CRÍTICA À AVALIAÇÃO PELA MODA DA DISTRIBUIÇÃO LOGNORMAL

ISABELA BECK DA SILVA GIANNAKOS (*)
Rua Dona Laura, 87, conj. 306
Porto Alegre, RS, Brasil - CEP : 90430-090
Fone/fax: (051) 222.5156

MANOEL LUIZ LEÃO (**)
Rua Miguel Tostes, 270
Porto Alegre , RS, Brasil - CEP: 90430-060
Fone: (051) 331-5618 - Fax: (051) 332.5833
E-mail: leao@VORTEX.UFRGS.BR

"Anyone who considers arithmetical methods of producing random numbers is, of course, in a state of sin."
(Matthews, 1995)

A Autora do presente trabalho levou a cabo, em 1994, a avaliação de um terreno em Porto Alegre, RS, por cometimento de instituição de âmbito nacional, utilizando amostra de 42 terrenos; como variável dependente adotou o valor unitário (em R$/m^2) e, como variáveis independentes, a área do imóvel, sua localização e a ordem cronológica das datas em que obtidos seus valores unitários. O modelo de regressão adotado decorria da transformação logarítmica da variável dependente e de uma das variáveis independentes.

Submetido o laudo, recebeu, do órgão técnico central da instituição, despacho demandando que "revisasse os cálculos", pois deveria promover a avaliação "pela moda", supostamente "o valor mais provável", dado que, alegadamente, adotara "a mediana".

Com o concurso do co-autor, foi contestado o surpreendente despacho, concluindo que nada havia a alterar no laudo e que não tinha qualquer legitimidade a invocação das características da distribuição lognormal, para, a partir das mesmas, adulterar o resultado da regressão.

Posteriormente, os Autores tiveram acesso ao artigo "EXAMINANDO OS MODELOS DE REGRESSÕES - (II)" assinado pelo autor do "software" do qual faz uso a instituição acima referida e no qual se propõe a "avaliação pela moda", sempre que, em busca da linearização do fenômeno, a variável dependente é transformada por logaritmação e os resíduos da regressão assumem forma de distribuição normal. Neste caso, a função antilogarítmica se torna assimétrica "à esquerda", isto é, a sua moda precede a mediana e esta, por sua vez, precede a média aritmética.

A essência do raciocínio é a seguinte: Após abordar a transformação logarítmica das variáveis de regressão, parte o autor do artigo da constatação básica de que a distribuição de uma variável aleatória é assimétrica (dita "lognormal") quando o seu logaritmo se distribui normalmente, diferindo, portanto, na distribuição da mesma, a moda da mediana e da média. Como a primeira é, segundo alega, "o valor mais provável", conclui que a estimativa do valor da referida variável, para o bem objeto da avaliação, se deva fazer "pela moda" e não "pela média" ou "pela mediana". Apresenta um exemplo em que o valor unitário do imóvel avaliado, "pela moda", é expresso por R$ 681,3412R$/m², quando a estimativa "pela média" forneceria R$ 761,2795/m². Daí sustentar que **"esta condição tem levado muitos avaliadores a superestimarem os valores dos bens avaliandos, não advertidos para o fato da assimetria da distribuição lognormal".**

Afirma ainda, o autor, no referido artigo: "...nem a média nem a mediana são indicativos do valor mais provável, que é a moda".

Isto posto, passa-se, agora, a comentar o procedimento proposto e adotado no referido "software", bem como algumas afirmações de seu autor:

1. A RUPTURA DAS PREMISSAS DA REGRESSÃO LINEAR.

A primeira objeção ao procedimento decorre da resposta à pergunta: É lícito invocar a "função original", que foi submetida a uma transformação conveniente, na busca pela linearidade, para, a partir dela, alterar a conclusão da regressão linear aplicada às variáveis transformadas?

Evidentemente, **NÃO**!

Não há porque, na regressão linear, invocar a equação dita "original", isto é, aquela em que, sem qualquer transformação, a variável dependente Y (no caso, valor por metro quadrado) é expressa em função das variáveis independentes, (no caso, acima descritas).

