Non-Normal Bivariate Distributions with Normal Marginals

CHARLES J. KOWALSKI*

Examples of non-normal bivariate (multivariate) distributions with normal marginals [15, 16, 18] can be used in the classroom to (a) contrast the correlation/regression structures of these distributions with that of the corresponding bivariate normal model [9, 10, 11, 12], (b) investigate the feasibility of employing coordinate transformations to normality as a prelude to analyses based on the assumption of joint normality [2, 13, 15, 16, 17, 18, 27] and (c) motivate the need for the development of multidimensional goodness-of-fit tests [2, 13, 16]. In addition, since most of the examples presented here are concerned with ways to generate distributions with *arbitrary* marginals, they can also be used to obtain a wide variety of non-normal bivariate (multivariate) distributions for these, and other, purposes [2, 11, 12, 13].

Fréchet's Class of Bivariate Distributions

Given the one-dimensional distribution functions F and G, a family of bivariate distributions having F and G as marginals has been defined by Fréchet [6] as

$$H(x, y) = w_1 H_0(x, y) + w_2 H_1(x, y), \qquad (1)$$

where $w_i \geqq 0$, $w_1 + w_2 = 1$, and

$$H(x, y) = \max\{0, F(x) + G(y) - 1\}, \qquad (2)$$

$$H(x, y) = \min\{F(x), G(y)\}. \qquad (3)$$

Both H_0 and H_1 are distributions with marginals F and G, and (1) is the family of mixtures of H_0 and H_1. Feller [5, p. 162] discusses example (3), which is the special case of (1) when $w_1 = 0$. It can be shown that H_1 is concentrated on the curve $F(x) = G(y)$ and is therefore singular. Similarly, H_0 concentrated on $F(x) + G(y) = 1$ is singular.

The special case of (standard) normal marginals corresponds to the case when $F = G = \Phi$ in the above expressions where, here and in the sequel, Φ denotes the standardized normal, $N(0, 1)$, distribution. The resulting distributions H, H_0 and H_1 have normal marginals, but the joint density of the variables does not exist. Thus there exist distributions with normal marginals for which the joint density of the variables is not defined [26]. If H is *any* bivariate distribution with marginals F and G it can be shown [6] that

$$H_0(x, y) \leqq H(x, y) \leqq H_1(x, y), \qquad (4)$$

a result of considerable theoretical interest. (4) can be used to gain some insight into the question of how non-normal a bivariate distribution with normal marginals can be. If H^* is any bivariate distribution with (standard) normal marginals,

$$\max\{0, \Phi(x) + \Phi(y) - 1\} \leqq H^*(x, y) \leqq \min\{\Phi(x), \Phi(y)\}, \qquad (5)$$

and $0 \leqq H^*(0, 0) \leqq 0.5$. Hence marginal normality does not restrict the form of H^* to any great extent in the region of highest probability density (assuming this joint density exists). One may also note that the bounds (5) cannot be improved in general since they are attained in the system (1) when $F = G = \Phi$.

Mixtures of Bivariate Normal Distributions

The bivariate density function

$$h(x, y) = w_1\phi_1(x, y) + w_2\phi_2(x, y), \qquad (6)$$

where the w's are probability weights and ϕ_i is the standard bivariate normal density function with correlation coefficient ρ_i, has normal marginals, but (6) is non-normal if $\rho_1 \neq \rho_2$. The random variables X and Y with density (6) have means 0, variances 1 and correlation $\rho = w_1\rho_1 + w_2\rho_2$, and it may be of interest to compare (6) with the bivariate normal density with these parameters, denoted here by ϕ.

The curves $h(x, y) = c$ are different from the elliptical equi-probable contours $\phi(x, y) = c$, and the departure may be considerable [18] for a proper choice of $(w_1, w_2, \rho_1, \rho_2)$. The conditional density of X given $Y = y$ in (6) is the mixture of the $N(\rho_1 y, 1 - \rho_1^2)$ and $N(\rho_2 y, 1 - \rho_2^2)$ densities with the same weights as in (6), and hence may differ considerably from the normal density with the same location and scale parameters. The conditional mean and variance are

$$E_h(X \mid y) = (w_1\rho_1 + w_2\rho_2)y = \rho y,$$

and

$$V_h(X \mid y) = 1 - (w_1\rho_1^2 + w_2\rho_2^2) + [(w_1\rho_1^2 + w_2\rho_2^2) - \rho^2]y^2,$$

so that $E_h = E_\phi$, but $V_h = V_\phi$ only when $|y| = 1$. The regression of X on Y in (6) is linear but not homoscedastic, $V_h(X \mid y)$ being a quadratic function of y.

