
Female	137	132	134	147	65,0	951545
Female	99	90	110	146	69,0	928799
Female	138	136	131	138	64,5	991305
Female	92	90	98	175	66,0	854258
Male	89	93	84	134	66,3	904858
Male	133	114	147	172	68,8	955466
Female	132	129	124	118	64,5	833868
Male	141	150	128	151	70,0	1079549
Male	135	129	124	155	69,0	924059
Female	140	120	147	155	70,5	856472
Female	96	100	90	146	66,0	878897
Female	83	71	96	135	68,0	865363
Female	132	132	120	127	68,5	852244
Male	100	96	102	178	73,5	945088
Female	101	112	84	136	66,3	808020
Male	80	77	86	180	70,0	889083
Male	83	83	86	.	.	892420
Male	97	107	84	186	76,5	905940
Female	135	129	134	122	62,0	790619
Male	139	145	128	132	68,0	955003
Female	91	86	102	114	63,0	831772
Male	141	145	131	171	72,0	935494
Female	85	90	84	140	68,0	798612

Gender	FSIQ	VIQ	PIQ	Weight	Height	MRI_Count
Male	103	96	110	187	77,0	1062462
Female	77	83	72	106	63,0	793549
Female	130	126	124	159	66,5	866662
Female	133	126	132	127	62,5	857782
Male	144	145	137	191	67,0	949589
Male	103	96	110	192	75,5	997925
Male	90	96	86	181	69,0	879987
Female	83	90	81	143	66,5	834344
Female	133	129	128	153	66,5	948066
Male	140	150	124	144	70,5	949395
Female	88	86	94	139	64,5	893983
Male	81	90	74	148	74,0	930016
Male	89	91	89	179	75,5	935863