Leviathan | Cadernos de Pesquisa Política N. 8, pp.66-95, 2014

Desvendando os Mistérios do Coeficiente de Correlação de Pearson: O retorno*

Dalson Britto Figueiredo Filho
Enivaldo Carvalho da Rocha
José Alexandre da Silva Júnior
Ranulfo Paranhos
Jorge Alexandre Barbosa Neves
Mariana Batista da Silva**

Resumo

O principal objetivo desse trabalho é aprofundar a discussão a respeito do conceito de correlação de Pearson. Para tanto, apresentamos as suas propriedades, aplicações e limites. Metodologicamente, utilizamos simulação básica para demonstrar como a presença de *outliers* pode tanto ocultar correlações reais quanto forjar associações inexistentes. Além disso, discutimos os perigos analíticos das correlações espúrias. Com esse artigo esperamos facilitar não só a interpretação, mas também a utilização da técnica de correlação na Ciência Política.

Palavras-chave: Correlação de Pearson; métodos quantitativos; Ciência Política.

Abstract

Unraveling the Mysteries of the Pearson Correlation Coefficient: The Return

The principal aim of this paper is to develop the discussion regarding Pearson correlation coefficient. To do so, we present its main characteristics, applications and limits. Methodologically, we employ basic simulation to demonstrate how outliers can both hide real correlation and make up false associations. In addition, we review the analytical shortcomings of spurious correlation. With this paper we hope to facilitate not only the interpretation, but also the application of correlation technique in Political Science.

Keywords: Pearson correlation; quantitative methods; Political Science.

* Esse trabalho contou com aporte financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Agradecemos também ao *Berkeley Initiative for Transparency in the Social Sciences* pelo treinamento recebido. Ver http://bitss.org/

** Dalson B. F. Filho é doutor em Ciência Política e professor do DCP-UFPE, email: dalsonbritto@yahoo.com.br. Enivaldo C. da Rocha é mestre em Estatística, doutor em Engª de Produção e professor do DCP-UFPE, email: enivaldocrocha@gmail.com. José Alexandre da Silva Júnior é doutor em Ciência Política e professor do ICS-UFAL, email: jasjunior2007@yahoo.com.br. Ranulfo Paranhos é doutor em Ciência Política e professor do ICS-UFAL, email: ranulfoparanhos@me.com. Jorge A. B. Neves é professor da UFMG e doutor em Sociologia, email: jorgeaneves@gmail.com. Mariana Batista da Silva é Cientista Política e pós-doutoranda DCP-UFPE, email: mariana.bsilva@gmail.com.

Introdução

Esse trabalho aprofunda a discussão intuitiva a respeito do coeficiente de Correlação de Pearson. Revisitamos o tema com o objetivo de apresentar uma discussão mais detalhada sobre dois pontos, apenas superficialmente abordados por Figueiredo Filho e Silva Junior (2009). São eles: (1) o papel dos *outliers* sobre a consistência do coeficiente de correlação de Pearson e (2) o problema das correlações espúrias. Em particular, a literatura sobre identificação de *outliers* é amplamente desenvolvida em áreas como Engenharia (Ben-Gal, 2005), Ciência da Computação (Hodge e Austin, 2004), Sistema de Informação (Chawla e Pei Sun, 2006), Estatística (Jensen e Ramirez, 1998), Saúde Pública (Seo, 2002), Mineração de Dados (*data mining*) (Kriegel, Kroger e Zimek, 2010), etc. É extremamente importante oferecer um guia básico sobre o papel dos *outliers* na pesquisa empírica dada a centralidade dos casos desviantes e a ausência de uma literatura mais intuitiva sobre o assunto.

Metodologicamente, revisamos a literatura especializada sobre o tema e apresentamos as propriedades, aplicações e limites do coeficiente de correlação de Pearson. Além disso, utilizamos simulação básica para demonstrar como a presença de *outliers* pode tanto ocultar correlações reais quanto forjar associações inexistentes. Discutimos também os perigos analíticos das correlações espúrias. Com esse artigo esperamos facilitar a interpretação e a utilização da correlação de Pearson.

O artigo está dividido da seguinte forma: a próxima seção apresenta as principais propriedades, aplicações e limites do coeficiente de correlação de Pearson. Depois disso, utilizamos simulação básica para ilustrar o efeito de observações desviantes sobre a consistência das estimativas. A meta é fornecer um guia prático e explicar como o pesquisador deve lidar com casos atípicos. A terceira seção apresenta exemplos de correlações espúrias. A última parte sumariza nossas conclusões.

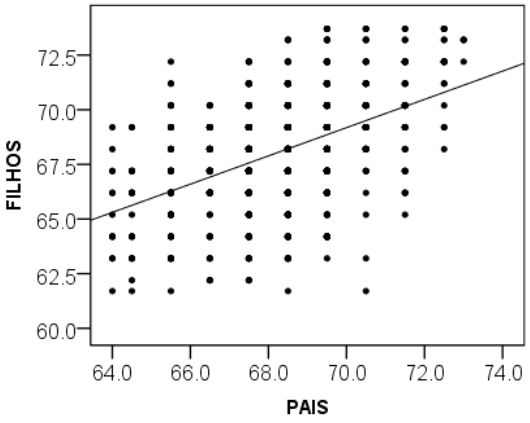
1. O Que é o Coeficiente de Correlação de Pearson?

Para Moore e McCabe (2003), "*correlation was introduced by the English gentleman-scientist Francis Galton (1822-1911) in 1888. Galton called r the index of co-relation and applied it to measurements such as the forearm lengths and the heights of a group of people*" (Moore e McCabe, 2003, p. 127). Stanton (2001) relata que

> the complete name of the correlation coefficient deceives many students into a belief that Karl Pearson developed this statistical measure himself. Although Pearson did develop a rigorous treatment of the mathematics of the Pearson Product Moment Correlation (PPMC), it was the imagination of Sir Francis Galton that originally conceived modern notions of correlation and regression (Stanton, 2001, p.1).

Dessa forma, apesar de levar o nome apenas do Karl Pearson, a origem desse coeficiente deve ser atribuída ao trabalho conjunto Pearson e Francis Galton (Stigler, 1989). Em seu trabalho, Galton coletou dados sobre a altura dos filhos e a altura média dos pais, definida como a média ponderada entre a média de altura do pai e 1,08 a altura da mãe[1]. A figura abaixo ilustra a correlação entre a altura média dos pais e dos filhos.

Figura 1 – Dispersão da altura média dos pais e dos filhos (Galton, 1883)



Fonte: Galton (1883).

Atualmente, o coeficiente de correlação de Pearson é a medida de associação mais utilizada em diferentes áreas da pesquisa científica (Chen e Popovich, 2002). Mas

[1] Informações disponíveis em: http://www.math.uah.edu/stat/data/Galton.html.

o que é correlação afinal? Para Jupp (2006), "*correlation refers to the linear relationship between variables. The correlation coefficient is a measure of the association between two numerical variables, usually denoted as x and y*" (Jupp, 2006, p.43). Moore e McCabe (2003) afirmam que "*a correlation measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r*" (Moore e Mccabe, 2003, p.127). Definimos correlação de Pearson como uma medida de associação linear entre variáveis quantitativas[2]. Depois de definir o conceito, o próximo passo é apresentar as suas principais características.

1.1 Principais propriedades

O coeficiente de correlação de Pearson varia entre -1 e 1. O sinal indica a direção da correlação (negativa ou positiva) enquanto que o valor indica a magnitude. Quanto mais perto de 1 mais forte é o nível de associação linear entre as variáveis[3]. Quanto mais perto de zero, menor é o nível de associação. Em particular, uma correlação de valor zero significa que as variáveis são ortogonais entre si (ausência de correlação). Uma correlação positiva indica que quando *x* aumenta, *y* também aumenta, ou seja, valores altos de *x* estão associados a valores altos de *y*. Por exemplo, peso e altura estão positivamente correlacionadas. Pessoas com altura acima da média, tendem a ter peso também acima da média. Uma correlação negativa indica que quando *x* aumenta, *y* diminui, ou seja, valores altos de *x* então associados a valores baixos de *y*[4]. A tabela abaixo apresenta a correlação entre dez variáveis simuladas.

[2] As variáveis podem estar associadas a partir de diversas formas funcionais (quadrática, cúbica, exponencial, etc.). Para Moore e McCabe (2003) associações lineares são "*particularly important because a straight line is a simple pattern that is quite common*" (Moore e Mccabe, 2003, p.126).

