EJEMPLOS Y APLICACIONES INSÓLITAS EN REGRESIÓN Y CORRELACIÓN

C. M. CUADRAS

Universitat de Barcelona

1. INTRODUCCIÓN

En las aplicaciones deterministas de las matemáticas, en las que se establece una relación no lineal entre dos conjuntos de variables $\boldsymbol{x}$ e $\boldsymbol{y}$, mediante una función f, normalmente no aparecen mayores dificultades que las propias de la naturaleza de la función (campo de variabilidad, existencia de derivadas, puntos singulares, etc.).

En estadística, sin embargo, se relacionan variables aleatorias, y si las variables $\boldsymbol{x}$ son de tipo control, entonces se hace intervenir un término de error aleatorio. Como en general no se conoce el tipo de modelo, se recurre a un modelo lineal

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

del que se estiman los parámetros, y se calcula el coeficiente de correlación múltiple R como medida de buen ajuste lineal, o en su caso el coeficiente de determinación R^2, que representa el grado de variabilidad de y que se explica a través de las variables $\boldsymbol{x}$.

A pesar de la simplicidad del modelo de regresión múltiple desde un principio se presentaron paradojas, que han podido ser explicadas más o menos con acierto.

Citemos, por ejemplo, el concepto de "regresión de la talla a la mediania", descubierto por F. Galton en sus estudios de la herencia padres-hijos. A padres altos corresponden hijos altos, pero menos altos que el padre en relación a la media, y a padres bajos corresponden hijos bajos, pero más bajos que el padre, también en relación con la media.

Esta aparecente anomalía, que parece forzar a las futuras generaciones a converger a una talla media única, se puede explicar fácilmente si tenemos en cuenta que la variabilidad de las tallas se mantiene estable si consideramos toda la población.

La llamada correlación espúrea, término introducido por K. Pearson, tiene su ilustración más famosa en la correlación entre el número C de cigüeñas y el número B de bebés nacidos en un cierto periodo de tiempo. Tal correlación, superior a 0.8 (según datos citados por J. Neyman) desaparece si consideramos la influencia de la variable M (tasa de mujeres) y correlacionamos C/M con B/M (véase Rios, 1991, pág. 140).

Estos y muchos otros ejemplos, son bien conocidos, se encuentran en la mayoría de libros, y el buen profesor debería mencionarlos para alertar al alumno de las malas interpretaciones en regresión y correlación.

En este artículo presentamos otros ejemplos y aplicaciones insólitas que por ser más recientes, son quizás menos conocidos, y asi los exponemos esperando sean de interés tanto para el profesor como para el usuario de la Estadística.

2. CORRELACIÓN NO SIEMPRE INDICA REDUNDANCIA

Indiquemos por Y la variable dependiente (también llamada respuesta, endógena) y por $X_1, \ldots, X_n$ las variables independientes (explicativas, exógenas).

Sean $r(Y, X_1) = r_1, r(Y, X_2) = r_2, \ldots, r(Y, X_n) = r_n$ los coeficientes de correlación entre la variable Y y las variables $X_1, \ldots, X_n$. Cuando todas las variables X están incorrelacionadas, el coeficiente de determinación es

$$R^2 = r_1^2 + \cdots + r_n^2$$

Pero, si las variables $\boldsymbol{x}$ están correlacionadas entre sí es usual encontrar

$$R^2 < r_1^2 + \cdots + r_n^2$$

pues "parte de la variabilidad de Y debida a cada X_i se puede explicar por la variabilidad mutua entre las variables correlacionadas X_i, X_j", es decir, la correlación implica una cierta redundancia entre las variables $\boldsymbol{x}$.

Sin embargo, no siempre es así, pudiéndose mostrar tanto ejemplos téoricos como prácticos en los que

$$R^2 > r_1^2 + \cdots r_n^2 \tag{1}$$

Tabla 1

Y	X_1	X_2
0	5	5
1	3	2
1	1	0
1	4	3
1	6	5
2	7	5
2	1	0

$r_{12} =$ 0.969087
$r_1 =$ -0.08846
$r_2 =$ -0.30429

Para los datos de la tabla 1 se cumple $R^2 = 0.7926, r_1^2 = 0.0078, r_2^2 = 0.0926$, luego

$$R^2 = 0.7926 > r_1^2 + r_2^2 = 0.1004$$

y la desigualdad se cumple.

