APLICAÇÔES DA CORRELAÇÂO E REGRESSÂO LINEAR

Nilo A de S. Sampaio [1]

RESUMO

Este artigo trata dos conceitos que envolvem Correlação e Regressão linear e suas aplicações em diversas áreas do conhecimento humano. Baseia-se em pesquisas bibliográficas e estudos de caso que ajudam a entender como essas ferramentas da estatística ajudam a compreender determinado estudo e na posterior tomada de decisão.

Palavras-chave: Correlação. Regressão. Estatística.

1 Introdução

Em experimentos que procuram determinar a relação existente entre duas variáveis, por exemplo, a dose de uma droga e a reação, concentração e densidade ótica, peso e altura, idade da vaca e a produção de leite, etc., dois tipos de situações podem ocorrer:

(a) uma variável (X) pode ser medida acuradamente e seu valor escolhido pelo experimentador. Por exemplo, a dose de uma droga a ser ministrada no animal. Esta variável é a variável independente. A outra variável (Y), dita variável dependente ou resposta, está sujeita a erro experimental, e seu valor depende do valor escolhido para a variável independente. Assim, a resposta (reação, Y) é uma variável dependente da variável independente dose (X). Este é o caso da Regressão. (b) as duas variáveis quando medidas estão sujeitas a erros experimentais, isto é, erros de natureza aleatória inerentes ao experimento. Por exemplo, produção de leite e produção de gordura medidas em vacas em lactação, peso do pai e peso do filho, comprimento e a largura do crânio de animais, etc. Este tipo de associação entre duas variáveis constitui o problema da Correlação. Atualmente, se dá à técnica de correlação uma importância menor do que a da regressão. Se duas variáveis estão correlacionadas, é muito mais útil estudar as posições de uma ou de ambas por meio de curvas de regressão, as quais permitem, por exemplo, a predição de uma variável em função de outra, do que estudá-las por meio de um simples coeficiente de correlação.

---

[1] Doutor em Engenharia Mecânica pela Unesp-SP. Professor da Associação Educacional Dom Bosco. Professor da UERJ-FAT. Professor de diversos cursos de Pós graduação.

e-mail: nilo.samp@terra.com.br

2 Desenvolvimento

Em estatística ou econometria, regressão linear é um método para se estimar a condicional (valor esperado) de uma variável *y*, dados os valores de algumas outras variáveis *x*. A *regressão*, em geral, trata da questão de se estimar um valor condicional esperado. A regressão linear é chamada "linear" porque se considera que a relação da resposta às variáveis é uma função linear de alguns parâmetros. Os modelos de regressão que não são uma função linear dos parâmetros se chamam modelos de regressão não-linear. Sendo uma das primeiras formas de análise *regressiva* a ser estudada rigorosamente, e usada extensamente em aplicações práticas. Isso acontece porque modelos que dependem de forma linear dos seus parâmetros desconhecidos, são mais fáceis de ajustar que os modelos não-lineares aos seus parâmetros, e porque as propriedades estatísticas dos estimadores resultantes são fáceis de determinar. Abaixo tem-se um gráfico com um modelo de regressão linear onde aparece a reta de regressão. (BUSSAB,2006)

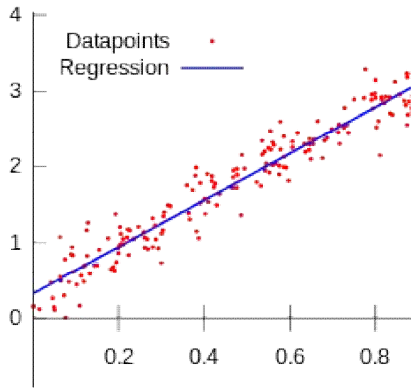

Figura-1: Exemplo de Regressão Linear

Para se estimar o valor esperado, usa-se de uma equação, que determina a relação entre ambas as variáveis.

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

(1)

Em que: Y_i - Variável explicada (dependente); é o valor que se quer atingir;

α - É uma constante, que representa a interceptação da reta com o eixo vertical;

β - É outra constante, que representa o declive(coeficiente angular)da reta;

X_i - Variável explicativa (independente), representa o fator explicativo na equação;

- Variável que inclui todos os fatores residuais mais os possíveis erros de medição. O seu comportamento é aleatório, devido à natureza dos fatores que encerra. Para que essa fórmula possa ser aplicada, os erros devem satisfazer

determinadas hipóteses, que são: serem variáveis normais, com a mesma variância σ^2 (desconhecida), independentes e independentes da variável explicativa X.

