

CORRELAÇÕES E CAUSALIDADE: BREVES REFLEXÕES NUMA PERSPECTIVA ECONÓMICA

Miguel Patrício¹

Índice: I – Breve Excurso Histórico sobre as Correlações e a Causalidade. II – Apontamentos sobre a Base Matemática das Correlações. III – Riscos e Limites das Correlações para a Economia



há mais de 100 anos, um economista de nome Warren Persons, escrevia algo que, já naquela altura, não era novidade, mas trazia uma importante sabedoria implícita. Dizia o autor na revista *American Statistical Association*: “*The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with en masse. Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate «other things remaining the same», with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than «such and such a cause tends to produce such and such an effect». Events can only be stated to be more or less probable.*”²

¹ Professor da Faculdade de Direito da Universidade de Lisboa.

² PERSONS, Warren M. – “The correlation of economic statistics”, in: *American Statistical Association*, 12 (92), 1910, p. 287.

O que pode parecer uma evidência, quer para o especialista, quer para o leigo, pode também, por razões que se depreenderão mais adiante, não o ser para alguns – o que assume uma especial gravidade quando na origem de um qualquer determinismo de natureza correlacional (ou de ordem causal) estão... especialistas.

Nessa medida, parece ser cada vez mais actual a necessidade de compreender o papel e a importância das correlações na Economia (e, de igual modo, no Direito), o que pode conseguir-se, ainda que em breves páginas, através da análise da sua história, dos seus avanços e, também, dos seus potenciais riscos e enviesamentos.

I – BREVE EXCURSO HISTÓRICO SOBRE AS CORRELAÇÕES E A CAUSALIDADE

A história moderna da *teoria da correlação* pode ser, para os efeitos do presente texto, sinteticamente resumida recorrendo aos contributos que foram sendo dados por Carl F. Gauß (1821-23), David Hume (1738/1748) e Stuart Mill (1843), Francis Galton (1877-78) e Karl Pearson (1895), Charles Spearman (1904) e Maurice Kendall (1938).

O *Princeps mathematicorum* deu inúmeros contributos nas mais diversas áreas científicas relacionadas com a matemática, mas interessará aqui salientar, em particular, o contributo excepcional dado com a *Theoria combinationis observationum erroribus minimis obnoxiae*, de 1821-23 (ainda que se deva notar que o referido texto é, no que respeita ao maior dos seus contributos para a análise estatística – o *método dos mínimos quadrados* –, uma versão refinada e desenvolvida de um texto de 1809³).

³ Há alguma disputa sobre a paternidade do método, uma vez que Adrien-Marie Legendre publica sobre o referido método em 1805, mas Gauß contra-argumentará que os seus avanços nesta matéria, ainda que não publicados, datam de, pelo menos, 1795. Por ter publicado antes de Carl F. Gauß, e também de uma forma independen-

Ainda que alguns autores, como é o caso de H. M. Walker (1929), refiram Carl Gauß como um dos principais precursores do desenvolvimento matemático da *teoria da correlação*, Karl Pearson, em 1920⁴, afirmaria que faltou a Gauß a ideia de “associação orgânica”, que seria a “concepção fundamental da correlação”. Seria, aliás, a falta desta “associação” que permitiria, no seu entender, definir a fronteira entre os “teóricos do erro” (os matemáticos pré-Galton), preocupados com uma *análise teórica do erro*, e a perspectiva *empírica*, elaborada por Francis Galton com recurso à *regressão estatística*.

Contudo, a par desta necessidade de analisar e mensurar o erro, outros autores ajudaram a entender a importância específica da correlação e da causalidade, bem como da relação que se pode estabelecer (ou não) entre ambas. É precisamente nesse contexto que se podem enquadrar as contribuições de David Hume e John Stuart Mill, ainda que as mesmas não contenham uma elaboração matemática.

O filósofo David Hume é um autor que dispensa quaisquer apresentações. E a sua importância actual para os estudos da causalidade, numa perspectiva económica, também é indiscutível, ainda que não tenha sido o primeiro a reflectir sobre tal assunto. Com efeito, como refere Kevin Hoover⁵, “*before Da-*

te, justifica-se mencionar o norte-americano Robert Adrain, com o artigo “Research concerning the probabilities of the errors which happen in making observations, &c.”, de 1808.

⁴ Disse K. Pearson, em “Notes on the history of correlation”, in: *Biometrika*, 13 (1), 1920, p. 27: “*Gauss’s treatment is almost the inverse of our modern conceptions of correlation. For him the observed variables are independent, for us the observed variables are associated or correlated. For him, the non-observed variables are correlated owing to their known geometrical relations with observed variables; for us the unobservable variables may be supposed to be uncorrelated causes, and to be connected by unknown functional relations with the correlated variables. In short there is no trace in Gauss’ work of observed physical variables being – apart from equations of condition – associated organically which is the fundamental conception of correlation.*”

⁵ HOOVER, Kevin D. – *Causality in Macroeconomics*. New York, Cambridge University Press, 2001, p. 2.

vid Hume's discussion of causality there was Aristotle and the scholastics and Nicolas Malebranche [...]. Yet Hume, the central figure of the philosophical school of British empiricism [...] stands at the headwaters of all modern discussions of [...] causality”.

Um dos principais contributos de David Hume consistiu na definição bipartida de causalidade que nos legou no seu *Human Understanding*, de 1748: “*We may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object has not been, the second never had existed*” (sect. VII, part II).

Em síntese: as relações causais (ou causa-efeito) apenas podem ser apreendidas através da experiência e não através de pensamento abstracto; e essas relações têm duas dimensões distintas: a regularidade, em termos probabilísticos, da sucessão (com natural prioridade temporal da causa sobre o efeito), e a dependência contrafactual.

Contudo, o *processo mental* subjacente à definição das relações causais assume-se, para Hume, como uma dimensão de relevo.

Como bem notam, a este respeito, Drakopoulos e Torrance⁶, citando *A Treatise on Human Nature*, de 1738: “*This [...] element, the regularly observed past constant conjunctions of ‘cause’ and ‘effect’, led Hume to identify what we now call the ‘problem of induction’. Although for all causal reasoning it is essential to assume ‘that instances, of which we have had no experience, must resemble those of which we have had experience, and that the course of nature continues always uniformly the same’ (Hume, p. 89), this crucial proposition cannot be established by either logical or empirical argument. Hume’s*

⁶ DRAKOPOULOS, Stravros A.; TORRANCE, Thomas S. – “Causality and determinism in economics”, in: *Scottish Journal of Political Economy*, 41 (2), 1994, p. 181.

'solution' to the problem that there appeared no justifiable basis for the formation of universal principles of order, was to make causality essentially subjective. He argued (Hume, p. 97) that the human mind tends to establish psychological connections between one observation and another and that these mental feelings, by a process of transference, are then read into the events themselves. Although Hume's final conclusion that causal relationships exist only subjectively did not find later support, the rest of his analysis provided the base from which subsequent analysis started."

John Stuart Mill iria seguir as pisadas de David Hume, no sentido em que tentaria desenvolver a ideia de que a *inferência indutiva*⁷ subjacente àquelas conexões pode ser formalizada em cinco *métodos de indução*, assim descritos no seu livro de 1843, *A System of Logic*⁸: 1) o método da *concordância directa* (a circunstância só pode ser considerada uma condição necessária se o efeito estiver presente); 2) o método da *diferença* (se a diferença entre ocorrer ou não um fenómeno depende apenas da remoção de uma circunstância, esta pode ser efeito, causa ou uma parte da causa); 3) o método agregado da *concordância* e da *diferença* (a circunstância que preenche os dois métodos anteriores em simultâneo, pode ser efeito, causa ou parte de uma causa); 4) o método do *resíduo* (se, através de induções anteriores, se tiver ligado todo um conjunto de factores, excepto um, a todo um conjunto de fenómenos, excepto um, o fenómeno remanescente pode ser considerado como efeito do factor remanescente); e 5) o método das *variações concomitantes* (se um de dois fenómenos varia de modo particular, concomitantemente com uma qualquer variação do outro, um

⁷ Ainda sobre o processo de *inferência causal* segundo David Hume, v., v.g.: GARRETT, Don – "Hume", in: BEEBEE, Helen; HITCHCOCK, Christopher; MENZIES, Peter (Eds.) – *The Oxford Handbook of Causation*. New York, Oxford University Press, 2009, pp. 75 ss..

