DEFINING THE GENERAL TRANSFORMATION TO NORMALITY: A PROPOSAL TO CORRELATE GENERAL NONNORMAL DISTRIBUTIONS

CHARLES B. MOSS

Abstract. Non-normality may be important in several instances in agricultural economics such as the valuation of crop insurance. This paper develops an extension of the inverse hyperbolic sine transformation to normality for modeling correlated non-normal variables. To demonstrate the overall technique, the paper estimates the non-normal transformation for crops in North Florida.

1. Introduction

The concept of transforming a random variable into normality using a flexible mapping function is not new to agricultural economics. Moss and Shonkwiler [2] use an inverse hyperbolic sine transformation to model nonnormality in corn yields using a stochastic trend model to model the changes in the mean of the yield distribution over time. More interestingly for our discussion here, Ramirez, Moss and Boggess [3] use the same transformation to model correlation among potentially nonnormal random variables. Both of these studies use a generalization of the inverse hyperbolic sine transformation introduced by Burbidge, Magee, and Robb [1]. Burbidge, Magee and Robb propose using the inverse hyperbolic sine to reduce the effect of outliers. This concept carries into the applications to model nonnormality in that the inverse hyperbolic sine transformation only admits leptokurtotic distributions (fat tails). In its original specification, the inverse hyperbolic sine transformation corrected for kurtosis, but did not modify skewness. The modification suggested by Moss and Shonkwiler introduced an additional parameter which allowed the distribution to be either positively or negatively skewed.

While the inverse hyperbolic sine transformation has several valuable properties, it is but one of an infinite class of valid transformation to normality. Specifically, any monotonic transformation can be used to transform one distribution into another distribution. This study examines a fairly general approach to define such a transformation. We develop a methodology for defining transformations to normality for the ease of modeling the correlation between potentially nonnormal random variables

Date: March 15, 2002.
1991 *Mathematics Subject Classification.* Primary 62E17, 15A15; Secondary 91B82, 91B30.
Key words and phrases. mathematical statistics, non-normality, transformation to normality.
The participants of SCC-76 Economics and Management of Risk in Agriculture and Natural Resources for their suggestions on a previous draft of this manuscript.

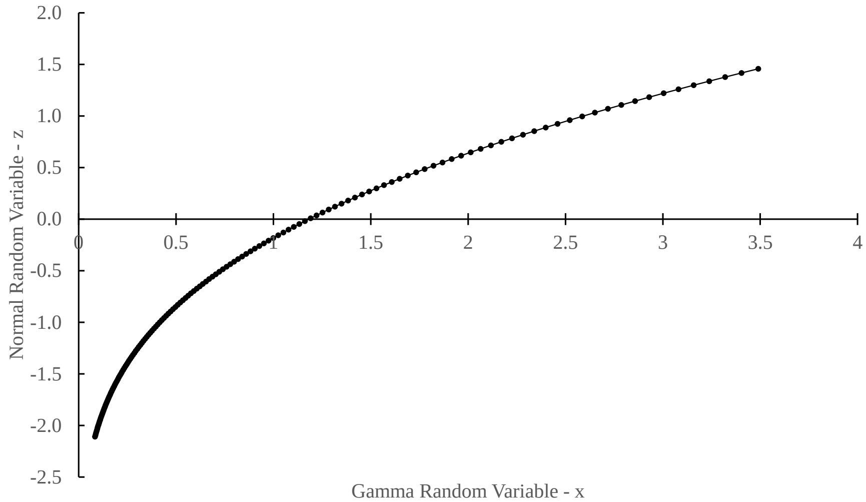

Figure 1. General Transformation between Gamma and Normality

2. A Heuristic Example

As a point of introduction, let us assume that we have two random variables with vastly different distributions a Gamma distribution and a Beta distribution. In addition, assume that we believe that the random variables are correlated and that this correlation is important for economic reasons. Maybe the distributions represent crop returns that the farmer can use to diversify risk. The concept is to develop a general approximation to each distribution based on a linear transformation to normality.

