Estudo das relações entre peso e altura de estudantes de estatística através da análise de regressão simples.

Wanessa Luana de Brito COSTA[1], Adriana de Souza COSTA[1.] Tiago Almeida de OLIVEIRA[1]

[1] Departamento de Estatística, Universidade Estadual da Paraíba-UEPB, Campus I, Campina Grande-PB. E-mail: Wanessaluanabc@hotmail.com..
[1] Departamento de Estatística, Universidade Estadual da Paraíba-UEPB, Campus I, Campina Grande-PB. E-mail: adriana_scsouza@hotail.com. [1] Departamento de Estatística, Universidade Estadual da Paraíba–UEPB, Campus I, Campina Grande-PB. E-mail: tadolive@cct.uepb.br

RESUMO

Este trabalho teve por objetivo estudar a relação entre duas variáveis, peso e altura de estudantes de estatística da UEPB, para medir o grau de influência que a variável independente altura tem sobre a variável dependente peso. Para o mesmo foi utilizada uma amostra de 30 alunos. Para verificar a relação entre as variáveis utilizou-se o método de regressão linear simples, que permite descrever a associação entre as respectivas variáveis do modelo. O método baseia-se inicialmente em medir a associação da variável dependente e a variável independente, a presença ou ausência de relação linear pode ser investigada sob dois pontos de vista: correlação e regressão. A amostra em estudo apresenta uma altura media de 1,66m e um peso médio de 62 kg. Com base nos valores analisados através do modelo, observou-se que existe correlação linear positiva entre as variáveis, ou seja, à medida que a altura cresce o peso aumenta. Com base nos resultados o modelo de regressão linear simples foi adequado aos dados.

PALAVRAS CHAVE: Correlação, regressão linear, resíduos.

1 INTRODUÇÃO

Para haver um desenvolvimento completo e saudável do corpo humano, desde criança é necessário ter uma boa alimentação, higiene, imunização contra doenças e cuidados com o meio ambiente. Esses fatores influenciam diretamente a vida, pois irá garantir que o sistema imunológico esteja mais preparado para combater doenças, e consequentemente favorecerá uma vida melhor.

A altura e o peso estão relacionados, não há um índice fixo, mas de forma geral, quanto maior a altura maior o peso do individuo, a relação entre peso e altura pode ser calculada de várias maneiras, um dos métodos utilizados é o índice de massa corporal (IMC) é um parâmetro bastante utilizado para classificar o individuo de acordo com seu peso e altura. Seu uso é disseminado principalmente entre profissionais que trabalham com o corpo, como médicos, fisioterapeutas e profissionais de Educação Física. O IMC é determinado pela divisão da massa do indivíduo pelo quadrado de sua altura, onde a massa está em quilogramas e a altura está em metros. Vigiar o seu peso tem uma importância critica para a saúde. Quanto mais peso excessivo tiver, maior é o risco de desenvolver vários problemas de saúde e maior é a probabilidade de o seu peso excessivo encurtar sua vida. (FERNANDES, 2011).

Em função dos fatores citados, estudar a relação entre altura e peso dos estudantes. Para conhecermos essa relação podemos utilizar o modelo de regressão linear simples (MRLS) que descreve a associação entre duas variáveis.

O objetivo desta pesquisa é estudar a relação entre duas variáveis quantitativas, (peso e altura), dos estudantes de estatística da UEPB 2009.1, através da regressão linear simples, para verificar até que ponto a variável altura explica à variável dependente peso.

2 METODOLOGIA

Foram utilizados dados provenientes de peso e idade de estudantes da turma de estatística (UEPB 2009.1 Manhã), utilizou-se uma amostra de 30 estudantes.

TABELA 1- MATERIAL E MÉTODOS

Nº	Altura em (m)	Peso em (Kg)
01	1,58	54
02	1,56	48
03	1,77	70
04	1,59	56
05	1,63	63
06	1,58	60
07	1,82	97
08	1,68	66
09	1,76	86
10	1,60	52
11	1,73	62
12	1,51	42
13	1,54	51
14	1,67	58
15	1,72	86
16	1,78	100
17	1,74	69
18	1,75	65
19	1,55	55
20	1,73	72
21	1,67	55
22	1,68	53
23	1,64	45
24	1,73	70
25	1,72	58
26	1,56	46
27	1,68	56
28	1,68	57
29	1,68	75
30	1,55	47

Fonte: (Alunos de estatística UEPB, 2009.1, manhã).

Para analisar o relacionamento entre as variáveis em estudo, procedeu-se um diagrama de dispersão sobre o conjunto de dados da Tabela 1, que podem ser representados na forma dos pares ordenados (x_1, y_1), (x_2, y_2),....,(x_i, y_i) afim de que se tenha ideia a respeito do tipo de relação existente entre as variáveis, da variabilidade associada a elas e da presença de pontos atípicos (outliers).

