ESTUDO SOBRE A CORRELAÇÃO E A REGRESSÃO LINEAR EM LIVROS DIDÁTICOS DO ENSINO SUPERIOR NO BRASIL

Study on correlation and linear regression in didactic books of higher education in brazil

Ailton Paulo de Oliveira Júnior
Daniel de Freitas Barros Neto
Gisele Cristiane Silva Alves

Resumo

O presente trabalho tem por objetivo analisar como são apresentados os conteúdos de Correlação Linear e Regressão Linear dentre os dez livros didáticos mais utilizados no Ensino Superior em todas as regiões do Brasil (OLIVEIRA JÚNIOR, 2011), considerando a referência à análise da significância aborda aspectos relacionados à relação entre as variáveis (grave ou moderada; direta ou inversa; etc.) e a identificação do tipo de atividades apresentadas nos livros didáticos indicando exemplos que façam a distinção entre os conceitos fundamentais da Correlação e Regressão, apresentar exercícios resolvidos e a serem resolvidos pelo aluno, analisar os contextos em que surgem os problemas-exemplo e as características de correlação. Observa-se que grande parte dos livros oculta à distinção entre dependência funcional e estatística. O problema da regressão é comumente tratado nos livros a partir de um diagrama de dispersão. E conjectura um modelo de função que melhor aproxima dos dados, sendo que este pode ser linear ou não.

Palavras-chave: Livros Didáticos; Correlação e Regressão linear; Ensino; Educação Superior; Brasil.

Abstract

This paper aims to analyze how the contents of Linear Correlation and Linear Regression are presented among the ten textbooks most used in Higher Education in all regions of Brazil (OLIVEIRA JÚNIOR, 2011), considering the reference to the analysis of significance addresses aspects. related to the relationship between the variables (severe or moderate; direct or inverse; etc.) and the identification of the type of activities presented in the textbooks indicating examples that distinguish between the fundamental concepts of Correlation and Regression, present solved exercises and to be solved by the student, analyze the contexts in which the example problems arise and the correlation characteristics. Most of the books hide the distinction between functional and statistical dependence. The problem of regression is commonly dealt with in books from a scatter diagram. And it conjectures a function model that best approximates the data, which can be linear or not.

Keywords: Didactic books; Correlation and linear regression; Teaching; Higher education; Brazil.

Introdução

Consideramos a importância e utilidade do ensino de correlação e regressão para alunos da Educação Superior e ainda destacamos que seu ensino e aprendizagem não estão isentos de problemas didáticos.

Destacamos, entre outras, as dificuldades de compreensão de alunos em torno dos conceitos de covariância e correlação, como não distinguir uma distribuição bidimensional de dois conjuntos de dados independentes; a concepção unidirecional da correlação (aceita apenas a correlação direta); a oposição entre o raciocínio numérico e o gráfico na estimação da correlação; ou a concepção causal (correlação confusa e causalidade) ao estimar uma correlação significativa (ESTEPA, 2008; ZIEFFLER, GARFIELD, 2009).

Além disso, sabe-se que os livros didáticos apresentam tradição no que envolve o uso deste recurso em sala de aula e em diferentes níveis de ensino, uma vez que é definido como uma invariante da escola como um material estável e mais duradouro na história da escola, embora sujeita a modificações e transformações (BRAGA; BELVER, 2016).

Consideramos alguns estudos que foram realizados na avaliação de livros didáticos que apresentam conteúdos de correlação e regressão:

1) Apresentação da correlação e regressão (SÁNCHEZ COBO, 1999; SÁNCHEZ COBO; ESTEPA; BATANERO, 2000, LAVALLE; MICHELI; RUBIO, 2006);
2) Análise em problemas de correlação e regressão (GEA, et al., 2013);
3) Análise da linguagem utilizada no ensino de correlação e regressão (GEA et al., 2014);
4) Análise do sentido de correlação e regressão, descrevendo os componentes da cultura e raciocínio estatístico (GEA; BATANERO; ROA, 2014);
5) Análise das definições dos conceitos ligados à correlação e regressão (BATANERO et al., 2016).

Do ponto de vista pedagógico o livro é uma ferramenta de ensino inerente ao processo de instrução que visa facilitar a aprendizagem por meio de uma linguagem que facilita a construção do conhecimento (PELLICER, 2007).

Para Braga e Belver (2016) na seleção dos livros didáticos deve-se considerar a representação do conhecimento científico e cultural a ser comunicada aos estudantes em uma situação histórica particular, envolvendo a transmissão de valores, ideologias etc.

E para facilitar a aprendizagem dos alunos, seria necessário assegurar uma apresentação correta dos conceitos relacionados à correlação e à regressão nos livros didáticos, o que se torna um recurso de grande importância para o aluno e para o professor (HERBEL, 2007; CORDERO; FLORES, 2007), sendo um assunto pouco investigado até hoje.

Portanto, o presente trabalho tem como objetivo realizar um estudo dos conceitos de correlação linear e regressão linear nos dez livros estatísticos mais utilizados no ensino superior (público e privado) do Brasil, identificados por Oliveira Júnior (2011) e indicar aspectos que são considerados essenciais para o seu ensino.

A correlação e regressão linear na Educação Superior

Para considerar se os livros didáticos atendem aos pressupostos teóricos da correlação e regressão, Batanero e Díaz (2008) destacam que uma das questões a qual procura-se responder em grande parte das pesquisas científicas é se existe um relacionamento, por exemplo, entre duas variáveis X e Y incluídas em um estudo. Se assim for, é imediato considerar a possibilidade de encontrar uma fórmula que expressa exatamente os valores de uma variável, dependendo da outra com um efeito preditivo.

De acordo com Gea et al. (2013), frequentemente realizamos associações e relações entre duas ou mais variáveis, buscando identificar como e se elas se correlacionam. A correlação e a regressão são conceitos estatísticos fundamentais que estendem a ideia de dependência funcional, relacionando-se com conceitos como a covariância, a distribuição, a centralização e a dispersão.

Para Jales (2014) a análise do coeficiente de correlação fundamenta-se no grau ou intensidade de associação entre variáveis, tratando-se de medida numérica que expressa a força da relação entre variáveis que representam dados quantitativos.