Se, na configuração original das variáveis, a dependência de Y em relação às variáveis independentes revelar-se não-linear, é legítima qualquer transformação nas variáveis (como, por exemplo, adotando $W=\ln(Y_c)$, $V_1=\ln(X_1)$, $V_2=X_2^{(-1)}$ e $V_3=X_3$), na busca de configuração linear para a dependência apontada pela amostra, tudo com o propósito de aumentar a qualidade do modelo, medida por seu "coeficiente de determinação". Entregues os valores de W, V_1, V_2 e V_3 aos procedimentos da regressão, eles se desencadearão sobre os mesmos, para determinar os coeficientes a, b_1, b_2 e b_3 que permitirão escrever a equação **linear**

$$W = a + b_1 * V_1 + b_2 * V_2 + b_3 * V_3, \qquad (1)$$

mediante a qual estima-se W quando conhecidos V_1, V_2 e V_3. Saliente-se, no entanto, que é inteiramente irrelevante para o resultado considerar que as variáveis utilizadas provenham de outras, convenientemente transformadas. O processo de regressão ignora, completamente, que W e V1 são logaritmos neperianos, que V_2 é um inverso e que V_3 é uma variável ligada ao tempo, elevada ao expoente +1. Para o método, todas as 4 variáveis são escalas numéricas "originais", cujo

(*) Engenheira Civil, com especialização em Edificações.
(**) Engenheiro Civil, Ex-Professor Titular, UFRGS-Escola de Engenharia e Faculdade de Ciências Econômicas, Métodos Quantitativos - Estatística e Pesquisa Operacional.

comportamento e cujo relacionamento linear é pesquisado e medido.

Assim sendo, todas as conclusões da regressão emanam das variáveis submetidas à sua investigação e a elas se aplicam; alcançado o desfecho (por exemplo, no caso, o logaritmo neperiano do valor unitário do terreno sujeito à avaliação), nada impede que se extraia o antilogaritmo deste valor, para expressar o dado na unidade original, a saber, Reais por metro quadrado. Além disto, o comportamento não-linear e assimétrico dos dados iniciais não precisa ser invocado para influir na conclusão da regressão linear, pois ele está, de qualquer forma, "embutido" no tratamento da versão linear, como se pode ver no item 5, abaixo.

2. OS DOIS COMPONENTES DA AVALIAÇÃO.

O resultado da avaliação por regressão linear (com recurso, portanto, à metodologia de mínimos quadrados) leva, sempre, a um resultado formado por dois elementos, um determinístico, ou **prescritivo**, outro probabilístico ou **aleatório**. O primeiro é resultado da introdução, em uma equação linear, dos valores que correspondem, em determinado imóvel a avaliar, às variáveis ditas independentes no modelo de regressão. Resolvida a equação para estes valores, ter-se-á o valor da variável dependente, isto é, a avaliação procurada para o imóvel em questão.

Esta avaliação, no entanto, invoca um pressuposto: "Na suposição de que a sua determinação decorra apenas do efeito das variáveis independentes e que o comportamento destas últimas "explique", completa e cabalmente, toda e qualquer variação no valor da variável dependente, tal como observado na amostra colhida."

Infelizmente, nunca este pressuposto é inteiramente atendido. Sempre resta, no valor dos próprios componentes da amostra, utilizados para o cálculo dos parâmetros da equação de estimativa, um resíduo, para mais ou para menos, "não explicado", isto é, **aleatório e determinado por causas, provavelmente em grande número, não assinaladas,** umas diminuindo, outras aumentando o valor do bem, em relação àquele decorrente da equação de regressão.

Este segundo componente "não determinístico" da avaliação, embora meramente "prospectivo", é essencial e não se pode confundir com o primeiro, pois, inclusive, fornece ao leitor do laudo (por exemplo, o próprio Juiz que tem a decisão da lide a seu cargo) uma indicação da qualidade do modelo de regressão linear adotado: Quanto maior o "coeficiente de determinação" do modelo, a saber, o quociente entre a "variância explicada" (a variância dos valores de W, fornecidos pela equação (1), em relação à média dos valores da variável dependente) e a variância total da mesma variável dependente, mais nítida a qualidade do modelo como determinante do valor do imóvel a avaliar e, conseqüentemente, mais robusto seu poder preditivo.