⁸ MILL, John Stuart – *A System of Logic. Vol. 1*. London, John W. Parker, 1843, pp. 454 ss..

destes será causa ou efeito do outro, ou ambos estão ligados por um factor causal comum).

Há questões que ficam por resolver neste enunciado de métodos, nomeadamente o processo que conduz à selecção das circunstâncias. Nada se diz, directamente, sobre esse processo. Contudo, parece ter existido uma razão, facultada pelo próprio J. Stuart Mill, para tal facto⁹: *“Nothing can better show the absence of any scientific ground for the distinction between the cause of a phenomena and its conditions, than the capricious manner in which we select from among the conditions that which we choose to denominate the cause. However numerous the conditions may be, there is hardly any of them which may not, according to the purpose of our immediate discourse, obtain that nominal pre-eminence.”*

Este entendimento de que não há uma base científica para a selecção causal teve (e tem) um eco significativo em correntes e autores contemporâneos, como é o caso, por exemplo, de David Lewis¹⁰: *“We sometimes single out one among all the causes of some event and call it ‘the’ cause, as if there were no others. Or we single out a few as the ‘causes’, calling the rest mere ‘causal factors’ or ‘causal conditions’. Or we speak of the ‘decisive’ or ‘real’ or ‘principal’ cause. We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about. I have nothing to say about these principles of invidious discrimination. I’m concerned with the prior question of what it is to be one of the causes (unselectively speaking). My analysis is meant to capture a broad and nondiscriminatory concept of causation.”*

Não há, aqui, qualquer necessidade de desenvolver as disputas filosóficas a este respeito (a escolha é arbitrária, se-

⁹ *Idem, ibidem*, p. 401.

¹⁰ LEWIS, David – “Causation”, in: *The Journal of Philosophy*, 70 (17), 1973, pp. 558-9.

gundo Hume; a escolha não é arbitrária porque existe previsibilidade, segundo Hart e Honoré; a escolha não é arbitrária mas é *singularista*, segundo Ducasse; a escolha não é arbitrária porque há um conceito amplo, ainda que impreciso, de selecção que orienta qualquer inferência causal, segundo Hart, Honoré e Lewis; a escolha pode não ser arbitrária se se souber qual o *contexto conversacional* em que o indivíduo se situa, segundo Schaffer¹¹)¹², mas retenha-se, deste breve excurso pela causalidade, que as condições para a definição de *associações* (e, portanto, correlações) eram filosoficamente estudadas com antecedência relativamente aos avanços estatísticos e econométricos dos finais do séc. XIX.¹³

Retomando a *linhagem* matemática, o próximo nome que não pode deixar de ser aqui mencionado é o de Sir Francis Galton, primo do (não menos famoso) naturalista e geólogo Charles Darwin.

Embora se fale no contributo de Auguste Bravais – que Karl Pearson salientou, em 1895, ao dizer que “*the fundamental theorems of correlation were for the first time and almost exhaustively discusse by Bravais nearly half a century ago. He*

¹¹ Esta ideia parece ser tributária dos contributos para a filosofia da linguística e da semântica dados por J. L. Mackie (1974) e Robert Stalnaker (1978).

¹² Uma análise aprofundada das diversas correntes pode ler-se, por ex., em: SCHAFFER, Jonathan – “Contrastive causation”, in: *The Philosophical Review*, 114 (3), 2005, pp. 314 ss..

¹³ Entre as correntes contemporâneas sobre a causalidade, uma há que, não tendo sido citada no *corpus* deste artigo, pode traduzir alguns dos problemas que adiante serão identificados: a teoria da causalidade como instrumento, de R. G. Collingwood (1937-38), i.e., a ideia de que certas causas só são causas para aqueles que têm um interesse prático/concreto relativamente a certo tipo de fenómenos: “*a cause is an event or state of things which it is in our power to produce or prevent, and by producing or preventing which we can produce or prevent that whose cause it is said to be*” (p. 305) ou, ainda, “*only a person who is pratically concerned with a certain kind of event can form an opinion about its cause. For a mere spectator there are no causes*” (p. 307). Ver: COLLINGWOOD, R. G. – “On the so-called idea of causation”, in: *Proceedings of the Aristotelian Society*, Vol. 38, 1937-38, pp. 85-112, reprint in: MORRIS, Herbert (Ed.) – *Freedom and Responsibility: Readings in Philosophy and Law*. Stanford, Stanford University Press, 1961, pp. 303-312.

deals completely with the correlation of two and three variables. [Even the Galton's correlation coefficient] appears in Bravais' work, but a single symbol is not used for it." –, K. Pearson acabaria por reconhecer, mais tarde, que “*Bravais has no claim, whatever, to supplant Francis Galton as the discoverer of the correlational calculus*”, chegando mesmo a afirmar que “*as far as Gauss and Bravais are concerned we must, i think, hold that they contributed nothing of real importance to the problem of correlation*”.¹⁴

O início da moderna teoria matemática da correlação começa, assim, com as *Typical Laws of Heredity in Man*, de 1877. É este um dos *marcos fundadores* e assume, aqui, um particular relevo, uma vez que é também neste *marco* que se poderá traçar a origem do conceito de *reversão/regressão*, hoje em dia tão abundantemente usado nos mais diversos estudos acadêmicos.

Apesar de primeiramente aflorada no texto de 1877 e, mais tarde, num artigo de 1886¹⁵, a definição de “*co-relação*” acabaria por surgir, de forma clara – como se pode ver pelas próprias palavras de Sir Francis Galton –, num texto de 1888¹⁶: “*Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. Thus the length of the arm is said to be correlated with that of the leg, because a person with a long arm has usually a long leg, and conversely. If the correlation be close, then a person with a very long arm would usually have a very long leg; if it be moderately close, then the length of the leg would usually be only*

¹⁴ PEARSON, Karl – “Notes on the history of correlation”, in: *Biometrika*, 13 (1), 1920, pp. 28 e 32.

¹⁵ GALTON, Francis – “Regression towards mediocrity in hereditary stature”, in: *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 1886, pp. 246-263.

¹⁶ GALTON, Francis – “Co-relations and their measurement, chiefly from anthropometric data”, in: *Proceedings of the Royal Society of London*, 45, 1888, pp. 135-6.

long, not very long; and if there were no correlation at all then the length of the leg would on the average be mediocre. It is easy to see that correlation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the correlation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be nil. Between these two extremes are a number of intermediate cases, and it will be shown how the closeness of correlation in any particular case admits of being expressed by a simple number.”

Num excerto revelador, T. J. Barnes dá-nos conta da razão da influência decisiva de F. Galton, bem como da razão da desvalorização do trabalho de Auguste Bravais por Karl Pearson no texto de 1920. Refere T. J. Barnes¹⁷: *“There had been other attempts to devise measures of ‘co-relation’ before Galton, but none were very successful [...]. What made the difference were the background assumptions in which Galton framed his work, and found in eugenics. Karl Pearson, the first Professor of Statistics in Britain, wrote two quite different histories of correlation and regression analyses separated by some twenty-five years. The first in 1895 attributed correlation to [...] Auguste Bravais, a mid-19th-century error theorist, interested in assessing the accuracy of astronomical measurements. [...]. By 1920, however, Pearson [...] had changed his mind, arguing that it was Galton who was the true originator of correlation because the problem to which Bravais applied his work was completely different from the one Galton was trying to solve. Bravais divided his equation in order to get ride of the amount of statistical variation of error around the true values of the variables, whereas for Galton it was precisely the statistical variation – the error – that needed to be kept. Once explained,*

¹⁷ BARNES, T. J. – “A history of regression: actors, networks, machines and numbers”, in: *Environment and Planning A*, 30 (2), 1998, pp. 213-4.

the variation could be made the source of intellectual progress: specifically for Galton's purposes, geniuses could be bred."

Não se poderá descurar, portanto, e em resumo, a relevância do trabalho feito por percursores de Galton como Bravais e Gauss ou, recuando ainda um pouco mais, como Robert-Simon Laplace e Giovanni Plana. Mas, em seu favor, F. Galton foi capaz de demonstrar a aplicação prática de conceitos que, até então, teimavam em pairar num plano teórico, isolados de qualquer aplicação empiricamente relevante, tanto no campo das ciências naturais como no campo das ciências humanas.