Sánchez Cobo (1998) considera ser preciso oferecer aos alunos situações de aprendizagem que mostrem a diversidade dos tipos de covariância (dependência, intensidade, sinal) de modo que possam eliminar concepções errôneas manifestadas.

Bussab e Morettin (2002) dizem que a análise da regressão modela a relação entre as variáveis, pois averigua a existência e o grau de dependência estatística entre as variáveis aleatórias. Em outras palavras, se constata um modelo (linear ou não) que descreve a relação entre variáveis, por meio do gráfico e da função que melhor representa a relação.

Razak et al. (2017) realizaram um estudo de caso com alunos que estavam sendo expostos a alguns conceitos teóricos dos tópicos de correlação e regressão para investigar sua capacidade de calcular e interpretar o coeficiente de correlação de Pearson e a inclinação da regressão. Os

resultados revelaram que um baixo percentual de alunos (19,43%) completou com sucesso a sua interpretação do coeficiente de correlação e 33,18% dos alunos conseguiram interpretar completamente o valor calculado da inclinação de regressão.

Portanto, para a compreensão correta desses conceitos é de vital importância dominar os processos de interpretação entre as representações (tabelas e diagramas de dispersão), bem como os resultados estatísticos (coeficiente de correlação).

Procedimentos metodológicos

Neste estudo, abordaremos a identificação dos problemas em livros didáticos, classificando, teoricamente, as principais problemáticas que dão sentido à Correlação e Regressão. Este estudo será realizado por meio da análise da significância da Correlação e Regressão, observando as principais questões elencadas por Estepa et al. (2012), ou seja: (1) Existe alguma relação entre as variáveis? (2) É grave ou moderada? (3) Direta ou inversa? (4) Posso usar uma variável para prever outra variável?

E um segundo aspecto será a identificação do tipo de atividades apresentadas nos livros didáticos, de acordo com Ortiz (1999): (1) Indicar exemplos que façam a distinção entre os diferentes aspectos da Correlação e Regressão; (2) Apresentar exercícios resolvidos; (3) Apresentar exercícios para serem resolvidos pelo aluno; (4) Analisar os contextos em que surgem os problemas-exemplo que mostram a aplicação dos conceitos estatísticos; (5) Analisar as características de correlação (intensidade, sinal, tipo) dos dados utilizados no livro texto como proposto por Sánchez Cobo (1999).

Assim, foram analisados os livros didáticos de Ensino Superior, identificados em Oliveira Júnior (2011) como os mais utilizados por 334 professores que ensinam conteúdos estatísticos nas universidades públicas e privadas, nas diversas áreas do conhecimento (Exatas, Saúde e Humanas), em todas as unidade federativas do Brasil. No caso deste trabalho ainda foi identificado aqueles livros didáticos que abordavam os conceitos de correlação e regressão.

Oliveira Júnior (2011) indicou que a seleção dos professores foi aleatória, de acordo com os seguintes procedimentos: (1) Identificar as Instituições de Ensino Superior (IES) do Brasil, divulgado no Cadastro das Instituições e Cursos de Educação Superior do Ministério da Educação INEP (Instituto Nacional de Estudos e Pesquisas Educacionais); (2) Visitar os sites de todas as instituições dos estados da federação para verificar quais cursos ofereciam disciplinas que focavam conteúdos estatísticos; (2) Buscar o endereço de e-mail dos cursos selecionados e enviamos mensagem solicitando participação dos professores na pesquisa; (3) Fazer contato por meio dos coordenadores de curso que enviaram a mensagem diretamente para seus professores ou diretamente pelo pesquisador quando o site disponibilizava a lista de e-mail destes professores ou o coordenador enviasse a lista de e-mail destes.

Assim, Oliveira Júnior (2011) solicitou, dentre outas questões, via questionário, que os professores citassem os livros didáticos que estes utilizam em sala de aula para ministrar suas aulas que abordam conceitos estatísticos e probabilísticos e assim pode-se identificar os mais utilizados em instituições públicas e provadas de educação superior.

Na Tabela 1 citamos os livros didáticos mais utilizados pelos professores de Estatística participantes desta pesquisa para o ensino da Correlação Linear e da Regressão Linear.

Tabela 1 - Frequência da utilização de livros didáticos por professores de instituições pública e privadas de Ensino Superior no Brasil, com conteúdo de Correlação e Regressão Linear.

Autores Livros Didáticos	Pública		Privada		Ambas	
	n*	%	n*	%	n*	%
Bussab e Morettin (2002)	14	17,7	16	9,5	3	10,7
Barbetta (1994)	3	3,8	13	7,7	2	7,1
Stevenson (1981)	-	0,0	17	10,1	1	3,6
Triola (1999)	5	6,3	12	7,1	1	3,6
Crespo (2009)	-	0,0	12	7,1	3	10,7
Vieira (2008)	3	3,8	10	6,0	1	3,6
Anderson, Sweeney e Willians (2007)	1	1,3	8	4,8	1	0,0
Callegari-Jacques (2003)	2	2,5	4	2,4	1	3,6
Larson e Farber (2010)	2	2,5	2	1,2	1	0,0
Toledo e Ovalle (1994)	1	2,5	3	1,2	1	0,0

* Número de ocorrências.
Fonte: Oliveira Júnior (2011).

Resultados

Como já apresentado neste trabalho, ao se estudar uma possível relação entre duas variáveis, consideramos as questões elencadas por Estepa et al. (2012) e Ortiz (1999). Tais perguntas derivam campos de problemas propostos nos livros analisados conforme destacado por Gea et al. (2013) e que nortearão as análises do que foi identificado como essencial para o ensino dos conceitos básicos de correlação e regressão.

Organização de dados bidimensionais e representação em registro gráfico e tabular

Considera-se que é necessário incluir exercícios ou problemas que conduzam os leitores a representar os dados, bem como a leitura de representações gráficas e tabulares associadas a dados bidimensionais.

No ensino e aprendizagem de Estatística, as informações representadas na forma de tabelas ou gráficos, é necessário que sejam exploradas a leitura e a interpretação dos dados ali representados. Desta forma busca-se que o aluno tenho um senso crítico dos conhecimentos estatísticos, como salienta Miranda (2008).

Oliveira Júnior e Alves (2017) dizem que o diagrama de dispersão e a tabela de frequência da distribuição binomial podem mostrar, de forma subjetiva, se existe ou não uma relação entre as variáveis, e assim introduzir o conceito de correlação.