Quanto maior a parcela "explicada" da variância total, menor será aquela "não explicada", isto é, indeterminada e aleatória, imune ao poder determinante das variáveis independentes. Isto porque as duas variâncias, a "explicada" e a "não explicada", por imposição do método de mínimos quadrados, somam, sempre, o valor da variância total da variável dependente.

Esta fração "não explicada" corresponde aos resíduos que se dispõem em torno da equação linear (uma reta, se houver uma única variável independente; um plano se forem duas as variáveis independentes e um hiperplano, se forem elas em número de três ou mais). No caso, como já referido, há três variáveis independentes, a área do imóvel, sua localização e a data aplicável ao dado.

O componente "não explicado" da avaliação, é, portanto, aleatório e não determinado **e assim deve permanecer, não só para informação do leitor como para preservar a integridade do próprio modelo de regressão.** Comunicar-lhe um valor determinado, seja ele qual for, é confundi-lo com o componente determinístico e prescritivo, fornecido pela equação de regressão, o que, por sua vez, como se verá a seguir, adultera esta última, viola as características básicas do método de mínimos quadrados, o fundamento da regressão linear, ou, alternativamente, adultera os valores da variável dependente Y, na amostra original.

3. O IMPASSE GERADO PELA ADOÇÃO DA MODA.

Na avaliação aqui referida, os resíduos não explicados pela equação de regressão, $\ln(Y)-\ln(Y_c)$, formam uma distribuição normal, com média 0 e desvio-padrão sigma=0,31141. Os valores "originais", não logaritmados, formam uma distribuição lognormal, assimétrica à esquerda, com mediana igual a 1 (ver item 6) e moda igual a Exp(-sigma^2), isto é, 0,9075776. Resta saber como determinar Z, o valor "corrigido" do bem, por influência da moda. Tem-se que $\ln(Y)-\ln(Y_c)=\ln(Y/Y_c)$. Segue-se $Z/Y_c = 0{,}9075776$ e

$Z = 0{,}9075776*Y_c$.

Na forma logarítmica,

$\ln(Z)=\ln(Y_c)+\ln(0{,}9075776)$, ou seja,

$\ln(Z)=\ln(Y_c)-0{,}0969762$

(sempre considerando o problema abordado neste artigo). Para obter-se, então, a avaliação "pela moda", cumpre adicionar ln(0,9075776) à equação de regressão, um valor numérico de natureza idêntica à do parâmetro linear "a" e que, por conseguinte, a este se adiciona, gerando, na equação, um novo parâmetro linear, a_1=a-0,0969762, e, portanto, **uma nova equação de regressão,** não mais aquela que resultou da amostra colhida, mas outra, que lhe é paralela, no espaço, pois permanecem inalterados seus parâmetros angulares, b_1, b_2 e b_3. Em lugar de

$$\ln(Y_c)=a+b_1*Ln(X_1)+b_2*X_2^{(-1)}+b_3*X_3, \quad (3)$$

(onde, depois, faz-se $Z=Y_c*0{,}9075776$),

passa-se a ter

$$Ln(Z)=a_1+b_1*Ln(X_1)+b_2*X_2^{(-1)}+b_3*X_3 \quad (4)$$

Mas este resultado gera um impasse:

Ou bem são mantidos inalterados os valores de Y da amostra colhida, respeitando-se o papel fundamental desta última, como testemunho do mercado e fonte básica de informação para todo o processo de avaliação, ou bem, alternativamente, modificam-se arbitrariamente os valores originais de Y, multiplicando-os, todos, pelo mesmo fator 0,9075776. No primeiro caso, a equação 3, acima, deixa de atender as exigências do método de mínimos quadrados: a soma dos quadrados dos resíduos deixa de ser mínima, o cálculo dos elementos discrepantes ("outliers") precisaria ser refeito e a média dos Y não mais se acha sobre a equação.