As stated above, the inverse hyperbolic sine transformation is but one of an infinite number of valid monotonic transformations. An exhaustive search is valid transformations is impossible, so another approach is to transform the data into a space to facilitate our search. I propose plotting the values of the transformed variables that yield the same probability. For example, I assume that given $x \sim \Gamma[\alpha, \beta]$ there exists a $z \sim \mathrm{N}[0, 1]$ that yields the same probability. In our example, I assume that $x \sim \Gamma[1.5, 1.0]$. For any x drawn from this distribution I can define

$$F^* = \int_0^x f\left(x|\,\alpha, \beta\right) dx \tag{2.1}$$

where $f\left(x|\,\alpha, \beta\right)$ is the probability density function for the Gamma distribution. Using this value it is possible to derive

$$z = G^{-1}\left(F^*\right) \ni: G\left(z\right) = \int_{-\infty}^{z} g\left(z|\,\mu, \sigma^2\right) dz \tag{2.2}$$

where $g\left(z|\,\mu, \sigma^2\right)$ is the probability density function for the normal distribution (initially we assume a standard normal distribution). Figure 1 presents the general form of this transformation to the standard normal distribution.

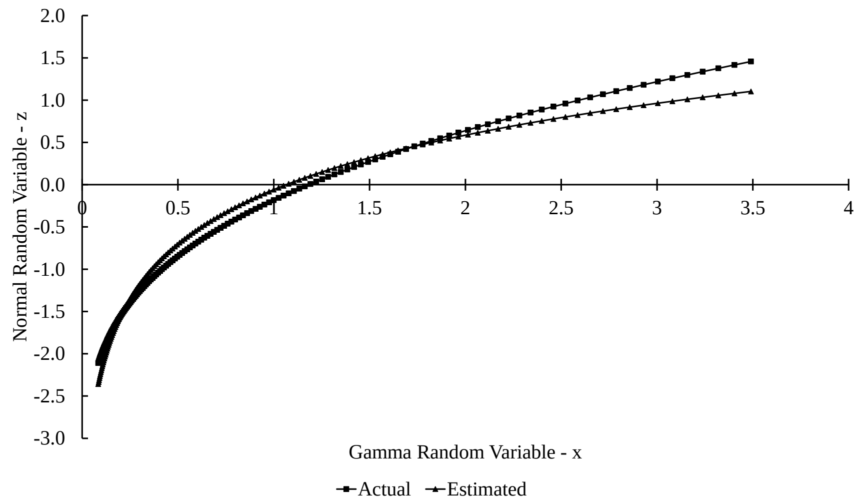

FIGURE 2. Comparison of Actual Mapping with the Estimated Mapping

The next step is to estimate a general monotonic mapping between the variables. In this case, a variant of the natural logarithm would seem appropriate

$$z = \gamma_0 + \gamma_1 \ln(x) \tag{2.3}$$

Estimating this transformation with ordinary least squares yields $\hat{\gamma}_0 = -0.0590$ (0.0129) and $\hat{\gamma}_1 = -0.9290$ (0.0103) (where the numbers in parentheses denote standard errors). This approximation is presented graphically in Figure 2. Figure 3 presents the true Gamma distribution and the approximation resulting from the transformation. The approximation could be improved by incorporating higher order log terms (i.e., quadratic or cubic terms) while maintaining the monotonicity of the transformation over the relevant range.

3. Empirical Example

Table 1 presents the observed yields for Cotton, Soybeans, and Potatoes in North Florida. Using this data, I computed the empirical cumulative density function defined as

$$\tilde{F}(x_{1i}) = \frac{1}{N} \sum_{x_{1j} \leq x_{1i}} 1. \tag{3.1}$$

Next, following Equation 2.3 I then compute the value of yield that would give the same cumulative density function value. Unfortunately, none of the data yields a marked depiction from normality. However, taking potatoes as an example, I apply the logarithmic form depicted in Equation 2.3. The result is are the estimates $\hat{\gamma}_0 = -1,289.1$ (278.6739) and $\hat{\gamma}_1 = 278.67.4$ (5.1965). Figure 4 presents actual and estimated values of $\hat{y}$ (i.e., the transformed variable that is normally distributed).