Observou-se que em todos os livros analisados neste trabalho se faz uso de tabelas e gráficos, sendo comum chamar a atenção para a interpretação do diagrama de dispersão, uma vez que essa forma de representação pode dar indícios do tipo de relação entre as variáveis.

Vale pontuar atividades que sugerem a construção da tabela ou do diagrama de dispersão, como indicado por Sánchez Cobo (1998). Destacamos Larson e Farber (2010) que sugerem exemplos como construir o diagrama de dispersão (Figura 1), pelo uso de tabelas por meio dos pares ordenados (x, y).

Figura 1 - Exemplo solicitando a construção de um diagrama de dispersão.

Tente você 2 Um gerente de marketing conduz um estudo para determinar se há uma relação linear entre a idade da pessoa e o número de revistas que cada pessoa assina. Os dados são mostrados na tabela a seguir. Mostre os dados no diagrama de dispersão e determine o tipo de correlação.

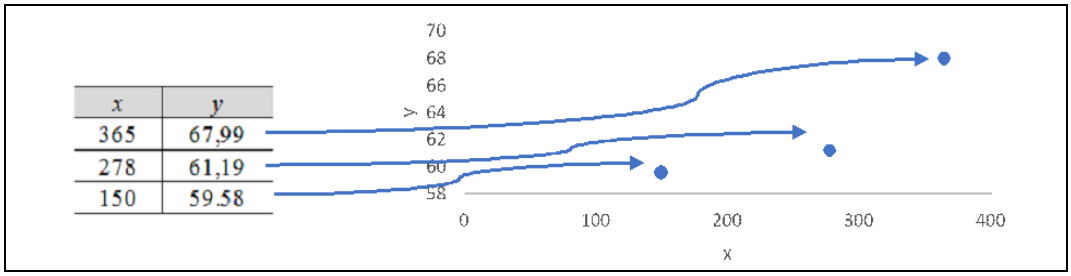
Idade, x	55	48	26	21	33	50	64	35
Assinaturas, y	2	3	0	4	3	0	6	1

a. Desenhe e nomeie os eixos x e y.
b. Faça o gráfico de cada par ordenado.
c. Parece haver uma correlação linear? Se sim, interprete a correlação no contexto dos dados

Fonte: Larson e Farber (2010, p.397).

Em Larson e Farber (2010) são discutidas técnicas estatísticas focadas em situações do mundo real. E destacamos também Barbetta (1994) que apresenta um esquema para a construção dos dados bidimensionais (Figura 2).

Figura 2 – Esquema apresentado no livro que explica como deve ser elaborado o diagrama de dispersão.

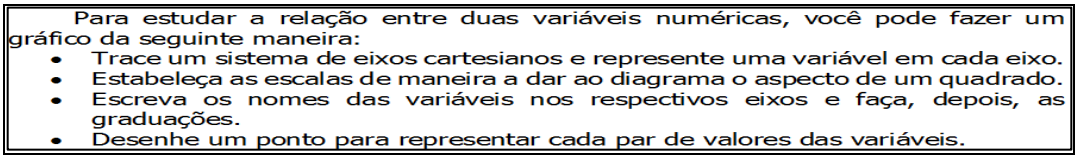


Fonte: Barbetta (1994, p. 252).

Bussab e Morettin (2002) e Barbetta (1994) apresentam histogramas como meio de representação quando é feita análise dos resíduos do modelo linear. E Bussab e Morettin (2002) trazem aplicações logo após as seções teóricas, além disso, aplica a teoria por meio dos pacotes computacionais: *Minitab* (*software* estatístico), *Excel* (planilha eletrônica) e *Splus* (pacote estatístico).

Destacamos Vieira (2008) que exibe um "passo-a-passo", figura 3, que orienta a construção do diagrama de dispersão por meio dos dados dispostos em uma tabela. E ainda apresenta, exercícios e problemas comentados voltados para área da Saúde. Os conceitos são transmitidos mais pela intuição do que por demonstração, sendo os exemplos simples e voltados à área da saúde exigindo pouco trabalho de cálculo.

Figura 3 - Orientações apresentadas no livro que indicam a forma como deve ser realizada a construção do diagrama de dispersão.



Para estudar a relação entre duas variáveis numéricas, você pode fazer um gráfico da seguinte maneira:

- Trace um sistema de eixos cartesianos e represente uma variável em cada eixo.
- Estabeleça as escalas de maneira a dar ao diagrama o aspecto de um quadrado.
- Escreva os nomes das variáveis nos respectivos eixos e faça, depois, as graduações.
- Desenhe um ponto para representar cada par de valores das variáveis.

Fonte: Vieira (2008, p.109).

Analisar a existência de relação entre variáveis

Após organizar os dados graficamente e por meio de tabelas, e que foi apresentado nos tópicos anteriores, Gea et al. (2013) explicam que é necessário refletir sobre a dependência funcional ou estatística dos dados, e assim verificar a dependência linear entre as variáveis, bem como a intensidade e a direção, de acordo com os aspectos relacionados a seguir. Assim, seguiremos com a apresentação dos resultados.

Definir/Analisar as variáveis que indicam um estudo estatístico bidimensional

Seguindo as indicações de Gea et al. (2013) buscaremos apresentar a necessidade em delimitar cada uma das variáveis que se vai estudar, sendo necessário saber qual a variável dependente (Y) e a independente (X), bem como identificar as diferenças de ambas.

Observamos que grande parte dos livros sugere na introdução dos capítulos de Correlação e Regressão, exemplos que solicitam ao aluno analisar a relação entre duas variáveis e que consideramos importante para a apreensão dos conceitos básicos de correlação e regressão.

Como exemplo destas relações, indicamos Vieira (2008) que propõe exercício solicitando a determinação do coeficiente de correlação entre a idade gestacional e o peso ao nascer de recém-nascidos (Figura 4).

Figura 4 - Tabela da idade gestacional, em semanas, e o peso ao nascer, em quilogramas de recém-nascidos.

Idade gestacional	Peso ao nascer
28	1,25
32	1,25
35	1,75
38	2,25
39	3,25
41	3,25
42	4,25

Fonte: Vieira (2008, p. 127).

