UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Maria da Vitória Dantas Silva

Análise de Regressão Múltipla em dados da PeNSE 2015 para João Pessoa-PB

Campina Grande - PB

Fevereiro de 2018

Maria da Vitória Dantas Silva

Análise de Regressão Múltipla em dados da PeNSE 2015 para João Pessoa-PB

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

Orientador: Prof. Gustavo Henrique Esteves

Campina Grande - PB

Fevereiro de 2018

É expressamente proibido 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 do trabalho.

S586a Silva, Maria da Vitória Dantas.
Análise de regressão múltipla em dados da PeNSE 2015 para João Pessoa - PB [manuscrito] : / Maria da Vitoria Dantas Silva. - 2018.
34 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2018.
"Orientação : Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística - CCT."

1. Regressão Linear Múltipla. 2. Análise de diagnóstico. 3. PeNSE.

21. ed. CDD 519.5

Maria da Vitória Dantas Silva

Análise de Regressão Múltipla em dados da PeNSE 2015 para João Pessoa-PB

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.

Trabalho aprovado em 09 de Fevereiro de 2018.

BANCA EXAMINADORA

Prof. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

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

Profa. Ms. Maria das Vitórias Alexandre Serafim
Universidade Estadual da Paraíba

Dedico esse trabalho aos meus pais Sebastiana Dantas e a José Soares da Silva e a meus irmãos em especial a João Antônio de Souza Silva (In Memorian)

Agradecimentos

Em primeiro lugar quero agradecer a Deus, pela sua infinita bondade de ter chegado até o fim, porque sem Ele não teria conseguido.

A minha mãe pelo carinho, amor, dedicação e compreensão que sem esse apoio não teria chegado onde cheguei.

Aos meus irmãos: Luzinalva Oliveira, José de Arimateia, Socorro Silva, Marlene Dantas, Sandro José do Nascimento e Luzeni Oliveira que sempre estavam do meu lado para que eu realizasse os meus sonhos.

Ao meu orientador Gustavo Esteves pelo compromisso e dedicação em sua orientação.

A Samara Rilda pelo companheirismo e ajuda em todos os anos do curso, sempre me incentivando para que terminasse.

A Melchior Batista e Dr. Raimundo que sempre me apoiaram e incentivaram quando falei que iria cursar Estatística.

Aos professores da banca, Tiago e Vitória, pela honra de assistir minha apresentação.

Aos professores do Departamento de Estatística da UEPB, que tivemos a honra de ter excelentes professores.

A minha amiga-professora Maria das Vitórias pelas ajudas e carinhos que sempre me tratou.

Aos meus amigos de curso: Carlos, Fátima, Ângela, Shirlley e Filipe que foi a melhor turma que a Universidade me deu, sem vocês não teria conseguido. Quero agradecer pela brincadeiras, risadas, brigas e união. A melhor turma que eu tive no curso.

Aos meus companheiro da Creche Wilson Pereira que sempre me ajudou.

Em especial quero agradecer a Carlos e Samara pela ajuda no trabalho de conclusão do curso, sempre prontos pra me ajudar.

Por fim quero agradecer a todos que me ajudaram no decorrer desses anos na UEPB.

"Dar o melhor de si é mais importante que ser o melhor."
(Mike Lermer)

Resumo

Neste trabalho foi estudado o modelo de regressão linear múltipla, desde o seu contexto histórico até um exemplo para aplicação da teoria. A regressão múltipla é uma ferramenta muito utilizada na estatística em determinadas áreas de conhecimentos. Para o exemplo de aplicação foi utilizado um banco de dados da Pesquisa Nacional de Saúde do Escolar - PeNSE - realizada em nível nacional no ano de 2015, mas aqui restrito ao município de João Pessoa - PB. Os dados foram ajustados utilizando o programa estatístico R na versão 3.1.1, que tem sido muito usado em trabalhos estatísticos. Com isso, o objetivo desse trabalho foi analisar as variáveis do banco de dados, para avaliar a prática de atividade física entre os alunos comparada com algumas variáveis explicativas. No geral, se observou que alunos do sexo masculino tendem a praticar mais atividade física do que aqueles do sexo feminino. Porém, no modelo de regressão linear múltipla, a análise gráfica de resíduos apontou alguns problemas nas pressuposições do modelo, e o gráfico dos quantis também apontou problemas com a normalidade dos dados, de onde se concluiu que esta abordagem não é adequada para estes dados e outros métodos estatísticos devem ser avaliados.

Palavras-chaves: Regressão Linear Múltipla, Análise de Diagnóstico, PeNSE.

Abstract

In this work the multiple linear regression model was studied, from its historical context to an application example to illustrate the theory. Multiple regression is a widely used statistical tool in certain areas of knowledge. For the application example, we used a database of the National Survey of School Health (*Pesquisa Nacional de Saúde do Escolar - PeNSE*) which was held all over the country in the year 2015, but restricted here to the João Pessoa - PB city. The data were adjusted using the statistical program R in its version 3.1.1, which has been widely used in statistical work. Therefore, the objective of this study was to analyze the variables of the database, to evaluate the practice of physical activity among students compared to some explanatory variables. Overall, it has been observed that male students tend to engage in more physical activity than those of females. However, in the multiple linear regression model, the graphical residual analysis pointed out some problems in the assumptions of the model, and the quantile chart also pointed out problems with the data normality, from which it was concluded that this approach is not adequate for these data and other statistical methods should be evaluated.

Key-words: Multiple Linear Regression, Diagnostic Analysis, PeNSE.

Lista de ilustrações

Figura 1 – *Boxplot* da variável prática de atividade física comparando gênero masculino contra feminino. 29
Figura 2 – *Boxplot* da variável prática de atividade física comparando idades. . . 29
Figura 3 – Gráfico dos resíduos contra os valores ajustados. 30
Figura 4 – Gráfico do *qqplot* para a distribuição normal. 31

Lista de tabelas

Tabela 1 – Análise de variância para o modelo de regressão linear simples. 17
Tabela 2 – Análise de variância para o modelo de regressão linear múltipla. 24
Tabela 3 – Características demográficas dos adolescentes do município de João Pessoa-PB, dados da PeNSE 2015. 27
Tabela 4 – Análise de variância para o modelo de regressão contendo as seis variáveis explicativas. 28
Tabela 5 – Resultados dos testes t de Student para os parâmetros do modelo ajustado com as três variáveis significativas do ajuste anterior. 28
Tabela 6 – Análise de variância para o modelo de regressão ajustado apenas com as três variáveis significativas. 28

Sumário

1 INTRODUÇÃO 11

2 FUNDAMENTAÇÃO TEÓRICA 12

2.1 Marco Histórico 12

2.2 Análise de Regressão Linear 13

2.3 Análise de Regressão Linear Simples 14

2.3.1 Estimação de Parâmetros 15

2.3.2 Decomposição da soma de quadrados 15

2.3.3 Coeficiente de determinação 16

2.3.4 Testes de Hipóteses 16

2.4 Análise de Regressão Linear Múltipla 18

2.4.1 Estimação de parâmetros 18

2.4.2 Soma de quadrados 21

2.4.3 Coeficiente de determinação múltipla 23

2.4.4 Testes de hipóteses 23

2.4.5 Intervalo de confiança para os coeficientes de regressão 24

2.4.6 Análise de resíduos 25

3 APLICAÇÃO 26

3.1 Resultados e discussão 26

4 CONCLUSÃO 32

REFERÊNCIAS 33

APÊNDICE A – ROTINA UTILIZADA NA ANÁLISE 34

1 Introdução

A prática regular de atividade física desde a infância e adolescência está associada a um benefício, tanto físico como psicológico, desde curto até longo prazo. Os hábitos associados à prática de atividade física adquiridos desde a adolescência podem aumentar ou até mesmo reforçar tais atividades na fase adulta, além de ajudar na manutenção de um estilo de vida saudável, como também na prevenção de doenças crônicas.