La interpretación de esta desigualdad ha sido dada por Hamilton (1987) para el caso $n = 2$ en términos de la multicolinearidad y la correlación parcial.

Hamilton prueba que esta desigualdad es equivalente a

$$r_{12}\left(r_{12} - \frac{2r_1r_2}{r_1^2 + r_2^2}\right) > o \tag{2}$$

donde $r_{12} = r(X_1, X_2)$ es la correlación entre las variables explicativas.

La (2) prueba que (1) tiende a cumplirse si r_{12} es un valor alto (hay multicolinearidad) y si $r_1^2 + r_2^2$ es pequeño (la correlación entre Y y las variables x es pequeña). Como señala Hamilton, el profesor de estadística no debe reflejar en sus enseñanzas la falsa creencia de que variables explicativas correlacionadas son siempre redundantes y que valores bajos de las correlaciones entre la variable dependiente y las variables explicativas (o el uso de diagramas de puntos, es decir, $x - y$ plots reflejando escasa relación) no deben usarse como criterio de eliminación de variables explicativas.

Routledge (1990) da una explicación interesante de (1). Supongamos que Y_1, Y_2 son las componentes principales calculadas a partir de la matriz de correlaciones entre las variables explicativas. Para $n = 2$ y suponiendo $r_{12} > 0$, las correlaciones entre Y y las componentes son

$$r(Y, Y_1) = \frac{r_1 + r_2}{[2(1 + r_{12})]^{1/2}}$$

$$r(Y, Y_2) = \frac{r_1 - r_2}{[2(1 - r_{12})]^{1/2}}$$

El coeficiente de determinación viene dado entonces por

$$R^2 = r^2(Y, Y_1) + r^2(Y, Y_2)$$

Routledge concluye que para que R^2 sea grande con respecto a r_1 y r_2 es necesario que: *(a)* Y esté altamente correlacionado con Y_2 pero no con Y_1, *(b)* La varianza de Y_2 sea mucho menor que la varianza de Y_1 (La conclusión es la misma si $r_{12} < 0$, bastando intercambiar las dos componentes principales). Obsérvese que como la varianza de Y_2 es proporcional a $(1 - r_{12})$, *(b)* significa que hay multicolinearidad. Como en tal caso la variabilidad de X_1, X_2 queda mayoritariamente explicada por Y_1, resulta finalmente que si se cumple (1) es que Y depende de la variabilidad "residual" de X_1, X_2.

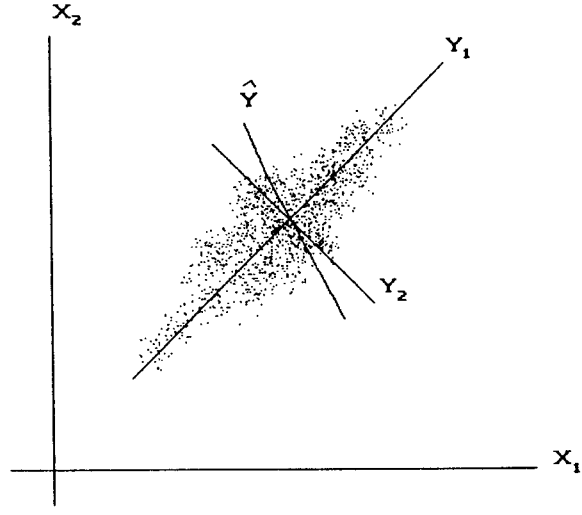

Figura 1. La fuerte correlación de Y con la segunda componente principal Y_2, cuya varianza es pequeña, induce a que la desigualdad $R^2 > r_1^2 + r_2^2$ se cumpla.