Nesse sentido, e como bem afirma Brian Clauser¹⁸, não sendo um extraordinário matemático ou estatístico, F. Galton abriu, contudo, as portas a um novo mundo de aplicação experimental e prática de conceitos-chave em matéria de *associações* ou *co-relações*: “*Galton was not a great mathematical statistician; he made no important contributions to that field. [...] [But] Galton deserves credit for our use of such basic analytic frameworks as percentile rank, correlation, and regression. He was not the first to describe the mathematical relationship represented by the correlation coefficient, but he re-discovered this relationship and demonstrated its application in the study of heredity, anthropology, and psychology. [Furthermore] he is responsible for the term correlation (from correlation), he discovered the phenomenon of regression to the mean, and he is responsible for the choice of r [for reversion and now regression] to represent the correlation coefficient. [In one word:] Galton developed statistical applications for the behavioral sciences.*”

Esta apreciação, honesta e globalmente correcta, em nada diminui, bem pelo contrário, a relevância e o impacto de F. Galton na história moderna da correlação.

¹⁸ CLAUSER, Brian E. – “The life and labors of Francis Galton: a review of four recent books about the father of behavioral statistics”, in: *Journal of Educational and Behavioral Statistics*, 32 (4), 2007, p. 440.

Como sintetizam Dominique Mari e Samuel Kotz¹⁹, “*by now, over a century later, contemporary scientists often take the correlation coefficient for granted. It is not appreciated that before Galton and Pearson, the only means to establish a relationship between variables was to deduce a causative connexion. There was no way to discuss – let alone measure – the association between variables that lack a cause-effect relationship.*”

Outros contributos importantes, como os de Karl Pearson (1895), C. Spearman (1904) e Maurice Kendall (1938), nomes já referidos, serão mais claramente entendidos na secção seguinte, relativa às bases matemáticas das correlações. Antes, contudo, uma breve nota para explicar a possível forma como os contributos de Galton e Pearson conduziram a uma crescente utilização da correlação em estudos económicos, e logo nas primeiras décadas do séc. XX.

Embora sem certezas, até porque nomes como Adam Smith, David Hume e John Stuart Mill já davam relevo à análise da causa em matérias económicas, e economistas como Stanley Jevons (1863) e Francis Edgeworth (1893) já apontavam para a vantagem do uso de conceitos e métodos estatísticos²⁰, pode dizer-se que foi, precisamente, com textos como o já citado artigo de Warren M. Persons que se iniciou a popularização do *coeficiente de correlação* (também) na Ciência Económica²¹; e, numa fase subsequente, com o desenvolvimento

¹⁹ MARI, Dominique Drouet; KOTZ, Samuel – *Correlation and Dependence*. London, Imperial College Press, 2001, pp. 25-6.

²⁰ Sobre a evolução no recurso a técnicas estatísticas pelos economistas do séc. XIX, ver, por exemplo: MITCHELL, Wesley C. – *Business Cycles: The Problem and Its Setting*. New York, National Bureau of Economic Research, 1927, pp. 193 ss..

²¹ Poderá, ainda, acrescentar-se o seu artigo de 1916: “Construction of a business barometer based upon annual data”, in: *The American Economic Review*, 6 (4), 1916, pp. 739-769. Também se costuma invocar a importância dos *Elements of Statistics* (1901) de Arthur Lyon Bowley (protegido de Alfred Marshall), que nesse texto acolheu os contributos de K. Pearson, Udny Yule e F. Y. Edgeworth. A propósito, não deixa de ser curioso notar que Alfred Marshall, apesar do notório gosto pela matemática, tentou refrear, numa carta de 1906, o entusiasmo de Bowley (for-

dos estudos econométricos (principalmente a partir dos anos de 1930, com, e.g., Ragnar Frisch, Irving Fisher e Jan Tinbergen²²), altura em que ocorre a popularização de ferramentas estatísticas mais avançadas em estudos económicos.

A ligação a uma dimensão experimental e de *teste de modelos* pode, contudo, ter o condão de, em certos casos, obnubilar as diferenças entre correlação e causalidade... como sucedeu no recente, famoso e controverso caso envolvendo Reinhart e Rogoff.

Note-se, contudo, em abono da verdade, que aquela indistinação já se fazia sentir, embora por outras razões, há mais de cem anos. P. ex., em 1897, K. Pearson parece ter-se sentido atraído pela ideia de associar a noção de causalidade à correlação *perfeita*.²³

II – APONTAMENTOS SOBRE A BASE MATEMÁTICA DAS CORRELAÇÕES

Antes de passar à descrição sumária das definições téc-

mado em matemática), como recordou Arthur Cecil Pigou (em *Memorials of Alfred Marshall*. London, Macmillan, 1925, p. 427): “(1) Use mathematics as a shorthand language, rather than as an engine of inquiry. (2) Keep them till you have done. (3) Translate into English. (4) Then illustrate by examples that are important in real life. (5) Burn the mathematics. (6) If you can't succeed in 4, burn 3.”

²² Contudo, é necessário notar que o próprio papel da correlação sofreu alterações nessa trintena de anos. Como refere Michel Armatte, em “Le statut changeant de la corrélation en économétrie (1910-1944)”, in: *Revue Économique*, 52 (3), 2001, p. 626: “la période 1930-1944 correspond assez précisément à la superposition des deux paradigmes, c'est-à-dire à la lente érosion de la corrélation comme problématique de économétrie, et à la difficile émergence de la théorie des tests dans le champ économique.” Com a *Cowles Commission* e o manifesto de Haavelmo (1944), pode, de facto, dizer-se que “la corrélation n'a [...] pas disparue, mais elle a changé de statut. Elle ne caractérise plus statistiquement les relations entre variables observables, mais la loi théorique des erreurs inobservables. La corrélation ne se donne plus par une mesure, elle est devenue le paramètre d'une modèle théorique.” (p. 629).

²³ Ver: ALDRICH, John – “Correlations genuine and spurious in Pearson and Yule”, in: *Statistical Science*, 10 (4), 1995, p. 365.

nicas de covariância ou de correlação, convirá ter presente que o que está em causa nas correlações é a existência de *relações estatísticas*. Não se trata, portanto, de apurar *relações determinísticas*.

No mesmo sentido, veja-se, por ex., o seguinte excerto²⁴: “*O termo relação, por ser empregado em diversas acepções, carece de ser precisado. Quando se diz que duas variáveis estão relacionadas pode querer significar-se que a conexão é bem definida e invariável, ou pode ter-se em mente uma ligação mais indefinida e mais vaga. Por exemplo, como é sabido, o perímetro de uma circunferência, C, e o raio do círculo correspondente, R, estão relacionados. A relação que liga essas duas variáveis é definida e inalterável, e pode exprimir-se por meio de uma expressão matemática: a relação funcional ou função é $C = 2\pi R$. Neste caso, diz-se que se tem uma relação determinística entre as duas variáveis. [Mas] quando se refere a relação que existe entre o PIB e as importações, ou entre a taxa de juro e a massa monetária [...] tem-se em mente um tipo de relações mais imprecisas. Nestes casos não se tem funções, no sentido matemático, mas meras relações estatísticas. [...]. Fundamentalmente, é da variação em média que se ocupam as relações estatísticas. Entre duas variáveis ligadas por uma relação estatística diz-se que existe correlação. Indica-se, assim, que os fenómenos não estão indissolúvelmente ligados, mas sim que a intensidade de um é acompanhada tendencialmente pela intensidade do outro, no mesmo sentido ou em sentido inverso.*”

Descontando agora, por necessidade de simplificação, o *processo mental* (seja ele *físico* ou *metafísico*) de selecção de variáveis, com recurso às chamadas cláusulas *cæteris paribus*, uma primeira ferramenta para a determinação da correlação se impõe: a *covariância*, como medida da variação conjunta de

²⁴ MURTEIRA, Bento *et alii* – *Introdução à Estatística*. Lisboa, Escolar Editora, 2010, pp. 39-40.

duas variáveis aleatórias.²⁵ Assim sendo, temos:

$$\sigma(x, y) = E [(x - E[x])(y - E[y])]$$

em que x e y são as variáveis escolhidas e $E[x]$ o valor da *expectância* de x . Por sua vez, este valor, para uma variável aleatória discreta x , com valores possíveis x_1, x_2, x_3, \dots , pode ser representado e calculado da seguinte forma:

$$E[X] = \sum_{i=1}^{\infty} x_i p(x_i)$$

Se $\sigma(x, y) > 0$, haverá uma dependência directa (positiva) entre as duas variáveis, i.e., a valores elevados de x corresponderão valores elevados de y ; se $\sigma(x, y) = 0$, não existe relação linear entre as variáveis escolhidas; e se $\sigma(x, y) < 0$, há uma dependência inversa (negativa), i.e., a valores elevados de x corresponderão valores reduzidos de y .