The correlation coefficient of X and Y with density (6) is $\rho = w_1\rho_1 + w_2\rho_2$, and this may be zero though the (marginally normal) variables X and Y are not independent [19]. This example illustrates the fact that linearity of regression is not a sufficient condition for the "goodness" of ρ as a measure of association.

Morgenstern's Bivariate Normal Distribution

This distribution is defined [23] by the distribution function

* Statistical Research Lab., Univ. of Michigan, 106 Rackham Bldg., Ann Arbor, Mich. 48104.

$$H(x, y) = \Phi(x)\Phi(y)\{1 + \alpha[1 - \Phi(x)][1 - \Phi(y)]\}, \tag{7}$$

where $|\alpha| \leqq 1$ is a parameter measuring association. Gumbel studied distributions of the same structure, but having exponential [11] and logistic [12] marginals. We note that $\alpha = 0$ corresponds to independence of X and Y, and that neither of Fréchet's boundary distribution functions are members of the family (7). The density corresponding to (7) is

$$h(x, y) = \phi(x)\phi(y)\{1 + \alpha[2\Phi(x) - 1][2\Phi(y) - 1]\}, \tag{8}$$

and the conditional density of X given $Y = y$ is

$$h(x \mid y) = \phi(x)\{1 + \alpha[2\Phi(x) - 1][2\Phi(y) - 1]\}. \tag{9}$$

The corresponding conditional expectation

$$E(X \mid y) = \alpha[2\Phi(y) - 1]\pi^{1/2} \tag{10}$$

is not a linear function of y, being linear instead in $\Phi(y)$, which shows that (10) can be linearized by use of normal probability paper.

The correlation coefficient is $\rho = \alpha/\pi$, and, since $|\alpha| \leqq 1$, ρ can assume values only in the range $|\rho| \leqq 1/\pi$ rather than the usual one $|\rho| \leqq 1$. This also follows from the more general observation that if $H_0(x, y) \leqq H(x, y) \leqq H_1(x, y)$ for all x, y then $\rho_0 \leqq \rho \leqq \rho_1$, using the obvious notation.

Farlie's Bivariate Normal Distribution

Farlie's family is a direct generalization of Morgenstern's family (7). Farlier [3, 4] defined

$$H(x, y) = \Phi(x)\Phi(y)\{1 + \alpha A[\Phi(x)]B[\Phi(y)]\}, \tag{11}$$

where the derivatives $d(\Phi A)/d\Phi$, $d(\Phi B)/d\Phi$ and the functions A, B are bounded with $A(1) = B(1) = 0$. Farlie noted that (11) was approximately bivariate normal with $\rho = \alpha$ for small values of α and x and y "not too large," provided that A and B satisfied the differential equations

$$d(\Phi A)/d\Phi = x, \qquad d(\Phi B)/d\Phi = y. \tag{12}$$

Since the derivatives (12) are unbounded, Farlie's requirement that x and y be "not too large" amounts to a truncation of the bivariate normal model.

Farlie [3, 4] examined the correlation structure of (11), studying the performance of various correlation coefficients (Spearman's rank correlation coefficient, Kendall's τ, the product-moment correlation coefficient and the probability of concordance) in his model. He gave conditions (on A and B) under which one or another of these coefficients provides an efficient test for association. In particular, he showed that the product-moment correlation coefficient, r, provides a fully efficient test for association only when the "disturbing" functions A and B are of the form

$$A[\Phi(x)] = \int_{-\infty}^{x} x\, d\Phi/\Phi(x),$$

$$B[\Phi(x)] = \int_{-\infty}^{y} y\, d\Phi/\Phi(y). \tag{13}$$

In the Morgenstern family (7), where $A = B = 1 - \Phi$ and (13) is not satisfied, rank correlation methods are more efficient than tests based on r.