Também em Larson e Farber (2010) são indicadas as variáveis: orçamento de 25 filmes da Fox com os ganhos brutos mundiais; solicitando o estudo da relação entre essas variáveis e em Bussab e Morettin (2002) sugerem o mesmo estudo de relação, trazendo como exemplo o peso e a idade das pessoas, consumo familiar e renda, entre outros.

Analisar a existência de uma dependência funcional ou estatística

Uma vez que se definem as variáveis, Gea et al. (2013) expõem a necessidade de identificar se temos uma dependência funcional ou uma dependência estatística. Este problema, mais adiante, irá motivar a análise da regressão.

Dos livros analisados somente Crespo (2009) e Vieira (2008) realizam explanações sobre as diferenças entre a dependência funcional e a estatística. Para tanto, Crespo (2009) cita a relação entre o perímetro (p) e o lado l de um quadrado: $2p = 4l$, para exemplificar uma relação funcional (ou determinística), uma vez que é possível determinar exatamente o valor de $2p$. Posteriormente é considerada a relação entre peso e estatura de um grupo de pessoas, para exemplificar a relação estatística.

Gea et al. (2013) inclui nesse campo de problemas a discriminação da correlação e causalidade. Assim, Barbetta (2014) e Callegari-Jacques (2003) realizam uma breve abordagem sobre o tema. Entretanto, destacamos Stevenson (1981), Triola (1999) e Larson e Farber (2010) que realizaram comentários importantes, pautando em exemplos para melhor compreensão dos leitores.

Para tanto mencionamos Stevenson (1981), um dos livros analisados neste trabalhos, que apresenta que quando duas variáveis estão correlacionadas, é possível predizer valores de uma delas com base no conhecimento de outra. Isso leva frequentemente à conclusão errônea de que uma variável é causa da outra. E isso é particularmente verdadeiro quando a variável "causal" precede a outra variável no tempo. Entretanto, o fato de haver um relacionamento matemático entre duas variáveis nada nos diz quanto a causa e efeito. Logo, há três explanações possíveis para a obtenção de uma correlação: existe uma relação de causa e efeito; ambas variáveis se acham relacionadas com uma terceira; ou a correlação é devida ao acaso.

Determinar a intensidade da relação

A partir do momento em que se identifica a relação entre as variáveis é preciso analisar a intensidade que variará a independência funcional. Assim, por meio do diagrama de dispersão podemos perceber esta análise de forma intuitiva, porém é por meio do coeficiente de correlação que podemos obter informações mais precisas da intensidade das variáveis relacionadas. Este campo inclui o cálculo do coeficiente de correlação linear e o cálculo do coeficiente de determinação, como destacado por Gea et al. (2013).

Todos os livros apresentam os cálculos de r_{xy}, porém muitos, num primeiro momento utilizam o diagrama de dispersão a fim de identificar, subjetivamente, se a correlação é forte, moderada ou fraca. É comum neste momento falar sobre imaginar uma linha imaginária (ascendente/descendente) sobre os diagramas, de forma que já prepare o aluno para a introdução do conceito de regressão.

O uso da tabela que orienta os cálculos de r (coeficiente de correlação), também é observado em todos os livros analisados. Gea et al. (2013) afirmam que esse tipo de representação é muito útil quando é

preciso realizar cálculos intermediários para obtenção de fórmulas, podendo ser adicionado tantas linhas e colunas quanto forem necessárias como apresentado em Toledo e Ovalle (1994), figura 5.

Figura 5 - Tabela auxiliar trazida pelo livro que organiza os dados.

	Renda (R$100) - Y	Poupança (R$1.000) - X	X²	Y²	XY
	10	4	16	100	40
	15	7	49	225	105
	12	5	25	144	60
	70	20	400	4900	1400
	80	20	400	6400	1600
	100	30	900	10000	3000
	20	8	64	400	160
	30	8	64	900	240
	10	3	9	100	30
	60	15	225	3600	900
Σ	**407**	**120**	**2152**	**26769**	**7535**

Fonte: Toledo e Ovalle (1994, p.417).

E, por meio do cálculo do coeficiente de correlação Callegari-Jacques (2003) apresenta uma tabela que orienta quanto à intensidade da relação, uma vez que r varia entre 0 e 1 (Figura 6), cujos valores próximos de -1 e +1 indicam forte correlação linear e próximos de 0 indicam ausência de correlação linear.

Figura 6 - Avaliação Qualitativa do grau de correlação entre duas variáveis.

|r|	A correção é dita
0	Nula
0 ─ 0,3	Fraca
0,3 ├─ 0,6	Regular
0,6 ├─ 0,9	Forte
0,9 ├─ 1	Muito Forte
1	Plena ou perfeita

Fonte: Callegari-Jacques (2003, p.90).

Consideramos ainda nesse campo a correlação espúria, onde apenas Barbetta (1994), Stevenson (1981), Vieira (2008) e Toledo e Ovalle (1994) mencionam:

> Quando duas variáveis X e Y forem independentes, o coeficiente de correlação será nulo. Entretanto, algumas vezes, isto não ocorre, podendo, assim mesmo, o coeficiente apresentar um valor próximo de ± 1. Neste caso a correlação é espúria (TOLEDO; OVALLE, 1994, p. 416).

Convém destacar a investigação do estatístico Gustav Fischer apresentado em Box, Hunter e Hunter (1978), mencionado como exemplo por Vieira (2008). Fischer apresentou um gráfico da população da cidade de Oldenburg em um período de sete anos (1930-1936) e observou o número de cegonhas em cada ano. Identificou uma correlação positiva entre o número de recém-nascidos e o número de cegonhas:

> A correlação entre essas duas variáveis é espúria: não indica relação de causa e efeito. Existe uma terceira variável, o crescimento da cidade, que implicava tanto no número de recém-nascidos (quanto maior a cidade, mais crianças nascem) quanto no número de casas com chaminés, perto das quais as cegonhas faziam seus ninhos. (VIEIRA, 2008, p. 120).

Outro aspecto é trazido por Sánchez Cobo (1998) quando ressalta que existem alunos que confundem o coeficiente de correlação com o de determinação. É comum em alguns livros restringir a definição do coeficiente de determinação como o quadrado do coeficiente de correlação, interpretado com uma medida descritiva da proporção da variação de Y que pode ser explicada por X, segundo o modelo especificado. (BARRETT, 2000 apud GEA et al., 2013).