Por outro lado, a análise de regressão linear múltipla é uma ferramenta muito importante em diversas áreas de conhecimento para construir e avaliar modelos que descrevem de maneira razoável relações entre várias variáveis explicativas e uma variável resposta, associadas a um determinado processo. Por exemplo, seria possível avaliar de que forma certas variáveis, tais como idade, raça e gênero, podem estar associadas com a prática de atividade física.

Neste sentido, o presente trabalho tem por objetivo verificar a relação entre prática de atividade física entre adolescentes do 9º ano do ensino fundamental e outras variáveis (gênero, idade, raça, corpo, tempo estimado, tempo total) em adolescentes no município de João Pessoa - PB, com o uso do método da análise de regressão múltipla. Os dados foram obtidos da Pesquisa Nacional de Saúde do Escolar - PeNSE - 2015, feita pelo Instituto Brasileiro de Geografia e Estatística (IBGE), com a colaboração dos Ministérios da Saúde e da Educação, feita em todas capitais e algumas cidades do interior dos estados brasileiros, com os adolescentes do 9º ano do ensino fundamental II.

O trabalho está dividido em 4 capítulos, onde o Capítulo 1 trata desta introdução; o Capítulo 2 é a fundamentação teórica, que fala sobre o marco histórico, a análise de regressão linear, tanto simples como múltipla, estimação de parâmetros, decomposição da soma de quadrados, coeficiente de determinação, testes de hipóteses, intervalos de confiança e análise de resíduos; o Capítulo 3 traz a aplicação, com a análise dos dados e os respectivos resultados e discussão; por fim o Capítulo 4 apresenta as conclusões e considerações finais.

2 Fundamentação Teórica

O conteúdo de desta seção relata os principais aspectos da utilização dos modelos de regressão linear múltipla, por meios de artigos e livros relacionados ao objetivo de pesquisa do trabalho.

2.1 Marco Histórico

Francis Galton era o mais novo de nove filhos de um próspero banqueiro e nasceu em uma família socialmente abastada. Aos 16 anos começou a aprender medicina, mas interessou-se pela matemática e obteve sua primeira formação acadêmica. Posteriormente foi considerado um dos personagens mais importante para evolução da estatística.

Galton produziu mais de 340 artigos e livros em toda sua vida, além de realizar várias pesquisas. Dentre essas pesquisas obteve uma de suma importância para a estatística, o conceito estatístico de correlação que foi amplamente tratado como regressão em direção à média. Em 1885 Galton verificou que, apesar da tendência de que pais altos tivessem filhos altos e pais baixos tivessem filhos baixos, a altura média dos filhos de um grupo de pais altos era menor que a de seus pais e a altura média de um grupo de filhos de pais baixos era maior que a de seus pais; em consequência, os filhos de pais altos e baixos "regrediam" igualmente à altura média de todos os homens. Nas palavras de Galton isso era uma "regressão à mediocridade".

Com isso a lei de Regressão Universal de Galton foi confirmada por Karl Pearson em 1903, que coletou mais de mil registros de alturas dos membros de grupos de famílias e verificou que a altura média dos filhos de um grupo de pais altos era inferior a altura de seus pais, e que altura média dos filhos de um grupos de pais baixos era superior a altura dos seus pais.(DEMÉTRIO; ZOCCHI, 2006)

De modo mais geral pode-se pensar no estudo da dependência de uma variável explicativa (Y) que se tem interesse em conhecer o comportamento em relação a uma ou mais variáveis ($X_1, X_2, ..., X_k$) com o único objetivo de estimar os valores de Y. O que torna possível modelar e tentar descrever como as variáveis estão relacionadas entre si, devendo ter sempre a preocupação de criar modelos estatísticos. Um dos métodos mais usados na estatística para se investigar a relação entres as variáveis é o método de regressão. (NAGHETTINI, 2007)

2.2 Análise de Regressão Linear

A análise de regressão tenta avaliar o efeito que as variáveis X_i, $i = 1, 2, 3, ..., k$, denominadas variáveis dependentes, auxiliares ou regressoras, têm sobre uma variável resposta Y, que é a variável de interesse. Logo, um modelo de regressão tenta determinar o quanto cada variável dependente X_i explica o comportamento de Y. Assim deseja-se encontrar a função g, tal que:

$$Y = g(X_1, X_2, ..., X_k).$$

Verifica-se que as relações entre as variáveis nem sempre são determinísticas, portanto é necessário escrever a variável Y sendo a soma de um quantidade determinística e uma quantidade aleatória. Por sua vez a parte aleatória é conhecida como sendo o erro, representando inúmeros fatores, que em conjunto, podem interferir na variável resposta. Logo, o modelo pode ser reescrito na seguinte forma

$$Y = g(X_1, X_2, ..., X_k) + \varepsilon,$$

tem-se que $g(X_1, X_2, ..., X_k)$ e ε são as partes sistemática e aleatória do modelo, respectivamente. Portanto o erro provoca distorção sobre a parte determinística. De modo geral, não se tem dados suficientes para estimar a função g diretamente, assim se assume que g tem uma forma mais simples. Geralmente são usados os modelos lineares, considerando tal função como linear.

Portanto o modelo de regressão linear é representado pela a seguinte equação

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

onde tem-se que $\beta_0, \beta_1, ..., \beta_k$ são os parâmetros desconhecidos. É necessário lembrar que o modelo é linear nos parâmetros e não necessariamente nas variáveis X_i. O modelo é considerado linear quando a derivada parcial do modelo em relação aos parâmetros não depende de nenhum dos parâmetros, caso contrário, o modelo é não linear. Portanto temos as seguintes equações que representam modelos lineares:

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

por outro lado os modelos abaixo são não lineares

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

De modo geral tem-se n observações em relação as variáveis Y e $X_1, X_2, ..., X_k$, assim é descrito o modelo matemático,

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \varepsilon, \ i = 1, 2, 3, ..., n.$$

A representação matricial para o modelo é utilizada para facilitar os cálculos e a notação, e é definida por:

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

com

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

2.3 Análise de Regressão Linear Simples

Considerando duas variáveis aleatórias X e Y, para saber que tipo de relação existe entre ambas, é necessário encontrar um modelo matemático que explique, se existir, a dependência entre a variável resposta Y em relação a variável explicativa X. (CHARNET et al., 2008)

Quando esta relação linear, o objetivo da análise de regressão é encontrar a reta que melhor explique a dependência dos dados, uma vez assim, estima previsões para o comportamento da variável Y em decorrer dos acontecimentos da variável X. O modelo de regressão linear simples é definido por:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i = 1, 2, 3, ..., n,$$

de modo que tem-se o seguinte:

- x_i representa o valor da variável explicativa (independente) na i-ésima observação;
- y_i representa o valor da variável resposta (dependente), que é aleatória, na i-ésima observação;
- β_0 e β_1 são os parâmetros do modelo, que devem ser estimados, e definem a reta de regressão;
- ε_i, é um erro aleatório, com $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ e $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall_i \neq j$;
- n é o tamanho da amostra.