TABLE 1. Yields for North Florida

Year	Original Data			Detrended Yields		
	Cotton	Soybeans	Potatoes	Cotton	Soybeans	Potatoes
1960	327	26	122	732.444	35.487	257.869
1965	353	26	148	720.903	34.609	271.288
1970	436	28	162	766.362	35.730	272.708
1975	346	24	194	638.821	30.852	292.127
1980	610	22	194	865.280	27.973	279.547
1985	693	26	226	910.739	31.095	298.967
1990	640	19	219	820.197	23.216	279.386
1995	472	26	210	614.656	29.338	257.806
2000	480	19	286	585.115	21.460	321.225
2005	762	32	273	829.574	33.581	295.645
2010	766	30	250	796.033	30.703	260.064
2014	914	43	240	914.000	43.000	240.000

TABLE 2. Transformed Distributions

Cumulative Distribution	Cotton		Soybeans		Potatoes	
	X_1	Z_1	X_2	Z_2	X_3	Z_3
0.009	529.099	534.389	21.460	22.284	233.838	219.955
0.027	585.115	580.394	23.216	24.139	240.000	231.368
0.045	609.132	604.668	25.230	25.117	242.192	237.389
0.064	614.656	622.014	25.811	25.816	242.516	241.692
0.082	628.623	635.833	26.392	26.374	245.870	245.121
0.100	638.821	647.496	27.527	26.844	246.676	248.014
0.118	668.508	657.696	27.919	27.255	247.257	250.544
0.136	676.082	666.838	27.973	27.623	249.032	252.812
0.155	676.936	675.179	27.987	27.960	250.256	254.881
0.173	688.887	682.894	28.379	28.271	256.774	256.795
0.191	692.590	690.107	28.405	28.561	257.806	258.584
0.209	696.395	696.909	28.689	28.836	257.869	260.272
0.227	700.411	703.371	29.271	29.096	257.902	261.875
0.245	702.804	709.547	29.338	29.345	260.064	263.407
0.264	704.640	715.482	29.744	29.584	262.290	264.879
0.282	709.607	721.212	29.798	29.815	263.548	266.301
0.300	720.903	726.765	30.446	30.039	263.934	267.678
0.318	732.444	732.169	30.703	30.257	265.322	269.019
0.336	739.558	737.444	30.852	30.470	265.579	270.327
0.355	744.706	742.608	31.041	30.678	265.772	271.609
0.373	760.541	747.680	31.095	30.882	270.999	272.867
0.391	761.214	752.672	31.203	31.084	271.288	274.105
0.409	761.428	757.600	31.284	31.282	272.708	275.328
0.427	766.362	762.475	31.325	31.479	277.740	276.537
0.445	766.525	767.308	31.500	31.674	278.580	277.736
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.991	1072.247	1028.946	43.000	42.221	342.709	342.641