Tal fato foi fortemente notado em Barbetta (1994), Toledo e Ovalle (1994), Triola (1999) e Anderson, Sweeney e Willians (2007). Convém ainda mencionar Crespo (2009) que não realiza nenhuma abordagem sobre R^2, que é uma das formas de avaliar a qualidade do ajuste do modelo, indicando

quanto esse foi capaz de explicar os dados coletados.

Entendemos que o uso de exemplos torna a explicação do coeficiente de determinação mais clara, como notamos em Vieira (2008):

> [...] Imagine que você quer comprar uma camiseta para uma criança. Você chega à loja e pede ajuda à vendedora. O que primeiro ela pergunta? A idade da criança, claro. Por quê? Porque o tamanho de uma criança é função da idade. Boa parte da variação do tamanho das crianças é explicada pela variação de suas idades - o que é medido pelo R^2. Portanto, saber a idade da criança ajuda na previsão do tamanho da sua camiseta (VIEIRA, 2008, p. 144).

Em Anderson, Sweeney e Willians (2007) é destacado a utilização de tecnologia computacional para aplicações em regressão linear e a determinação da equação de regressão e o coeficiente de determinação:

> Em aplicações modernas de análise de regressão, um software de computador é quase sempre usado para fazer os cálculos necessários para determinar a equação de regressão estimada e o valor do coeficiente de determinação. No entanto, quando resolvemos um problema pequeno com uma calculadora, podemos obter eficiências computacionais usando fórmulas alternativas SST (total da soma dos quadrados) e SSR (soma dos quadrados devida à regressão). Ilustramos esse procedimento com o exemplo da Armand's Pizza Parlors (ANDERSON; SWEENEY; WILLIANS, 2007, p. 456).

A resolução de exemplos ou exercícios resolvidos que envolvam o cálculo de R^2, seguido de uma interpretação do resultado, torna-se conveniente no que se refere à compreensão do significado deste coeficiente, como em Vieira (2008), figura 7.

Figura 7 - Exemplo para o cálculo e interpretação do coeficiente de determinação.

Exemplo 7.9. Coeficiente de determinação.
Calcule o coeficiente de determinação para os dados apresentados na Tabela 7.2 e na Tabela 7.5 e discuta cada uma delas.
Usando os cálculos intermediários já apresentados na Tabela 7.3, é possível obter $R^2 = 0{,}994$. Isto significa que 99,4% da variação da quantidade de procaína hidrolisada no plasma se explica pelo tempo decorrido após sua administração. Em outras palavras, se você souber o tempo que decorreu depois que a procaína foi colocada no plasma, poderá justificar 99,4% da variação de procaína que hidrolisou.
Para dados da Tabela 7.5, com a ajuda de um computador (ou de seu professor) é possível obter, $R^2 = 0{,}282$, um valor baixo. Se fosse alto, a explicação seria de que, dado o peso de um homem, a pressão arterial seria altamente previsível. No entanto, fatores como idade, vida sedentária, hereditariedade e certos hábitos, como o hábito de fumar e consumo abusivo de sal devem ser, também, importantes.

Fonte: Vieira (2008, p.145).

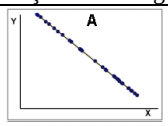
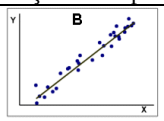
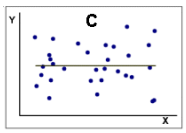
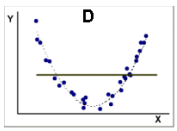
Após a determinação da reta de regressão, é preciso verificar se os dados amostrais podem ser descritos pelo modelo da equação, sendo necessário determinar a parcela de variabilidade amostral, que pode ser explicada a partir desta reta (NAGHETTINI; PINTO, 2007). E, para realizar esta verificação, considera-se o Coeficiente de Determinação que é a relação da variação explicada com a variação total (LARSON; FARBER, 2010).

Determinar a direção da relação entre variáveis

Neste campo pretende-se verificar se as variáveis estão correlacionadas direta (quando uma cresce a outra também cresce) ou inversamente (quando uma cresce a outra decresce ou diminui). Incluiremos também a relação curvilínea.

Usar o diagrama de dispersão para verificar a direção da relação entre as variáveis se torna muito comum nos livros analisados. Alguns estendem essa representação incluindo um tipo de relação não linear, como Larson e Farber (2010) (Figura 8).

Figura 8 - Diagramas de dispersão das direções da relação entre as variáveis apresentadas no livro.

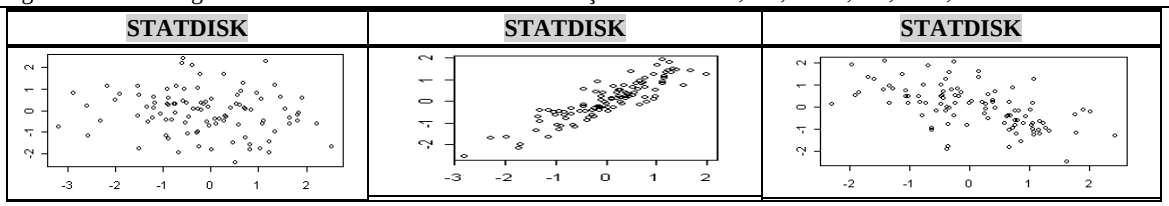
Correlação linear negativa	Correlação linear positiva	Não há Correlação	Correlação não linear
 Conforme x aumenta, y tende a decrescer.	 Conforme x aumenta, y tende a aumentar.		

Fonte: Larson e Farber (2010, p.395).

Como podemos observar, há uma tendência compatível nos livros, de que, por meio do diagrama de dispersão é possível, intuitivamente, verificar se existe ou não correlação, bem como a direção e a intensidade dela. Nesse sentido Triola (1999) apresenta a tarefa inversa por meio de uma atividade (Figura 9), isto é, associar a partir de três coeficientes de correlação qual o diagrama que melhor representa.

Figura 9 – Exemplo de problema de identificação de diagramas de dispersão utilizando calculadora científica.