Segundo Triola (2008) o coeficiente linear de Pearson r mede a intensidade da relação linear entre os valores quantitativos emparelhados x e y em uma amostra. O mesmo foi aplicado e é dado por:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Testou-se a hipótese de o coeficiente de correlação linear ser igual à zero ($H_{01}: \rho = 0$) vs a hipótese do coeficiente de correlação linear ser diferente de zero ($H_{02}: \rho \neq 0$)

A estatística de teste é dada por:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Onde, t_0 é a estatística do teste, n é o tamanho da amostra e r é a estimativa do coeficiente de correlação linear. A estatística do teste, t_0, segue uma distribuição t de Student com (n-2) graus de liberdade, sob a plausibilidade da hipótese nula $H_0: \rho = 0$. A hipótese nula é rejeitada se:

$$\text{Valor P} < \alpha$$

Onde α é o nível de significância adotado previamente (α = 0,05).

O modelo adotado para descrever a relação entre uma variável explanatória x e a variável resposta y foi o modelo de regressão linear simples. O modelo faz a seguintes suposições, em ordem decrescente de importância: o valor médio da variável resposta é uma função linear de x; a esperança dos erros é igual a zero; a variância dos erros é constante; os erros são independentes; $\varepsilon_i \sim N(0, \sigma^2)$.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

Em que,

y_i: Valor da variável dependente (resposta) para o i-ésimo elemento da amostra, X: Valor (conhecido) da variável independente para o i-ésimo elemento da amostra, β_0 e β_1: São parâmetros desconhecidos (a serem estimados) e ε_i é o Erro amostral.

Segundo (ET AL CHARNET 2008) o método dos mínimos quadrados analisa as n-diferenças para cada reta candidata entre cada valor y e o valor na reta, correspondentes ao respectivo valor x. A reta selecionada é a reta que apresenta a menor soma de quadrados de tais diferenças. O método foi utilizado têm-se os estimadores para os parâmetros de interesse:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{e} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

Obtendo-se, portanto a equação da reta ajustada dada por:

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

Após o ajuste do modelo de regressão linear, utilizou-se a análise de variância para verificar a significância do MRLS. Tem-se a seguinte relação:

De acordo com TRIOLA (2008) a soma de quadrado total (SQT), pode ser quebrada em componentes de SQreg (Regressão) e SQres (Resíduo ou Erro). Tem-se a seguinte relação:

$$SQT = SQ\,\mathrm{Re}\,g + SQ\,\mathrm{Re}\,s$$

A soma de quadrados total de Y(SQT) mede a variabilidade dos valores observados em torno de sua média, cuja soma de quadrados é dada por:

$$SQT = \sum (y - \bar{y})^2$$

a soma de quadrados dos resíduos (SQres.), representa a variação que se supõe comum a todas as populações consideradas, expressa através de:

$$SQ\,\mathrm{Re}\,s = \sum (y - \hat{y})^2 = \sum (y - \hat{\beta}_0 - \hat{\beta}_1 x)^2 = \Sigma e^2$$

Modelo resultante das distâncias entre os valores do modelo e a média:

$$SQ\,\mathrm{Re}\,s = \Sigma (y - \bar{y})^2 - \Sigma (y - \hat{y})^2 = \Sigma (\hat{y} - \bar{y})^2$$

O Coeficiente de determinação (R^2**)** é uma medida de qualidade do modelo em relação à sua habilidade de estimar corretamente os valores da variável resposta Y, e é dado pela seguinte equação:

$$R^2 = SQ\,\mathrm{Re}\,s / SQT$$

Procedeu-se um teste para verificar a adequação do modelo. Para testar estatisticamente a falta de ajuste do MRLS, deve-se ter pelo menos dois valores da variável resposta para alguns valores da variável regressora e pode-se testar duas hipóteses de interesse:

H_{02}: O MRLS é adequado ou H_{12}: O MRLS é não adequado.

Utilizou-se a decomposição da soma de quadrados de resíduos (SQRes), dada por:

SQRes= SQFa + SQEp

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{k} n_i(y_{ij}-\bar{y}_i)^2,$$

Em que, n_i é o número de repetições da variável y. Análise de variância completa é dada na tabela 2.

Tabela 2: Esquema geral para análise de variância.

Fonte de Variação	GL	SQ	QM	F_{cal}
Altura	p-1	SQReg	SQReg/ p-1	QMReg/QMRes
Resíduo	n-2	SQRes	SQRes/n-2	-
(Falta de Ajuste)	(k-2)	SQFa	SQFa/k-2	QMFa/QMEp
(Erro Puro)	(n-k)	SQEp	SQEp/n-k	
Total	n-1	SQT		

$$H_{01}: \rho = 0 \;\; vs \;\; H_{02}: \rho \neq 0$$

Segundo (ET AL CHARNET 2008) Os resíduos de um modelo de regressão linear têm uma relação muito forte com a qualidade do ajuste feito, bem como a confiabilidade dos testes estatísticos sobre os parâmetros do modelo.