[3] Cohen (1998) apresenta a seguinte classificação no que diz respeito à magnitude do coeficiente: 0,10<r<0,29 = pequeno; 0,30<r<0,49 = médio e r>0,50 = grande. Para Dancey e Reidy (2006) valores até 0,30 devem ser considerados fracos, entre 0,40 e 0,60 moderados e acima de 0,70 fortes.

[4] Algebricamente, o coeficiente de correlação de Pearson é calculado a partir da seguinte fórmula:

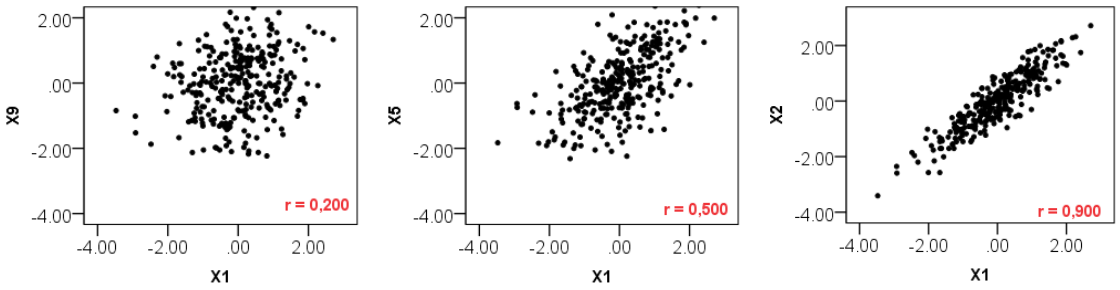
Tabela 1 – Variáveis simuladas

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,00	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10
p-valor	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,084)
	Número de observações = 300									

Fonte: elaboração própria

Simulamos dez variáveis com um nível de associação linear decrescente. Todas as variáveis são normais, com média zero e desvio padrão igual a um. O grau de associação entre X_1 e X_2 é de 0,900 com p-valor abaixo de 1%. Já a correlação entre X_1 e X_{10} é de 0,100 com p-valor significativo a 10%[5]. Analiticamente, gráficos de dispersão são especialmente úteis para visualizar correlações bivariadas. Para Jupp (2006), "*in addition to calculating the correlation coefficient, it is also advisable to view a scatter diagram to gain a better appreciation of any relationship, or its absence*" (Jupp, 2006, p.44). A figura abaixo ilustra correlações com diferentes níveis de associação linear.

Figura 2 – Correlações simuladas (n=300)



Fonte: elaboração própria.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Tanto a média quanto a variância desempenham um papel central na estimação desse coeficiente. A literatura sempre enfatiza que essa medida deve ser utilizada para medir o grau de associação linear entre variáveis quantitativas discretas e/ou contínuas. E que para variáveis ordinais o pesquisador deve utilizar a correlação de Spearman. No entanto, geralmente tanto a correlação de Pearson quanto a de Spearman tendem a apresentar escores bastante similares.

[5] Em geral, os artigos científicos adotam três diferentes padrões de significância estatística: 1%, 5% e 10%. É comum em Ciências Humanas trabalhar com até 10%, enquanto em Ciências Naturais o limite de 5% é mais usual.

Para inspecionar um gráfico de dispersão deve-se projetar uma linha reta passando pela origem dos dois eixos (*X* e *Y*) que represente a tendência de associação linear entre as variáveis. Quanto mais próximo de uma linha reta, mais forte é a linearidade da relação. Quanto mais contíguos forem os pontos em relação à reta, maior é a magnitude da correlação entre as variáveis. Por exemplo, a correlação entre X_1 e X_2 (r = 0,900) é claramente melhor representada por uma linha reta com pontos contíguos do que a associação entre X_1 e X_5 (r = 0,500). Ao se considerar a relação entre X_1 e X_9, fica difícil observar sequer uma tendência já que a distribuição dos dados mais parece uma nuvem aleatória de pontos em um espaço bidimensional.

É importante ainda definir as principais propriedades do coeficiente de correlação de Pearson.

1) Diferente da regressão, o coeficiente de correlação não diferencia entre variáveis independentes (*x*) e dependentes (*y*). Ou seja, a correlação entre *x* e *y* é a mesma entre *y* e *x*. Lembrando o mantra: correlação é diferente de causalidade. Toda causalidade pressupõe correlação, mas nem toda correlação é sinônimo de causalidade.

2) Para estimar a correlação de Pearson, todas as variáveis devem ser quantitativas discretas e/ou contínuas. Isso porque o coeficiente é estimado a partir da média e da variância das variáveis. Como não é possível calcular média e variância para variáveis qualitativas, o coeficiente de correlação de Pearson não deve ser utilizado para analisar o padrão de associação entre variáveis ordinais e/ou nominais[6].

3) Por ser uma medida padronizada (z), o coeficiente de correlação de Pearson não se altera ao se modificar a unidade de medida das variáveis. Ele é adimensional, ou seja, desprovido de unidade física. O tempo pode ser calculado em segundos, minutos, dias, etc. O peso pode ser calculado em gramas, quilos, toneladas. Não importa. O coeficiente continua o mesmo.

4) É muito comum, principalmente em congressos, ouvir pesquisadores relatando o valor do coeficiente em termos percentuais. Isso é errado. Uma correlação de 0,5 não significa 50% de associação. Uma correlação de 0,8 não representa o dobro de uma correlação de 0,4. Para Chen e Popovic (2002), "*Pearson's r does to refer to a proportion, nor does it represent the proportionate strength of a relationship"* (Chen e Popovic, 2002, p.12).

5) A correlação de Pearson apenas detecta relações lineares. Para Moore e McCabe (2003), "*correlation does not describe curved relationships between variables, no matter how strong they are*" (Moore e McCabe, 2003, p.128). Diante de uma relação não linear, esse coeficiente não será capaz de descrever adequadamente o padrão de associação entre as variáveis.

[6] Para estimar a correlação entre variáveis ordinais e nominais deve-se optar pelo coeficiente de correlação de Sperman ou Kendall's tau-b. Ao se deparar com uma variável quantitativa e uma variável categórica dicotômica é possível utilizar a *Point-Bisserial Correlation* (Chen e Popovic, 2002).

6) O coeficiente de correlação sempre varia entre -1 e 1. Quanto mais próximo de zero, menor é o nível de associação linear entre as variáveis. A magnitude da correlação aumenta na medida em que os valores se aproximam de um, independente do sinal. Valores muito próximos a um sugerem que a dispersão se aproxima de uma linha reta.

7) Deve-se evitar a utilização desse coeficiente para estimar a relação entre variáveis em amostras pequenas já que "*the sampling distribution of Pearson's r does not approximate normality for small samples. When n increases, distributions will aproximate normal more slowly when ρ≠0 then when ρ=0*" (Chen e Popovic, 2002, p.15)[7].

8) Deve-se evitar também a utilização da correlação sempre que alguns de seus pressupostos forem violados. Se, mesmo assim, o pesquisador optar pela sua utilização, ele deve reportar os procedimentos utilizados para corrigir os problemas e/ou justificar como a violação dos pressupostos afeta a credibilidade dos resultados.

9) Tanto a direção quanto a magnitude do coeficiente são afetadas pela presença de observações destoantes. *Outliers* podem gerar correlações espúrias e/ou esconder associações reais.

10) Deve-se evitar tirar conclusões substantivas a partir da análise de correlações bivariadas. Alguns pesquisadores acabam inferindo, precipitadamente, relações de causalidade a partir do exame de correlações entre duas variáveis. A realidade social é complexa e, dificilmente, é possível explicar o funcionamento de algum fenômeno político a partir de uma única variável.

1.2 Aplicações

Nas Ciências Sociais brasileira, o coeficiente de Pearson (r) é timidamente utilizado. Na verdade, esse é o padrão para qualquer outra técnica quantitativa (Soares, 2005). Para os propósitos desse artigo, pesquisamos quatro importantes revistas (Rev. Bras. De Ciências Sociais, Revista Dados, Rev. de Sociologia e Política e Rev. Opinião Pública), perfazendo um total de 823 artigos no período entre 2000 e 2009. A tabela abaixo sumariza essas informações.