O parâmetro β_0 é chamado **intercepto** e representa o ponto em que a reta regressora corta o eixo dos y's. Já β_1 é o parâmetro que representa a inclinação da reta regressora e é dito **coeficiente de regressão** ou **coeficiente angular**. Assim, tem-se que para o aumento de uma unidade na variável x, o valor $E(Y|X)$ muda o correspondente a β_1 unidades.

2.3.1 Estimação de Parâmetros

O foco principal deste trabalho é abordar a regressão linear múltipla, portanto não serão realizadas demonstrações para regressão linear simples, mas é importante explicar alguns resultados da regressão simples para facilitar o entendimento do método de regressão múltipla.

Ao estimar os valores de β_0 e β_1, encontra-se a reta que melhor representa a relação entre X e Y. Conforme (TAVEIRA, 2015), com a estimação dos parâmetros é possível reduzir a soma das distâncias entre a função linear e os pontos observados na amostra. Há vários métodos que auxiliam na obtenção das estimativas dos parâmetros, o método de mínimos quadrados, conhecido também como mínimos quadrados ordinários e o método de máxima verossimilhança.

O método de mínimos quadrados tem a função de encontrar as 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 aplicar o método de mínimos quadrados os erros devem ser distribuídos aleatoriamente, na qual a distribuição seja normal e independente.

Ao aplicar o método obtém-se a equação matemática que minimizar os erros e explica melhor a dependência entre as variáveis:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

logo,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

são os estimadores de máxima verossimilhança de β_0 e β_1, onde

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad \text{e} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Através das suposições do modelo, as seguintes relações são satisfeitas, onde β_0,β_1 e $\hat{y}_i$ possuem distribuição normal, com seus respectivos parâmetros dados por:

$$\begin{aligned}
\hat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \\
\hat{\beta}_0 &\sim N\left(\beta_0; \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right), \\
\hat{y} &\sim N\left[\beta_0 + \beta_1; \sigma^2\left(\frac{1}{n} + \frac{\left((x_i - \bar{x})^2\right)}{S_{xx}}\right)\right].
\end{aligned}$$

2.3.2 Decomposição da soma de quadrados

Os valores observados na amostra podem ser definidos pelas seguintes somas de quadrados.

- A variância total de y:

$$SQtotal = \sum_{t=1}^{n}(y_i - \bar{y})^2 \Rightarrow SQtotal = S_{yy}.$$

- Soma de quadrados de regressão, que é a variância explicativa pela regressão:

$$SQreg = \hat{\beta}_1 \sum_{t=1}^{n}(x_1 - \bar{x})(y_i - \bar{y}) \Rightarrow SQreg = \hat{\beta}_1 S_{xy}$$

- Soma de quadrados de resíduo ,que a variância não explicativa pela regressão:

$$SQreg = \sum_{t=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{t=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y}) \Rightarrow SQreg = SQtotal - SQreg$$

Do que foi apresentado acima, é possível se escrever $SQtotal = SQreg + SQres$, que é conhecido como a decomposição de quadrados total. Também é possível se mostrar que as somas de quadrados tem distribuição χ^2 ,com os seguintes graus de liberdade:

$$\frac{SQtotal}{\sigma^2} \sim \chi^2_{n-1},$$
$$\frac{SQreg}{\sigma^2} \sim \chi^2_{1},$$
$$\frac{SQres}{\sigma^2} \sim \chi^2_{n-2}.$$

2.3.3 Coeficiente de determinação

Uma das formas de avaliar a qualidade do ajuste do modelo é através do coeficiente de determinação, que indica o quanto o modelo foi capaz de explicar os dados observados. Tal coeficiente é definido por:

$$R^2 = \frac{SQreg}{SQtot} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}},$$

que indica a proporção da variação de y que é explicada pela regressão. O valor de R^2 pertence ao intervalo $[0, 1]$, quanto mais próximo de um, melhor é o ajuste do modelo.

2.3.4 Testes de Hipóteses

Na regressão linear pode ser verificado dois testes, o teste t de Student, e o teste F de Snedecor. Para o teste t é possível testar se há evidência de regressão linear entre as variáveis, e é preciso contrastar as seguintes hipóteses:

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

Lembrando que $\hat{\beta}_1 \sim N(\beta_1; \frac{\sigma^2}{S_{xx}})$, é possível se definir a seguinte estatística

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}},$$

em que, em geral, o valor de σ^2 não é conhecido e deve ser substituído pelo seu estimador S^2, de modo que a estatística acima segue distribuição t de Student com $n-2$ graus de liberdade. Então, ao nível de significância α, compara-se o valor calculado para a estatística Z, digamos z, com o valor tabelado $t_{\frac{\alpha}{2}}$, e decide-se que:

- Se $-t_{\frac{\alpha}{2}} <$ z $< t_{\frac{\alpha}{2}}$, aceita-se H_0 e conclui-se que ao nível de significância α não há regressão linear.
- Se $|z| > t_{\frac{\alpha}{2}}$, rejeita-se H_0 e conclui-se que ao nível de significância α, há indício de regressão linear entre X e Y.

É importante se observar que o mesmo procedimento, com as devidas adaptações, para se testar hipóteses acerca do parâmetro β_0, intercepto do modelo, tais como

$$\begin{cases} H_0: & \beta_0 = b_0 \\ H_1: & \beta_0 \neq b_0 \end{cases}.$$

Para o teste F de Snedecor, é importante observar que as hipóteses a serem testadas são as mesmas do teste t para β_1, ou seja, as hipóteses de interesse são

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

e neste caso, constrói-se uma tabela conhecida como tabela da análise de variância, representada na Tabela 1.

Tabela 1 – Análise de variância para o modelo de regressão linear simples.

Fonte de variação	G.L	S.Q	Q.M	F
Regressão	1	$\hat{\beta}_1 S_{xy}$	$\hat{\beta}_1 S_{xy}$	$\hat{\beta}_1 S_{xy}$
Resíduo	$n-2$	$S_{yy} - \hat{\beta}_1 S_{xy}$	$\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} = S^2$	-
Total	$n-1$	S_{yy}	-	

A estatística F representada na última coluna da Tabela 1 segue distribuição F de Snedecor com 1 e $n-2$ graus de liberdade. Logo, após fixar o nível α de significância e encontrar o ponto crítico para a distribuição $F_{1,n-2}$, denotado por $F_{tabelado}$, decide-se que:

- Se $F_{calculado} > F_{tabelado}$, rejeita-se H_0 ao nível de significância α, e se conclui que parece existir regressão linear.
- Se $F_{calculado} < F_{tabelado}$, aceita-se H_0, ao nível de significância α, e se conclui que não existe indício de regressão entre X e Y.