Para se calcular os parâmetros α e β usam-se as seguintes expressões:

$$\hat{\alpha} = \frac{\sum X^2 \sum Y - \sum (XY) \sum X}{n \sum X^2 - (\sum X)^2} \quad (2)$$

$$\hat{\beta} = \frac{n \sum (XY) - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad (3)$$

Definindo $\overline{X} = \frac{\sum X}{n}$ e $\overline{Y} = \frac{\sum Y}{n}$, temos que $\hat{\alpha}$ e $\hat{\beta}$ se relacionam por:

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$$

Com a reta de regressão calculada podemos estimar valores futuros de y em função de possíveis valores de x, como por exemplo calcular a produção estimada para o mês número 5 (5º. Mês), etc.

O método utilizado para o cálculo destes parâmetros ($\hat{\alpha}$ e $\hat{\beta}$) chama-se método dos mínimos quadrados e Credita-se Carl Friedrich Gauss como o desenvolvedor das bases fundamentais do mesmo, em 1795, quando Gauss tinha apenas dezoito anos. Entretanto, Adrien-Marie Legendre foi o primeiro a publicar o método em 1805, em seu *Nouvelles méthodes pour la détermination des orbites des comètes.* Gauss publicou suas conclusões apenas em 1809. Este método é o procedimento de estimação dos parâmetros de um modelo de regressão por meio da minimização da soma dos quadrados das diferenças entre os valores observados da variável resposta em uma amostra e seus valores preditos pelo modelo. Possui aplicações em áreas como biologia, engenharia, estatística, física matemática, entre outras, principalmente aquelas que objetivam relacionar uma variável dependente (Y) em função de variáveis explicativas (X1,...,Xk). Em teoria da probabilidade e estatística, correlação, também chamada de coeficiente de correlação, indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. No uso estatístico geral, *correlação* ou co-relação se refere a medida da relação entre duas variáveis, embora correlação não implique causalidade. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados. Vários coeficientes são utilizados para situações diferentes. O mais conhecido é o coeficiente de correlação de Pearson, o qual é obtido dividindo a covariância de duas variáveis pelo produto de seus desvios padrão. Apesar do nome, ela foi apresentada inicialmente por Francis Galton. A correlação falha em capturar dependência em algumas instancias. Em geral é possível mostrar que há pares de variáveis aleatórias com forte dependência estatística e que no entanto apresentam correlação nula. Para esse caso devem-se usar outras medidas de dependência (CRESPO, 1999).

Em estatística descritiva, o coeficiente de correlação é a medida do grau e da direção de uma relação linear entre duas variáveis. O símbolo r representa o coeficiente de correlação amostral. O coeficiente de correlação populacional é representado por ρ .(LARSON, R, FARBER, B, 2004).

Este coeficiente, normalmente representado por r ou ρ assume apenas valores entre -1 e 1.

- $\rho = 1$ Significa uma correlação perfeita positiva entre as duas variáveis.
- $\rho = -1$ Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.
- $\rho = 0$ Significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado $\rho = 0$ deve ser investigado por outros meios.

  calcula-se o coeficiente de correlação de Pearson ou coeficiente de regressão populacional segundo a seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}} \quad (4)$$

  onde $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ são os valores medidos de ambas as variáveis. Para além disso

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad (5)$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \quad (6)$$

  são as médias aritméticas de ambas as variáveis. Uma interpretação muito utilizada para este coeficiente é a seguinte:

- 0.70 para mais ou para menos indica uma forte correlação.
- 0.30 a 0.7 positivo ou negativo indica correlação moderada.
- 0 a 0.30 Fraca correlação.

Para o cálculo do r (coeficiente de regressão amostral) uma das muitas fórmulas usadas é a seguinte:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] \times \left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]}} \quad (7)$$

Existem inúmeras aplicações da regressão e correlação linear nas ciências sociais, na engenharia, na área biomédica etc, aqui seram mostradas algumas delas na forma de estudo de caso para a melhor compreensão por parte de leitores menos habituados com estudos estatísticos.

Em um primeiro estudo de caso será analisada uma situação hipotética de uma companhia que tem gastos com propaganda e esta situação encontra-se cadastrada na Tabela 1 abaixo.