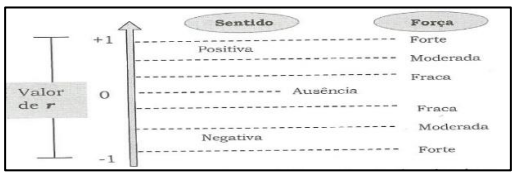
4. Identificação de Diagramas de Dispersão Dão-se, abaixo, três diagramas de dispersão do STAT DISK. Associe os diagramas com os seguintes valores de coeficientes de correlação linear: r = 0,857, r = -0,658, r = 0, 012.



Fonte: Triola (1999, p.423).

E, complementando, após identificar a existência de relação entre as variáveis e abordar sobre o coeficiente de correlação, Barbetta (1994), procurara estabelecer os seguintes aspectos: i) existe correlação entre as variáveis; ii) correlação positiva e negativa; iii) correlação forte e fraca (Figura 10).

Figura 10 - Esquema que representa sentido e força da correlação em função do valor de r (coeficiente de correlação).



Fonte: Barbetta (1994, p. 258).

O estudo do comportamento da relação entre variáveis também é realizado por meio da determinação dos coeficientes de correlação que são métodos estatísticos para se medir as relações entre variáveis e o que elas representam. Busca entender como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas. Embora não implique em causalidade, o coeficiente de correlação exprime em números essa relação, ou seja, quantifica a relação entre as variáveis.

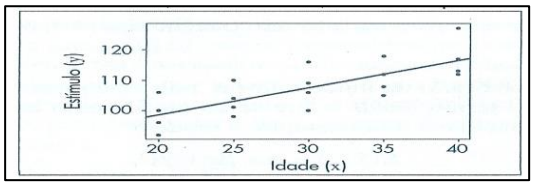
Predizer uma variável em função de outra

Uma vez identificada a existência de uma relação entre as variáveis (X, Y), devemos encontrar uma função que permite usar uma variável para prever outra variável. Este é o problema da regressão, que de acordo com Gea et al. (2013) pode ser decomposto de acordo com os seguintes aspectos: Analisar o ajuste linear entre variáveis e Fazer estimações mediante o ajuste linear entre variáveis.

Analisar o ajuste linear entre variáveis

Nos livros analisados nota-se um padrão na apresentação dos conteúdos da regressão linear, o qual primeiramente trabalha o ajuste da reta por meio do diagrama de dispersão como em Bussab e Morettin (2002), figura 11); segundo apresenta a fórmula e os cálculos para estimação dos parâmetros da reta, por meio do Método dos mínimos quadrados; terceiro é realizado algumas estimações (Interpolação e extrapolação), e por fim, alguns livros aplicam o teste de hipótese para comprovação do modelo.

Figura 11 - Diagrama de dispersão de idade e reação ao estímulo, com reta ajustada apresentado no livro como exemplo de relação entre variáveis.

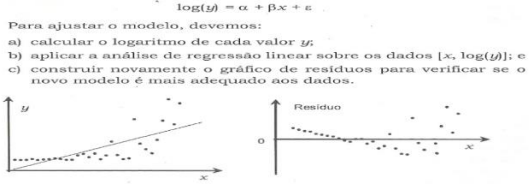


Fonte: Bussab e Morettin (2002, p.437).

Entretanto é necessário discutir se uma reta seria um bom modelo de ajuste dos dados. Mencionamos Bussab e Morettin (2002) e Barbetta (1994) que realizam uma abordagem sobre a transformação logarítmica, seguido de exemplos como em Barbetta (1994) (Figura 12).

Porém, somente Toledo e Ovalle (1994) trabalham com diversos modelos matemáticos não lineares. Dentre eles: Função Potência (Transformação logarítmica), a função exponencial, função hipérbole I e II e o ajuste por meio da parábola. Todos exemplificados em contextos, representação gráfica e tabular.

Figura 12 – Exemplo do livro indicando uma relação exponencial, ou seja, não linear.


$\log(y) = \alpha + \beta x + \varepsilon$

Para ajustar o modelo, devemos:

a) calcular o logaritmo de cada valor y;
b) aplicar a análise de regressão linear sobre os dados $[x, \log(y)]$; e
c) construir novamente o gráfico de resíduos para verificar se o novo modelo é mais adequado aos dados.

Fonte: Barbetta (1994, p.281).

Porém, vale destacar que Toledo e Ovalle (1994) perde a oportunidade de trabalhar a comparação dos modelos de ajuste, bem como questões que solicite ao aluno eleger qual o melhor dentre os modelos ao resolver um problema por meio de um exemplo ou exercício, uma vez que é dado ao aluno o modelo a ser utilizado.

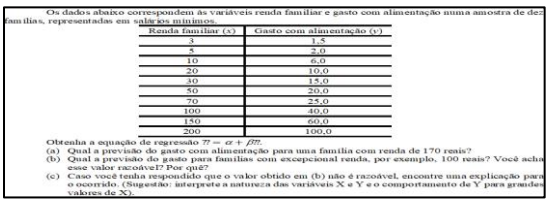
Em Anderson, Sweeney e Willians (2007) é apresentado uma representação gráfica comparativa de modelo linear e não linear. Tal visualização se faz importante para que se perceba o comportamento de tais modelos.

O diagrama de dispersão é importante no processo ensino e aprendizagem de correlação e regressão, pois trata-se de uma representação gráfica que analisa a relação entre duas variáveis quantitativas — uma de causa e uma de efeito. Quando você tem uma hipótese do que causou algo, mas ainda deseja comprová-la por meio de uma análise mais aprofundada. Esse tipo de diagrama torna visível se o que acontece em uma variável causou interferência na outra.

Fazer estimações mediante o ajuste linear entre variáveis

Calculada a reta de regressão, Gea et al. (2013) pontuam a necessidade de apresentar exercícios de estimação. O cálculo da estimação implica numa reflexão em torno do valor esperado e o valor real uma vez que pode haver difenrentes valores observados, como em Bussab e Morettin (2002), figura 13.

Figura 13 - Exercício apresentado no livro para estimar Y (variável independente) tomando valores de X (variável dependente).


Os dados abaixo correspondem às variáveis renda familiar e gasto com alimentação numa amostra de dez famílias, representadas em salários mínimos.