2.4 Análise de Regressão Linear Múltipla

Regressão múltipla é uma coleção de técnicas estatísticas para construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo. A diferença entre a regressão linear simples e a múltipla é que na múltipla são tratadas duas ou mais variáveis explicativas simultaneamente. De acordo com Sassi et al. (2011) a regressão múltipla pode ser usada no intuito de melhorar o modelo desenvolvido para explicar o comportamento das variáveis do banco de dados que estão sendo estudadas.

Admite-se que X_1,X_2,...,X_k sejam as variáveis independes e Y a variável dependente, com uma amostra de n observações $(x_{1i}, x_{2i}, .., x_{ki}, y_i)$, $i = 1, 2, 3, ..., n$. Logo o modelo de regressão linear múltipla é definido por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \varepsilon_i, \ i = 1, 2, 3, ..., n.$$

Deste modo temos a forma matricial do modelo de regressão linear múltipla, sendo definida por:

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon,$$

com

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{22} & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

As suposições sobre o modelo de regressão linear múltipla são dadas a seguir:

- Existe relação linear entre y e x_j, $j = 1, 2, ..., k$;
- os valores dos x_j são fixos, não são variáveis aleatórias;
- as variáveis aleatórias ε_i têm distribuição normal;
- $E(\varepsilon) = \mathbf{0}$, onde $\mathbf{0}$ representa o vetor nulo;
- $var(\varepsilon_i) = \sigma^2$, $\forall$ os valores de $i = 1, 2, ..., n$;
- Os erros são não correlacionados dois a dois.

2.4.1 Estimação de parâmetros

Conforme Morais (2010) e Almeida (2014), para estimar os parâmetros do modelo de regressão linear múltiplo, é possível recorrer ao método dos mínimos quadrados, que

permite encontrar uma reta que minimize a distância entre os pontos observados e a reta, fazendo, em média, a soma dos desvios quadráticos ser igual a zero.

Dado $\hat{\beta}$ e ϵ, vetores das estimativas e dos desvios do modelo, assim logo são definido como:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{e} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Assim:

$$\epsilon = Y - X\hat{\beta}.$$

A soma dos quadrados dos erros é dado por:

$$\epsilon'\epsilon = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon_1\epsilon_1 + \epsilon_2\epsilon_2 + ... + \epsilon_n\epsilon_n = \sum_{i=1}^{n} \epsilon_1^2$$

Ainda, a soma de quadrados pode ser definida como:

$$Z = \epsilon'\epsilon = (Y' - \beta'X')(Y - \beta X) = Y'Y - Y'X\beta - \beta'X'\beta - \beta'X'X\beta,$$

mas note que as matrizes $Y'X\beta$ e $\beta'X'\beta$ uma é transposta da outra, portanto são equivalentes, ambas possuem um único elemento, assim renomeando a soma de quadrados para Z tem-se que:

$$Z = Y'Y - 2\beta'X'\beta + \beta'X'X\beta,$$

e, para se obter o ponto de mínimo para os valores de β, a função Z deve ser diferenciada e igualada à zero, assim tem-se:

$$\begin{aligned} \frac{\partial Z}{\partial \beta} &= \partial(Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}) = 0 \\ &\Rightarrow \partial(Y'Y) - 2\partial(\hat{\beta}'X'Y) + \partial(\hat{\beta}'X'X\hat{\beta}) = 0 \\ &\Rightarrow -2(\partial\hat{\beta}')X'Y + (\partial\beta')X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}(\partial\beta') = 0. \end{aligned}$$

Temos que $(\partial\hat{\beta}')X'X\hat{\beta} = \hat{\beta}'X'X(\partial\hat{\beta})$ são matrizes simétricas com um único elemento, a equação acima pode ser reescrita na seguinte forma $-2(\partial\hat{\beta}')X'Y + 2(\partial\hat{\beta}')X'X\beta = 0$, assim temos que;

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

Para que; $\partial\hat{\beta}' = 0$ é necessário que:

$$X'X\hat{\beta} = X'Y \tag{2.1}$$

Na equação (2.1) temos um sistema de equações normais, cuja solução fornece estimativas dos parâmetros constituintes do vetor $\hat{\beta}$. Multiplicando os dois lados da igualdade pela a matriz $(X'X)^{-1}$, a qual é inversa da matriz $(X'X)$, considerando a matriz como não-singular, tem-se o resultado definido por:

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

e uma vez solucionado as matrizes, aplicando o produto das mesmas, temos:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{2.2}$$

Também é possível se observar que

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{21} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{j=1}^{n} x_{1j} & \sum_{j=1}^{n} x_{2j} & \cdots & \sum_{j=1}^{n} x_{kj} \\ \sum_{j=1}^{n} x_{1j} & \sum_{j=1}^{n} x_{1j}^2 & \sum_{j=1}^{n} x_{1j}x_{2j} & \cdots & \sum_{j=1}^{n} x_{1j}x_{kj} \\ \sum_{j=1}^{n} x_{2j} & \sum_{j=1}^{n} x_{1j}x_{2j} & \sum_{j=1}^{n} x_{2j}^2 & \cdots & \sum_{j=1}^{n} x_{2j}x_{kj} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{n} x_{jk} & \sum_{j=1}^{n} x_{1j}x_{kj} & \sum_{j=1}^{n} x_{2j}x_{kj} & \cdots & \sum_{j=1}^{n} (x_{kj})^2 \end{bmatrix} = (X'X)',$$

de modo que se verifica que $X'X$ é uma matriz simétrica, ou seja $X'X = (X'X)'$. Além disso, tem-se que

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} x_{1j}y_j \\ \sum_{j=1}^{n} x_{2j}y_j \\ \vdots \\ \sum_{j=1}^{n} x_{kj}y_j \end{bmatrix}.$$

Quando a matriz X for de posto incompleto, é necessário usar a inversa generalizada condicional, $(X'X)^{-}$, de mesmo modo para todas as estimações feitas utiliza-se a inversa trivial, $(X'X)^{-1}$, ocorrerá quando as variáveis explicativas forem qualitativas.

Ainda podemos obter outros resultados importantes, por meio da equação (2.1). Assim, tem-se:

$$X'X\hat{\beta} - X'Y = 0,$$

onde zero representa o vetor nulo, portanto,

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

logo, rescrevendo a equação acima da seguinte maneira:

$$X'\epsilon = 0,$$

tem-se que ϵ' é o vetor de erros do modelo. E da relação dada temos que $\sum_{i=1}^{n} \epsilon_j = 0$ e $\sum_{i=1}^{n} x_{ij}\epsilon_i = 0$

Com a soma dos desvios nula, define-se que:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i, \tag{2.3}$$

e assim,

$$\sum_{i=1}^{n} \hat{y}_i = X\hat{\beta}.$$

Com W uma matriz simétrica e idempotente, temos que:

$$\begin{aligned} W &= X(X'X)^{-1}X' \\ X\hat{\beta}\hat{Y} &= X(X'X)X'Y \\ \hat{Y} &= WY. \end{aligned}$$

Na sequência, segue a demonstração que $\hat{\beta}$ é um estimador não tendencioso para o vetor de parâmetros β.