Não obstante, esta seria a opção mais aceitável, pois tudo o que se pretende é invocar o comportamento assimétrico da distribuição do antilogaritmo para "retificar" a estimativa fornecida pela equação de regressão. Fiel a esta invocação, o valor "corrigido", Z, não se pode situar no centro dos resíduos, prevalecendo, portanto, a equação 3 como resultado da regressão. No entanto, o "software" em exame prefere a opção pela equação 4, onde se deslocam todos os dados do problema para, então, com nova equação de regressão, alcançar o valor pretendido, fruto da moda da distribuição lognormal, agora como elemento central da estimativa, quando, obviamente, no problema inicial, tinha posição excêntrica.

A solução de impor uma estimativa desviada do centro deixaria patente a violação das premissas da regressão linear e do método dos mínimos quadrados, além de evidenciar que a moda **subestima** o valor do bem, pois, recalculados os resíduos, desta feita em relação aos valores de $\ln(Z)=\ln(Y_C)+\ln(0{,}9075776)$, verifica-se, no problema aqui focalizado, que 23 dos 42 valores ficarão acima do plano de referência e apenas 19 deles se situarão abaixo deste último. A soma dos valores absolutos dos resíduos positivos e negativos representa, respectivamente, 68,15% e 31,85% das 42 ocorrências. Quer em número, pois, quer em valor agregado dos desvios, a amostra de 42 elementos aponta para a proporção abaixo referida, no item 5, segundo a qual a moda é superada por 62,23% das ocorrências da distribuição lognormal do problema em tela.

Na segunda hipótese, adotada no procedimento praticado pelo "software" em exame, a equação 4 mantém o cumprimento dos requisitos da regressão linear, ao preço, porém, de se alterarem todos os valores colhidos no campo, o que constitui inaceitável distorção dos dados de mercado.

Tudo se passa como se não mais se tratasse do problema inicial, mas de outro, **incidente sobre uma amostra adulterada, onde os 42 valores iniciais da variável dependente são multiplicados, todos, por 0,9075776!**

Em suma, a intenção inicial era a de "corrigir" a avaliação proveniente da aplicação da regressão linear a uma determinada amostra concreta, colhida no campo, em decorrência da descabida consideração da distribuição assimétrica dos resíduos, numa equação não-linear e não contemplada na regressão.

Para tanto, desfez-se a distinção essencial entre os dois componentes da avaliação, o prescritivo e o aleatório; e terminou-se por adotar, como avaliação do bem objeto da perícia, a conclusão obtida de um problema diverso do original, onde todos os dados foram deslocados. Agora, porém, aceita-se a avaliação pela equação de regressão linear, antes contestada, porque, desta feita, graças a esta artificiosa e arbitrária violação, o valor numérico resultante é, precisamente, o da moda da distribuição antilogarítmica, "original", dos resíduos! Uma situação de inversão hierárquica à qual, sem dúvida, se aplica a versão norteamericana de nosso "carro adiante dos bois": *"The tail is wagging its dog..."*

O mais curioso deste cortejo de estranhas inovações é o seu corolário: Substituída a avaliação inicial por outra, fruto de dados alterados, que não exprimem mais a situação do mercado, subestimando-o, volta-se, tal como da primeira vez, a estimar "pela mediana", isto é, adota-se, como valor do imóvel a avaliar, aquele que resulta da nova equação de regressão, quando nela se introduzem os valores das variáveis independentes que caracterizam o bem a avaliar. Agora, porém, o proce-dimento adotado parece não nutrir mais preocupação com o componente aleatório; ele é subitamente esque-cido e desprezado, ficando-se, apenas, com o compo-nente "prescritivo"!

Neste caso, não há como deixar de fazer com que prevaleça o mesmo argumento invocado para impugnar a avaliação inicial, qual seja, que a moda da nova distribuição lognormal (que se obtém multiplicando a anterior, novamente, por 0,9075776) é "o valor mais provável", recomeçando o ciclo, que se perpetuará, culminando por reduzir a zero o valor do bem objeto da perícia...