Tabela 2 - Períodicos amostrados

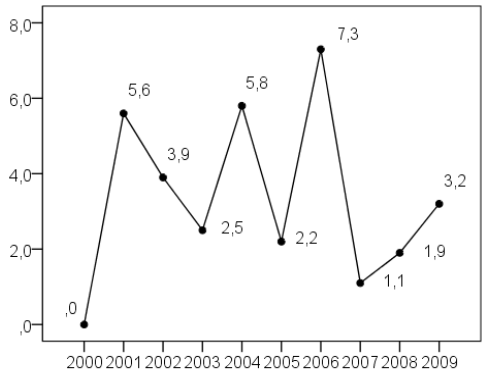
Revistas	***Números***	***Artigos***
Rev. Bras. De Ciências Sociais	29	262
Rev. Dados	38	229
Rev. de Sociologia e Política	21	202
Rev. Opinião Pública	19	130
Total	*107*	*823*

Fonte: elaboração própria

[7] Sugerimos três principais alternativas: (1) tentar coletar mais observações, ou seja, aumentar o tamanho da amostra; (2) utilizar técnicas não paramétricas de estimação e (3) na ausência de normalidade, empregar aproximações usualmente utilizadas por modelos lineares generalizados (Cordeiro, 1986; Gill, 2001).

A figura abaixo ilustra a utilização de correlação de modo geral e Pearson (r), em particular, durante o período analisado.

Figura 3 – Utilização de Correlações e do Coeficiente de Pearson (r) (%)



Fonte: elaboração própria.

Quanto ao uso do coeficiente de Pearson, a maior performance foi registrada em 2006 (7,3%). Em alguns anos o percentual não atingiu 2% do que foi publicado (2000, 2007 e 2008). Portanto, o que é mais marcante é a escassa utilização dessa estatística. Por ser básica, esses dados corroboram os achados de Werneck Vianna *et al* (1989), Valle e Silva (1999) e Santos e Coutinho (2000): a utilização de técnicas quantitativas de pesquisa é limitada nas Ciências Sociais no Brasil.

É importante apresentar alguns dados sobre como a técnica está sendo reportada. O foco repousa sobre quatro elementos básicos: (1) número de casos; (2) magnitude do coeficiente; (3) parâmetros de avaliação do coeficiente e (4) significância estatística. Sem exceção, todos os manuais consultados apontam esses critérios como chaves para compreender o teste (Dancey e Reidy, 2005; Pollock, 2006; Pallant, 2007; Ho, 2009). A tabela abaixo sumariza essas informações.

Tabela 3 - Apresentação de Elementos do Pearson (r)[8] (n = 61)

	Não	
Elementos	**n**	**%**
Número de Casos	35	57,4
Magnitude do Coeficiente	5	8,2
Parâmetro de Avaliação	42	68,9
Significância Estatística	29	47,5

A omissão de elementos básicos é preocupante. A maior ausência é de parâmetros de avaliação da magnitude do coeficiente reportado (68,9%). Em alguns, os autores se contentam em ler apenas o sentido da correlação, em outros classificam os coeficiente com adjetivos imprecisos como "impressionante", "inesperada" e "extraordinária". Da forma como foi reportado é impossível avaliar a força da correlação observada.

Outro ponto grave é a ausência de significância estatística, já que em 47,5% dos casos ela não é reportada. Essa informação é fundamental para avaliar o potencial de generalização dos resultados. Vale dizer, na maioria dos casos de omissão desse elemento não existe qualquer advertência quanto à incapacidade de generalização da análise. Essa falha é ainda mais grave quando não se reporta o número de casos, a segunda omissão mais comum (57,4%). Por fim, em 8,2% dos casos, o coeficiente sequer é reportado, não sendo citado todos os outros elementos, o que coloca em xeque a própria realização do teste. Afinal, de que vale dizer que duas variáveis estão correlacionadas se não se apresenta a direção e a magnitude da relação, bem como o número de casos e a significância estatística? Na nossa opinião: nada.

2. *Outliers:* o que são, como identificá-los e como lidar com eles?

Apresentamos aqui apenas uma discussão básica sobre os diferentes métodos de identificação e manejo de *outliers*. Os leitores interessados em abordagens mais avançadas devem seguir as referências bibliográficas. Em particular, Chandola,

[8] O exame da presença dos elementos considerou o parágrafo, gráfico e tabelas que antecede ou segue a citação da correlação de Pearson (r) encontrada.

Banerjee e Kumar (2007) e Hodge e Austin (2004) apresentam um *survey* dessa literatura. Berton e Liang (2011) discutem a detecção de *outliers* em redes complexas. Ben-gal (2005) apresenta uma introdução à lógica de detecção de observações atípicas. Seo (2002) compara diferentes metodologias de detecção de *outliers* em dados univariados. Comparativamente, Barnett e Lewis (1994) apresentam uma das mais completas abordagens sobre o tema.

2.1 O que são?

Para Moore e McCabe (2003), "*an outlier is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals*" (Moore e Mccabe, 2003, p.162). Para Rodrigues e Paulo (2007), "as observações atípicas (ou *outliers*) são observações com uma combinação única de características identificáveis, sendo notavelmente diferentes das outras observações (parecem ser inconsistentes com o restante da amostra)" (Rodrigues e Paulo, 2007, p.27). Hawkins (1980) define *outlier* "*as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*" (Hawkins, 1980, p.15). O quadro abaixo sumariza diferentes definições.

Quadro 1 – Definições de *outliers*

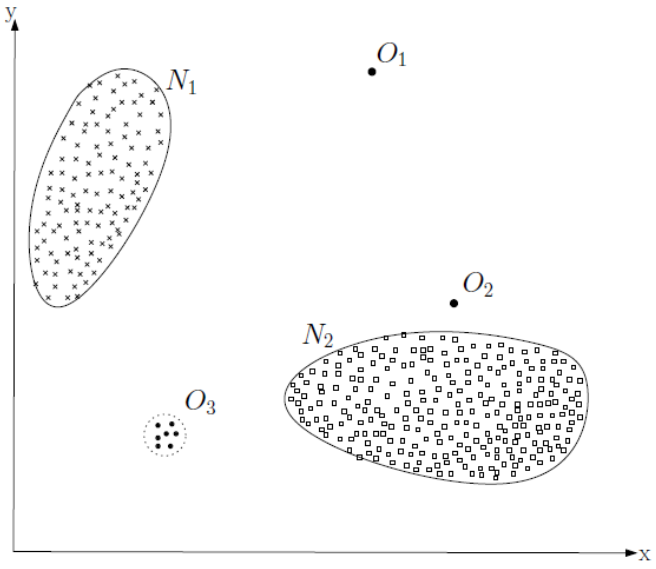
AUTOR (ANO)	***DEFINIÇÃO***
Grubbs (1969)	an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which occurs
Hawkins (1980)	an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism
Johnson (1992)	An observation in a data set which appears to be inconsistent with the remainder of that set of data
Mendenhall *et al* (1993)	Observations whose values lies very far from the middle of the distribution in either direction

Barnett e Lewis (1994)[9]	Indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
Pyle (1999)	An outlier is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable

Fonte: Elaboração dos autores a partir de Dol e Verhoog (2010).

Depois de definir o conceito, o próximo passo é compreender como essas observações se diferenciam dos outros casos da amostra. A figura abaixo ilustra a dispersão de *outliers* em um espaço bidimensional.

Figura 4 – Exemplo de *outliers* em um espaço bidimensional



Fonte: Chandola, Banerjee e Kumar (2007).

O gráfico da figura acima tem duas regiões normais (N_1 e N_2). Por sua vez, O_1 e O_2 representam duas observações atípicas enquanto O_3 é uma região de *outlier*, ou seja, que se distancia das regiões normais. Observa-se que o caso com menor valor no eixo X do grupo N_2 está muito distante dos escores dos casos agrupados em O_3.

[9] Os autores afirmam que "*thus the outlier is that observation whose removal from the sample effects the greatest reduction in the internal scatter of the data set*" (Barnett e Lewis, 1994, p.215).

De acordo com Chandola, Banerjee e Kumar (2007), existem quarto principais motivos que explicam a presença de casos desviantes: (1) atividade maliciosa; (2) erro de instrumento; (3) mudança no meio ambiente e (4) erro humano.

O exemplo típico da atividade maliciosa como gerador de casos desviantes é quando a central do cartão de crédito liga para o indivíduo para confirmar uma determinada compra. Sempre que as observações se distanciarem muito bruscamente da média, é um sinal de que algo anormal pode estar acontecendo.

O *outlier* causado por erro de instrumento é aquele que recebe uma determinada característica/atributo muito diferente das demais observações por falha do instrumento. Imagine uma balança desregulada ou um termômetro defeituoso. Se o pesquisador depende desses instrumentos para atribuir valores para os seus casos, ele corre o risco de incluir uma observação destoante em sua amostra.

Mudanças abruptas no meio ambiente, como uma tempestade, tendem a produzir observações muito diferentes do esperado. É comum observar esse tipo de ocorrência em reportagens sobre os índices pluviométricos, quando o jornalista afirma que em determinada semana choveu mais do que o esperado para todo o mês.