Tabela-1: Vendas da Companhia x Gasto com Propaganda

Gastos com Propaganda em milhares de dólares,x	Vendas da companhia em milhares de dólares,y	x.y	x^2	y^2
2,4	225	540	5,76	50.625
1,6	184	294,4	2,56	33.856
2,0	220	440	4	48.400
2,6	240	624	6,76	57.600
1,4	180	252	1,96	32.400
1,6	184	294,4	2,56	33.856
2,0	186	372	4	34.596
2,2	215	473	4,84	46.225
$\sum$ x= 15,8	$\sum$ y = 1634	$\sum$ x.y=3289,8	$\sum x^2$=32,44	$\sum y^2$=337.558

Fonte: LARSON, R, FARBER, B. Estatística Aplicada ,pág 338.

Substituindo os dados da Tabela-1 na expressão (7) obtemos um r = 0,913, ou seja como o valor de r está muito próximo de 1, há uma forte correlação linear positiva. À medida que aumenta a quantia gasta em propaganda, crescem também as vendas da companhia. Usando o mesmo exemplo da Tabela-1 pode-se criar um segundo estudo de caso que envolva o conceito de regressão linear. Para tanto a idéia seria estimar as vendas da companhia para um gasto com propaganda na ordem de 4,0 milhões de dólares. Esse estudo passa inicialmente pelo cálculo da inclinação, ou coeficiente angular, aqui representado por: β e do intercepto ou coeficiente linear, aqui representado por: α.

Substituindo os dados nas (2) e (3) dadas por:

$$\hat{\beta} = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

e

$$\hat{\alpha} = \frac{\sum X^2 \sum Y - \sum(XY) \sum X}{n\sum X^2 - (\sum X)^2}$$

Obtemos:

= 50,7287 e α = 104,0608, obtendo assim a reta de regressão dada por:

Y = 50,7287 .X + 104,0608 e com essa situação podemos calcular o valor das vendas da companhia para um gasto com propagandas na ordem de 4 milhões, ou seja: X=4, que será de Y = 50,7287 .(4) + 104,0608 = 306,9756 milhões de dólares.

Uma outra aplicação da Regressão Linear é na estimação de uma demanda futura de um determinado produto ou serviço. Conforme Ritzman e Krajewski (2004), uma previsão é uma avaliação de eventos futuros, usada para fins de planejamento, sendo que as mudanças nos negócios resultantes da concorrência global, ou alguma mudança tecnológica acelerada e as preocupações ambientais crescentes exercem forte pressão sobre a capacidade de uma organização gerar previsões precisas. Mas apesar disso, as previsões são necessárias para auxiliar na determinação de que recursos serão necessários, e para a programação dos recursos existentes ou ainda da aquisição de recursos adicionais se necessário. A previsão de demanda integrada ao planejamento da produção tem como propósito fornecer as informações sobre a demanda futura dos produtos, para que a produção dessa forma possa ser estimada com antecedência, permitindo que os recursos produtivos estejam disponíveis no momento exato, na quantidade e principalmente na qualidade adequada. A previsão de demanda envolve um estudo acerca de informações sobre a demanda futura para um determinado mercado, por meio desta é possível as organizações ajustar o seu planejamento de recursos para atender aos clientes e reduzir custos (matéria-prima, mão de obra, insumos etc.) em suas tarefas operacionais. Este tema envolve um processo que busca estimar a necessidade ou comportamento do mercado para que seja ajustado o planejamento de produção da empresa. Russomano (2000) define previsão de demanda como um processo sistemático e racional na busca de informações acerca das possíveis vendas futuras dos produtos ou serviços de uma organização.Dessa forma, Tubino (2007) salienta que a previsão de demanda é a base para o planejamento estratégico tanto da produção, vendas ou finanças de qualquer empresa, isso porque de algum modo as atividades dessa empresa são direcionadas conforme o rumo em que elas acreditam que o seu negócio andará, sendo que esse rumo é geralmente feito com base em previsões, sendo a previsão de demanda a principal delas. Para se obter uma previsão, existe uma série de métodos disponíveis, mas pode-se subdividi-los em dois grandes grupos: os métodos qualitativos e os quantitativos. Os métodos qualitativos, também chamados de métodos de julgamento, conforme Reid e Sanders (2005), são aqueles em que a previsão é feita de maneira subjetiva pelo responsável, nos quais as ocorrências levantadas pelos especialistas são baseadas na intuição, no conhecimento e até mesmo na experiência dessa pessoa na área. E ainda destacam que, como esse tipo de método é realizado com base no critério e opinião humana, essas previsões podem ser tendenciosas, isso porque elas podem estar relacionadas com uma motivação pessoal, disposição ou convicção de alguma coisa. No que se referem aos métodos quantitativos, na concepção de Moreira (2008), estes compreendem como aqueles que fazem uso de modelos matemáticos para se atingir os valores previstos, os quais permitem um controle do erro, mas exigem informações quantitativas preliminares. Esse método de previsão pode ser dividido em duas categorias: modelos de séries temporais e modelos de séries causais. A análise de modelos temporais exige basicamente o conhecimento de valores passados da demanda, ou de maneira geral da variável que se pretende prever, enquanto que nos modelos causais a demanda de um produto ou conjunto de produtos a uma ou mais variáveis internas ou externas à organização. Alguns dos modelos quantitativos de previsão como