No entanto, para medir o grau de associação linear entre variáveis, a covariância apresenta limitações evidentes. Bastará pensar-se que “*o valor que vai assumir depende não só da variação conjunta das variáveis como também dos valores absolutos que estas variáveis assumem.*”²⁶

Karl Pearson é, habitualmente, considerado o autor da

²⁵ A covariância amostral pode ser dada pela seguinte fórmula, sendo j e k as variáveis, N o número de observações das mesmas, e $E(x)$ a média da população:

$$q_{jk} = \frac{1}{N} \sum_{i=1}^N (x_{ij} - E(x_j))(x_{ik} - E(x_k))$$

Se $E(x)$ for desconhecida, então, assumindo-se \bar{x} como a média da amostra de um vector aleatório x , a estimativa de covariância será definida como:

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

²⁶ MURTEIRA, Bento *et alii* – *Introdução à Estatística*. Lisboa, Escolar Editora, 2010, p. 40.

formulação matemática da correlação, ou, de forma mais rigorosa, o autor do denominador, por uns, *coeficiente de correlação linear de Bravais-Pearson* e, por outros, *coeficiente de correlação produto-momento de Pearson*. O referido coeficiente é tradicionalmente apresentado do seguinte modo²⁷:

$$r_p = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

em que σ_{xy} designa a covariância entre as variáveis x e y , e $\sigma_x \sigma_y$ designam os *desvios-padrão*.

Por sua vez, estes correspondem à raiz quadrada da *variância* (i.e., da média dos quadrados das diferenças entre cada valor da variável e a média aritmética da distribuição) e poderão ser calculados da seguinte forma:

$$\sigma = \sqrt{E((X - E(X))^2)} = \sqrt{E(X^2) - (E(X))^2}$$

em que X é a variável aleatória e $E[x]$ o valor da esperança de X .²⁸

²⁷ O coeficiente de Pearson amostral (r) é um simples desdobramento do coeficiente *supra* apresentado e a forma pela qual costuma ser apresentado é a seguinte:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

²⁸ O desvio-padrão amostral (s_N) pode ser obtido através da seguinte fórmula, sendo $\{x_1, x_2, x_3, \dots, x_n\}$ os valores observados na amostra, N o tamanho da amostra, e \bar{x} a média dos valores observados:

$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Este desvio-padrão (dito *não-corrigido*) é um *estimador* consistente, sendo útil para amostras de grande dimensão, uma vez que para $N > 75$, a margem de enviesamento

O coeficiente só pode variar entre -1 e +1 (i.e., por outras palavras, a *distância de Pearson* está compreendida no intervalo [0, 2]). O valor +1 significa que a equação linear descreve perfeitamente a relação entre x e y , com todos os pontos observados a coincidirem com a linha através da qual se verifica que a aumentos de x correspondem aumentos de y ; o valor 0 significa que não há uma correlação linear entre as variáveis escolhidas; e o valor -1 significa que a equação linear descreve perfeitamente a relação entre x e y , com todos os pontos observados a coincidirem com a linha através da qual se verifica que a aumentos de x correspondem diminuições de y .²⁹

Interpretações de valores intermédios, nos intervalos]-1, 0[e]0, +1[, devem ser feitas com cautela, em especial quando se aplica este coeficiente a estudos na área das ciências *sociais* (dados os factores exógenos não-observáveis ou de difícil mensuração).

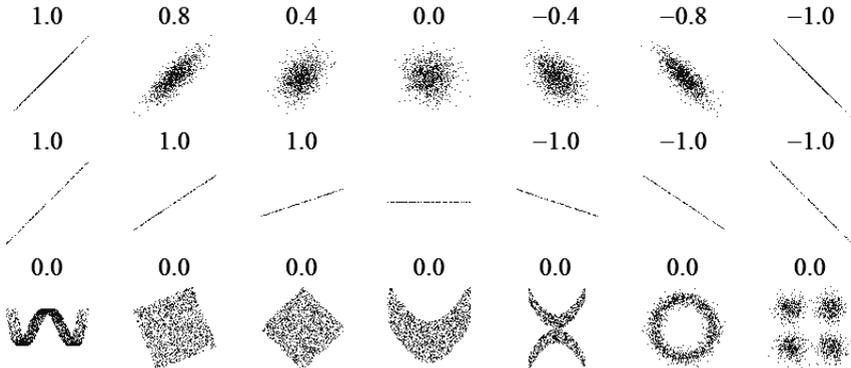
A projecção gráfica de alguns dos valores que se poderão obter pelo coeficiente de Pearson surge resumida no seguinte quadro de diagramas de dispersão³⁰:

é inferior a 1%. Se o enviesamento for corrigido (ainda que não eliminado), pode apresentar-se pela seguinte fórmula:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

²⁹ Duas notas: 1) os valores -1 e +1 correspondem, respectivamente, a correlações negativas e positivas *perfeitas*; e 2) o valor 0 não significa que as variáveis escolhidas sejam independentes entre si, dado que podem existir associações não-lineares entre as mesmas. A este respeito, ver, v.g.: GUIMARÃES, Rui Campos; CABRAL, José Sarsfield – *Estatística*. Lisboa, Verlag Dashöfer, 2.ª Edição, 2011, p. 105.

³⁰ Convém notar que o diagrama central não tem um valor atribuído porque, sendo a variância de valores observados no eixo Y igual a zero, o coeficiente de correlação fica indefinido.



Para lá de lineares (e não-lineares ou curvilíneas³¹), ou de positivas, negativas e de valor zero, as correlações podem também ser classificadas como *simples*, *múltiplas* ou *parciais*. Assim: as correlações *simples* são aquelas em que apenas duas variáveis são consideradas para efeito de cálculo; já as correlações *múltiplas* são aquelas em que uma variável é relacionada, simultaneamente, com um conjunto de outras variáveis; e as correlações *parciais* são aquelas em que apenas duas variáveis são consideradas (como nas correlações simples), mas presumindo-se a existência de outras variáveis constantes durante o cálculo.

Para além do *canónico* coeficiente de Pearson, outros coeficientes de correlação que podem ser apontados são, por ex., os seguintes³²:

A) O coeficiente de correlação intra-classe, desenvolvi-

³¹ Sobre os coeficientes de correlação não-lineares, ver, por exemplo: WANG, Ting; ZHANG, Shiqiang – “Study on linear correlation coefficient and nonlinear correlation coefficient in mathematical statistics”, in: *Studies in Mathematical Sciences*, 3 (1), 2011, pp. 58-63.

³² Há uma natural simplificação, porque não se mostra necessário, aqui, um aprofundamento. No entanto, desenvolvimentos destes coeficientes, bem como a ilustração de outros – como, por ex., o coeficiente de correlação mediana de Nils Blomqvist ou “Blomqvist β ” (1950), o coeficiente de correlação de Leo Goodman e William Kruskal ou “Goodman-Kruskal G ” (1954 a 1972), ou o coeficiente de correlação absoluto de Christopher Bradley (1985) – podem ser encontrados em textos avançados de Estatística.

do por Ronald Fisher em 1954, e que, ao contrário da correlação (inter-classe) de Pearson, visa demonstrar como unidades de um mesmo grupo se assemelham às unidades de outro grupo (por ex., por razões relativas a hereditariedade). Assim, v.g., no caso de estarem em causa pares de dois valores ($x_{n,1}$, $x_{n,2}$), a média \bar{X} do grupo e o coeficiente r de Ronald Fisher serão definidos, respectivamente, por:

$$\bar{x} = \frac{1}{2N} \sum_{n=1}^N (x_{n,1} + x_{n,2})$$

$$r = \frac{1}{N s^2} \sum_{n=1}^N (x_{n,1} - \bar{x})(x_{n,2} - \bar{x})$$

B) Os coeficientes de correlação de *ranking* ou de correlação ordinal, que visam detectar e medir o grau de similaridade entre duas ordenações da mesma variável ou de duas variáveis, e de que são exemplos famosos o coeficiente de correlação de Charles Spearman ou “Spearman ρ ” (1904), e o coeficiente de correlação de Maurice Kendall ou “Kendall τ ” (1938). Vejamos, brevemente, o que caracteriza cada um deles.