Por fim, o caso destoante gerado por erro humano é um dos mais recorrentes na pesquisa científica. Muitos bancos de dados são efetivamente elaborados por estudantes de graduação e pós-graduação com diferentes níveis de treinamento técnico. Mesmo pesquisadores experientes muitas vezes cometem erros durante o processo de coleta e/ou tabulação de dados. Para Barnett e Lewis (1978),

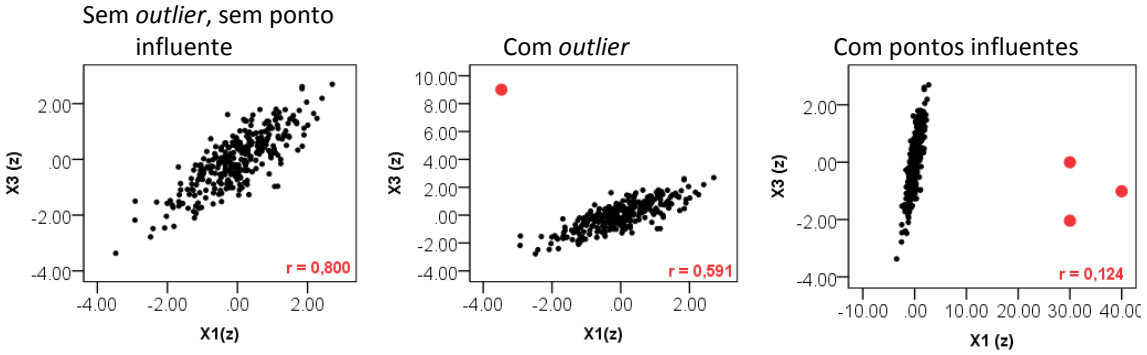
> there is a class of situations where outliers are readily handled, where the manner of dealing with them is obvious and non-controversial. Such is the situation when human errors lead to blatantly incorrect recording of data, or where lack of regard to practical factors results in serious misinterpretation" (Barnett e Lewis, 1978, p.6).

Antes de analisar os dados, muito tempo, talento e energia devem ser alocados em *data cleaning*, ou seja, o processo pelo qual o banco de dados será sistematicamente verificado em busca de eventuais erros. Para Chandola, Banerjee e

Kumar (2006), "*the importance of outlier detection is due the fact that outliers in data translate to significant (and often critical) information in a wide variety of application domains*" (Chandola, Banerjee e Kumar, 2006, p.1).

A literatura geralmente diferencia *outliers* de pontos influentes. Analiticamente, *outliers* são destoantes no *y* (variável dependente), enquanto pontos influentes são destoantes no *x* (variável independente). A figura abaixo ilustra a mesma distribuição, com um *outlier* e um ponto influente, respectivamente[10].

Figura 5 – Correlação, *outliers* e pontos influentes



Fonte: elaboração própria

Para se ter uma ideia do efeito devastador de um caso destoante, a correlação passou de 0,800 para 0,591 ao se incluir um único *outlier*. Ou seja, a presença de um único caso atípico foi suficiente para subestimar a correlação observada. Similarmente, bastou a inclusão de três pontos influentes para que uma associação forte praticamente desaparecesse, como demonstra o gráfico da direita (r = 0,124). Do ponto de vista puramente estatístico, tanto *outliers* quanto pontos influentes são sempre indesejáveis. No limite, eles podem tanto criar uma falsa correlação entre variáveis que são de fato independentes quanto esconder uma associação entre variáveis que estão, na realidade, associadas.

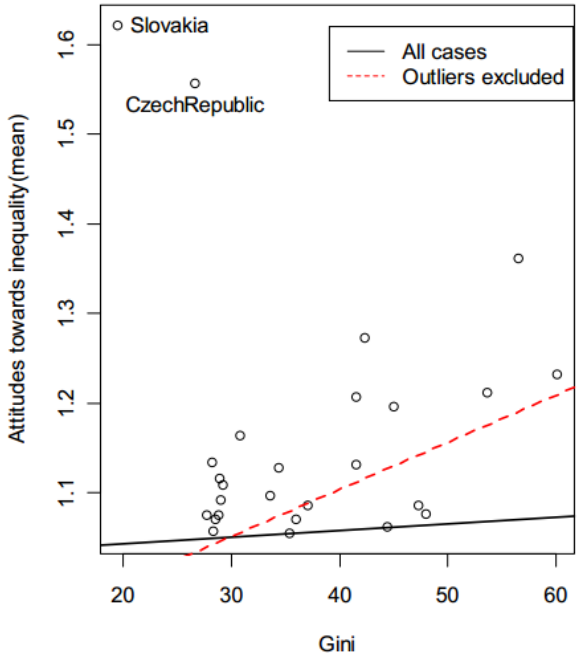
Para Jacoby (2005), "*more importantly, separated points can have a strong influence on statistical models—deleting outliers from a regression model can*

[10] Nesse artigo utilizamos *outliers*, pontos influentes, observações destoantes e casos atípicos como sinônimos.

sometimes give completely different results" (Jacoby, 2005, p.1). A presença de observações atípicas influencia não só o valor da constante do modelo de regressão, mas também a magnitude e o sinal dos coeficientes. Para Tabachnick e Fidell (2007), "*outliers are found in both univariate and multivariate situations, among both dichotomous and continuous variables, among both IVs and DVs, and in both data and results analyses. They lead to both Type I and Type II errors*" (Tabachnick e Fidell, 2007, p.28). Apesar disso, Chandola, Banerjee e Kumar (2006) afirmam que "*outliers exist in almost every data set*" (Chandola, Banerjee e Kumar, 2006:2). Por esse motivo é importante compreender os procedimentos básicos que devem ser utilizados tanto para identificar casos destoantes, quanto para lidar apropriadamente com eles.

Outliers são particularmente prejudiciais em amostras pequenas. O gráfico na figura abaixo ilustra esse argumento.

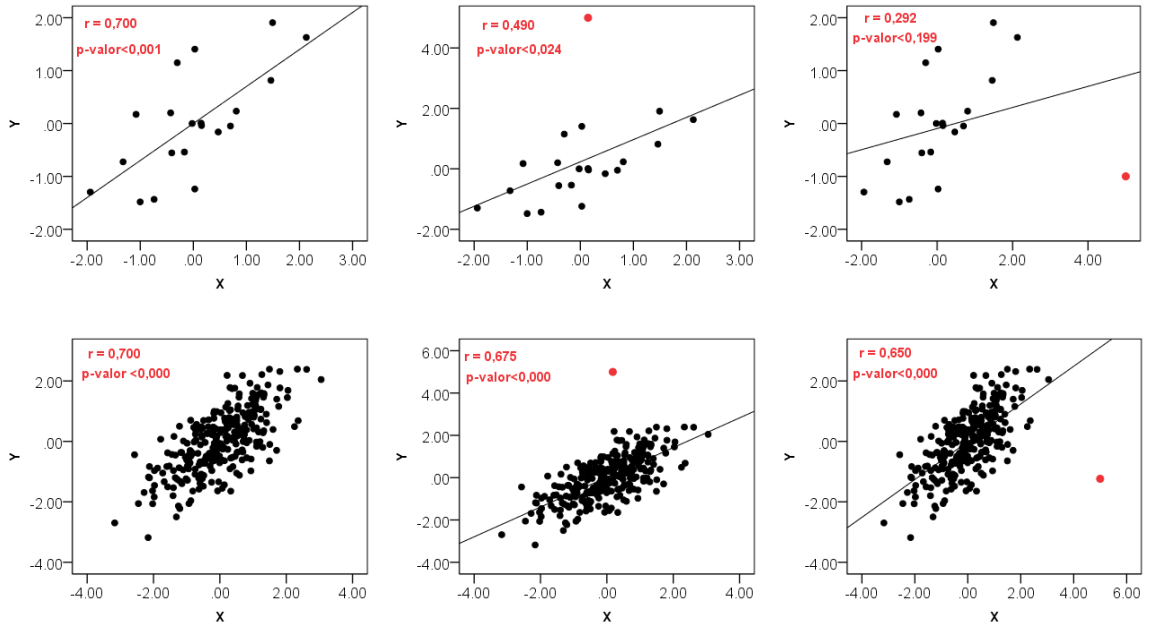
Figura 6 – Impacto dos *outliers* em amostras pequenas



Fonte: Jacoby (2005)

No caso acima, *Slovakia* e *Czech Republic* estão "puxando" a linha de regressão para baixo. Em termos mais técnicos, isso quer dizer que a inclusão desses casos na análise subestima a magnitude da correlação entre as variáveis. Para se ter uma ideia do impacto dos *outliers*, o coeficiente de determinação (R^2) do modelo passa de 0,175 com todos casos para 0,462 com a exclusão das observações atípicas[11]. O coeficiente de regressão do modelo com todos os casos é de 0,0007 (p-valor<0,27). Ao se excluir os *outliers*, tem-se uma estimativa de 0,0053 (p-valor<0,0005). Se o pesquisador não remover esses casos, ele chegaria à conclusão de que não existe associação entre as variáveis, quando na verdade existe. A figura abaixo ilustra um exemplo simulado da presença do mesmo *outlier* em amostras de tamanhos diferentes.