The density corresponding to (11) is of the form

$$h(x, y) = \phi(x)\phi(y)\{1 + [\alpha d(\Phi A)/d\Phi][d(\Phi B)/d\Phi]\}. \tag{14}$$

When (12) is satisfied, this reduces to $h(x, y) = \phi(x)\phi(y)[1 + \alpha xy]$, which is nearly normal only for small α, x and y, as mentioned earlier. The corresponding conditional density of X given $Y = y$ is

$$h(x \mid y) = \phi(x)[1 + \alpha xy],$$

with $E(X \mid y) = \alpha y$, $V(X \mid y) = 1 - \alpha^2 y^2$, so that the regression is linear but not homoscedastic. The departure from normal theory in this case is, however, slight as long as α, x and y are small.

In the general case (14) the regression function is $E(X \mid y) = k\alpha d(\Phi B)/d\Phi$, where k is a constant. It is possible to choose B to obtain quadratic, triangular, and discontinuous regression functions [3, 4].

Plackett's Bivariate Normal Distribution

This distribution was introduced by Plackett [24] and was subsequently studied by Mardia [20, 21, 22] and Steck [25]. The distribution function H was defined by Plackett as the solution of the equation

$$[\Phi(x) - H][\Phi(y) - H] = \psi[1 - \Phi(x) - \Phi(y) + H]H \tag{15}$$

which satisfied the condition (5). He showed that for any fixed $\psi > 0$, (15) has a single solution satisfying (5) which is a valid joint distribution function with normal marginals. This family includes Frechet's boundary distribution functions, and is the only known family of bivariate distributions with normal marginals having this property and containing $\Phi(x)\Phi(y)$. When $\psi = 0$ the distribution is concentrated on $x = -y$, corresponding to (3). Independence of X, Y corresponds to $\psi = 1$. When $\psi = \infty$, the distribution is concentrated on $x = y$.

Mardia [20] gave the solution of (15), and the corresponding density function was studied by Mardia [20, 21, 22] and Steck [25]. Plackett compared the distribution defined by (15) with the normal distribution function $\Phi(x, y \mid \rho)$ with

$$\rho = -\cos[\pi\psi^{1/2}/(1 + \psi^{1/2})],$$

and found agreement over much of the xy-plane to within a few units of the third decimal place. Mardia showed that with

$$\rho = 2\sin[\pi(\psi^2 - 1 - 2\psi\ln\psi)/6(\psi - 1)^2]$$

the agreement was even better, with the exception of a neighborhood of the origin. The conditional distributions also agree closely with normal theory [24], but the correlation coefficient has the complicated form

$$\rho(\psi) = \frac{1}{\psi - 1} \int_0^\infty \int_0^\infty \{\psi + 1 - A[u(x), u(y)] - A[1 - u(x), u(y)]\}\, dx\, dy,$$

where

$$u(x) = (2\pi)^{-1/2} \int_0^x \exp(-t^2/2)\, dt,$$

and

$$A(u, v) = \{\psi + [(\psi - 1)(u + v)]^2 - 4\psi(\psi - 1)uv\}^{1/2}.$$

It can be shown that $\rho(\psi) \lesseqqgtr 0$ according as $\psi \lesseqqgtr 1$, that $\rho(\psi) = -\rho(1/\psi)$, and that $\rho(0) = -1$, $\rho(\infty) = 1$. Mardia and Steck also considered the moments of a general (arbitrary marginals) distribution of this form.

Vaswani's Bivariate Normal Distribution

Suppose that T is uniformly distributed on the interval $[0, 1]$ and that X and Y are such that $X = \Phi^{-1}(T)$ with

$$Y = \begin{cases} -\Phi^{-1}(T + \tfrac{1}{2}), & \text{for } 0 \leqq T \leqq \tfrac{1}{2}, \\ -\Phi^{-1}(T - \tfrac{1}{2}), & \text{for } \tfrac{1}{2} < T \leqq 1. \end{cases} \tag{16}$$

where Φ^{-1} is the inverse of the standard normal distribution function. Then X and Y are both $N(0, 1)$ [28, 29]. The graph of Y against X resembles the two branches of the hyperbola $XY =$ constant. Since $P(XY > 0) = 1$, $\rho > 0$; in fact, $\rho = 0.3225$, even though Y decreases as X increases and conversely. The distribution (16) is singular, but Vaswani pointed out that a non-singular one with the same properties could be constructed by considering instead the joint distribution of X and $Y + \epsilon$, where ϵ is a normally distributed "error" independent of X and Y.