El caso $n > 2$ ha sido estudiado por Cuadras (1992a), dando fórmulas explícitas que reflejan el grado de influencia de Y con las componentes principales $Y_1, Y_2, \ldots, Y_n$. Tales fórmulas prueban que para cumplirse (1), Y debe estar muy correlacionada con las componentes principales de varianza menor que 1 (en análisis de componentes principales, las últimas componentes tienen varianza menor que 1). La desigualdad (1) proporciona un sencillo procedimiento para advertir al usuario que:

(a) La llamada regresión por componentes principales, donde las variables explicativas se sustituyen por componentes apropiadas, puede ser improcedente,

(b) en técnicas de ordenación y representación de datos a lo largo de los primeros ejes principales, los cuales se interpretan mediante correlaciones con variables externas (muy al uso en Ecología y Ciencias Sociales), puede resultar inapropiada esta práctica, pues si (1) es cierta, tales variables externas dependen sobretodo de los últimos ejes principales, que se ignoran por interpretarse como "ruido" o como parte de la variabilidad de las variables explicativas que no se considera en la representación.

3. AÑADIENDO VARIABLES CORRELACIONADAS NO NECESARIAMENTE SE MEJORA EL AJUSTE

Supongamos una regresión múltiple de Y sobre $X_1, \ldots, X_n$ y que Y está correlacionada positivamente con m de las n variables, siendo $m \leq n$, es decir,

$$\begin{aligned} r(Y, X_i) &> 0 \qquad i = 1, \ldots, m \\ r(Y, X_i) &= 0 \qquad i = m+1, \ldots, n \end{aligned}$$

Fijado n, parece natural que el coeficiente de determinación R^2 crezca con m, manteniendo fijas las correlaciones simples $r_{ij} = r(X_i, X_j)$, pues aumenta la relación global de las variables $\boldsymbol{x}$ sobre Y. Sin embargo no es así, como se desprende de una contribución de Tiit (1984):

Supongamos las variables $\boldsymbol{x}$ equicorrelacionadas

$$r(X_i, X_j) = c \qquad i, j = 1, \ldots, n$$

y que también las correlaciones positivas de Y con $X_1, \ldots, X_m$ son iguales

$$r(Y, X_i) = r \qquad i = 1, \ldots, m$$

Entonces R^2 viene dado por

$$R^2 = \frac{mr^2[1 + (n - m - 1)c]}{[1 + (n-1)c](1-c)} \tag{3}$$

que, curiosamente, no es una función estrictamente creciente en m. Por ejemplo, tomemos $c = 0.3$, $r = 0.5$, $n = 8$. La tabla 2 ilustra R^2 como una función de m.

Tabla 2

m	R^2	R
0	0	0
1	0.3226	0.5688
2	0.5760	0.7589
3	0.7604	0.8720
4	0.8756	0.9357
5	0.9217	0.9601
6	0.8988	0.9481
7	0.8065	0.8981
8	0.6452	0.8032

Podemos observar que $R^2 = 0.9217$ con $m = 5$ variables correlacionadas con Y; no obstante $R^2 = 0.6452$ con $m = 8$, es decir, con todas las n variables correlacionadas con Y. El paso de 5 a 8 variables correlacionadas (pero manteniendo $n = 8$), no mejora, sino que empeora sustancialmente, el ajuste de Y sobre las variables $X_1, \ldots, X_n$.

4. CURIOSO PROCEDIMIENTO PARA MEJORAR UNA ESTIMACIÓN: LA PARADOJA DE STEIN

Supongamos que $X_1, \ldots, X_n$ son variables normales independientes, donde cada X_i es $\mathrm{N}(\mu_i, 1)$. Dada una observación x_i de cada $\mathrm{N}(\mu_i, 1)$ parece razonable tomar como estimación de μ_i

$$\hat{\mu}_i = x_i \qquad i = 1, \ldots, n \tag{4}$$

Es bien conocida en la literatura la paradoja de Stein: la estimación $\hat{\mu}_i$ de μ_i es inadmisible bajo función de pérdida cuadrática. Es decir, si lo que se trata es minimizar

$$\sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)^2 \tag{5}$$

resulta que para $n \geq 3$, cualquier estimador de la forma

$$\hat{\mu}_i = \left(1 - \frac{c}{A}\right) x_i \tag{6}$$

produce una función de riesgo (= valor esperado de (5)) que es menor, siendo

$$A = \sum_{i=1}^{n} x_i^2$$

y c una constante tal que $0 < c < 2(n-2)$. El mejor valor para c es $n-2$. (6) es el llamado estimador de James-Stein. Una explicación a tal peculiaridad se encuentra postulando una distribución prior normal para μ, es decir, $\mu_1, \ldots, \mu_n$ es una "muestra" de una población normal $N(0, \tau^2)$. Tal argumento lleva al llamado estimador de Efron-Morris

$$\hat{\mu}_i = \overline{x} + \left(1 - \frac{c}{nS^2}\right)(x_i - \overline{x}) \tag{7}$$

que también mejora la función de riesgo para $n \geq 4$, siendo

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

y c tal que $0 < c < 2(n-3)$. El mejor valor para c es $n-3$.