Figura 7 – Mesma correlação e mesmos *outliers* em amostras com tamanhos diferentes



Fonte: elaboração própria

Simulamos duas variáveis normais com média zero e desvio padrão igual a um. O nível de correlação entre elas é de 0,700. No primeiro exemplo, com uma amostra de vinte casos, incluímos um *outlier* em *y* com valor 5. Observa-se que a correlação

[11] Para uma introdução ao papel do R^2 em Ciência Política ver Luskin (1984), King (1986) e Figueiredo Filho, Silva Júnior e Rocha (2011).

passou de 0,700 (p-valor<0,000) para 0,490 (p-valor<0,024). O sinal permaneceu o mesmo (+), mesmo havendo forte redução na magnitude da associação, prejudicando a significância estatística. Incluímos aleatoriamente um caso desviante na distribuição da variável *x* (valor 5). O coeficiente de correlação de Pearson passou de 0,700 (p-valor<0,000) para 0,292 (p-valor<0,199), ou seja, como a significância estatística está fora dos limites usualmente empregados (10%, 5% e 1%), o pesquisador seria levado a não rejeitar a hipótese nula, quando, na realidade, deveria rejeitá-la. Repetimos esse mesmo procedimento agora para uma amostra de 300 casos. Observa-se que a magnitude do coeficiente foi apenas marginalmente alterada enquanto a significância estatística permaneceu consistente. A principal conclusão desse exercício de simulação é que amostras grandes são mais resistentes à presença de *outliers*[12].

2.2 Como detectar outliers?[13]

De acordo com Barnett e Lewis (1978), o primeiro teste objetivo para detectar observações atípicas foi elaborado pelo astrônomo norte-americano Benjamin Peirce (1852)[14]. Para Ben-Gal (2005),

> one of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is

[12] Outro efeito negativo produzido pela presença de observações atípicas é a distorção gráfica. Para o professor Jacoby (2005), "*outliers can affect visual resolution of remaining data in plots (forces observations into clusters and the temporary removal of outliers, and/or transformations can spread out clustered observations and bring in the outliers (if not removed)*" (Jacoby, 2005, p.3).

[13] Chandola, Banerjee e Kumar (2006) afirmam que "outliers detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains" (Chandola, Banerjee e Kumar, 2006, p.1).

[14] Para Peirce, observações destoantes *k* em uma amostra *n* devem ser rejeitadas se "*the probability of the system errors obtained by retaining them is less than that of the system errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations*" (*apud* Barnnett e Lewis, 1978, p.18).

> therefore important to identify them prior to modeling and analysis (Ben-Gal, 2005, p.1).

A identificação de *outliers* pode ser realizada através de técnicas univariadas ou multivariadas. Outra clivagem apontada pela literatura é a distinção entre técnicas paramétricas e não-paramétricas[15].

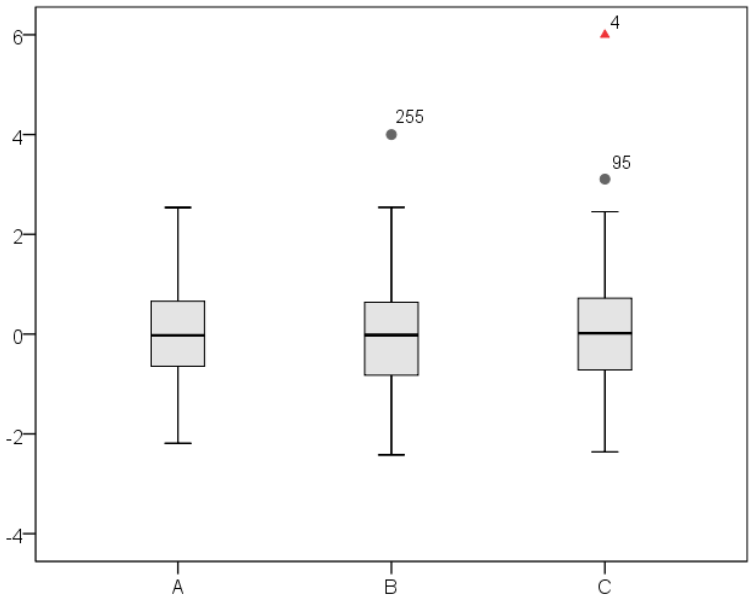
O procedimento mais elementar para identificar casos destoantes é a estatística descritiva. Primeiramente, o pesquisador pode observar a magnitude do desvio padrão em relação à média. Uma distribuição muito heterogênea pode ser um indicativo da presença de uma observação destoante. Além disso, é possível utilizar outras estatísticas para avaliar a dispersão dos dados[16], além da análise gráfica. Em particular, tanto o gráfico de dispersão quanto o *boxplot* são eficientes para identificar a presença de casos destoantes. Para interpretar o *boxplot* é necessário entender como ele funciona. A distribuição é representada por uma caixa retangular conectada às linhas. O tamanho da caixa representa a amplitude interquartílica e contém 50% de todas as observações. A linha transversal dentro da caixa representa a mediana. Os limites das linhas superiores e inferiores delimitam os valores máximo e mínimo, respectivamente. Eventuais círculos acima ou abaixo desses limites podem ser considerados como potenciais *outliers*. É possível ainda identificar a presença de pontos extremos, eles estarão sempre acima ou abaixo dos pontos destoantes. Para Pallant (2007), "*in addition to providing information on outliers, a boxplot allows you to inspect the pattern of scores for your various groups. It provides an indication of the*

[15] Para Jupp (2006), "*non-parametric tests can be used with skewed data. They are not influenced by outliers, which are extremes scores that can affect parametric testes because they increase the variance of the data set. Non-parametric tests are less sensitive than the parametric equivalents, and do not permit one to test the significance of the interactions between independent variables*" (Jupp, 2006, p.214). Mais informações sobre como identificar casos atípicos podem ser encontradas em: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[16] Por exemplo, tanto a *skewness* quanto a *kurtosis* podem ser utilizadas para melhor compreender a distribuição das variáveis. Uma distribuição perfeitamente normal apresenta *skewness* e *kurtosis* com valor zero. Como regra geral, quanto mais essas estatísticas se distanciarem de zero, tanto pior. *Skewness* positiva sugere observações concentradas a esquerda e com valores baixos. *Skewness* negativa indica concentração de observações a direita e com valores altos. *Kurtosis* positiva sugere concentração no centro com caldas longas. *Kurtosis* negativa indica concentração de casos nos extremos. Analiticamente, consideramos que a inspeção gráfica do histograma também deve ser realizada no sentido de conhecer a distribuição das variáveis. Em geral, no entanto, o pesquisador não precisa se preocupar muito com essas estatísticas quando trabalha com grandes bases de dados.

variability in scores within each group and allows a visual inspection of the differences between groups" (Pallant, 2007, p.77). O gráfico da figura abaixo ilustra a distribuição de três variáveis simuladas com o objetivo de demonstrar como o *boxplot* pode ser útil na identificação de *outliers*.

Figura 8 – *Boxplot* para três variáveis simuladas



Fonte: elaboração própria.

Todas as variáveis foram simuladas como normais, com média zero e desvio padrão igual a um. Depois disso, selecionamos, aleatoriamente, um caso da variável B para assumir valor 4 (ID 255) e outro caso da variável C para assumir valor 6 (ID 4). O *boxplot* identifica rapidamente que observações se distanciam das demais. Sugerimos que ele deva ser utilizado mais durante o processo de limpeza do banco e/ou em estudos puramente descritivos do que propriamente como opção de apresentação gráfica.

Além do *boxplot*, é possível utilizar o histograma e o gráfico de dispersão para examinar visualmente a distribuição dos dados. O pesquisador também pode observar a variação do coeficiente de determinação (r^2) entre o ajuste dos modelos: um modelo

com todos os casos e outro excluindo os casos destoantes. Quanto maior a diferença entre eles, maior é o impacto dos *outliers* (ver exemplo de Jacoby (2005) acima).

Uma medida usualmente empregada para capturar o efeito dos *outliers* e/ou pontos influentes sobre a consistência das estimativas é o *hat-values* (h_i) (Fox, 2008)[17]. O objetivo é determinar quanto cada observação de *y* se distancia dos seus respectivos valores preditos ($\hat{y}$).