Renda familiar (x)	Gasto com alimentação (y)
3	1,5
5	2,0
10	6,0
20	10,0
30	15,0
50	20,0
70	25,0
100	40,0
150	60,0
200	100,0

Obtenha a equação de regressão $?? = \alpha + \beta??$.
(a) Qual a previsão do gasto com alimentação para uma família com renda de 170 reais?
(b) Qual a previsão do gasto para famílias com excepcional renda, por exemplo, 100 reais? Você acha esse valor razoável? Por quê?
(c) Caso você tenha respondido que o valor obtido em (b) não é razoável, encontre uma explicação para o ocorrido. (Sugestão: interprete a natureza das variáveis X e Y e o comportamento de Y para grandes valores de X).

Fonte: Bussab e Morettin (2002, p.471).

Vale ainda ressaltar que somente Triola (1999) e Larson e Farber (2010) apresentam exercícios/exemplos que sugerem o uso da calculadora. Ambos os livros apresentam exercícios a serem resolvidos com alguma ferramenta tecnológica, porém deixa a critério do aluno a escolha dessa ferramenta.

Destacamos que a análise de regressão estuda a relação entre uma variável chamada variável dependente e outra variável chamada variável independente. A relação entre elas é representada por um modelo matemático, que associa a variável dependente com as variáveis independentes que é realizado para possibilitar a estimação de valores a partir do modelo criado.

Considerações Finais

A principal conclusão é que o estudo da correlação e regressão é complexo devido à diversidade de objetos matemáticos envolvidos, que também são inter-relacionados uns com os outros e com muitos outros objetos estatísticos e matemáticos. Ou seja, a correlação e regressão são baseadas em conhecimento prévio de muitos objetos, tais como variáveis estatísticas, frequências e tipos, medidas de dispersão e posição central e gráficos estatísticos.

Consideramos que na apresentação dos conteúdos em livros didáticos que abordem a correlação e a regressão quando é analisado a possível relação entre variáveis de interesse, onde não deve se basear apenas o seu estudo ajustando os dados a um padrão pré-definido, como um ajuste linear, mas deve-se estudar vários modelos e então os comparar para determinar qual a distribuição que melhor se ajuste àquela distribuição dos dados. Deve-se também estudar a significância estatística dos parâmetros utilizados para descobrir se a relação entre variáveis deve-se, ou não, ao acaso, etc.

Assim, após o estudo sugerimos que "Estatística Básica" de Tolledo e Ovalle (1994) indica os elementos teóricos necessários e mais próximos do estudo. Entretanto é apresentada uma baixa quantidade de exercicios/exemplos, bem como tarefas que sugerem o uso de tecnologias. Nesse sentido acreditamos que "Estatística Aplicada" de Larson e Farber (2010) e Anderson, Sweeney e Willians (2007) podem complementar a escassez de exercicios e o uso da tecnologia observado em Toledo e Ovalle (1994). Portanto, em nosso entendimento os três livros se tornam complementares no que se refere aos aspectos teóricos e pedagógicos do ensino da Correlação e Regressão.

De acordo com Ortiz (1999) os livros matemáticos se diferenciam no tema, na relação entre o autor e os leitores e na formação dos argumentos. Estes aspectos correspondem a três "metafunções" da linguagem: ideacional, interpessoal e textual. A função ideacional se refere ao tipo de objetos que participam numa atividade matemática, isto é, a forma que a linguagem expressa às categorias da própria experiência de mundo. A função interpessoal expressa as relações sociais e pessoal entre o autor e os leitores. A função textual é o que faz a linguagem operacionalmente relevante em seu contexto e diferencia uma mensagem viva de um dicionário.

Com base nas considerações de Ortiz (1999), entendemos que os resultados obtidos nesta pesquisa devam ser interpretados com precaução, uma vez que o impacto do livro depende não só do próprio livro, mas do leitor e do professor.

Portanto, carecemos de estudos no campo da análise do conhecimento estatístico para abrir uma discussão e contribuir para um processo ensino e aprendizagem mais eficaz e que assegure que a distância que separa o conhecimento científico e a sua transmissão deva ser mais didático, não comprometendo a apreensão dos conceitos. Isto levanta a necessidade de incorporar a formação de professores, por exemplo, espaços para a análise de livros didáticos para o ensino dos diversos conteúdos.

Referências

ANDERSON, D. R.; SWEENEY, D.; WILLIAMS, T. A. **Estatística Aplicada à Administração e Economia.** São Paulo: Thomson Learning, 2007.

BARBETTA, P. A. **Estatística aplicada às ciências sociai**s. Florianópolis: Ed. UFSC, 1994.

BATANERO, C., GEA, M. M., LÓPEZ-MARTÍN, M. M., ARTEAGA, P. Análisis de los conceptos asociados a la correlación y regresión en los textos de bachillerato. **Didacticae**, v. 1, p. 60-76, 2016.

BATANERO, C.; DÍAZ, C. **Análisis de datos con Statgraphics**. Granada: Departamento de Didáctica de la Matemática, 2008. Disponível em: <http://www.ugr.es/~batanero/pages/ARTICULOS/anadatos.pdf>. Acesso em: 21 mar. 2017.

BOX, G. E. P., HUNTER, W. G., HUNTER, J. S. **Statistics for experimenters**. New York: Wiley, 1978.

BRAGA, G.; BELVER, J. L. El análisis de libros de texto: una estrategia metodológica en la formación de los profesionales de la educación**. Revista Complutense de Educación**, *v. 27, n.* 1, p. 199-218, 2016.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica.** São Paulo: Saraiva, 2002.

CALLEGARI-JACQUES, S. M. **Bioestatística:** princípios e aplicações. Porto Alegre: ArtMed, 2003.

CORDERO, F., FLORES, R. El uso de las gráficas en el discurso matemático escolar. Un estudio socio epistemológico en el nivel básico a través de los libros de texto. **Revista Latinoamericana de Matemática Educativa,** v. 10, n. 1, p. 7-38, 2007.

CRESPO, A. A. **Estatística** Fácil. São Paulo: Saraiva, 2009.

DÍAZ-LEVICOY, D.; GIACOMONE, B.; LÓPEZ-MARTÍN, M. de M.; PIÑEIRO, J. L. Estudio sobre los gráficos estadísticos en libros de texto digitales de educación primaria española. ***Profesorado - Revista de Currículum y Formación de Profesorado,*** v. 20, n. 1, p. 133-156, enero-abril 2016.