Una explicación sencilla e interesante de la paradoja de James-Stein ha sido dada por Stigler (1990), en términos de la teoría de la regresión. Como dice Stigler: "información sobre el precio de las manzanas en Washington y sobre el precio de las naranjas en Florida puede ser utilizada para mejorar la estimación del precio del vino en Francia, cuando tales precios se suponen independientes".

El planteamiento de Stigler da una justificación clara de porqué debe ser $n \geq 3$. Adoptando una perspectiva galtoniana, el autor imagina una regresión de μ sobre x, es decir, ajustar una recta de regresión a la "nube" de puntos $(x_1, \mu_1), (x_2, \mu_2), \ldots, (x_n, \mu_n)$, donde obviamente los valores $\mu_1, \ldots, \mu_n$ son desconocidos.

La estimación de μ_i en tales términos es

$$\hat{\mu}_i = \overline{\mu} + \hat{\beta}(x_i - \overline{x})$$

Como la estimación obvia para $\overline{\mu}$ es la media $\overline{x}$, el problema se reduce a encontrar $\hat{\beta}$ en

$$\hat{\mu}_i = \overline{x} + \hat{\beta}(x_i - \overline{x}) \tag{8}$$

Suponiendo $\mu_1, \ldots, \mu_n$ distribuidos independientemente con alguna distribución con media μ y varianza finita, y que $x = \mu + e$, donde e es N(0,1) (esto es, los x_i

son medidas de μ con un error normal), entonces un fácil desarrollo prueba que

$$\hat{\beta} = \left(1 - \frac{n-1}{nS^2}\right)$$

y asi (8) se convierte en (7) para $c = n - 1$.

Si ahora suponemos que el término constante en la regresión de μ sobre x es 0, es decir, $\hat{\mu}_i = \hat{\beta} x_i$, entonces se obtiene

$$\hat{\beta} = \left(1 - \frac{n}{A}\right)$$

que es el estimador (6) para $c = n$. La figura 2 ilustra este planteamiento, bajo el cual, naturalmente, necesitamos tener $n > 2$ puntos.

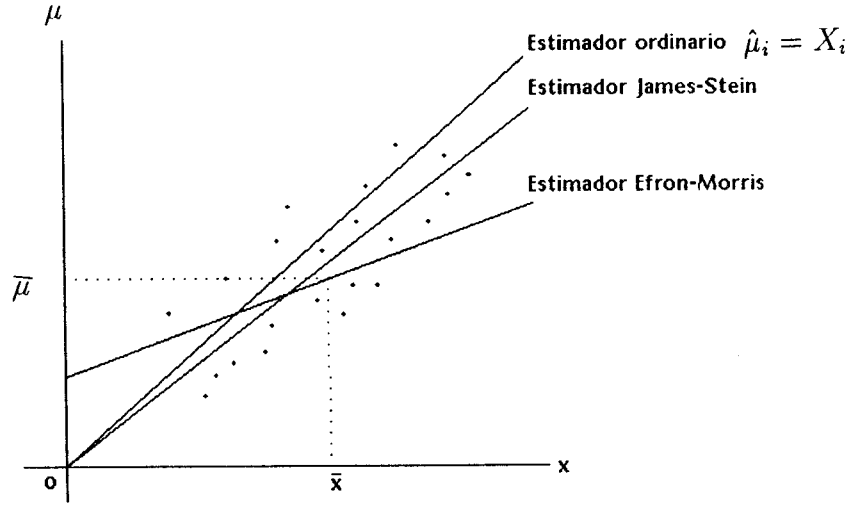

Figura 2. Ilustración de la paradoja de Stein en términos de la regresión de μ_i (parámetro desconocido) sobre x_i.