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X)^{-1}X'\varepsilon. \end{aligned}$$

Em ambos os lados da equação aplicamos a esperança matemática, assim:

$$E(\hat{\beta}) = E(\beta) + E[(X'X)^{-1}X'\varepsilon],$$

e como X não é variável aleatória e $E(\varepsilon) = 0$, segue o resultado da equação:

$$E(\hat{\beta}) = E(\beta) + 0 \Rightarrow E(\hat{\beta}) = \beta.$$

2.4.2 Soma de quadrados

As mesmas somas de quadrados definidas para o modelo de regressão linear simples também podem ser definidas, com as devidas adaptações, para o caso múltiplo.

- Soma de Quadrado de Resíduos (SQ_{Res})

Para obter o cálculo da soma de quadrados do resíduo é necessário relembrar as equações (2.1) e (2.2), logo pode ser escrita na forma:

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

- Soma de Quadrados Total (SQ_{total})

A soma de quadrados total é definida como medida da variabilidade total dos dados, é dada pela a expressão:

$$\begin{aligned}
SQ_{total} &= \sum_{j=1}^{n}(y_j - \hat{y})^2 \\
&= \sum_{j=1}^{n} y_j^2 - \frac{(\sum_{j=1}^{n} y_j)^2}{n} \\
&= Y'Y - \frac{(\sum_{j=1}^{n} y_j)^2}{n}.
\end{aligned}$$

- Soma de Quadrados de Regressão (SQ_{Reg})

A soma do quadrado da regressão é a medida da quantidade de variação da variável aleatória dependente explicada pela equação de regressão linear múltipla. Logo e dada por:

$$\begin{aligned}
SQ_{Reg} &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n} \hat{y_i}^2 \\
&= \sum_{i=1}^{n} \hat{Y_i}^2 - \frac{\left(\sum_{i=1}^{n} \hat{Y_i}\right)^2}{n} \\
&= \hat{Y_i}'\hat{Y_i} - \frac{\left(\sum_{i=1}^{n} \hat{Y_i}\right)^2}{n} \\
&= (X\hat{\beta})'X\hat{\beta} - \frac{\left(\sum_{i=1}^{n} \hat{Y_i}\right)^2}{n} \\
&= \hat{\beta}'X'X\hat{\beta} - \frac{\left(\sum_{i=1}^{n} \hat{Y_i}\right)^2}{n} \\
&= \hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} \hat{Y_i}\right)^2}{n}.
\end{aligned}$$

Sabemos que SQtotal = SQreg + SQres pode se escrever a soma qui quadrados de resíduos da seguinte forma:

$$SQ_{Res} = SQ_{total} - SQ_{Reg}.$$

2.4.3 Coeficiente de determinação múltipla

O coeficiente de determinação é uma estatística usada para medir a proporção da soma de quadrados total que é explicada pela regressão. Tal coeficiente pode ser definido através da seguinte expressão:

$$R^2 = \frac{SQ_{Reg}}{SQ_{total}} = 1 - \frac{SQtotal - SQres}{SQ_{total}}.$$

E para aperfeiçoar essa medida corrige-se o coeficiente dividindo pelos graus de liberdade, de modo que:

$$R_c^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

2.4.4 Testes de hipóteses

Os testes de hipóteses são aplicados para testar se há regressão linear entre as variáveis do modelo. Da mesma forma que no caso da regressão simples, serão abordados aqui os testes F de Snedecor e t de Student.

- Teste de significância para o modelo de regressão (teste F de Snedecor)

O teste F é empregado para verificar se as variáveis independentes conjuntamente contribuem para explicar a variação da variável resposta, logo as hipóteses são dadas por:

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

A estatística teste é definida por:

$$F = \frac{QM_{Reg}}{QM_{Res}} \sim F_{(k,n-p)},$$

logo temos que k é o numero de variáveis independentes e $p = k + 1$ é o número de parâmetros do modelo. Assim, após encontrar o valor F calculado, o valor de F tabelado, atribuído ao nível de significância α, tem–se que:

- Se $F_{calculado} > F_{tabelado}$, rejeita-se a hipótese H_0, conclui-se ao nível α de significância que há regressão;
- Se $F_{calculado} < F_{tabelado}$, aceita-se a hipótese H_0, conclui-se ao nível α de significância que não há indícios de regressão;

Por questões de visualização e praticidade o procedimento acima pode ser expresso na tabela da análise de variância (ANOVA), que está representada na Tabela 2, e que contém a estatística de teste F na sua última coluna.

Tabela 2 – Análise de variância para o modelo de regressão linear múltipla.

Fonte de variação	G.L	S.Q	Q.M	F
Regressão	k	SQ_{Reg}	$\frac{SQ_{Reg}}{k}$	$\frac{QMreg}{QMres}$
Resíduo	$n-p$	SQ_{Res}	$\frac{SQ_{Res}}{n-p}$	-
Total	$n-1$	SQ_{total}	-	

- Teste de significância para os coeficientes de regressão (teste t de Student)

Para o pesquisador muitas vezes é interessante testar as hipóteses acerca dos coeficientes de regressão individualmente, para determinar o potencial de cada um no modelo. Assim tem-se os testes para as hipóteses:

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0. \end{cases}$$

Neste caso, a estatística teste é dada por:

$$T = \frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)} \sim t_{n-p},\ j = 1, 2, .., k, \tag{2.4}$$

onde $S(\hat{\beta}_j)$ é a estimativa do desvio padrão de $\hat{\beta}_j$. E então a regra de decisão fica dada por:

- Se $t_{calculado} > t_{tabelado}$, rejeita-se a hipótese H_0, e se conclui que ao nível α de significância, a variável associada ao parâmetro β_j não pode se eliminada do modelo, pois ajuda a explicar a regressão linear.
- Se $t_{calculado} < t_{tabelado}$, aceita-se a hipótese H_0, e assim tem-se que ao nível α de significância, conclui-se que a variável pode ser eliminada do modelo sem tanto dano para a explicação da regressão linear.

2.4.5 Intervalo de confiança para os coeficientes de regressão

Por meio do da construção de intervalos de confiança também se pode avaliar os parâmetros do modelo. Tem-se que ao encontrar um intervalo de confiança que contenha o verdadeiro valor do parâmetro β_j, com $j = 1, 2, ..., k$, a um certo nível de significância α, que se desejar.

Na equação abaixo tem-se uma expressão para a construção de um intervalo de $100 \times (1-\alpha)\%$ de confiança para o coeficiente de regressão β_j, com $j = 1, 2, .., k$

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

2.4.6 Análise de resíduos

Para que os resultados de uma análise de regressão sejam confiáveis, tanto no modelo de regressão linear simples, como no modelo de regressão múltipla é fundamental que as suposições do modelo ajustado sejam válidas.