Outro procedimento comumente recomendado pela literatura é padronizar as variáveis e considerar como *outliers* aquelas observações com valores acima ou abaixo de três desvio padrão em relação à média (Rodrigues e Paulo, 2007). A literatura em Estatística desenvolveu diferentes testes para identificar a presença de *outliers* em uma determinada distribuição. Especificamente em relação aos testes que assumem uma distribuição aproximadamente normal, a maior parte deles utiliza alguma distância entre os casos e a média. Por exemplo, o teste de Grubbs é recomendado para testar a presença de um único caso destoante (Grubbs, 1950, 1969). O teste de Tietjen-Moore é uma generalização do teste de Grubbs para mais de um caso (Tietjen e Moore, 1972). O teste *generalized extreme studentized deviate* requer apenas o estabelecimento de um limite superior em relação ao número estimado de *outliers* e é adequado para casos em que o pesquisador não sabe a quantidade de observações destoantes presentes em seu banco de dados.

Por fim, é importante também destacar a distância de Cook como método desejável para identificar a presença de observações atípicas. No original, "*a diagnostic*

[17] Diferentes pacotes estatísticos fornecem estimativas variadas para não só identificar, mas também avaliar o impacto dos *outliers*. Por exemplo, *o Statistical Package for Social Sciences* (SPSS), versão 20, oferece as seguintes opções: (1) DfBeta(s). The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant; (2) Standardized DfBeta. Standardized difference in beta value. The change in the regression coefficient that results from the exclusion of a particular case. You may want to examine cases with absolute values greater than 2 divided by the square root of N, where N is the number of cases. A value is computed for each term in the model, including the constant; (3) DfFit. The difference in fit value is the change in the predicted value that results from the exclusion of a particular case; (4) Standardized DfFit. Standardized difference in fit value. The change in the predicted value that results from the exclusion of a particular case. You may want to examine standardized values which in absolute value exceed 2 times the square root of p/N, where p is the number of parameters in the model and N is the number of cases e (5) Covariance ratio. The ratio of the determinant of the covariance matrix with a particular case excluded from the calculation of the regression coefficients to the determinant of the covariance matrix with all cases included. If the ratio is close to 1, the case does not significantly alter the covariance matrix (SPSS Help, versão 20).

statistic called Cook's distance can be used to summarize essential information about the influence about the influence of each case on the estimated regression coefficients. Cook's distance is a mathematical measure of the impact of deleting a case. It is not intended for use as statistical test" (Cook e Weisberg, 1999, p.357). Ela representa a importância de cada observação para os coeficientes de regressão quando um caso específico é retirado da análise (Figueiredo Filho *et al*, 2011).

2.3 Como lidar com os outliers?

A literatura aponta diferentes procedimentos para lidar com os casos destoantes[18]. Uma possibilidade é utilizar de modelos não lineares ou corrigir a distribuição das variáveis utilizando alguma transformação (Figueiredo Filho *et al*, 2011). Comparativamente, tanto o log quanto a transformação de Box-Cox (Weisberg, 2005) são amplamente utilizadas. Iglewicz e Hoaglin (1993) sugerem três principais abordagens: (1) *labeling*; (2) *acomodation* e (3) *identification*. Substantivamente, a rotulação (*labeling)* consiste em conduzir uma investigação adicional com o objetivo de melhor compreender os *outliers*. Dependendo do número de casos destoantes e dos recursos logísticos, o pesquisador pode implementar um estudo de caso e/ou uma análise comparativa de poucos casos (*Small n*).

Por sua vez, a acomodação consiste em utilizar diferentes técnicas estatísticas com o objetivo de minimizar os efeitos adversos produzidos pelas observações destoantes sobre a consistência das estimativas. O pesquisador pode considerar a alteração da forma funcional, transformação das variáveis e/ou empregar técnicas menos sensíveis à presença de *outliers*. Alguns manuais sugerem a recodificação do caso destoante por um valor menos abrupto como forma de minimizar os efeitos negativos dos *outliers*. Particularmente, sugerimos que os pesquisadores evitem fazer isso. E, se optarem por fazê-lo, é importante reportar, exatamente, os critérios utilizados. Advertimos ainda que o pesquisador não pode simplesmente deletar uma

[18] Para uma discussão bastante didática ver Barnett e Lewis (1978).

observação destoante do banco de dados. Na dúvida, ele deve consultar os manuais aplicados e/ou buscar ajuda de pesquisadores mais experientes.

Por fim, a identificação consiste em utilizar diferentes métodos para detectar quais são as observações destoantes e tentar estimar o seu impacto sobre a consistência e eficiência das estimativas. Evidentemente, esses procedimentos apenas devem ser empregados quando o pesquisador descartar a hipótese de erros no banco de dados, ou seja, os *outliers* de fato existem na realidade e não são um artefato produzido por imperícia humana e/ou outro processo gerador.

Independente da abordagem escolhida pelo pesquisador, é importante seguir a recomendação de Afifi e Clark (1999): realizar duas análises multivariadas, uma com todos os casos e outra sem as observações destoantes. Quanto mais diferentes forem os resultados, pior, ou seja, maior é o impacto dos *outliers*. Contrariamente, quanto mais semelhantes forem as estimativas, melhor, ou seja, menor é o efeito das observações atípicas sobre a consistência das estimativas.

3. Cuidado com as Correlações Espúrias

Essa seção apresenta exemplos pitorescos de correlações espúrias. O objetivo é alertar os pesquisadores a respeito dos perigos de realizar inferências a partir da análise de correlações bivariadas. É conhecida a tendência do cérebro humano em identificar padrões, mesmo quando diante de distribuições aleatórias. Para Sagan (1985), "*humans are good, she knew, at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent*" (Sagan, 1985, p.15). É igualmente comum, tanto na academia quanto fora dela, ouvir inferências causais a partir do exame de relações bivariadas. Dada a complexidade inerente à realidade social, dificilmente algum fenômeno relevante pode ser analisado a partir de relações bivariadas. O quadro 2 sumariza exemplos de correlações espúrias.

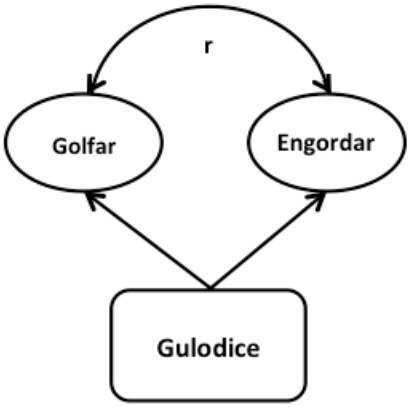
Quadro 2 – Exemplos de correlações espúrias

Correlação espúria observada	*Direção*	*Variável omitida*
Quantidade de sorvete e mortes por afogamento (Moore, 1993)	+	Temperatura (estação do ano). Consumo de sorvete e mortes por afogamento tendem a ser maiores nos períodos mais quentes do ano
Tamanho das duas mãos	+	Genética
Salário dos ministros e preço de vodka	+	Área (urbana ou rural). Os salários e os preços de vodka são maiores em regiões urbanas
Tamanho do pé e desempenho em leitura	+	Idade. Quanto mais velho, maior é a habilidade da criança em ler.
Número de policiais e quantidade de crimes	+	Densidade populacional
Consumo de chá e câncer de pulmão	–	Tabagismo

Fonte: http://www.southalabama.edu/coe/bset/johnson/oh_master/Ch11/Tab11-02.pdf

Comparativamente, considere a correlação entre golfar e engordar. Isso porque existe a crença de que bebê que golfa muito, ganha peso mais rápido. A figura abaixo ilustra a correlação entre golfar e engordar.

Figura 9 – Correlação entre golfar e engordar



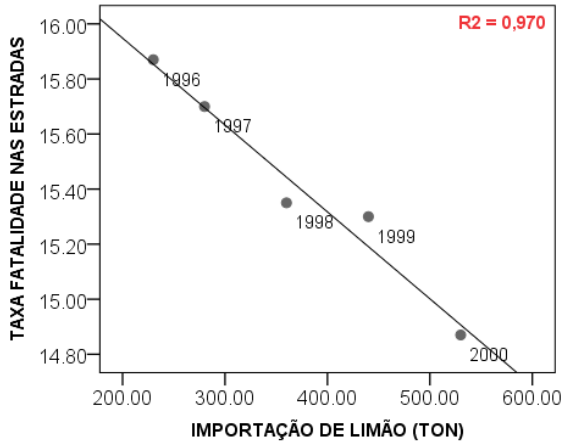
Fonte: elaboração própria.