O coeficiente de Spearman pode definir-se como uma medida não-paramétrica de dependência estatística entre duas variáveis. Assim sendo, neste coeficiente estão em causa funções monotónicas (e não lineares, como sucedia no coeficiente de Pearson). O intervalo de resultados possíveis é, contudo, igual $[-1, +1]$ e a interpretação dos mesmos idêntica. No fundo, trata-se simplesmente de um coeficiente de Pearson aplicado entre variáveis ordenadas³³. A fórmula é a seguinte:

³³ Ver, por exemplo: FIELLER, E. C.; HARTLEY, H. O.; PEARSON, E. S. – “Tests for rank correlation coefficients. I.”, in: *Biometrika*, 44 (3/4), 1957, p. 470.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

em que d_i corresponde às diferenças entre x_i e y_i entre as ordens ou *rankings* de cada observação de duas variáveis x e y , e n o número de pares de *rankings* associados.

O coeficiente de correlação de M. Kendall (“Kendall τ ”) também é uma medida não-paramétrica. Contudo, se o “Spearman ρ ” era, como se disse, um coeficiente de Pearson para variáveis ordenadas, o τ de Kendall representa, antes, uma probabilidade³⁴. Para esse efeito, a equação de Maurice Kendall faz uso da noção de *concordância* (entre pares associados):

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

No denominador está representado o total de pares de combinações. Os pares são *concordantes* se, suponha-se, os *rankings* para ambos os elementos dos pares (X_1, Y_1) e (X_2, Y_2) forem na mesma *direcção* (1). São *discordantes* se forem em *direcção* oposta (2). Colocando em *função sinal*:

- (1) $\text{sgn}(X_2 - X_1) = \text{sgn}(Y_2 - Y_1)$
- (2) $\text{sgn}(X_2 - X_1) = -\text{sgn}(Y_2 - Y_1)$

A finalizar esta parte, justifica-se, agora, uma definição

³⁴ Como salienta Mirella Lapata, em “Automatic evaluation of information ordering”, in: *Computational Linguistics*, 32 (4), 2006, p. 474: “Kendall’s τ can be interpreted as a simple function of the probability of observing concordant and discordant pairs (Kerridge 1975). In other words, it is the difference between the probability that in the observed data two variables are in the same order versus the probability that they are in different orders (the probability is rescaled to range from -1 to 1 as is customary for correlation)”.

do conceito de regressão e uma breve exposição de alguns modelos de regressão.

Podemos definir-se regressão linear (ou simples) como o método matemático por via do qual se modela a relação entre uma variável dependente Y e outra independente (já se for entre aquela e um conjunto de outras variáveis independentes, denomina-se análise de regressão linear múltipla³⁵). Representa-se do seguinte modo:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

em que Y_t é a variável dependente (ou explicada/*regressando*), X_1, X_2, \dots, X_p as variáveis independentes (ou explicativas/*regressores*), $\beta_1, \beta_2, \dots, \beta_p$ os parâmetros (que medem a influência das variáveis independentes sobre o *regressando*), β_0 a *intersecção* (uma constante que representa a intersecção da recta de regressão com o eixo vertical), ε a perturbação aleatória, e p o número de parâmetros independentes considerados.

Num modelo de regressão linear com uma só variável explicativa (X_i), só há dois parâmetros e o hiperplano corresponderá a uma recta definida pela seguinte fórmula:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

em que ε_i é o erro associado à medição de X_i , e considerando-se preenchidos os pressupostos de *normalidade* da *per-*

³⁵ A regressão linear múltipla expressa-se através da seguinte fórmula:

$$Y_i = \beta_0 + \sum \beta_i X_{ip} + \varepsilon_i$$

De notar que a regressão pode ser não-linear. Nesse caso, com aproximação a uma função linear (o que introduz enviesamento estatístico e justifica, portanto, uma ainda maior cautela interpretativa):

$$f(x_i, \beta) \approx f^0 + \sum_j J_{ij} \beta_j$$

turbação ε (ou seja: erros normalmente distribuídos e mutuamente independentes, variância constante ou *homoscedasticidade*, e expectância igual a zero³⁶), i.e.: $\varepsilon_i \sim N(0, \sigma^2)$.

As *rectas de regressão* são as que melhor se ajustam às *nuvens de pontos* (ou diagramas de dispersão) gerados por uma distribuição binomial. Por seu lado, as *linhas de tendência* são uma das decorrências práticas comuns da análise de regressão linear.

A cautela com a validação dos resultados obtidos (com exercícios de validade e análise de *resíduos*) e com a interpretação dos mesmos não deve ser descuidada. A título de aviso, vejam-se estes quatro casos distintos que, contudo, partilham a mesma *linha*³⁷:

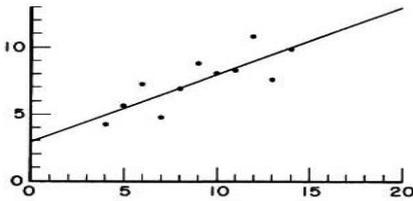


Figure 1

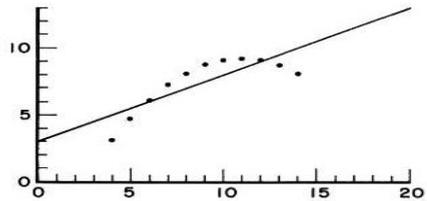


Figure 2

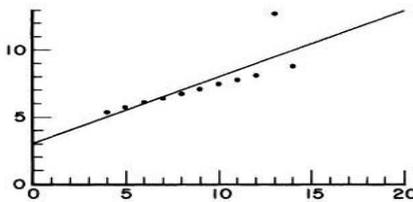


Figure 3

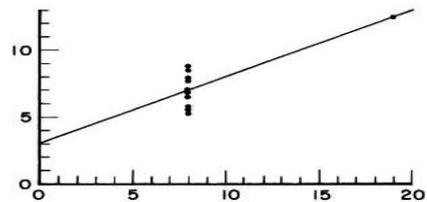


Figure 4

³⁶ Ver, e.g.: MURTEIRA, Bento *et alii* – *Introdução à Estatística*. Lisboa, Escolar Editora, 2010, p. 348.

³⁷ ANSCOMBE, Francis J. – “Graphs in statistical analysis”, in: *The American Statistician*, 27 (1), 1973, pp. 19-20.

Sendo a previsão (para além do teste de hipóteses) um dos objectivos essenciais da análise de regressão, justifica-se terminar esta breve nota técnica com uma referência ao *coeficiente de determinação* (R^2), i.e., com a definição de uma medida que permite perceber se a relação linear entre y e x é suficientemente forte para que a *recta* estimada possa permitir a previsão de valores de y .³⁸ Assim, tendo em consideração que y_i são os valores observados, a média (\bar{y}) das observações será definida por:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Sabendo-se o valor desta média, e sabendo-se, ainda, que f_i (ou \hat{y}) são os valores previstos, SS_{tot} corresponde ao somatório dos quadrados das diferenças entre a variável dependente (y_i) e a média, e SS_{res} é o somatório dos quadrados das estimativas de ε_i :

$$SS_{tot} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_{i=1}^n (\varepsilon_i)^2$$

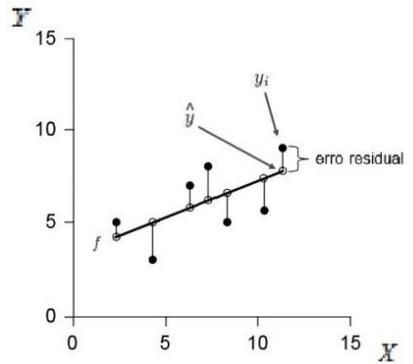
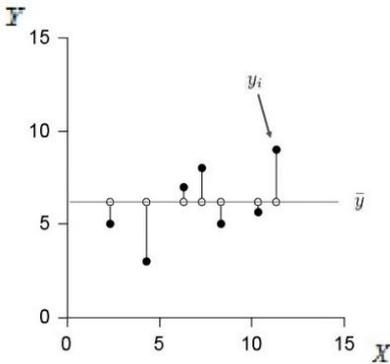
A partir destes obtém-se a definição algébrica do *coeficiente de determinação*³⁹:

³⁸ Ver, por ex.: PINTO, José Castro; CURTO, José Dias – *Estatística para Economia e Gestão*. Lisboa, Edições Sílabo, 2.^a Ed., 2010, pp. 414-6. Se o coeficiente atingir o valor máximo (1), tal significa que a variação total de y pode ser totalmente explicada pelas variações de x .