A análise de resíduos desempenha um papel fundamental neste sentido, pois oferece técnicas que nos ajudam a verificar a presença de indícios de adequabilidade do modelo por meio de análises, principalmente gráficas, dos resíduos. Como já visto anteriormente, o vetor de resíduos é definido por

$$\epsilon = Y - X\hat{\beta}.$$

Lembrando de alguns resultados importantes, tínhamos que

$$\begin{aligned} y &\sim N(X\beta, \sigma^2 I) \\ \hat{\beta} &\sim N[\beta, \sigma^2 (X'X)^{-1}] \\ \hat{y}_i = x_i'\beta &\sim N[x_i'\beta, \sigma^2 x_i'(X'X)^{-1}x_i]. \end{aligned}$$

Além disso, a esperança e a variância dos resíduos são definidas por

$$E(\epsilon) = E[Y - X\hat{\beta}] = 0$$

e

$$\begin{aligned} var(\epsilon) &= var[Y - X\hat{\beta}] \\ &= \sigma^2[I - X(X'X)^{-1}X'], \end{aligned}$$

logo, a distribuição do vetor de resíduos fica dada por

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

Segundo Hoffman (2006) a matriz $X(X'X)^{-1}X'$ é considerada matriz de projeção H, em que, é simétrica e idempotente e os valores da diagonal principal dessa matriz são h_{ii}, com $0 < h_{ii} < 1$, $i = 1, 2, ..., n$. Onde h_{ii} é o valor observado da influência de x_i a x.

A condição de normalidade pode ser verificada usando um gráfico de probabilidade normal, que podem ser divididos em dois tipos:

- *PPplot*: probabilidade acumulada esperada para distribuição normal, em função da probabilidade observada acumulada dos resíduos.
- *QQplot*: quantil de probabilidade esperado para distribuição normal em função dos resíduos.

3 Aplicação

Os dados analisados neste trabalho são provenientes da PeNSE (Pesquisa Nacional de Saúde do Escolar), que foi realizada no ano de 2015, com uma parceria entre os Ministérios da Saúde e da Educação com o IBGE (Instituto Brasileiro de Geografia e Estatística). A amostra estudada aqui está restrita aos indivíduos (adolescentes) do município de João Pessoa, situada no litoral paraibano, capital do Estado da Paraíba, embora a pesquisa tenha sido executada em nível nacional.

A pesquisa PeNSE ocorreu em três edições, nos anos de 2009, 2012 e 2015, e a abordagem deste estudo refere-se ao ano de 2015. A pesquisa incluiu todas as capitais dos estados brasileiros e o Distrito Federal, além de algumas cidades do interior, foram entrevistados 102.301 alunos do 9° do Ensino Fundamental II, no qual os dados foram coletados entre os meses de abril e setembro de 2015.

Ao todo no município de João Pessoa-PB, 68 escolas e 89 turmas de estudantes participaram da PeNSE, com um total de 3.036 alunos matriculados no 9° ano no ensino fundamental II, dentre os quais 2.423 alunos responderam ao questionário.

3.1 Resultados e discussão

Para começar o estudo de regressão linear entre as variáveis do banco de dados, é necessário definir a variável dependente, também conhecida como variável resposta.

A partir dos dados coletados na PeNSE no ano de 2015 em João Pessoa-PB, deseja-se saber quais fatores contribuem para explicar a prática de atividade física entre os alunos que participaram do estudo. Para isso foi definida a variável dependente como sendo o tempo total de atividade física praticada pelos alunos no período de uma semana (anterior ao momento da entrevista). As variáveis utilizadas como independentes ou explicativas para o modelo de regressão linear foram: tempo estimado de atividade física na semana, gênero, idade, raça/cor, trabalho e percepção do corpo. Maiores detalhes sobre as variáveis podem ser obtidos diretamente no site do IBGE.

Assim, para chegar aos resultados da pesquisa foi utilizado o *software* estatístico R na sua versão 3.1.1 (R Core Team, 2014), onde os dados foram ajustados pelo programa. Do total de 3.036 alunos matriculados no 9° do ensino fundamental II em João Pessoa-PB, 2.423 estavam presentes no momento da entrevista e responderam ao questionário, e para esse estudo foram trabalhadas com 2.417 observações.

O modelo de regressão testado para este caso ficou dado por:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \varepsilon_i, \ i = 1, 2, ..., 2.414,$$

onde y_i é o tempo total de atividade física praticada na última semana antes da entrevista, x_1 é o tempo estimado de atividade física de acordo com as tarefas dos alunos, x_2 é a idade, x_3 é o gênero, x_4 é a raça/cor, x_5 designa se o aluno trabalha ou não e x_6 é a percepção corporal do aluno, ε_i é o erro para cada observação e os β_j são os parâmetros associados às variáveis explicativas do modelo de regressão com $i = 1, 2, ..., 2.414$ e $j = 1, 2, ..., 6$.

Na Tabela 3 podemos observar as descrições das características demográficas, se trabalha ou não e como considera seu corpo para os participantes da PeNSE 2015 em João Pessoa-PB. Onde a maioria dos alunos era do sexo feminino (52,67%) com faixa etária de 14 anos (48,53%) e dar cor parda (43,70%), sendo que a maioria dos adolescentes não trabalha (91,59%) e considera seu corpo normal (54,33%).

Tabela 3 – Características demográficas dos adolescentes do município de João Pessoa-PB, dados da PeNSE 2015.

Variáveis	Tamanho das observações (% válidos)
Demográfica	
Sexo ($n = 2417$)	
Feminino	1273 (52,67%)
Masculino	1144 (47,33%)
Idade ($n = 2388$)	
$\leq$ 13 anos	569 (23,83%)
14 anos	1159 (48,53%)
$\geq$ 15 anos	660 (27,64%)
Raça/cor ($n = 2416$)	
Branca	782 (32,37%)
Preta	270 (11,18%)
Amarela	190 (7,86%)
Parda	1058 (43,70%)
Indígena	116 (4,80%)
Trabalha ($n = 2415$)	
Sim	203 (8,41%)
Não	2212 (91,59%)
Como considera seu corpo ($n = 2393$)	
Muito Magro	122 (5,10%)
Magro	487 (20,35%)
Normal	1300 (54,33%)
Gordo	424 (17,72%)
Muito Gordo	60 (2,50%)

O modelo de regressão, contendo seis variáveis explicativas, apresentado no início desta seção foi ajustado para estes dados e para se verificar a existência de regressão entre as variáveis foi gerada uma ANOVA, que está representada na Tabela 4. Podemos observar

os graus de liberdades, somas de quadrados, quadrados médios, e onde iremos observar as variáveis independentes do modelo de regressão. Neste caso, observa-se que apenas as variáveis tempo estimado, idade e gênero foram fortemente significativas para o modelo.

Tabela 4 – Análise de variância para o modelo de regressão contendo as seis variáveis explicativas.