A correlação observada entre golfar e engordar pode ser explicada na medida em que as variáveis tem a mesma causa: gulodice[19]. Ao se controlar pelo efeito da gulodice, a correlação entre as variáveis desaparece, caracterizando a espuriosidade da relação. Para Moore e McCabe (2003),

> beware the lurking variable is good advice when thinking about an association between two variables. The observed association between the variable x and y is explained by a lurking variable z. Both z and y change in response to changes in z. this common response creates an association even though there may be no direct causal link between x and y (Moore e Mccabe, 2003, p.181)[20].

Lembrando que "we can safely omit control variables, even if they have a strong influence on the dependent variable, as long as they do not vary with the included explanatory variable" (King, Keohane e Verba, 1994, p.169). Ou seja, a exclusão de variáveis independentes apenas é um problema quando elas se correlacionarem com as demais variáveis explicativas incluídas no modelo. Para os propósitos desse artigo, examinaremos visualmente diferentes correlações espúrias.

Figura 10 – Correlação entre importação de limão e taxa de fatalidade nas estradas (EUA)



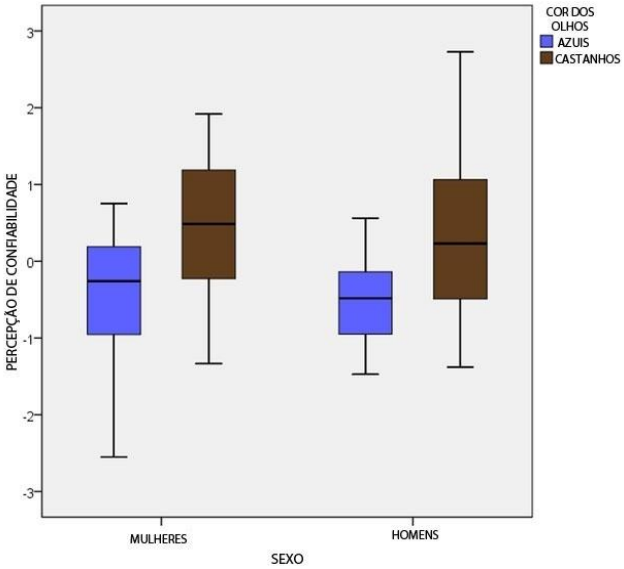
Fonte: traduzido de http://pipeline.corante.com/archives/2009/04/01/mexican_lemons_to_the_rescue.php

[19] De acordo com o Houssais, o termo gulodice é proveniente da alteração da palavra gulosice e remonta ao século XV. No Nordeste brasileiro a palavra gulodice é usualmente empregada para descrever indivíduos que comem de forma demasiada.

[20] Amaral (2010) afirma que "a omissão de uma variável importante pode causar correlação entre o erro e as variáveis explicativas, o que pode gerar viés e inconsistência em estimadores MQO" (Amaral, 2010, p.3).

A partir da inspeção gráfica desses dados o pesquisador não deve concluir que a importação de limão tem algum efeito causal sobre a taxa de fatalidade nas estradas. Lembremos que o método científico começa com a teoria. É a partir dela que hipóteses são elaboradas e testadas. Duas variáveis podem estar relacionadas porque uma cresceu, enquanto a outra reduziu no tempo. O pesquisador também deve evitar criar explicações ad hoc com o objetivo de justificar correlações dessa natureza. Devemos evitar, a todo custo, a máxima de torturar os dados até que eles confessem.

Figura 11 – Correlação entre cor dos olhos e percepção de confiabilidade



Fonte: traduzido de http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0053285

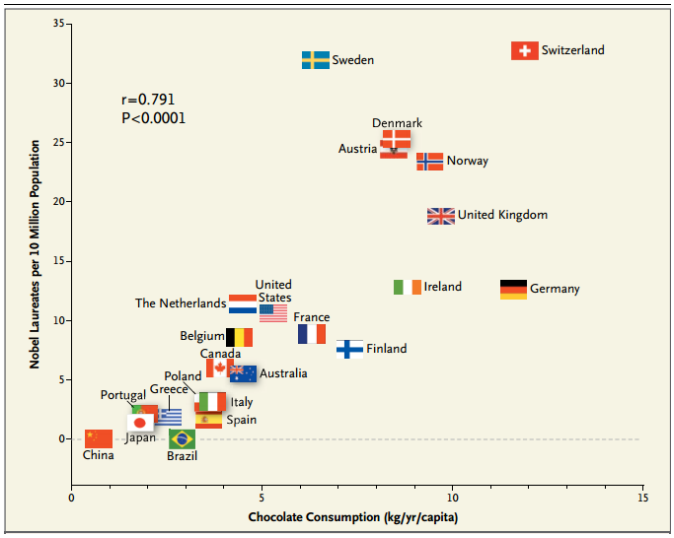
Em uma pesquisa recente, Kleisner *et al* (2013) detectaram uma correlação entre nível de confiabilidade e cor dos olhos, independente do sexo. Para os homens (F = 6,72 e p-valor=0,014) e para as mulheres (F = 11,01 e p-valor = 0,002). Em particular, pessoas de olhos castanhos foram consideradas mais confiáveis do que pessoas de olhos azuis (ver gráfico da figura acima). Os autores observaram, no entanto, que a cor dos olhos também está correlacionada com o formato do rosto. Após incluírem essa

última variável na análise Kleisner *et al* (2013) concluiram que "although the brown-eyed faces were perceived as more trustworthy than the blue-eyed ones, it was not brown eye color per se that caused the stronger perception of trustworthiness but rather the facial features associated with brown eyes" (Kleisner *et al*, 2013, p.1).

O problema de variável omitida é recorrente em diferentes desenhos de pesquisa. De acordo com Moore e McCabe (2003), a variável omitida pode afetar dramaticamente a consistência das estimativas. No que diz respeito especificamente ao coeficiente de correlação, "lurking variables sometimes create nonsense correlations between x and y. They can also hide a true relationship between x and y" (Moore e McCabe, 2003, p.164).

Por mais estranhas que possam parecer, as correlações espúrias, assim como os *outliers*, são recorrentes em diferentes áreas do conhecimento. Optamos por examinar a correlação reportada por Messerli (2012), no artigo *Chocolate Consumption, Cognitive Function and Nobeal Laureates,* publicado no *New England Journal of Medicine*. Os resultados sugerem que "*there was a close, significant linear correlation (r=0.791, P<0.0001) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries*" (Messerli, 2012, p.1562). O gráfico da figura abaixo ilustra essa correlação.

Figura 12 – Correlação entre consumo de chocolate e prêmio Nobel



Fonte: Messerli (2012)

Messerli (2012) afirma que a cada aumento de 0,4kg no consumo de chocolate *per capita* por ano, em média, deve-se esperar um prêmio Nobel. O autor conclui que "chocolate *consumption enhances cognitive function, which is a sine qua non for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country*" (Messerli, 2012, p.1564). Essas inferências devem ser tomadas com extrema cautela, para dizer o mínimo. Primeiro, a amostra de países analisados não é aleatória. Segundo, o critério de inclusão dos casos foi a conveniência de oferta de dados, que exerce um efeito desastroso sobre a consistência das estimativas, ou seja, os coeficientes serão enviesados. Terceiro, a amostra é pequena. Tanto o efeito dos *outliers* quanto a probabilidade de surgimento de correlações espúrias aumentam quando a amostra é reduzida. Quarto, o autor não incluiu nenhuma variável de controle. Provavelmente, a variável omitida aqui é renda/educação. Espera-se que ao se controlar o efeito dessas variáveis, a correlação entre consumo de chocolate e prêmio Nobel desapareça, caracterizando a espuriosidade da relação.

Considerações Finais

O avanço computacional e o compartilhamento de informações revolucionaram as possibilidades de análise de dados. Análises que antes demoravam semanas ou meses para serem realizadas em computadores institucionais podem ser realizadas em poucos segundos em computadores pessoais. No entanto, antes de extrair conclusões substantivas dos dados é necessário identificar e corrigir a presença de *outliers*.

O principal objetivo desse trabalho foi aprofundar a discussão a respeito do coeficiente de correlação de Pearson utilizando uma abordagem intuitiva. Em particular, examinamos o papel de observações atípicas sobre a consistência do coeficiente, além de alertar os pesquisadores a respeito do perigo das correlações espúrias. Foi demonstrado que *outliers* são especialmente problemáticos em amostras pequenas. Eles podem tanto esconder a presença de uma correlação verdadeira

quanto forjar uma associação inexistente. Na medida em que o número de casos cresce, menor é o efeito dos casos destoantes sobre a consistência das estimativas.