Para melhor analisar os resíduos levou-se em conta sua variabilidade, obtiveram-se os resíduos transformados dados por:

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}, i = 1, 2, ..., n$$

$$d_i^* = \frac{e_i}{\sqrt{\hat{\sigma}_{(i)}^{\,2}(1-h_{ii})}}, i = 1, 2, ..., n$$

Neste caso, temos $\sigma^2_{(i)}$ como a soma de quadrados média dos resíduos do MRLS, ajustando se utilizarmos a i-ésima observação. Com essa transformação, temos que a variável aleatória d_i tem distribuição t de Student com (n-1-2) graus de liberdade (sob a hipótese de o MRLS ser adequado). A normalidade dos resíduos foi verificada por meio da função da probabilidade observada acumulada dos erros (Normal P-P Plot).

3 RESULTADOS E DISCUSSÃO

De início foi realizada uma análise descritiva dos dados onde foi observada a altura e o peso médio dos alunos (Tabela 3).

TABELA 3 - Análise descritiva para as variáveis peso e altura.

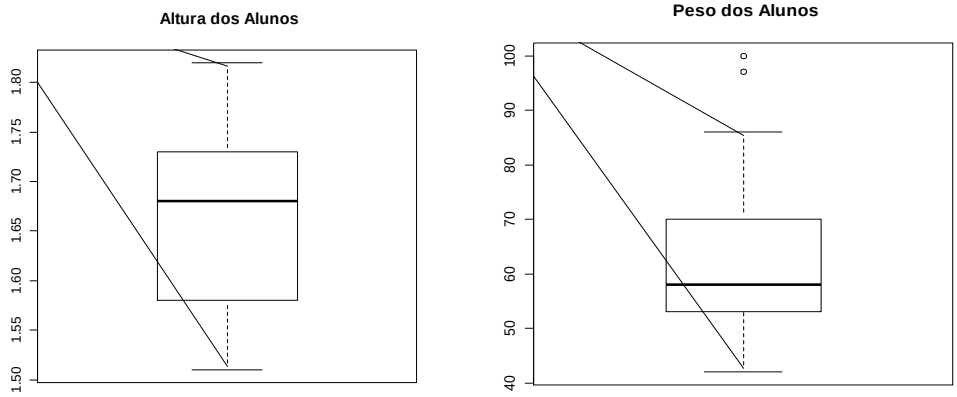
Variáveis	Valor Mínimo	1º Quartil	Mediana	Media	3º Quartil	Maximo
Altura	1,510	1,582	1,680	1,663	1,730	1,820
Peso	42	53,25	58.00	62.47	69.75	100

Fonte: (Alunos de estatística UEPB, 2009.1, manhã).

Na tabela 3 observa-se uma altura media de 1,66m e um peso médio de aproximadamente 62 kg.

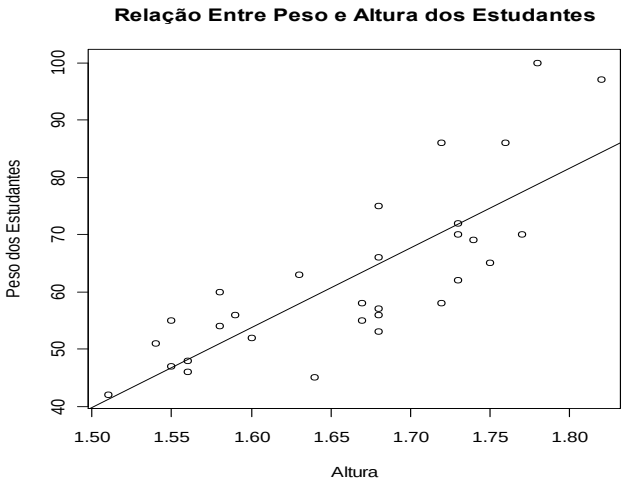
No Box Plot para as variáveis em estudo percebe-se a ocorrência de dois pontos atípicos, porém os mesmos não alteram as suposições do modelo.

Figura 1 – Box Plot para as variáveis Peso e Altura.



Fonte: (Alunos de Estatística UEPB, 2009.1- Manhã).

Figura 2 – Gráfico de dispersão com a reta de regressão ajustada para a relação entre peso e altura dos alunos.



A figura 2 mostra que existe uma correlação positiva entre as variáveis, pois à medida que a altura dos estudantes aumenta o seu respectivo peso também aumenta.