5. CÓMO DETECTAR CLUSTERS MEDIANTE VARIABLES INCORRELACIONADAS

Sean X_1, X_2 dos variables aleatorias con una distribución que podemos suponer normal (aunque no es necesario para el desarrollo que sigue). Supongamos

que π_1 y π_2 son dos poblaciones, y que X_1, X_2 están incorrelacionadas en cada una de las poblaciones

$$r(X_1, X_2) = 0 \qquad \text{en} \quad \pi_i \qquad i = 1, 2 \tag{9}$$

Si consideramos una población global $\pi = \pi_1 \cup \pi_2$, entonces podemos interpretar π_1 y π_2 como dos "clusters" diferentes. Supongamos que las proporciones de π_1 y π_2 son p y q respectivamente, siendo $p + q = 1$. Entonces, a pesar de la correlación nula en cada π_i, la correlación sobre π viene dada por

$$r(X_1, X_2) = \frac{pq(\mu_1 - \overline{\mu}_1)(\mu_2 - \overline{\mu}_2)}{a \cdot b} \quad \text{en } \pi = \pi_1 \cup \pi_2 \tag{10}$$

donde $\mu_1, \mu_2, \sigma_1, \sigma_2$ son las medias y desviaciones típicas en π_1 y $\overline{\mu}_1, \overline{\mu}_2, \overline{\sigma}_1, \overline{\sigma}_2$ son los mismos momentos en π_2. La desviación típica de X_1 es a siendo

$$a^2 = p\sigma_1^2 + q\overline{\sigma}_1^2 + pq(\mu_1 - \overline{\mu}_1)^2$$

y la desviación típica de X_2 es b, cuya expresión es análoga.

Pero ahora la correlación (10) ya no es cero (Figura 3), salvo que:

1) $p = 1$ y entonces tenemos un único cluster π_1.

2) $q = 1$, y el único cluster es π_2.

3) $\mu_i = \overline{\mu}_i$, y entonces los dos clusters se confunden en uno.

(El caso $\mu_1 = \overline{\mu}_1, \mu_2 \neq \overline{\mu}_2$ también anula (10), pero no lo discutimos aqui).

En otras palabras, si sabemos que dos variables están incorrelacionadas y procedemos a calcular la correlación sobre una amplia muestra de una población, resulta que si la correlación obtenida no es realmente cero, estamos frente a dos (o tal vez más) poblaciones o clusters.

Téngase en cuenta que el coeficiente de correlación es un concepto ligado a una *única población* y que cuando mezclamos varias poblaciones, con valores medios distintos para las variables, pueden aparecer falsas correlaciones. Y así ocurrió en un estudio realizado sobre una población germánica, donde se correlacionó la pigmentación de la piel con el índice encefálico, resultando una correlación notable que extrañó a los antropólogos. Pero un análisis posterior de los datos reveló que había en realidad dos poblaciones: una alpina, con mucha pigmentación y elevado índice encefálico, y otra nórdica, en la que ocurría lo contrario. La correlación, de por si baja en cada grupo, aumentaba artificialmente al considerar ambos conjuntamente.

Resumiendo, una correlación no nula o más alta de lo esperado puede indicarnos que existen dos o más clusters en nuestros dados.

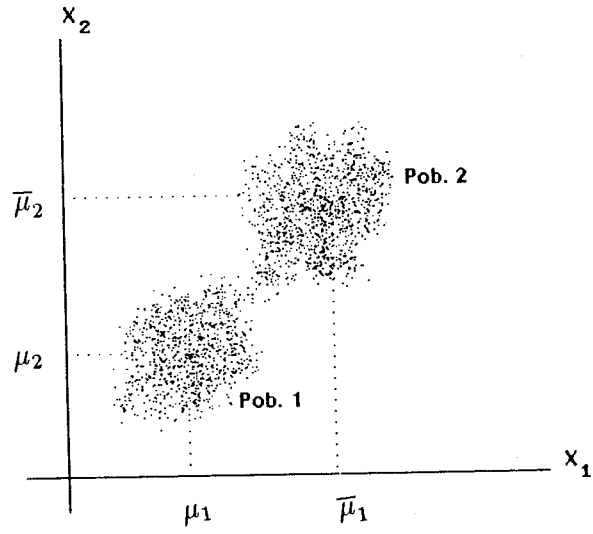

Figura 3. La correlación para las variables es cero en π_1 y en π_2, pero deja de ser cero si juntamos π_1 con π_2.