ESTEPA, A. Interpretación de los diagramas de dispersión por estudiantes de Bachillerato. ***Enseñanza de las Ciencias***, Barcelona, v. 26, n. 2, p. 257-270, 2008.

ESTEPA, A.; GEA, M. M.; CAÑADAS, G. R.; CONTRERAS, J. M. Algunas notas históricas sobre la correlación y regresión y su uso en el aula. ***Número****s*, Tenerife, v. 81, p. 5-14, 2012.

GEA, M. M., BATANERO, C., ARTEAGA, P., CAÑADAS, G. R., CONTRERAS, J. M. Análisis del lenguaje sobre la correlación y regresión en libros de texto de bachillerato. ***SUMA,*** v. 76, 37-45, 2014.

GEA, M. M., BATANERO, C., CAÑADAS, G. R., CONTRERAS, J. M. Un estudio empírico de las situaciones-problema de correlación y regresión en libros de texto de bachillerato. In: BERCIANO, A., GUTIÉRREZ, G., ESTEPA, A., CLIMENT, N. (Eds.). **Investigación en Educación Matemática** *XVII*. Bilbao: Sociedad Española de Investigación en Educación Matemática, 2013. p. 293-300.

GEA, M. M., BATANERO, C., ROA, R. El sentido de la correlación y regresión. **Números,** Tenerife, España, v. 87, p. 25-35, 2014.

GEA, M.; BATANERO, C.; CAÑADAS, G.; ARTEAGA, P. La organización de datos bidimensionales en libros de texto de Bachillerato. Memorias das **Jornadas Virtuales de Didáctica de la Estadística, Probabilidad y Combinatoria - SEIEM, Granada, España,** v. 1, p. 1-8, 2013.

HERBEL, B. A. From intended curriculum to written curriculum: Examining the" voice" of a mathematics textbook. **Journal for Research in Mathematics Education**, v. 38, n. 4, p. 344-369, 2007.

JALES, P. C. ***A importância do Ensino de Regressão Linear Simples no Ensino Médio: um estudo com alunos do 3º ano do Ensino Médio – IFMA - Imperatriz.*** 2014. 59 f. Dissertação (Mestrado em Matemática) - Programa de Pós-Graduação em Matemática, da Universidade Federal do Piauí, 2014.

LARSON, R.; FARBER, B. **Estatística Aplicada**. 4. ed. São Paulo: Prentice Hall, 2010.

LAVALLE, A. L., MICHELI, E. B., RUBIO, N. Análisis didáctico de regresión y correlación para la enseñanza media. **Revista Latinoamericana de Matemática Educativa,** v. 9, n. 3, p. 383-406, 2006.

MIRANDA, L. M. **Correlação e Regressão em curso de engenharia**: uma abordagem com foco na leitura e interpretação de dados. 2014. 160 f. Dissertação (Mestrado Profissional em Ensino de Ciências e Matemática) – Mestrado Profissional em Ensino de Ciências e Matemática da Pontifícia Universidade Católica de Minas Gerais, 2008.

NAGHETTINI, M. PINTO, E. J. A. **Hidrologia Estatística**. Serviço Geológico do Brasil. CPRM. Ministério de Minas e Energia. 2007.

OLIVEIRA JÚNIOR. A. P. de. Reflexão sobre as Características Sociodemográficas, Educacionais, do uso de Tecnologias e das Práticas Docentes de Professores de Estatística no Ensino Superior no Brasil. ***Bolema,*** Rio Claro (SP), v. 24, n. 39, p. 387-412, ago. 2011.

OLIVEIRA JÚNIOR. A. P. de.; ALVES, G. C. S. A correlação e a regressão linear em livros didáticos nos cursos de graduação no Brasil. In: Congreso Iberoamericano de Educación Matemática, 8., 2017. ***Anais...*** Madrid, Espanha, 2017.

ORTIZ, J. J. **Significado de los conceptos probabilísticos elementales en los textos de Bachillerato**. Tesis Doctoral. Universidad de Granada, 1999.

PELLICER, A. Calidad de los textos escolares. In: MINEDUC (Ed.). **Primer Seminario Internacional de Textos Escolares.** *SITE 2006*. Santiago: MINEDUC, 2007. p. 279-285.

RAZAK, F. A., BAHARUN, N, DERAMAN, N. A., ISMAIL, R. P. Assessing students' abilities in interpreting the correlation and regression analysis. **Journal of Fundamental and Applied Sciences**, Algeria, v. 9, n. 5S, p. 644-661, 2017.

SÁNCHEZ COBO, F. T. **Significado de la correlación y regresión para los estudiantes universitarios**. Tesis doctoral. Universidad de Granada, 1999.

SÁNCHEZ COBO, F. T., ESTEPA, A., BATANERO, C. Un estudio experimental de la estimación de la correlación a partir de diferentes representaciones. **Enseñanza de las Ciencias,** *v. 18, n.* 2, p. 297-310, 2000.

STEVENSON, W. J. **Estatística aplicada à administração.** São Paulo: Harbra, 1981.

TOLEDO, G. L.; OVALLE, I. I**. Estatística Básica.** São Paulo: Editora Atlas, 1984.

TRIOLA, M. F. **Introdução à Estatística**. Rio de Janeiro: LTC, 1999.

VIEIRA, S. M**.** ***Introdução à Bioestatística.*** 4.ed. São Paulo: Campus. 2008.

ZIEFFLER, A.; GARFIELD, J. Modeling the growth of students' covariational reasoning during an introductory statistics course. **Statistics Education Research Journal**, *v.* 8, n. 1, p. 7-31, 2009.

Ailton Paulo de Oliveira Júnior: Doutor e Pós-Doutor em Educação, Professor Universidade Federal do ABC/UFABC, Santo André, SP, ailton.junior@ufabc.edu.br

Daniel de Freitas Barros Neto: Mestre em Ensino e História das Ciências e da Matemática, Universidade Federal do ABC/UFABC, Santo André, SP, danielfbn@gmail.com

Gisele Cristiane Silva Alves: Graduado em Matemática, Universidade Federal do Triângulo Mineiro/UFTM, Uberaba, MG, giselealialves@yahoo.com.br