o caso do modelo da regressão linear, o método da média móvel simples, bem como a média móvel ponderada, o método exponencial móvel e ainda a exponencial móvel com tendência (MARTINS; LAUGENI, 2005).

Diante de tais considerações este estudo apresenta uma metodologia quantitativa usando um método causal ou explicativo que é a regressão linear simples. Em um terceiro estudo de caso será analisado uma situação também hipotética do número de viagens em função do valor da carga transportada que aparece na Tabela 2 abaixo.

Tabela-2: Número de viagens por ano x Valor da carga transportada (x R$ 10^3)

Ano	Valor da carga transportada (x R$ 10^3),x	Número de viagens por ano ,y	x.y	x^2	y^2
1	2,5	264	660,0	6,25	69,696
2	1,3	116	150,8	1,69	13,456
3	1,4	165	231,0	1,96	27,225
4	1,0	101	101,0	1,00	10,201
5	2,0	209	418,0	4,00	43,681
TOTAIS	$\sum$ x= 8,2	$\sum$ y = 855	$\sum$ x.y=1560,8	$\sum x^2$=14,9	$\sum y^2$=164,259

De posse da tabela acima e usando-se as fórmulas de regressão linear tem-se:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$$ = 171 e $$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$ = 1,64

Substituindo os dados nas (2) e (3) dadas por:

$$\hat{\beta} = \frac{n \sum (XY) - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

e

$$\hat{\alpha} = \frac{\sum X^2 \sum Y - \sum (XY) \sum X}{n \sum X^2 - (\sum X)^2}$$

Obtemos:

= 109,230 e α = - 8,137, obtendo assim a reta de regressão dada por: Y = 109,230 .X - 8,137, e com essa situação podemos estimar o valor o número de viagens por exemplo para uma carga que custe R$ 4.000,00 que seria

4.10^3,00 e portanto ficaria: Y = 109,230 .(4) - 8,137 = 428,783 que seria aproximadamente 429 viagens neste ano.

De uma forma geral a reta de regressão funciona como uma ferramenta para estimar com uma precisão muito boa uma relação que existe, e pode ser observada graficamente e assim proceder através de uma função que é linear uma estimação para a variável dependente (Y) em função da variável independente (X). Ainda de acordo com um quarto estudo de caso Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um aditivo (X). Para isto, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados seguem na tabela 3 abaixo:

Tabela-3: octanagem da gasolina x adição de um aditivo

X(Quantidade % de aditivo)	Y(Octanagem da gasolina)
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85

Plotando-se os dados em um gráfico pode-se perceber que a relação pode ser aproximada por uma reta de regressão e assim fica fácil perceber que pode-se utilizar o método dos mínimos quadrados para calcular a reta de regressão e assim proceder futuras estimações.