6. CÓMO CORRELACIONAR VARIABLES CON CUALQUIER DISTRIBUCIÓN MARGINAL

Supongamos que disponemos de dos muestras: $x_1, x_2, \ldots, x_n$ para la variable x, e $y_1, y_2, \ldots, y_n$ para la variable y. Ambas muestras, entendidas como distribuciones marginales, se representan a lo largo de los ejes 0X, 0Y de la figura 4.

¿Podemos construir una muestra bivariante

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

que nos dé un coeficiente de correlación r tal que $0 < r < 1$?

La respuesta es afirmativa. Empecemos suponiendo media 0 y varianza 1 y la misma distribución muestral, es decir:

$$x_1 = y_1,\ x_2 = y_2, \ldots, x_n = y_n \qquad \overline{x} = 0 \quad s_x = 1$$

y que $r = n_1/n$ (o bien n_1 es tal que la fracción n_1/n es muy próxima a r). Dividimos la muestra en n_1 y n_2 puntos, donde $n = n_1 + n_2$. Permutamos las x_i y las y_i para formar $x'_1, \ldots, x'_n$ e $y'_1, \ldots, y'_n$. A continuación situamos n_1 puntos sobre la recta $y = x$.

$$(x'_1, x'_1), (x'_2, x'_2), \ldots, (x'_{n_1}, x'_{n_1}) \tag{11}$$

de modo que la permutación se ha obtenido con la restricción

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(x'_1)^2 = 1$$

Los siguientes n_2 puntos se situan al azar, sin reflejar ninguna tendencia,

$$(x'_{n_1+1}, y'_{n_1+1}), \ldots, (x'_n, y'_n) \tag{12}$$

aunque suponiendo que

$$\sum_{i=n_1+1}^{n} x'_i y'_1 = 0$$

La correlación para la nube de puntos (11) es 1 y para la nube de puntos (12) es un valor que no influye. Es fácil ver entonces que la correlación de (11) + (12) es $r = n_1/n_2$.

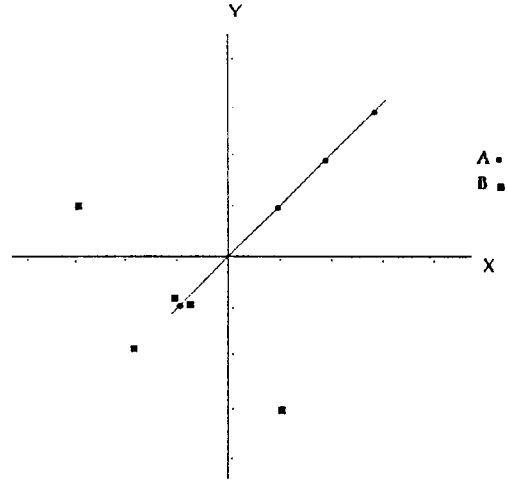

Figura 4. Representación de la muestra bivariante A: (1,1), (1,1), (2,2), (3,3), (-1,-1), y B: (-1,-1), (-1,-1), (-2,-2), (-3,1), (1,-3), construidas a partir de -3, -2, -1, -1, -1, 1, 1, 1, 2, 3 de modo que $r = 0.5$.

Como ilustramos en la figura 4, una proporción r de puntos A están sobre la recta $y = x$, mientras que otra proporción $1 - r$ de puntos B están distribuidos al azar. Al juntar ambas nubes de puntos obtenemos la correlación r.
Si las muestras $(x_1, \ldots, x_n), (y_1, \ldots, y_n)$ son distintas, tomaríamos entonces una proporción r de puntos (x, y) alineados y procederíamos análogamente con el resto. Sin embargo, el campo de posibles valores para r es más restringido.

Esta construcción es también válida para funciones de distribución marginales cualesquiera (la condición de media 0 y varianza 1 no es restrictiva).