Other Miscellaneous Examples

Anderson [1, p. 37] considers a bivariate normal distribution and shifts a probability mass from Quadrant IV to Quadrant I and the same from Quadrant II to Quadrant III. This is done in such a way as to preserve marginal normality, but the resulting bivariate distribution is clearly non-normal. Variations on this theme were given by Freund [7, p. 306] and Kendall and Stuart [14, p. 368].

Feller [5, p. 99] considers an odd, continuous function $u(\cdot)$, vanishing outside the unit interval and such that $|u(\cdot)| \leqq (2\pi e)^{-1/2}$. The density $h(x, y) = \phi(x)\phi(y) + u(x)u(y)$ is non-normal, but has normal marginals. This example is a variant of the Morgenstern and Farlie families.

Gumbel [11] considers the family defined implicitly by the equation

$$\{-\ln H(x, y)\}^m = \{-\ln \Phi(x)\}^m + \{-\ln \Phi(y)\}^m, \tag{17}$$

where $m \geqq 1$. When $m = 1$, the joint distribution reduces to the product $\Phi(x)\Phi(y)$. As m increases, $H(x, y)$ approaches H_1, as given in (3) with $F = G = \Phi$. Thus for the system (17) the bounds (5) may be sharpened to give $\Phi(x)\Phi(y) \leqq H(x, y) \leqq \min\{\Phi(x), \Phi(y)\}$ and, hence, the correlation coefficient is never negative. The expression (17) is so cumbersome as to limit its utilization for most practical purposes; Gumbel [8] has indicated, however, that it may be useful in the study of bivariate extreme value theory.

Concluding Remarks

We have presented a list of the more widely known examples of bivariate distributions with arbitrary marginals, specialized these families to the normal-marginal case, and thus obtained a number of examples of non-normal bivariate systems with normal marginals. Other examples can be generated, e.g., by mixing, shifting appropriate probability masses, or choosing different functions A and B in (11). This list illustrates several kinds of departures from bivariate normality when the marginal distributions are normal, and it should be useful for a variety of purposes in the classroom.

Acknowledgments

I would like to thank Samuel Kotz for suggesting that I publish this collection of examples, and the referees, who made many valuable suggestions facilitating the revision of the original manuscript.

REFERENCES

[1] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York (1958).

[2] Andrews, D. F., Gnanadesikan, R. and Warner, J. L. Transformations of multivariate data. *Biometrics* 27 (1971): 825–40.

[3] Farlie, D. J. G. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 47 (1960): 307–23.

[4] Farlie, D. J. G. The asymptotic efficiency of Daniels' generalized correlation coefficient. *J. Royal Statist. Soc., B,* 23 (1961): 128–42.

[5] Feller, W. *An Introduction to Probability Theory and Its Applications.* Vol. II. Wiley, New York (1966).

[6] Fréchet, M. Sur les tableaux de corrélation dont les marges sont donées. *Ann. Univ. de Lyon* 14 (1951): 53–77.

[7] Freund, J. E. *Mathematical Statistics.* Prentice-Hall, Englewood Cliffs, N.J. (1962).

[8] Gumbel, E. J. *Statistics of Extremes.* Columbia Univ. Press, New York (1958).

[9] Gumbel, E. J. Distributions á plusieurs variables dont les marges sont données. *Comptes Rendus de l'Académie des Sciences,* Paris, 246 (1958): 2717.

[10] Gumbel, E. J. Multivariate distributions with given marginals. *Revista da Faculdade de Ciencias,* Lisbon, 7 (1959): 179–218.

[11] Gumbel, E. J. Bivariate exponential distributions. *J. American Statist. Assn.* 55 (1960): 698–707.