A correlação linear de Pearson foi de r=0,795 (79,5%), e com um índice de confiança de 95%, pode-se afirmar que existe uma correlação linear positiva entre as variáveis, isto é x e y (peso e altura) variam no mesmo sentido à medida que a altura cresce o peso aumenta.

Ajustou-se o modelo aos dados e estimou-se os parâmetros β_0 e β_1 e foram obtidos os seguintes resultados:

$$\widehat{\beta_0} = -168{,}7 \text{ e } \widehat{\beta_1} = 139{,}1$$

Portanto, tem-se a equação da reta ajustada dada por:

$$\hat{y} = -168{,}7 + 139{,}1 \times \quad (1)$$

TABELA 3 - Análise de Variância para peso e altura dos estudantes de estatística.

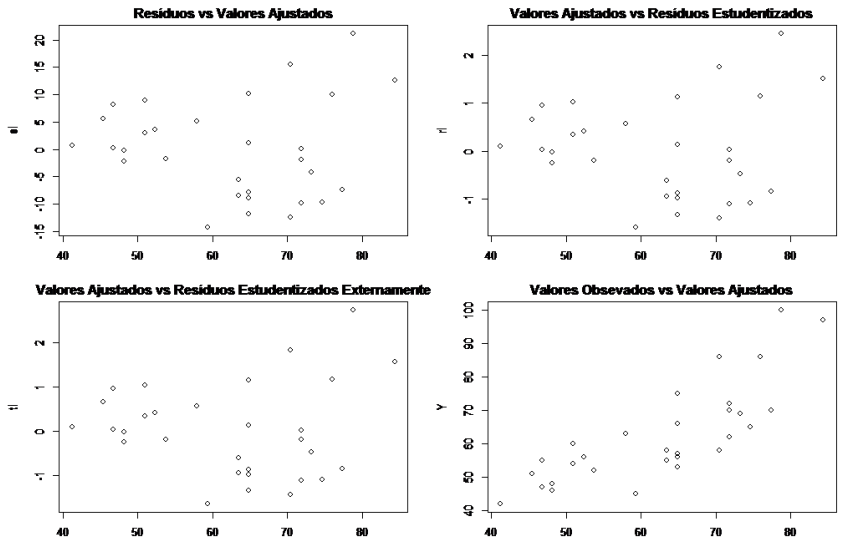
Fonte de Variação	GL	SQ	QM	F_{cal}	Valor P
Altura	1	3948,1	3948.1	48.034	< 0.001
Resíduo	28	2301,4	82.2		
Falta de Ajuste	17	1471,7	86,6	1,1477	0,4182
Erro Puro	11	829,7	75,4		
Total	29	6249,5			

Fonte: (Alunos de estatística UEPB, 2009.1, manhã).

Na tabela 3, rejeita-se $H0_1$, ou seja, o coeficiente de inclinação da reta difere de zero. E aceita-se $H0_2$ indicando um bom ajuste do modelo aos dados. O coeficiente de determinação foi de 63,17%(R^2=0,6317), ou seja, o modelo ajustado explicou 63,17% da variação na variável resposta Y(peso dos estudantes).

Por meio da figura 3, podemos perceber uma aleatoriedade entre os resíduos, ou seja, existe homogeneidade nas variâncias. Validando o modelo de regressão ajustado.

Figura 3 – Resíduos para peso e altura dos estudantes de estatística.



CONCLUSÃO

Com base nos valores analisados através do modelo de regressão linear simples, observou-se que existe correlação linear positiva entre as variáveis (peso e altura), à medida que a altura cresce o peso aumenta. O coeficiente de determinação R^2 nos mostra que o modelo ajustado explicou 63,17% da variação sobre a variável peso dos estudantes. Conclui-se que o modelo de regressão linear simples ajustado foi adequado aos dados.

REFERÊNCIAS

TRIOLA, F, M**. Introdução á Estatística,** 10 ed. Rio de janeiro, RJ: LTC, 2008.

TAKIUT, D. A. **O Desenvolvimento da criança nos primeiros anos de vida,** acesso em 20/10/2012, disponível em **http://www2.videolivraria.com.br/pdfs/14871.pdf.**

CHARNET,R., FREIRE,C.A.L., CHARNET,E.M.R., BONVINO,H., **Análise de Modelos de Regressão Linear,** Campinas, SP: Ed. UNICAMP, 2008.

MARTINS, A, G, **Estatística Geral e Aplicada**. 3 ed. São Paulo, SP: Atlas, 2008.

DEMÉTRIO, G.B. CLARICE. Modelos de regressão. Pg.19-21, 107-120.

FERNANDES, J. **A importância do Controle do Peso** acesso em 14/11/2011, disponível em **http://nutricionista.com.pt/artigos/a-importancia-do-controlo-do-peso.jhtml.**