	GL	SQ	QM	F	P valor
Tempo Estimado	1	41877283	41877283	1896,418	<0,002
Idade	1	369822	369822	16,7474	<0,004
Gênero	1	473947	473947	21,4628	<0,003
Raça	4	288501	72125	3,2662	0,011
Trabalho	1	8	8	0,0004	0,98494
Corpo	4	71527	17882	0,8098	0,51877
residual	2375	52445475	22082	-	-

Desta forma, resolveu-se reajustar o modelo usando apenas estas três variáveis significativas. Assim, a Tabela 5 apresenta os resultados dos testes t de Student para os parâmetros, onde se pode verificar que há evidência do efeito para todas as três variáveis independentes em estudo com exceção do intercepto, que não mostrou significância estatística, ou seja, as três variáveis parecem de fato significativas para o modelo de regressão linear múltipla.

Tabela 5 – Resultados dos testes t de Student para os parâmetros do modelo ajustado com as três variáveis significativas do ajuste anterior.

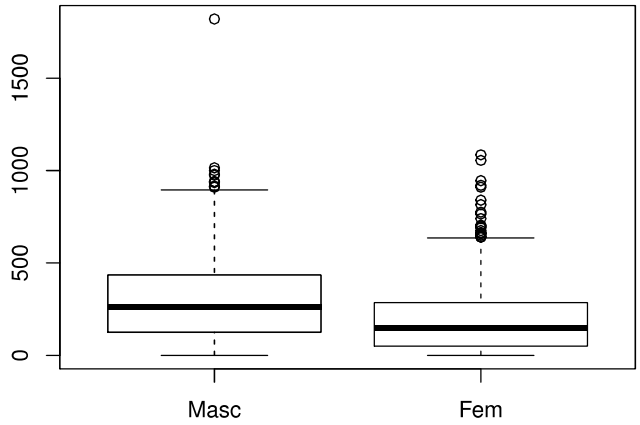
	Estimativas	Erro padrão	Valor t	$Pr(>\|t\|)$
Intercepto	-9,53546	40,89775	-0,233	0,815662
Tempo estimado	0,92729	0,02299	40,338	<0,002
Idade	10.34686	2,82501	3,663	0,000255
Gênero feminino	-29,13180	6,37065	-4,573	<0,005

Na Tabela 6 tem-se a ANOVA para o modelo ajustado apenas com as três variáveis, observa-se que todas as variáveis para compor o modelo ajustado, foram estatisticamente significativo ao nível de 5% de significância para o modelo de regressão linear, reforçando os resultados da tabela anterior.

Tabela 6 – Análise de variância para o modelo de regressão ajustado apenas com as três variáveis significativas.

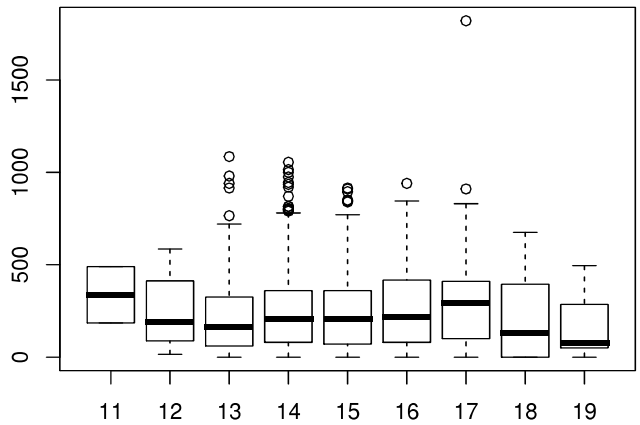
	GL	SQ	QM	F	P valor
Tempo Estimado	1	42014250	42014250	1900,244	<0,002
Idade	1	371297	371297	16,793	<0,004
Gênero	1	462332	462332	20,911	<0,005
Residual	2402	53108032	22110	-	-

Figura 1 – *Boxplot* da variável prática de atividade física comparando gênero masculino contra feminino.



Para a interpretação dos parâmetros que deram significativos, especialmente aqueles associados à variáveis categóricas, é interessante se analisar os gráficos do tipo *boxplot*. Neste sentido, observando a Figura 1, nota-se que a distribuição da variável prática de atividade física comparando sexo feminino e masculino podem ser consideradas simétricas em torno do valor mediano e onde se observa uma maior prática de atividade física entre os alunos do sexo masculino do que os do sexo feminino. Verifica-se também que a altura da caixa no sexo feminino é menor do que no sexo masculino, indicando que o sexo feminino é mais homogêneo quanto as variáveis, apesar de contar com valores atípicos em ambos.

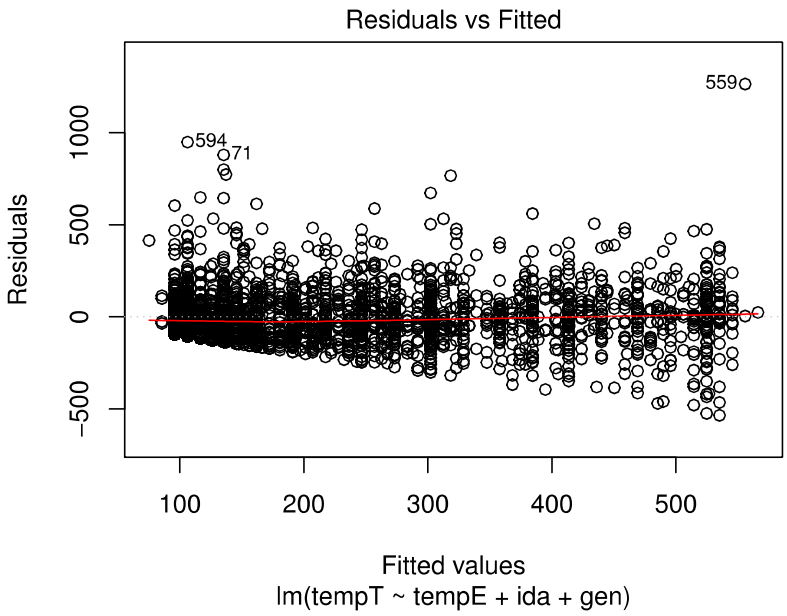
Figura 2 – *Boxplot* da variável prática de atividade física comparando idades.



Observando a Figura 2, nota-se que a distribuição da variável prática de atividade física em relação as idades dos alunos em geral podem ser consideradas simétricas em torno do valor mediano. Verifica-se também que a altura da caixa de 17 anos é maior do que a de 19 anos que é menor a altura da caixa de idades, indicando que a prática de atividade física parece mais homogênea para a idade de 17 anos, apesar de contar com valores atípicos. Com relação a diferença dos valores médios, observa-se que os alunos de 11 anos aparentemente praticam mais atividade física, que se mantém ligeiramente estável entre 12 e 16 anos, voltando a aumentar aos 17 e diminuindo para os menores níveis entre os 18 e 19 anos, tendo uma distribuição bastante assimétrica para os 19 anos.