Sean F_X, F_Y las funciones de distribución marginales. Si $F_X = F_Y$, basta tomar la distribución bivariante (X, X) con probabilidad r y la distribución bivariante (X, Y), donde Y es independiente de X, con probabilidad $1 - r$. La distribución conjunta obtenida es tal que la correlación global es igual a r.

Si la distribución es simétrica (X y $-X$ están igualmente distribuidas), la construcción que dé lugar a una correlación r negativa es análoga. Basta tomar $(X, -X)$ con probabilidad $|r|$. Pero si $F_X \neq F_Y$, la construcción es más complicada (ver Cuadras, 1991, pág. 55; Cuadras, 1992b).

Resumiendo, dada la distribución marginal de X y la distribución marginal de Y, que pueden ser cualesquiera, y un coeficiente de correlación r (cuyo valor máximo dependerá de la naturaleza de las marginales) es posible obtener una distribución bivariante con tales marginales dadas, proporcionando una correlación r entre X e Y.

7. BIBLIOGRAFÍA

[1] **Cuadras, C.M.** (1991). *Métodos de Análisis Multivariante.* Barcelona: PPU.

[2] **Cuadras, C.M.** (1992a). "Interpreting an inequality in multiple regression". *Manuscript.*

[3] **Cuadras, C.M.** (1992b). "Probability distributions with given multivariate marginals and given dependence structure". *J. of Multivariate Analysis,* **41**, in press.

[4] **Hamilton, D.** (1987). "Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Correlated variables are not always redundant". *The American Statistician,* **41**, 129–132.

[5] **Routledge, R.D.** (1990). "When stepwise regression fails: correlated variables some of which are redundant". *Int. J. Math. Educ. Sci. Technol.,* **21**, 403–410.

[6] **Rios, S.** (1991). *Iniciación Estadística.* Madrid: Paraninfo.
[7] **Stigler, S.M.** (1990). "The 1988 Neyman Memorial Lecture: A Galtonian perspective on shrinkage estimators". *Statistical Science,* **5**, 147–155.
[8] **Tiit, E.M.** (1984). "Formal computation of regression parameters". *Proceedings Sixth Symposium* COMPSTAT 1984 (T. Havranek, Z. Sidak, M. Novak, eds.), 497–502, Physica-Verlag, Vienna.

ENGLISH SUMMARY:

UNUSUAL EXAMPLES AND APPLICATIONS IN REGRESSION AND CORRELATION

C. M. Cuadras

1. INTRODUCTION

In the deterministic applications of Mathematics, a non-linear relationship between two sets of variables x and y, say, is set up by means of a function f. Usually, the only difficulties found are related to the type of function (variability field, derivatives, singularity points).

In statistics, nevertheless, one relates random (or statistical) variables and, because the true model is not always known, a linear model like

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

is used, the parameters are estimated and the multiple correlation coefficient R is computed as a measure of linear fitting, or R^2, the coefficient of determination, which measures the degree of variability of y which is accounted for by x variables.

In spite of the simplicity of this model, many complications and paradoxes have arisen since the first applications of regression techniques (Galton's regression to the mean, the spurious correlation, etc.)

In this paper some unusual and relatively new applications and paradoxes are presented, in order to motivate both the teacher and the user of statistics.

2. CORRELATED VARIABLES ARE NOT ALWAYS REDUNDANT

Let Y and $X_1, \ldots, X_n$ be a response variable and n explanatory variables, respectively. Let us indicate by $r_1 = r(Y, X_1), \ldots, r_n = r(Y, X_n)$, the simple correlations between Y and $X_1, \ldots, X_n$.

In the applications, one hopes that R^2 will be larger than $r_1^2 + \cdots + r_n^2$, reflecting the fact that correlated variables contain only redundant information about Y. However, this belief is erroneous. It is possible to find theoretical and practical examples where the opposite inequality (1) is accomplished (see Table 1).

In the case $n = 2$, Hamilton (1987) provides a convincing interpretation of inequality (1), showing that (1) is equivalent to (2), where r_{12} is the simple correlation between X_1 and X_2. Inequality (2) shows that the disturbing inequality (1) arises when r_{12} is high (multicollinearity) and $r_1^2 + r_2^2$ is at the same time small. As Hamilton points out, statistics teachers should not suggest that correlated explanatory variables are always redundant and should make it clear that small simple correlations or $x - y$ scatterplots (showing a poor relation) are dangerous ways of discarding variables.