[12] Gumbel, E. J. Bivariate logistic distributions. *J. American can Statist. Assn.* 56 (1961): 335–49.
[13] Kempthorne, O. Multivariate responses in comparative experiments. In: *Multivariate Analysis*, P. R. Krishnaiah. Academic Press, New York (1966): 521–40.
[14] Kendall, M. G. and Stuart, A. *The Advanced Theory of Statistics*. Vol. I. Griffin, London (1958).
[15] Kowalski, C. J. and Tarter, M. E. Co-ordinate transformations to normality and the power of normal tests for independence. *Biometrika* 56 (1969): 139–48.
[16] Kowalski, C. J. The performance of some rough tests for bivariate normality before and after coordinate transformations to normality. *Technometrics* 12 (1970): 517–44.
[17] Kowalski, C. J. The OC and ASN functions of some SPRT's for the correlation coefficient. *Technometrics* 13 (1971): 883–41.
[18] Kowalski, C. J. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statist.* 21 (1972): 1–12.
[19] Lancaster, H. O. Zero correlation and independence. *Australian Journal Statist.* 1 (1959): 53–6.
[20] Mardia, K. V. Some contributions to contingency-type bivariate distributions. *Biometrika* 54 (1967): 235–49.
[21] Mardia, K. V. The performance of some tests of independence for contingency-type bivariate distributions. *Biometrika* 56 (1969): 449–51.
[22] Mardia, K. V. *Families of Bivariate Distributions*. Griffin, London (1970).
[23] Morgenstern, D. Einfache Beispiele zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematische Statistik* 8 (1956): 234–5.
[24] Plackett, R. L. A class of bivariate distributions. *J. American Statist. Assn.* 60 (1965): 516–22.
[25] Steck, G. P. A note on contingency-type bivariate distributions. *Biometrika* 55 (1968): 262–4.
[26] Strassen, V. The existence of probability measures with given marginals. *Ann. Math. Statist.* 36 (1965): 423–39.
[27] Tarter, M. E. and Kowalski, C. J. A new test for, and class of transformations to, normality. *Technometrics* 14 (1972): 735–44.
[28] Vaswani, S. P. A pitfall in correlation theory. *Nature* 160 (1947): 405–6.
[29] Vaswani, S. P. Assumptions underlying the use of the tetrachoric correlation coefficient. *Sankhyā* 10 (1950): 269–76.

Goodness of Fit in Generalized Least Squares Estimation

A. BUSE*

Standard econometric textbooks do not develop a goodness of fit statistic when the generalized least square estimation procedure is developed.[1] Given that generalized least squares (GLS) is a form of weighted least squares one would expect that the various results on partitioning of sums of squares which are used in the definition of R^2 with ordinary least squares have parallel expressions in terms of weighted variables. Such is indeed the case and this is demonstrated below. The relevance of the generalized R^2 for applied work is examined and note is made of some not uncommon pitfalls that occur when the data are transformed for ordinary least squares estimation.

Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1}$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations on the dependent variable, $\mathbf{X}$ is an $(n \times k)$ matrix of observations on k explanatory variables which is fixed in repeated samples and has rank $k \leq n$, $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown parameters and $\mathbf{u}$ is an $(n \times 1)$ vector of disturbances. The disturbance vector is assumed to have the following properties

$$E(\mathbf{u}) = \mathbf{0}, \tag{2}$$

$$E(\mathbf{uu}') = \mathbf{V}, \tag{3}$$

where $\mathbf{V}$ is a known positive definite matrix. If GLS is applied to (1) it is not difficult to show that

$$\mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y} = \hat{\mathbf{Y}}'\mathbf{V}^{-1}\hat{\mathbf{Y}} + \hat{\mathbf{u}}'\mathbf{V}^{-1}\hat{\mathbf{u}}, \tag{4}$$

where

$$\hat{\mathbf{u}} = \mathbf{Y} - \hat{\mathbf{Y}},$$

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

This is the usual partitioning of sums of squares in the linear model but the results are in terms of generalized sums of squares.

The normal equations associated with the estimator of $\boldsymbol{\beta}$ are defined by

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{5}$$

which if multiplied out give

$$\begin{pmatrix} \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_2 \cdots \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_k \\ \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2 \cdots \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_k \\ \cdots & \cdots \\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_2 \cdots \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_k \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \cdot \\ \cdot \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{Y} \\ \mathbf{X}_2'\mathbf{V}^{-1}\mathbf{Y} \\ \vdots \\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{Y} \end{pmatrix}, \tag{6}$$

where the matrix $\mathbf{X}$ has been partitioned by columns

$$\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2 : \cdots : \mathbf{X}_k) \tag{7}$$

The jth equation of (6) gives

$$(\mathbf{X}_j'\mathbf{V}^{-1}\mathbf{X}_1)\hat{\beta}_1 + (\mathbf{X}_j'\mathbf{V}^{-1}\mathbf{X}_2)\hat{\beta}_2 + \cdots$$

* Dept. of Economics, Univ. of Alberta, Edmonton, Canada. This note evolved out of a technical report on qualitative dependent variables financed by the Alberta Human Resources Research Council. Errors and ambiguities should be attributed to the author and not Don Hester who, on very short notice, commented on the penultimate draft of this note.

[1] See for example Goldberger [2].