³⁹ Sobre este coeficiente ver, também, por ex.: RODRÍGUEZ, Elena Martínez – “Errores frecuentes en la interpretación del coeficiente de determinación lineal”, in: *Anuario Jurídico y Económico Escurialense*, 38, 2005, pp. 318 e ss.; e DEVORE, Jay L.; BERK, Kenneth N. – *Modern Mathematical Statistics with Applications*.

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$$

Quanto mais próximos da linha de regressão linear estiverem os valores, mais próximo estará R^2 de 1 (os resultados deste *coeficiente* podem situar-se no intervalo $[0, 1]$)⁴⁰. *Infra* podem ver-se, graficamente, à esquerda, a distância dos valores observados relativamente à média (\bar{y}) e, à direita, os *erros residuais* face a uma regressão linear f :



III – RISCOS E LIMITES DAS CORRELAÇÕES PARA A ECONOMIA

Na parte inicial deste texto, relativa ao excuro histórico sobre as correlações e a causalidade, deixou-se subentendido o abuso, intencional ou inadvertido, em estabelecer nexos causais a partir de correlações (mesmo que correctamente calculadas).

Há uma infinidade de problemas (para lá dos meramen-

Belmont, CA, Thomson, 2007, pp. 619 e ss..

⁴⁰ Valores negativos podem ocorrer em caso de aplicação do coeficiente a funções não-lineares. De notar que valores elevados de R^2 não significam que o modelo de regressão tem utilidade preditiva.

te técnicos) relacionados com a interpretação das correlações. Nesse sentido, procurar-se-á fazer aqui um resumo de alguns dos mais significativos problemas interpretativos.

Sistematizando, serão tratadas, de forma sintética, as seguintes questões: 1) a da distinção entre *relação determinística* e *estatística*; 2) a do *problema da identificação*; 3) a da correlação sem causa, da causa sem correlação e da correlação sem linearidade; 4) a do papel do factor tempo na correlação, na causalidade e, de um modo mais amplo, na própria Ciência Económica; 5) a dos riscos práticos de enunciar “*falsos positivos*”.

Quanto à primeira das questões, relativa à distinção entre *relação determinística* e *estatística*, já se referiu, anteriormente, que não deve haver confusão entre ambas, sob pena de estarmos a tentar extrair dos números aquilo que eles não nos podem dar: uma explicação causal de fenómenos económicos e sociais complexos, que são mais densos do que quaisquer fórmulas, funções ou outra *parafernália* matemática. Para a Ciência Económica, a Estatística e a Matemática são úteis se forem utilizadas com parcimónia e honestidade, e se conjugadas, sempre que necessário, com ciências como a História, o Direito, a Filosofia, a Geografia ou a Sociologia. Utilizar a Estatística ou a Matemática como um alicerce elaborado para veicular uma mera “crença” (seja ela admitida ou não) é pouco ou nada do domínio da ciência (seja ela ciência *exacta* ou *social*) e muito mais do domínio do “*charlatanismo pseudo-científico*”.⁴¹

A Ciência Económica, para se afirmar, de pleno direito, como ciência *social*, não precisa de determinismos. Precisa, antes, de seriedade e humildade metodológica. Nesse sentido – e como já avisavam, entre outros, Alfred Marshall ou L. Walras⁴² (insuspeitos, dada a receptividade que mostraram ao de-

⁴¹ Ver: ARAÚJO, Fernando – *Introdução à Economia*. Coimbra, Almedina, 3.ª Ed., 2005, pp. 95 ss..

⁴² Este último, aliás, com ataques pessoais a F. Edgeworth, o que demonstra a sua animosidade para com certas “*fantasias [matemáticas] estereis*” (como lhes chamou

envolvimento do *método matemático*) –, a Matemática ou a Estatística são, quando muito, *ferramentas*, nunca *chaves explicativas*. Aliás, múltiplos investigadores contemporâneos têm vindo a salientar, quer no contexto da Ciência Económica, quer no contexto de outras ciências *sociais*, que a etapa de um certo “deslumbramento” com a aplicação de (alegadamente) sofisticado instrumentário das (ditas) *ciências exactas* corresponde a uma vaga que já faz parte do passado.⁴³

Apesar do alerta, parece que aquele “deslumbramento” continua, por infortúnio, a “fazer escola” em algumas Faculdades de Economia pelo mundo inteiro e em muitos eixos de poder (político e económico), com as consequências que se adivinham (e, por vezes, se sentem). Esta pode ser uma das razões para a quase indiscutida aceitação de certas teses, como sucedeu, por algum tempo, com a citada tese de Reinhart e Rogoff⁴⁴.

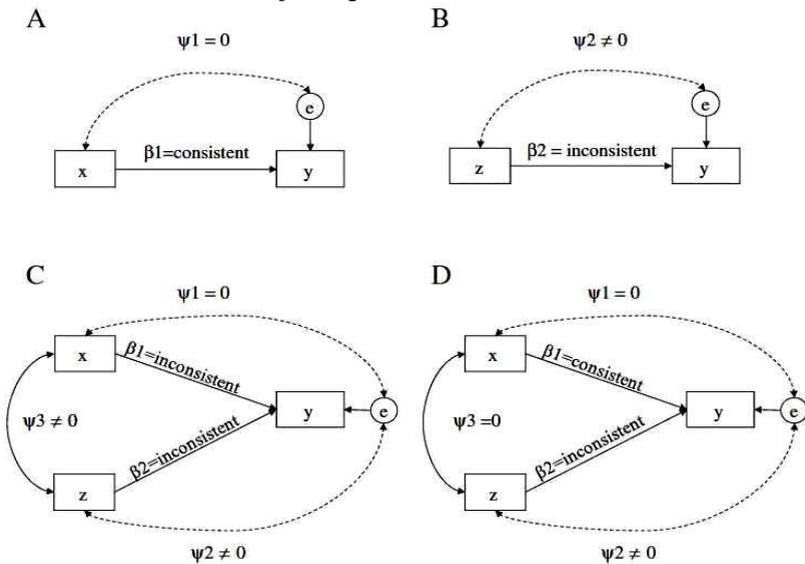
Referiu-se, há pouco, o *problema da identificação*, que pode ser sinteticamente definido como um problema geral de

Léon Walras em duas cartas dirigidas a L. Bortkiewicz em 1889) que poderiam desacreditar o uso ponderado da Matemática na Economia. Como refere, a este propósito, Donald Walker, em “Edgeworth versus Walras on the theory of tatonnement”, in: *Eastern Economic Journal*, 13 (2), 1987, p. 155: “Walras had hoped, he confided to Charles Gide, that the mathematical method in economics would eliminate the posturing and charlatanism Edgeworth displayed (Walras to Gide, November 3, 1889, L 933). Walras thought him a humbug and a puffist who was quite capable of discrediting the new method of mathematical economics (Walras to Gide, April 11, 1891, L 1000)”.

⁴³ Veja-se, por exemplo: BOETTKE, Peter J.; LEESON, Peter T.; SMITH, Daniel J. – “The evolution of economics: where we are and how we got here”, in: *The Long Term View*, 7 (1), 2008, p. 15. No mesmo sentido, observa Jean Tirole, conhecido economista da chamada *escola de Toulouse* (em entrevista à revista *Sciences et Avenir*, n.º 776, de Outubro de 2011, p. 52): “[L’Économie] est incontestablement une science sociale, au même titre que la psychologie, la sociologie, le droit, la science politique... pour la bonne raison qu’elle décrit des comportements humains. Jusqu’à la fin du XIX^e siècle, l’économie était très ouverte [...]. [...] depuis dix à vingt ans, on observe un retour aux origines.”

⁴⁴ Pelo que se sabe, não houve, sequer, *peer review* do famoso artigo de 2010 publicado na *American Economic Review*. Terá sido um mero lapsos ou acreditou-se que o texto era à prova de crítica?

detecção e de selecção de variáveis explicativas de um fenómeno.⁴⁵ Ligado a este, um outro problema deve também ser considerado: o da *endogeneidade* ou, dito de outro modo, o problema da existência de variáveis omitidas sem as quais não é consistente explicar ou sequer enunciar efeitos de uma variável sobre outra(s). Como ilustram, de forma gráfica, John Antonakis *et alii*⁴⁶, a *endogeneidade* de uma variável pode retirar a *consistência* às relações que tenham sido estabelecidas:



⁴⁵ Charles F. Manski, em “Identification problems in the social sciences and everyday life”, in: *Southern Economic Journal*, 70 (1), 2003, p. 11, tem uma definição vulgar mas bem elucidativa deste problema: “Suppose that you observe the almost simultaneous movements of a person and of his image in a mirror. Does the mirror image cause the person’s movements, does the image reflect the person’s movements, or do the person and image move together in response to a common external stimulus? Empirical observations alone cannot answer this question. Even if you were able to observe innumerable instances in which persons and their mirror images move together, you would not be able to logically deduce the process at work. To reach a conclusion requires that you understand something of optics and of human behavior.”