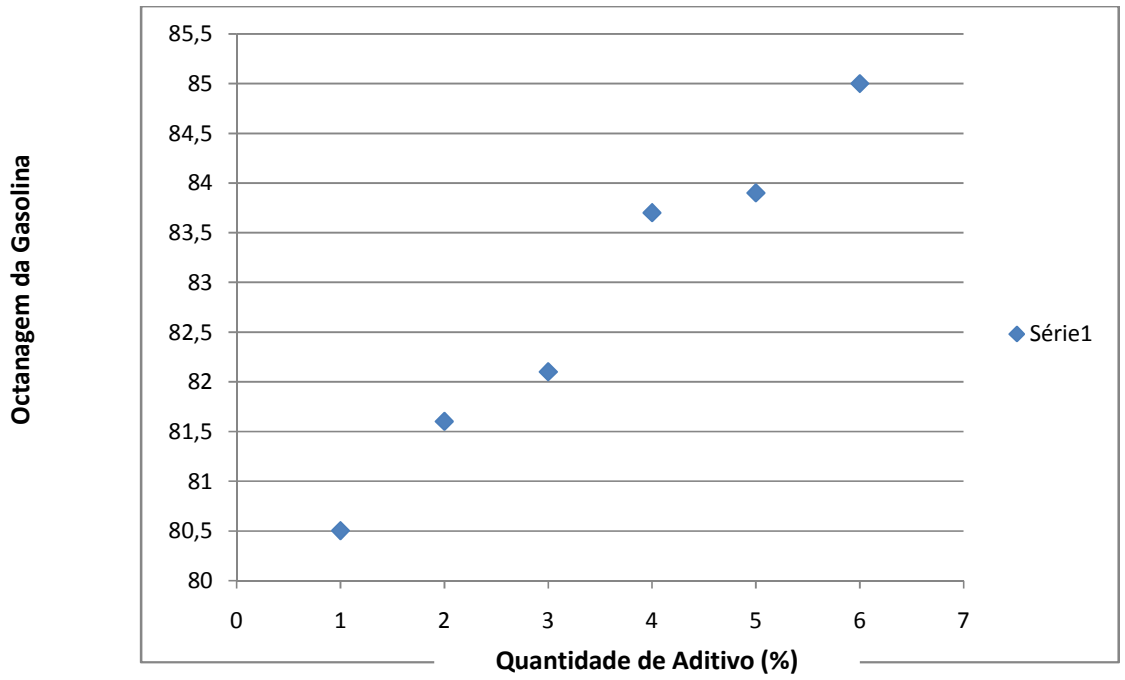

Figura-2: octanagem da gasolina x adição de um aditivo

De posse dos dados da tabela 3 e novamente utilizando-se as equações (2) e (3) obtém-se: β = 0,886 e α = 79,7, obtendo assim a reta de regressão dada por: Y = 0,886 .X + 79,7 que é a reta de regressão e permite estimar uma determinada octanagem da gasolina em função da quantidade de aditivo.

Uma observação importante é que o coeficiente de correlação pode estudar e estabelecer relações entre variáveis, predizer o comportamento de uma variável em função do comportamento de outra variável (predição), fornece informação sobre o grau de relação entre variáveis, porém ele não permite estabelecer uma relação de causa e efeito. Para exemplificar este óbice existe uma estudo de caso clássico que é o quinto estudo de caso: O matemático Arthur Engel recolheu, na cidade alemã de. Oldenburg no período de 1930 a 1936, os dois conjuntos de dados que se apresentam na tabela 4 abaixo:

Tabela-4: habitantes x casais de cegonhas

ANO	X(casais de cegonhas)	Y(habitantes)
1930	132	55400
1931	142	58400
1932	166	65000
1933	188	67700
1934	240	69800
1935	250	72300
1936	252	76000

A hipótese seria que o aumento do número de casais de cegonhas é a causa do número de habitantes, o cálculo do coeficiente de regressão produz r = 0,95, ou seja um valor muito alto que indica que existe uma forte correlação positiva porém é sabido que isso não indica causa.

3 Conclusões

Dos resultados dos estudos de caso e da revisão de bibliografia é possível observar que a análise de regressão e correlação linear é uma ferramenta estatística poderosa e capaz de estimar parâmetros desconhecidos e ainda prever a correlação entre variáveis dependentes e uma ou várias variáveis independentes, para a posterior tomada de decisão.

REFERÊNCIAS

- BUSSAB, Wilton. *Estatística Básica*. Saraiva. 5a edição 2006. 540p. ISBN 85-02-03497-9.
- CRESPO, Antônio Arnot. Estatística Fácil.17ed. . São Paulo: Saraiva,1999.
- LARSON, R, FARBER, B. Estatística Aplicada. 2ª edição. São Paulo: Pearson - Prentice Hall, 2004.
- MARTINS, P. G.; LAUGENI, F. P. Administração da produção. 2. ed. São Paulo: Saraiva, 2005.
- MOREIRA, D. A. Administração da produção e operações. 2. ed. São Paulo: Cengage Learning, 2008.
- REID, R. D.; SANDERS, N. R. Gestão de operações. Rio de Janeiro: LTC, 2005.
- RITZMAN, L. P.; KRAJEWSKI, L. J. Administração da produção e operações. São Paulo: Prentice Hall, 2004.
- RUSSOMANO, V. H. PCP: planejamento e controle da produção. 6. ed. São Paulo: Pioneira, 2000.
- Stigler, S. M..*The History of Statistics: The Measurement of Uncertainty before 1900*. [S.l.]: Harvard University Press, 1986. 410 p.
- TUBINO, D. F. Planejamento e controle da produção: teoria e prática. São Paulo: Atlas, 2007.