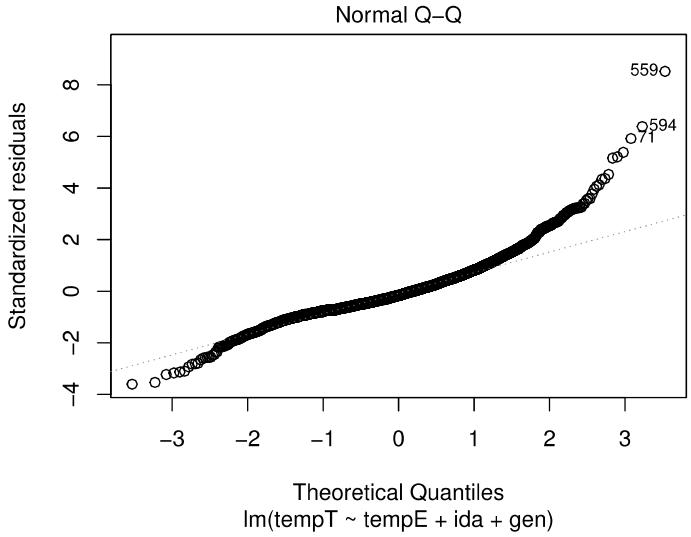
Para a validação do ajuste do modelo, o gráfico dos resíduos versus os valores ajustados é utilizado para verificar a pressuposição de que os resíduos são aleatoriamente distribuídos e tem variância constante. Percebe-se na Figura 3 que os resíduos encontram-se situados aproximadamente em torno da faixa centrada em zero, onde tem uma boa indicação de validação de independência, pois não apresenta configurações especiais como presença de resíduos positivos ou negativos. Porém verifica-se a presença de pontos discrepantes (*outliers*). Verifica-se também na figura que a variabilidade dos resíduos aparentam aumentar ao longo dos valores ajustados, além de apresentar alguns pontos muitos distantes dos demais o que leva a suspeitar que a suposição de homocedasticidade pode estar violada.

Figura 3 – Gráfico dos resíduos contra os valores ajustados.



Na Figura 4 está representado o gráfico dos quantis, também chamado de *qqplot*. Este gráfico compara os quantis amostrais dos resíduos do modelo com os quantis teóricos de uma distribuição normal, e pode ser usado para avaliar a suposição de normalidade. Assim, quando a configuração dos pontos no gráfico se aproxima de uma reta, a suposição de normalidade é sustentável, o que não acontece aqui e indica que os resíduos do modelo ajustado não parecem seguir distribuição normal, principalmente nas caudas. Que podem ser um problema na distribuição dos dados.

Figura 4 – Gráfico do *qqplot* para a distribuição normal.



4 Conclusão

A análise de regressão linear múltipla é muito utilizada na tentativa de se explicar a associação entre variáveis, com essa técnica estatística é possível escrever um modelo associando uma variável dependente com outras variáveis independentes, através de uma função linear.

A aplicação desse banco de dados da PeNSE 2015, com o uso da análise de regressão múltipla, para o ajuste de uma função linear envolvendo seis variáveis explicativas na tentativa de explicar a prática de atividade física entre estudantes do 9° ano do Ensino Fundamental II na cidade de João Pessoa-PB, onde foi observado que três destas variáveis foram significativas (tempo estimado, idade e gênero). Porém, os gráficos de resíduos apontaram alguns problemas na distribuição dos dados, de modo que o gráfico *qqplot* não parece seguir uma distribuição normal.

Neste trabalho foi usado o modelo de regressão linear múltipla, para os dados da PeNSE 2015 no município de João Pessoa-PB, onde foram verificadas as variáveis tempo estimado, gênero, raça/cor, trabalho e corpo para a prática de atividade física na última semana que antecedeu a pesquisa. Com isso o modelo foi ajustado usando métodos computacionais implementados no programa R (R Core Team, 2014), onde três das variáveis independentes foram significativas. Pela parte gráfica dos resíduos contra valores ajustados percebe-se que resíduos encontram-se situados aproximadamente em torno da faixa centrada em zero, mas apresentam alguma evidência de heterocedasticidade. Já o gráfico *qqplot* evidenciou problemas acerca da distribuição dos resíduos, que não parecem seguir normalidade.

Deste modo, verificou-se que a modelagem usada não parece adequada para este banco de dados e os resultados obtidos não são confiáveis e passíveis de interpretação. Como trabalhos futuros seria importante que algumas transformações fossem avaliadas ou até mesmo outros métodos estatísticos, como os modelos lineares generalizados fossem avaliados aqui na busca de outras opções mais satisfatórias para o presente banco de dados.

Referências

ALMEIDA, H. M. *Análise de Regressão Linear Múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma industria de calçados* . 2014. Citado na página 18.

CHARNET, R. et al. *Análise de Modelos de Regressão Linear com Aplicação*. 2. ed. Campinas, SP: UNICAMP, 2008. Citado na página 14.

DEMÉTRIO, C. G.; ZOCCHI, S. *Modelos de Regressão*. [S.l.: s.n.], 2006. Citado na página 12.

HOFFMAN, R. *Análise de Regressão - Uma introducão á econometria*. 4ª. ed. [S.l.]: Hucitec, 2006. v. 378. Citado na página 25.

MORAIS, N. F. *Análise de Regressão Linear com estudo de caso em acidentes de trânsito*. 2010. Citado na página 18.

NAGHETTINI, M. A. *Hidrologia Estatistica*. [S.l.: s.n.], 2007. Citado na página 12.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Citado 2 vezes nas páginas 26 e 32.

SASSI, C. P. et al. *Modelos de regressão linear múltipla utilizando os softwares R e Statistica: uma aplicação a dados de conservação de frutas*. 2011. Disponível em: <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_377.pdf>. Acesso em: 02 de fevereiro de 2018. Citado na página 18.

TAVEIRA, A. P. *Uso de Regressão Linear Múltipla para caracterizar a relação funcional entre o tamanho do cerébro e o nível de inteligência*. 2015. Citado na página 15.

APÊNDICE A – Rotina utilizada na análise

Este apêndice apresenta a rotina da análise dos dados utilizada no programa R.

```
## Carregando os dados já salvos
load("dados_Vitoria.RData")

## Ajustando o modelo de regressão logística para sex ~ beb + fum + dro
fit.model <- lm(tempT ~ tempE + ida + gen + rac + tra + cor,
data=dados.Vitoria)

## Verificando o ajuste do modelo
summary(fit.model)

## Tabela de ANOVA
anova(fit.model)

## Reajustando o modelo com as variáveis significativas
fit.model <- lm(tempT ~ tempE + ida + gen, data=dados.Vitoria)

## Verificando o ajuste do modelo
summary(fit.model)

## Tabela de ANOVA
anova(fit.model)

## Diagnóstico do modelo
attach(dados.Vitoria)
dados.Vitoria
summary(dados.Vitoria)
plot(fit.model)

## boxplot por gênero
boxplot(tempT~gen, data=dados.Vitoria)

## boxplot por idade
boxplot(tempT~ida, data=dados.Vitoria)

## gráfico de dispersão entr tempE e tempT
plot(tempT~tempE, data=dados.Vitoria)
```