4. A MODA É "O VALOR MAIS PROVÁVEL"?

Há, nesta afirmação, uma imprecisão e um erro semântico. **A moda não corresponde a uma probabilidade**, pois uma ordenada da função de densidade de probabilidade de uma variável aleatória contínua é, apenas, tal como exprime sua denominação, uma "densidade de probabilidade". Ao conceito de probabilidade corresponde, por definição, **uma área** sob a curva da função de densidade. Ora, um ponto no eixo (ou eixos) das abscissas não define uma área; impõe-se um intervalo, isto é, **uma faixa da variável**, para definir uma probabilidade, expressa pela área, sob a curva, correspondente à faixa delimitada pelos seus pontos extremos. Não faz sentido, pois, falar em "probabilidade de um ponto" no eixo das abscissas.

Não há, portanto, tal coisa como $P(Y/Y_C=\text{Moda})$, se Y/Y_C é uma variável contínua, como o é, no caso, o valor unitário do terreno. Faz sentido, isto sim, dizer, por exemplo, que

$P(a<=Y/Y_C<=b)=0{,}10$,

que se lê como "a probabilidade de Y/Y_C, variável aleatória contínua, ser maior ou igual a "a" e menor ou igual a "b" é 0,10", em que os valores "a" e "b" definem uma faixa na função de densidade.

Assim, quando se diz que "a moda é o valor mais provável", o que se quer realmente é afirmar que, se a moda de uma distribuição se situa dentro do intervalo a,b, a probabilidade de Y/Y_C, obtido ao acaso, cair neste intervalo, é a maior dentre as probabilidades correspondentes a todos os demais intervalos **de mesma amplitude b-a**, na respectiva função de densidade.

No caso presente, onde a função lognormal tem Y/Y_C variando entre zero e +infinito e a moda é igual 0,9075776, se tomada, por exemplo, uma faixa de largura 0,1, de 0,85 a 0,95, a probabilidade de se obter um valor de Y/Y_c nela contido será de 0,133708, isto é, somente 13,37% das ocorrências de Y/Y_c cairão neste intervalo. Para qualquer outra faixa, **de mesma amplitude**, a probabilidade de Y/Y_c nela recair será fatalmente menor. Este é o exato sentido da expressão "a moda indica o valor mais provável da variável". Mas ela também significa que, extraído ao acaso um valor do quociente Y/Y_c, ele **incidirá em outra faixa que não a modal**, com a probabilidade esmagadoramente predominante de 0,866292. Em outras palavras, 86,63% das ocorrências surgem **fora** da faixa de largura 0,10 onde se acha a moda!

Dependendo da faixa arbitrada, o resultado do sorteio de um valor de Y/Y_c, expresso em dois eventos: "Contém a moda" e "Não a contém", pode indicar, como no caso da faixa de 0,10, que a resposta negativa é mais de seis vezes mais provável. Se mais estreita a faixa, como, por exemplo, de 0,90 a 0,91, a probabilidade do quociente Y/Y_c nela se achar é de apenas 0,013446 (1,34%), contra a probabilidade de 0,986554 (98,65%) de ali **não** se encontrar, muito embora continue prevalecendo que nenhuma outra faixa de mesma largura (0,01) apresente probabilidade maior que aquela assinalada para a faixa modal. Se, por outro lado, a faixa for suficientemente ampla para que se verifique:

$$P(a <= Y/Yc <= b) > 0,5,$$

de modo que a probabilidade de Y/Y_c recair **fora** da mesma não alcance 0,5, ter-se-á uma amplitude tão alargada que a significância do valor modal nela se dilui. Ademais, ela passará a abranger tanto a moda como a mediana e a média, ficando-se sem saber qual das três medidas de posição, afinal, é "o valor mais provável"...

Vê-se, assim, que é temerário substituir toda a distribuição da variável aleatória lognormal pelo valor de sua moda (nem sequer por uma faixa em torno da mesma), a qual, a rigor, por ser a ordenada do ponto máximo da função de densidade, corresponde, nas abscissas, a um ponto sem dimensão e, portanto, tem, formalmente, probabilidade **nula** de ocorrência! Isto quanto à falácia do "valor mais provável".