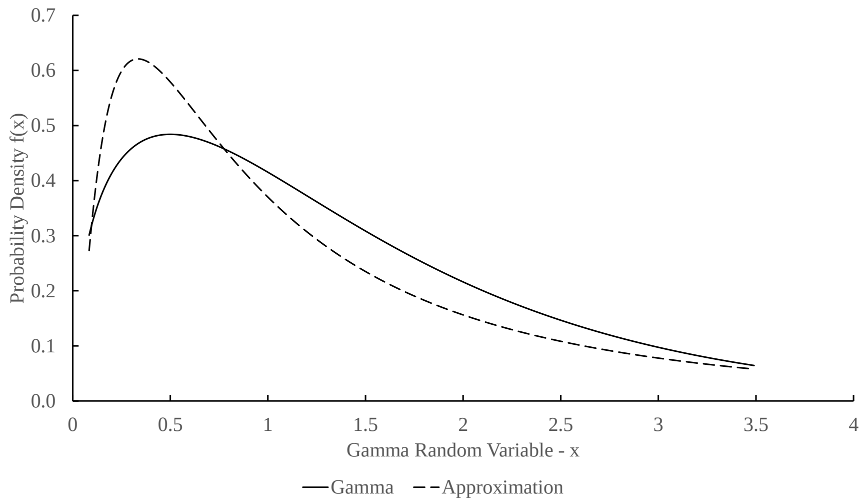

FIGURE 3. Comparison of Actual and Approximated Gamma Distribution

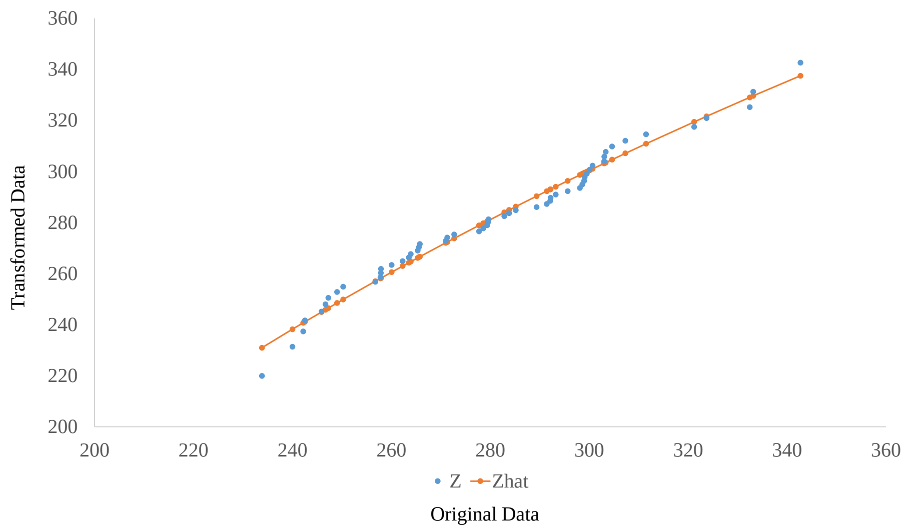

FIGURE 4. Estimated Transformed Potato Yields

These transformed variables can be used to compute the correlation between cotton and potato yields depicted in Table 3. In this case, the transformation is almost linear for the entire range of potato yields (i.e., potato yields are probably normally distributed). Hence, the correlation coefficient for the transformed yields and the untransformed yields are almost identical at -0.093.

Next, assume that I conclude that cotton and soybeans are normally distributed while potatoes are non-normally distributed under the logarithmic transformation

Table 3. Transformed Potato Yields Paired with Cotton Yields

Year	Cotton	Potatoes	Transformed Potatoes
1960	732.444	257.869	258.223
1965	720.903	271.288	272.360
1970	766.362	272.708	273.815
1975	638.821	292.127	292.984
1980	865.280	279.547	280.718
1985	910.739	298.967	299.434
1990	820.197	279.386	280.557
1995	614.656	257.806	258.155
2000	585.115	321.225	319.445
2005	829.574	295.645	296.320
2010	796.033	260.064	260.585
2014	914.000	240.000	238.211

in Equation 2.3. The parameters of transformation along with the variance covariance matrix for yields can be estimated using maximum likelihood

$$\begin{aligned} z_{1i} &= x_{1i} \\ z_{2i} &= x_{2i} \\ z_{3i} &= \gamma_0 + \gamma_1 \ln\left(x_{3i}\right) \\ f\left(z, \gamma_0, \gamma_1, \sigma^2\right) &\propto \left|\Omega\right|^{-N/2} \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(z_i - \mu\right)' \Omega^{-1} \left(z_i - \mu\right)\right] \frac{\gamma_1}{x_{3i}} \end{aligned} \tag{3.2}$$

where γ_1 / x_{3i} is the Jacobian of the transformation. Following the general approach from [3], I can maximize the natural logarithm of the likelihood function.

4. Discussion

This paper outlines a generalization of the approach used by Moss and Shonkwiler [2] to model correlated non-normal random variables such as yields. In this paper, I consider a general mapping function with the only restriction that the mapping be positively monotonic. To demonstrate the concept, I use cotton, soybean and potato data for North Florida. Unfortunately, each of these distributions appear to be normal.

References

[1] J.B. Burbidge, L. Magee, and A.L. Robb. Alternative transformation to handle extreme values of the dependent variable. *Journal of the American Statistical Associateion*, 83(401):123–127, March 1988.

[2] C. B. Moss and J.S. Shonkwiler. Estimating yield distributions using a stochastic trend model and nonnormal errors. *American Journal of Agricultural Economics*, 75(5):1056–1062, November 1993.

[3] O. A. Ramirez, C. B. Moss, and W. G. Boggess. Estimation and use of the inverse hyperbolic sine transformation to model nonnormal correlated random variables. *Journal of Applied Statistics*, 21:289–304, 1994.

(Charles B. Moss) 1175 McCarty Hall
Gainesville, FL 32611-0240 line 2
E-mail address, Charles B. Moss: cbmoss@ufl.edu
URL: http://ricardo.ifas.ufl.edu