⁴⁶ ANTONAKIS, John *et alii* – “On making causal claims: a review and recommendations”, in: *The Leadership Quarterly*, 21 (6), 2010, p. 1090.

Neste exemplo gráfico, pode perceber-se que a consistência da relação é definida pela existência de *endogeneidade* (representada aqui pela variável “*e*”). Assim, no caso A, “*e*” não se correlaciona com *x* ($\psi_1 = 0$), pelo que β_1 é consistente; no caso B, ocorre precisamente o contrário; no caso C, embora *x* seja variável exógena, a relação entre *x* e *y* é inconsistente porque *x* sofre, indirectamente, do enviesamento gerado pelas relações de *z* com *x* ($\psi_3 \neq 0$) e com “*e*” ($\psi_2 \neq 0$); finalmente, no caso D, a relação entre *x* e *y* é consistente dada a ausência de correlação entre *x* e *z* (mesmo que esta última variável se correlacione com “*e*” e, por essa razão, tal possa afectar, indirectamente, a consistência referida).

Note-se, contudo, que este é apenas um de múltiplos outros problemas, que bem são sistematizados por Fernando Araújo⁴⁷: “*como é que a identificação de uma variável causal autoriza a abstracção da sua condição de variável causada? com que rigor se atribui estatuto causador a uma variável, se vem a verificar-se que ela é inteiramente dependente [ou seja, endógena], nas suas manifestações, de uma terceira variável que foi omitida? como é que se adverte a hipótese de ambas as variáveis serem resultados de uma terceira, e não causa uma da outra? como é que se representa, num universo de duas variáveis apenas [ou mesmo de múltiplas variáveis, recorrendo, para o efeito, a instrumentos estatísticos mais ou menos sofisticados], um nexos causal que resulta da combinação [sob forma incompreendida ou dificilmente compreensível] de um grande número de variáveis?*”

Acrescente-se que o *problema da identificação* pode ainda revestir dificuldades temporais. Vejam-se, a título de exemplo, os seguintes riscos: supor-se que a sequência temporal entre fenómenos determina uma relação de tipo causa-efeito (sofisma *post hoc ergo propter hoc*), ou que a simples sobrepo-

⁴⁷ ARAÚJO, Fernando – *Introdução à Economia*. Coimbra, Almedina, 3.ª Ed., 2005, pp. 75-6.

sição temporal de dois fenómenos permite inferir a existência de ligação entre os mesmos (sofisma *cum hoc ergo propter hoc*).

Perante tamanhas dificuldades, e não desvalorizando o papel que estatísticas ou matemáticas (mais ou menos avançadas) podem ter na eliminação ou, pelo menos, na diminuição de alguns dos riscos enunciados, chega-se a uma conclusão que gera alguma insegurança: a escolha de variáveis (e a respectiva formalização) tem, na sua base – para além de uma margem de aleatoriedade associada –, razões insondáveis para os próprios investigadores (quando estas não são cínicas...).⁴⁸ Acresce que a realização de “testes de fiabilidade” sobre as variáveis escolhidas permite, somente, se for bem feita, atestar que tais variáveis não podiam deixar de ser escolhidas... mas pouco mais para lá disto.⁴⁹

⁴⁸ D. McCloskey, no seu estilo habitualmente provocatório, foi mais longe na denúncia das (potenciais) razões subjacentes a uma sempre crescente e, em muitos casos, já excessiva formalização: “*economists think that science involves axiomatic proofs of theorems and then econometric tests of the QED, which therefore will test the axioms. In truth, the physicists could care less about mathematical proofs. Even the theoreticians in physics spend much of their time reading the physical equivalent of agricultural economists or economic historians. Pencil-and-paper guys are uncommon in physics departments. [...] The economists, to put it another way, have adopted the intellectual values of the Math Department - not the values of the Department of Physics [...] they admire from afar. [...] the large scale of formalization in economics [...] seems mean spirited. It probably is. I feel mean about it. But someone has to say it, because everyone knows that it is true: a dominant coalition of the formalizers are not themselves tolerant of science. It is an open secret that they want economics to become a branch of the Math Department. [...] One economics department after another has been seized by the formalists and marched off to a Gulag of hyperspace searching.*” [Vd. “Economics science: a search through the hyperspace of assumptions?”, in: *Methodus*, 3 (1), 1991, pp. 8 e 14.]

⁴⁹ Com efeito, este ainda frequente recurso à formalização poderá, paradoxalmente, denunciar uma certa “preguiça mental”, uma tentativa de simplificação abstracta que permita evitar um “confronto” denso e prolongado com os dados que se podem extrair da realidade. Transparecendo o receio desse confronto, veja-se, e.g., o seguinte excerto de um artigo de D. W. Katzner, precisamente em resposta a McCloskey: “[*parafraseando Granger*] «if the axiomatizations proposed by ... the economist are of any use for the progress of [economic] science, it is not because they appear to ape the constructions of mathematics. It is because they offer to rational

Especificamente no que diz respeito à correlação, é necessário ter em conta as várias falácias interpretativas que a podem acompanhar. Nesse sentido, foram nomeadas três, aliás já indirectamente analisadas, que permitem entender como é delicada a tarefa de um economista quando assenta o seu discurso em correlações.

Com efeito, ainda que exista relação causal entre variáveis correlacionadas, não é condição para haver uma relação causal a existência de uma correlação (seja ela *forte* ou *menos forte*). Por outro lado, se através do cálculo do valor da correlação se concluir que as variáveis escolhidas são independentes, tal também não impede que as mesmas possam estar causalmente ligadas por variáveis simplesmente ignoradas.⁵⁰ De facto, os casos de causa sem correlação directa podem existir (presença de *variáveis intermédias* entre as *variáveis principais*, com as quais se estabelecem correlações *fortes*), pelo que, dada a provável existência de *terceiras variáveis* (com efeitos de *mediação*, *moderação* ou *supressão*), a questão decisiva impõe-se: como evitar correlações *espúrias*?⁵¹

thought the sole means of escaping from the attractions of data derived from experience» and one of the few ways of extracting positive structures from the complex and confounding jumble of reality.” [Vd. “In defense of formalization in economics”, in: *Methodus*, 3 (1), 1991, p. 22.]

⁵⁰ Pela mesma razão (e também pelo facto de poderem haver causas que operam em simultâneo com os respectivos efeitos, ou ainda causas que não produzem alterações visíveis), também se mostra duvidosa a aplicabilidade geral da denominada “causalidade de Wiener-Granger” (1969), i.e., a extracção de uma ideia de causalidade a partir do facto de se observar que o comportamento actual/passado de uma série temporal permite prever o comportamento de outra série temporal (unidireccionalidade da causa). Não há dúvida de que o próprio Clive Granger reconheceu as limitações da “*G-causality*” (propondo-a, apenas, como um possível teste empírico em contexto de modelos de regressão linear), mas pode ser que nem todos os economistas as reconheçam... Sobre a referida causalidade e as suas aplicações (que vão muito além dos domínios económicos), ver, por exemplo: GUO, Shuixia; LA-DROUE, Christophe; FENG, Jianfeng – “Granger causality: theory and applications”, in: FENG, Jianfeng; FU, Wenjiang; SUN, Fengzhu (Eds.) – *Frontiers in Computational and Systems Biology*. London, Springer, 2010, pp. 83-111.

⁵¹ Pode recorrer-se a correlações *parciais* para aumentar o grau de certeza quanto à associação entre duas (de mais) variáveis, mas não para impedir *erros de especifica-*

Quanto à existência de correlação sem linearidade, a mesma ficou evidente pelos gráficos de Francis Anscombe, já *supra* reproduzidos. Esse exemplo permite perceber que a linearidade entre variáveis não se obtém através de um coeficiente de correlação e que a observação gráfica (para lá da observação directa) é imprescindível para perceber a força e características da relação subjacente ao valor da correlação.

Outra limitação que não poderá ser ignorada prende-se com a consideração do papel e da complexidade do factor temporal na correlação, no eventual estabelecimento de nexos causais, e até num contexto mais amplo de fidelidade do *discurso* económico à realidade detectada e detectável.