5. A VARIÁVEL TRANSFORMADA CONTÉM A INFORMAÇÃO NECESSÁRIA À DESCRIÇÃO DA VARIÁVEL "ORIGINAL".

Os resíduos da variável logaritmada, $D=Ln(Y)-Ln(Y_c)$, embora não pareçam se ocupar do comportamento do dado de origem, $Q=Y/Y_c$, mesmo assim, descrevem o comportamento deste último. A distribuição de D é normal, com média 0; seus extremos acham-se no infinito negativo e positivo. Logo, o antilogaritmo de D, isto é, Q, se estende de 0 a +infinito, com mediana em 1,0. Dividindo a amplitude de Q em seis intervalos, 5 com amplitude de 0,5, mais um que se estende de 2,5 a + infinito, pode-se obter a probabilidade de $Q=Y/Y_c$ se encontrar em cada um destes intervalos, simplesmente adotando os logaritmos neperianos de 0,5, 1,0, 1,5, 2,0 e 2,5, promovendo, a seguir, a integração da função normal nos intervalos definidos por estes logaritmos. Pelo princípio da comparabilidade dos eventos, estas mesmas probabilidades valem para os intervalos homólogos da variável Q.

Assim, pode-se, no problema aqui referido, obter a seguinte tabela de valores:

Classes de Q	Probabilidade	Prob. Cumulativa
0 - 0,5	0,01301	0,01301
0,5 - 1,0	0,48699	0,50000
1,0 - 1,5	0,40355	0,90355
1,5 - 2,0	0,08344	0.98699
2,0 - 2,5	0,01138	0,99837
2,5 - +inf.	0,00163	1,00000

De posse desta informação, poderá o perito, se assim o entender, transmitir a seus constituintes e ao Juiz da lide que, no problema em exame, o componente aleatório da avaliação é tal que, se obtidos valores de $Q=Y/Y_c$ da população de integrantes do modelo, cerca de 13 em mil quocientes se situarão entre 0 e 0,5; em torno de 487 em mil se acharão entre 0,5 e 1,0, de tal forma que a metade se encontre nas duas primeiras classes da distribuição (pois 1,0 é a mediana); cerca de 40% estarão entre 1,0 e 1,5; 8%, aproximadamente, recairão entre 1,5 e 2,0; pouco mais de 11 por mil serão encontrados entre 2,0 e 2,5; finalmente, os 1 ou 2 por mil restantes exibirão valor superior a 2,5.

Poderá optar por outras descrições, como, por exemplo: "O componente aleatório da avaliação de um terreno é tal que o quociente entre o valor real de Y e o valor estimado, Y_c, tem distribuição assimétrica, que se estende de 0 a +infinito, com mediana em 1; ademais, os quocientes compreendidos entre 0,85 e 1,15 ocorrerão na proporção de cerca de 37,23%".

Algo, em suma, assaz diferente da afirmação segundo a qual é lícito substituir todo este complexo comportamento aleatório simplesmente pela adoção de um único valor supostamente o "mais provável", a moda da distribuição...

6. A MODA NÃO É ESTIMADOR SATISFATÓRIO

A opção pela moda **perigosamente subestima** o valor do componente aleatório.

Já foi referida a afirmação do autor do software e do artigo aqui enfocados, segundo a qual os peritos que, nestas circunstâncias, não avaliarem "pela moda", estarão **superestimando** o valor do bem. Esta afirmação, possivelmente, é responsável pelo rechaço sofrido, na instituição-cliente, pela avaliação mencionada no início do presente trabalho, cuja autora foi intimada a "revisar seus cálculos"...

Na realidade, a audaciosa afirmação inverte os fatos: subestima o valor do bem aquele que avalia "pela moda".