Routledge (1990) shows that inequality (1) is satisfied when the response variable Y is highly correlated with the second principal component obtained from the x variables, that is, when Y depends on the "residual variability" of the explanatory variables (see Fig. 1).

The case $n > 2$ can be studied by relating the response variable Y to the principal components with small variance.

3. ADDING CORRELATED VARIABLES DOES NOT NECESSARILY IMPROVE THE FIT

Suppose that Y is equally correlated with $X_1, \ldots, X_n$ and uncorrelated with $X_{m+1}, \ldots, X_n$, where $m \leq n$. When n is fixed, one has the natural belief that by increasing m (the number of correlated variables), the coefficient of determination R^2 also increases. However, this is not always true, as proved by Tiit (1984). When the explanatory variables are also equicorrelated, then R^2 is given by (3), which is not a strict increasing function of m.

For $n = 8$, Table 2 shows that R^2 is larger when $m = 5$ variables are considered than when all $n = 8$ variables are taken into account in the multiple regression.

4. A CURIOUS PROCEDURE TO IMPROVE AN ESTIMATE: THE PARADOX OF STEIN

Let $X_1, \ldots, X_n$ be independent random variables, each X_i distributed as $N(\mu_i, 1)$, $i = 1, \ldots, n$. When an observation x_i of $N(\mu_i, 1)$ is obtained, a fair estimate of μ_i is given by (4). However, C. Stein showed that this "ordinary" estimator is inadmissible. It turns out that, for $n \geq 3$, any estimator of the form (6), called the James-Stein estimator, improves upon the previous one.

Stein, James, Lindley and others give explanations of this surprising fact. An interesting approach, from "the Bayesian point of view" leads to estimator (7), called the Efron-Morris estimator.

Why $n \geq 3$? Stigler (1990) provides a very interesting clarification from the Galtonian perspective, that is, from the point of view of regression. He imagines a regression of μ on x, by fitting a regression line to the "points" $(x_1, \mu_1), \ldots, (x_n, \mu_n)$, where, of course, $\mu_1, \ldots, \mu_n$ are unknown. This approach, after some easy computations, leads to (8), which yields equation (7) for a suitable $\hat{\beta}$ and, therefore, the Efron-Morris estimator is found. In addition, if a regression line with zero intercept $y = \beta x$, is considered, then the James-Stein estimator is obtained.

An illustration is given in Fig. 2, where we can immediately understand that $n > 2$ is necessary to perform a regression.

5. RECOGNIZING CLUSTERS BY MEANS OF UNCORRELATED VARIABLES

Let X_1, X_2 be two uncorrelated random variables defined on a population π_1, and assume the same for another population π_2, that is, (9) holds.

By considering a "global" population $\pi = \pi_1 \cup \pi_2$, where π_1 and π_2 are understood as two different clusters, then, in spite of the zero correlation inside

each π_i, the global correlation is given by (10), which is not zero, in general, when the two clusters are really different.

In other words, if it is known that two random variables are uncorrelated, it turns out that when we find a non-zero correlation, we could acknowledge that we really have two or more clusters (see Fig. 3).

6. HOW TO CORRELATE VARIABLES WITH ANY MARGINAL DISTRIBUTION

Suppose that $x_1, x_2, \ldots, x_n$ is a sample of a variable X and $y_1, y_2, \ldots, y_n$ is a sample of a variable Y.

It is possible to obtain a bivariate sample

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

such that the sample correlation between X and Y is as close as possible to a given $0 < r < 1$. Let us write $r = n_1/n$ and $n = n_1 + n_2$. Next, we set n_1 points on the straight line $y = x$ and the remaining points at random, but preserving the marginal distributions. Then we find correlation 1 for n_1 points and correlation 0 for n_2 points, where $n = n_1 + n_2$. It can be proved (by imposing some conditions) that the correlation obtained by gathering the n points is equal to r (see Fig. 4).

This procedure can be extended to any random variables X and Y with probability distribution functions F_X and F_Y, respectively. That is, it is possible to construct a bivariate distribution, where the correlation r (satisfying some restrictions) is given and F_X, F_Y are also given.