Acreditamos que a Ciência Política nacional pode se beneficiar bastante da utilização de técnicas quantitativas de pesquisa. No entanto, para que os métodos possam efetivamente auxiliar a construção do conhecimento científico é necessário compreender as potencialidades e limitações de cada ferramenta de pesquisa. Com esse artigo esperamos facilitar não só a interpretação, mas também a utilização da técnica de correlação na Ciência Política.

Referências Bibliográficas

Afifi, Abdelmonem A. and Clark, Virginia. 1999. *Computer-Aided Multivariate Analysis*, 2nd ed. London: Chapman and Hall Limited.

Barnett, Vic e Lewis, Toby. 1994. *Outliers in Statistical Data*. 3rd. Edition. John Willey and Sons.

Ben-Gal, Irad. 2005. "Outlier detection". In: Maimon, O.; Rockach, L. (eds). *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. New York, p. 1-16.

Berton, Lilian e Zhao, Liang. 2011. "Caracterização de Classes via Otimização em Redes Complexas", In: VIII Encontro Nacional de Inteligência Artificial (ENIA2011), 2011, Natal, Anais do VIII Encontro Nacional de Inteligência Artificial (ENIA2011). , v. 1, pp. 548-559.

Chandola, Varun; Banerjee, Arindam and Kumar, Vipin. 2007. "Anomaly Detection - A Survey", *ACM Computing Surveys* 41 (2007), 15.

Chawla, Sanjay and Sun, Pei. 2006. "Slom: A new measure for local spatial outliers". *Knowledge and Information Systems*, Vol. 9, No. 4, pp. 412–429.

Chen, Paula Y. and Popovich, Peter M. 2002. *Correlation: parametric and nonparametric measures.* London: Sage, 95p.

Cohen, Jacob. 1998. *Statistical Power Analysis for the Behavioral Sciences, Second Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, R. Dennis. and Weisberg, Sanford. 1994. *An Introduction to Regression Graphics*. New York: Wiley.

Cordeiro, Gauss M. 1986. *Modelos Lineares Generalizados*. Campinas: UNICAMP, 286p .

Dancey, Christine P. e Reidy, John. 2006. *Estatística Sem Matemática para Psicologia: Usando SPSS para Windows*. Porto Alegre, Artmed.

Figueiredo Filho, Dalson. B. e Silva Júnior, Joé. A. Da. 2009. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). *Revista Política Hoje*, v. 18, n. 8, p. 115-146.

Figueiredo Filho, Dalson. B.; Silva Junior, José. A da; Rocha, Enivaldo. C. Da. 2011. What is R2

all about? *Leviathan - Cadernos de Pesquisa Política*, v. 3, p. 60-68.

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models*. Second Edition, Sage Publications.

Gill, Jeff. 2001. "Generalized linear models: a unified approach. Sage University Paper Series on Quantitative Applications in the Social Sciences", 07-134, Thousand Oaks, CA.

Grubbs, Frank E. 1950. "Sample Criteria for Testing Outlying Observations". *Annals of Math. Statistcs,* 21 , 27-58.

Grubbs, Frank. E. 1969. "Procedures for detecting outlying observations". *Technomctrics*, 11, 1-21.

Hawkins, Douglas. 1980. *Identification of Outliers*, Chapman and Hall, London.

Ho, Robert. 2006. "Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS." *Journal of Statistical Software*, July 2006, Volume 16, Book Review 4.

Hodge, Victoria. and Austin, Jim. 2004. "A survey of outlier detection methodologies." *Artificial Intelligence Review*, 22:85–126.

Jensen, Donald. R. and Ramirez, Donald. E. 1998. "Detecting outliers with cook's D_1 statistic". *Computing Science and Statistics* 29(1), 581-586.

Johnson, Richard A. 1992. *Applied Multivariate Statistical Analysis*. Prentice Hall.

Jupp, Victor. 2006. *The Sage Dictionary of Social Research Methods.* London: Sage.

Kriegel, Hans-Peter, Kroger, Peer, and Zimek, Arthur. 2009. "Clustering high-dimensional data: A sur- vey on subspace clustering, pattern-based clustering, and correlation clustering." *ACM Trans. Knowl. Discov. Data*, 3:1:1–1:58.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science*, 30:666-687.

King, Gary, Keohane, Robert e VERBA, Sidney. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton*. N.J.: Princeton University Press.

Kleisner, Karel; Priplatova, Lenka; Frost, Peter and Flegr, Jaroslav. 2013. "Trustworthy-Looking Face Meets Brown Eyes". *PLOS ONE*, jan/13, Volume 8.

Luskin, Robert C. 1984. "Looking for R2: Measuring Explanation outside OLS". *Political Methodology* 10: 513-32.

Mendenhall William.; Reinmuth James. E. and Beaver, Robert J. 1993. *Statistics for Management and Economics*. Belmont, CA: Duxbury Press.

Messerli, Franz. H. 2012. "Chocolate consumption, cognitive function, and Nobel laureates". *The New England Journal of Medicine,* Oct.18. *In:* http://scottbarrykaufman.com/wp-

content/uploads/2012/10/Messerli-2012.pdf.

Moore, David S. and Mccabe, George P. 2003. *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Pallant, Julie. 2007. *SPSS Survival Manual*. Open University Press.

Pearson, Karl. 1914. "On the probability that two independent distributions of frequency are really samples of the same population, with special reference to recent work on the identily of trypanosome strains". *Biometrika*, v.10, p.85-143.

Pollock III, Philip H. 2006. *A Stata Companion to Political Analysis*. Washington, DC: CQ Press.

Pyle, Dorian. 1999. *Data Preparation for Data Mining*. Morgan Kaufmann

Rodrigues, Adriano e Paulo, Edilson. 2007. "Introdução à análise multivariada". In. Corrar, Luiz J.; Paulo, Edilson; Dias-Filho, J. Maria (Coord.). *Análise multivariada: para os cursos de administração, ciências contábeis e economia*. São Paulo: Atlas, p. 1-72.

Sagan, Carl. 1985. *O romance da Ciência.* Rio de janeiro, Francisco Alves.

Santos, Maria Helena C. e Coutinho, Marcelo. 2000. "Políticacomparada: estado das artes e perspectivas no Brasil". *BIB,* 54: 3-146.

Seo, Songwon. 2006. *A review and comparison of methods for detecting outliers in univariate data sets*. Dissertation of M. Sc., Graduate School of Public Health, University of Pittsburgh, Pittsbur- gh, Pennsylvania, US.

Soares, Gláucio A. D. 2005. "O Calcanhar Metodológico da Ciência Política no Brasil". *Sociologia, Problemas e Práticas.* Lisboa, v. 2, n. 48, p. 27-52.

Stanton, Jeffrey M. 2001. "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors". *Journal of Statistics Education*: Syracuse University. 9.

Stigler, Stephen. 1989. "Francis Galton's account of the invention of correlation". *Statistical Science*, *4*(2), 73–79.

Tabachnick, Barbara. and Fidell, Linda. 2007. *Using multivariate analysis*. Needham Heights: Allyn & Bacon.

Tietjen, Gary and Moore, Roger. 1972. "Some Grubbs-types statistics for detection of several outliers". *Technometrics*, 14 , 583-597.

Valle e Silva, Nelson. 1999. *Relatório de Consultoria sobre Melhoria do Treinamento em Ciência Social Quantitativa e Aplicada no Brasil*. Rio de Janeiro, Laboratório Nacional de Computação Científica.

Weisberg, Sanford. 2005. *Applied linear regression.* Hoboken NJ: John Wiley.

Werneck Vianna, Luiz; Carvalho, Maria A. R.; MELO, Manuel P. C. e BURGOS, Marcelho B. 1998. "Doutores e teses em ciências sociais". *Dados*, 41, 3: 453-515.

Sites visitados

Amaral, Ernesto F. L. 2010. "Aulas 27 e 28: Problemas Adicionais de Especificação e de Dados". Disponível em: http://www.ernestoamaral.com/docs/dcp030d-101/Aulas27-28.pdf Acesso em: 17 fev/2013.

Jacoby, William G. 2005. "Lecture 11: Outliers and Influential data". Disponível em: http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/11.Outliers.pdf Acesso em: 18 fev/2013.

http://pipeline.corante.com/archives/2009/04/01/mexican_lemons_to_the_rescue.php Acesso em: 10 fev, 2013.

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0053285 Acesso em: 03 fev, 2013.

Tramitação do artigo na revista
Submetido: 10/06/2013
Revisões requeridas: 15/08/2013
Versão revista: 10/10/2013
Aceito: 15/04/2014