Ao adotar o valor proposto pela equação de regressão linear, o perito, como acima referido, estará informando o Juiz a quem se dirige que o valor pelo qual avaliou o bem é dado por Y_c; adicionalmente, há um componente aleatório, de carater aditivo ou subtrativo, com determinado desvio-padrão, cujo resultado, porém, tanto excederá o valor Y_c, como lhe ficará aquém, com a mesma probabilidade de 50%. Decorre isto do princípio dos "eventos comparáveis"; o perito avaliou, na realidade, o logaritmo neperiano de Y_c; os resíduos aleatórios são medidos como $Ln(Y)-Ln(Y_c)$, onde $Ln(Y_c)$ é o valor central de uma distribuição normal e, portanto, sua mediana. Conseqüentemente, seu homólogo, antilogaritmo, necessariamente, marcará, na distribuição lognormal de Y/Y_c, também a mediana. No caso, como o antilogaritmo de 0 é a unidade, a mediana de Y/Y_c terá valor 1,0.

Já quem avalia pela moda de Y/Y_c estará adotando no caso, um valor que, na distribuição lognormal, estará, sempre, à esquerda da mediana, isto é, **será fatalmente menor que 1,0**. Se assim é, a área à esquerda da moda, isto é, a probabilidade de ocorrer um quociente Y/Y_c menor que esta última, no problema em tela, será de 0,37774, isto é, apenas 37,77% das ocorrências se situam **aquém** da moda. Conseqüentemente, 62,23% delas serão **maiores** que a mesma. Qual o sentido, então, em fazê-la substituir a avaliação por Y_c? Quem assim avaliasse, deveria, obrigatoriamente, advertir o Juiz destinatário de seu laudo de haver adotado um proce-dimento que grosseiramente viola os paradigmas da regressão linear, que substitui uma variável aleatória por um único valor fixo, sob questionável alegação de ser ele "o mais provável" e que, finalmente, comete condenável deformação da conclusão, por subestimar o valor do componente aleatório da avaliação.

7. A OPÇÃO PELA MÉDIA

O artigo "EXAMINANDO OS MODELOS DE REGRESSÕES - (II)", deixa clara a preferência de seu autor pela "avaliação pela Moda", quando aplicável a distribuição lognormal, embora, à certa altura, admita que "tanto a estatística moda, como a mediana e/ou a média são indicativos".

Por que, então, não adotar a média aritmética? Como representativa da variável aleatória, ela é mais significativa que a moda (cuja qualidade de "valor mais provável" é "misleading", como visto); a média aritmética é o "valor esperado" da variável, aquele para o qual tende o quociente entre a soma dos valores observados e o número deles; é, ainda, o valor, no eixo das abscissas, que identifica o centro de gravidade da distribuição, isto é, o ponto que, colocado sobre o bisel de uma balança, equilibraria o formato da função densidade de probabilidade. Mas oferece os mesmos impedimentos já apontados para a moda, nos itens 1 e 2, acima, a saber: introduz, na regressão linear, como fator de decisão, as características da função dita "originária", não-linear, transformada em logarítmica precisamente para alcançar linearidade; viola os pressupostos do método de mínimos quadrados, fundamento da regressão, ou, alternativa-mente, equivale a adulterar a amostra original, multiplicando, no caso presente, todos os seus valores por 1,04968. Além disto, substitui, arbitrariamente, o componente aleatório da avaliação por valor determinado, que distorce o resultado visado. Finalmen-te, onde a moda subestima o valor dos quocientes Y/Y_c, a média o superestima, embora em escala menos pronunciada.

Aliás, nunca é demais repetir, a razão principal para rejeitar toda e qualquer tentativa de avaliar seja "pela Moda", "pela Média" ou "pela Mediana" está na interpretação implícita, de que se estaria avaliando a variável Y ou os quocientes Y/Y_c, quando, na realidade, está-se lidando tão somente com a variável W (ver equação 1), sem qualquer preocupação com sua origem ou "paternidade".

No "paper" do autor da proposta aqui discutida figura a seguinte declaração (pontuação e sintaxe como no original): *A média, é um indicativo generalizado, justamente pelo fato de ser uma estatística que se confunde com a mediana e a moda em razão de simetrias de distribuições de freqüências e, por isso mesmo, serve-se para citações espirituosas do tipo: "o nadador morreu afogado em um lago com 20 cm de profundidade média."* Não se evidencia exatamente o que possa o autor pretender ao dizê-lo, mas, ao que parece, se trata de salientar que o uso da média aritmética como indicativo da variável aleatória pode oferecer riscos mortais, talvez insinuando que o mesmo "defeito" não surja no caso das outras medidas de posição. Mas não é exato dizer que a única razão para a média ser um "indicativo generalizado" está em que, nas distribuições simétricas, ela se confunde com a moda e a mediana e que, "por isso mesmo" se presta a ditos espirituosos, como o citado, sobre o infortunado nada-dor.