Como salienta, a esse respeito, Francisco Louçã⁵², “*depois de um século de auge do paradigma do tempo linear e das trajectórias dos acontecimentos num mundo completamente determinista, o positivismo refugiou-se numa reconstrução mecânica da noção do tempo cíclico e repetitivo. [...] A Economia, que se queria considerar a rainha das ciências sociais, procurou assim estabelecer as condições necessárias para a indução e para a refutação, imitando os preceitos da física. No entanto, os guardiões da ortodoxia limitavam-se de facto a imitar as técnicas e os pressupostos da física energética de meados do século XIX, ignorando simultaneamente os desenvolvimentos. Assim, [...] foi-se aceitando a estática como realização da dinâmica e os processos estacionários como a aproximação adequada aos processos evolucionistas, isto é, que o tempo era uma variável redutível. [...] [Só que] o tempo não*

ção (resultantes de “*omitted-variable bias*” ou *OVB*). Ainda a propósito da causalidade sem correlação, das causas que não provocam alterações, das *cadeias causais* e das *causas partilhadas*, ver: SHUGAN, Steven M. – “Causality, unintended consequences and deducing shared causes”, in: *Marketing Science*, 26 (6), 2007, pp. 731-741.

⁵² LOUÇÃ, Francisco – “A flecha e o alvo do tempo - paradoxos económicos”, in: *Revista Crítica de Ciências Sociais*, 46, 1996, pp. 136, 144 e 152. No mesmo sentido, leia-se, ainda: LOUÇÃ, Francisco – *Turbulência na Economia*. Lisboa, ISEG-UTL, 1996, pp. 32 ss. (polic.).

pode ser representado naturalmente por uma variável endógena [...] nem tão-pouco pode ser representado por uma variável exógena: mesmo admitindo que a tendência estatística de longo prazo medisse a cronologia do sistema, existiria sempre uma estrutura prevalecte em todas as variáveis, que é a sua dimensão temporal irreduzível. O tempo é, portanto, parte, senão mesmo a parte constitutiva essencial da complexidade. [...] esta análise da complexidade é simultaneamente mais completa e mais prudente do que a alternativa imposta pelo paradigma positivista.”

A terminar, deve recordar-se que o risco de “*falsos positivos*” não é raro nem, muito menos, inócuo: bastará aqui lembrar o impacto da *supra* citada tese de Reinhart e Rogoff para se ter uma ideia das consequências reais que os referidos “*falsos positivos*” podem ter na Economia (e nas Leis) dos Estados.



REFERÊNCIAS BIBLIOGRÁFICAS

- ALDRICH, John – “Correlations genuine and spurious in Pearson and Yule”, in: *Statistical Science*, 10 (4), 1995.
- ANSCOMBE, Francis J. – “Graphs in statistical analysis”, in: *The American Statistician*, 27 (1), 1973.
- ANTONAKIS, John *et alii* – “On making causal claims: a review and recommendations”, in: *The Leadership Quarterly*, 21 (6), 2010.
- ARAÚJO, Fernando – *Introdução à Economia*. Coimbra, Almedina, 3.^a Ed., 2005.
- ARMATTE, Michel – “Le statut changeant de la corrélation en économétrie (1910-1944)”, in: *Revue Économique*, 52 (3), 2001.

- BARNES, T. J. – “A history of regression: actors, networks, machines and numbers”, in: *Environment and Planning A*, 30 (2), 1998.
- BOETTKE, Peter J.; LEESON, Peter T.; SMITH, Daniel J. – “The evolution of economics: where we are and how we got here”, in: *The Long Term View*, 7 (1), 2008.
- CLAUSER, Brian E. – “The life and labors of Francis Galton: a review of four recent books about the father of behavioral statistics”, in: *Journal of Educational and Behavioral Statistics*, 32 (4), 2007.
- COLLINGWOOD, R. G. – “On the so-called idea of causation”, in: *Proceedings of the Aristotelian Society*, Vol. 38, 1937-38.
- DEVORE, Jay L.; BERK, Kenneth N. – *Modern Mathematical Statistics with Applications*. Belmont, CA, Thomson, 2007.
- DRAKOPOULOS, Stravros A.; TORRANCE, Thomas S. – “Causality and determinism in economics”, in: *Scottish Journal of Political Economy*, 41 (2), 1994.
- FIELLER, E. C.; HARTLEY, H. O.; PEARSON, E. S. – “Tests for rank correlation coefficients. I.”, in: *Biometrika*, 44 (3/4), 1957.
- GALTON, Francis – “Co-relations and their measurement, chiefly from anthropometric data”, in: *Proceedings of the Royal Society of London*, 45, 1888.
- GALTON, Francis – “Regression towards mediocrity in hereditary stature”, in: *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 1886.
- GARRETT, Don – “Hume”, in: BEEBEE, Helen; HITCHCOCK, Christopher; MENZIES, Peter (Eds.) – *The Oxford Handbook of Causation*. New York, Oxford University Press, 2009.
- GUIMARÃES, Rui Campos; CABRAL, José Sarsfield – *Estadística*. Lisboa, Verlag Dashöfer, 2.^a Edição, 2011.

- GUO, Shuixia; LADROUE, Christophe; FENG, Jianfeng – “Granger causality: theory and applications”, in: FENG, Jianfeng; FU, Wenjiang; SUN, Fengzhu (Eds.) – *Frontiers in Computational and Systems Biology*. London, Springer, 2010.
- HOOVER, Kevin D. – *Causality in Macroeconomics*. New York, Cambridge University Press, 2001.
- KATZNER, D. W. – “In defense of formalization in economics”, in: *Methodus*, 3 (1), 1991.
- LAPATA, Mirella – “Automatic evaluation of information ordering”, in: *Computational Linguistics*, 32 (4), 2006.
- LEWIS, David – “Causation”, in: *The Journal of Philosophy*, 70 (17), 1973.
- LOUÇÃ, Francisco – “A flecha e o alvo do tempo - paradoxos económicos”, in: *Revista Crítica de Ciências Sociais*, 46, 1996.
- LOUÇÃ, Francisco – *Turbulência na Economia*. Lisboa, IS-EG-UTL, 1996.
- MANSKI, Charles F. – “Identification problems in the social sciences and everyday life”, in: *Southern Economic Journal*, 70 (1), 2003.
- MARI, Dominique Drouet; KOTZ, Samuel – *Correlation and Dependence*. London, Imperial College Press, 2001.
- McCLOSKEY, D. – “Economics science: a search through the hyperspace of assumptions?”, in: *Methodus*, 3 (1), 1991.
- MILL, John Stuart – *A System of Logic. Vol. 1*. London, John W. Parker, 1843.
- MITCHELL, Wesley C. – *Business Cycles: The Problem and Its Setting*. New York, National Bureau of Economic Research, 1927.
- MURTEIRA, Bento *et alii* – *Introdução à Estatística*. Lisboa, Escolar Editora, 2010.
- PEARSON, Karl – “Notes on the history of correlation”, in:

- Biometrika*, 13 (1), 1920.
- PERSONS, Warren M. – “Construction of a business barometer based upon annual data”, in: *The American Economic Review*, 6 (4), 1916.
- PERSONS, Warren M. – “The correlation of economic statistics”, in: *American Statistical Association*, 12 (92), 1910.
- PIGOU, Arthur Cecil – *Memorials of Alfred Marshall*. London, Macmillan, 1925.
- PINTO, José Castro; CURTO, José Dias – *Estatística para Economia e Gestão*. Lisboa, Edições Sílabo, 2.^a Ed., 2010.
- RODRÍGUEZ, Elena Martínez – “Errores frecuentes en la interpretación del coeficiente de determinación lineal”, in: *Anuario Jurídico y Económico Escurialense*, 38, 2005.
- SCHAFFER, Jonathan – “Contrastive causation”, in: *The Philosophical Review*, 114 (3), 2005.
- SHUGAN, Steven M. – “Causality, unintended consequences and deducing shared causes”, in: *Marketing Science*, 26 (6), 2007.
- WALKER, Donald A. – “Edgeworth versus Walras on the theory of tatonnement”, in: *Eastern Economic Journal*, 13 (2), 1987.
- WANG, Ting; ZHANG, Shiqiang – “Study on linear correlation coefficient and nonlinear correlation coefficient in mathematical statistics”, in: *Studies in Mathematical Sciences*, 3 (1), 2011.