A média é um importante descritor da variável aleatória; é o seu "valor esperado" e o centro de gravidade da função densidade de probabilidade. Mas ela é apenas parte da caracterização da variável, que se complementa com a informação de sua dispersão (ou, alternativamente, pelo conhecimento da função densidade de probabilidade). Na realidade, o que vitimou o nadador foi a ignorância deste importante detalhe; por esta razão, jogou-se ele de ponta-cabeça em local demasiado raso - e fraturou a coluna cervical - ou mergulhou em ponto demasiado profundo - e morreu afogado porque, afinal, não era tão bom nadador quanto imaginava. Portanto, o que realmente importaria saber era a **probabilidade** com que um ponto qualquer do lago apresentasse profundidade demasiado grande ou muito pequena. O conhecimento seja da média, da mediana ou da moda, para este fim, de pouco lhe serviria.

Também importa salientar o fato de ser o mencionado "dito espirituoso" extremamente oportuno, no contexto do que ora se discute, pois, o que realmente vitimou o banhista (que ao menos dispunha do atenuante de, provavelmente, ser leigo em probabilidade e estatística...) foi o perigo, sempre mortal, a que se expõe quem assume o risco de substituir uma variável aleatória por um único de seus valores, ainda que seja aquele que, enganosamente, se apresente como "o mais provável", tal como ocorre, exatamente, com a avaliação "pela moda"...

8. À GUISA DE CONCLUSÃO

Os computadores hoje oferecidos no mercado, somados às poderosas planilhas eletrônicas também disponíveis, bem como "softwares" de matemática aplicada, de amplo alcance, tornam triviais muitos problemas de regressão linear e cálculo de probabilidades. Fatalmente, este quadro afetará o mercado para "softwares" locais destinados a peritos e avaliadores, o que, por sua vez, poderá desencadear, entre seus autores, uma atitude de ansiosa busca de "algo mais" a oferecer, perseguindo "avanços que os outros não têm", com o risco de se proporem inovações insólitas, duvi-dosas ou infundadas, cujo resultado, a médio prazo, poderia afetar a credibilidade dos profissionais que as utilizassem.

Em decorrência, os Autores recomendam que se institua, no Brasil, pelo organismo representativo dos profissionais engajados na atividade pericial, um sistema de verificação e credenciamento de "softwares" destinados ao uso de peritos, para que sua liberação para a comercialização somente ocorra após avaliados por especialistas, entre eles professores universitários de notório saber, com Doutorado em Métodos Quantitativos, Probabilidade e Estatística ou Matemática Aplicada, para evitar o surgimento de procedimentos infundados, abusivos e temerários, além de constrangedores para os profissionais, onde os clientes, segundo sua conveniência ou interesse, venham a exigir que seus peritos e assistentes técnicos optem por abordagens que ora subestimem, ora superestimem a avaliação de um bem.

BIBLIOGRAFIA

Chou, Ya-lun. Statistical Analysis With Business and Economic Applications. Holt, Rinehart and Winston, Inc., N. York, 1969.

Hald A. Statistical Theory with Engineering Applications. John Wiley & Sons, Inc. N. York, 1952.

Hamburg, Morris. Statistical Analysis for Decision Making, 2nd. Edition. Harcourt Brace Jovanovich, Inc. N. York, 1977.

Kmenta, Jan. Elementos de Econometria - Teoria Econométrica Básica, 2 volumes. Editora Atlas, São Paulo, 1994.

Matthews, Robert. "It's a lottery", New Scientist, 22 July 1995, N° 1987, pp. 38-42.

Wolfram, Stephen. Mathematica - A System for Doing Mathematics by Computers, 2nd. Edition. Addison-Wesley Publishing Co. Reading, Massachusetts, 1991.