Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 17

11-1-2007

Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data

L. W. Huson
Biostatistics Group, F. Hoffman-La Roche

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 530-536

1538 – 9472/07/$95.00

Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data

L. W. Huson
Biostatistics Group, F.Hoffman-La Roche

Zero-clustered data occur widely in medical research and are characterised by the presence of a group of observations of value zero in a distribution of otherwise continuous non-negative responses. A simulation study was conducted to investigate the properties of a number of correlation coefficients applied to samples of zero-clustered data.

Key words: zero-clustered data, Pearson correlation, Spearman correlation, weighted rank correlation.

Introduction

The defining characteristic of zero-clustered data is the presence of a group of observations of value zero in a distribution of otherwise continuous non-negative responses. This type of data is regularly encountered in a wide variety of medical and clinical applications (see e.g. Lachenbruch 1976; 2001a, 2001b, 2002).

Delucchi and Bostrom (2004) discussed a number of endpoints often used in psychiatric studies which typically exhibit zero-clustering, and Berk (2002) gave, as further examples of zero-clustered data, the antibody response to a vaccine, levels of alcohol consumption, severity rating of side-effects, and intensity of pain during labour. In the field of Health Economics, Buntin and Zaslavsky (2004) commented on the "spike of zero values" that is often seen in otherwise non-negative observations in data on health care costs or resource usage, and Chang and Pocock (2002) discussed a specific example of such data in their analysis of numbers of hours of personal care services received by a group of elderly patients. Other terms which have appeared in the literature to describe this type of data are semi-continuous (e.g. Schafer & Olsen 1999) and zero-inflated (e.g. Tu, 2002). Specifically excluded from consideration here are zero-inflated count data, which constitute a separate and widely studied phenomenon.

Some authors note that in the analysis of zero-clustered data, it may be appropriate to bear in mind the different possible origins of the zero values. Zeros may arise, for instance, by the deliberate censoring of any negative values and the setting of such values to zero. An example of such an endpoint is the ACRn score widely used in studies of rheumatoid arthritis (van Riel & van Gestel, 2000). Alternatively the zeros may arise from an unintentional censoring process, such as an imprecise or insensitive measuring device, where small values of an endpoint cannot be detected and response is therefore recorded as zero (see e.g. Moulton & Curriero, 2002). Finally, the zeros may be genuine and accurate values properly representing a patient's response (e.g. Chang & Pocock, 2002).

The proportion of zero values seen in practice in this type of data is variable from one type of endpoint to another. Delucchi and Bostrom (2004), for example, analysing data on addiction severity scores, reported proportions of zeros in different data sets ranging from 6% to 77%. Tu and Zhou (1999) cited data on

Dr. Les Huson has worked as a Consultant Medical Statistician in the pharmaceutical and biotechnology industries for 25 years, specializing in the design and analysis of controlled clinical trials. He obtained his PhD in Statistics at Imperial College, London, and currently holds an appointment as a Visiting Consultant Medical Statistician with the Statistical Advisory Service, Imperial College, London.

hospital in-patient charges in which approximately 75% of the values are zero. In many applications, however, the proportion of zeros would be expected to be smaller – Lachenbruch (2001a), for example, studied cases in which 10% or 20% of the values were zeros.

A further characteristic of zero-clustered data is that the distribution of non-zero part of the data is often skewed, with a long tail of high values. Models often suggested as appropriate for the non-zero part of the data are the lognormal or log-gamma distributions (see e.g. Lachenbruch, 2001a; Moulton & Curriero, 2002).

Although methods of analysis of zero-clustered data have been studied in the literature (see e.g. Lachenbruch 1976; 2001a, 2001b, 2002), the problem of measuring the degree of correlation between two samples of zero-clustered data has not previously been investigated. This article describes the results of a simulation study designed specifically to examine the performance of a number of different measures of correlation when applied to zero-clustered data. The study reported here was split into two parts. In the first simulation study the performance of two conventional correlation measures – the Pearson and Spearman correlation coefficients – was studied in the context of application to zero-clustered data. The second simulation study investigated the performance of three little known weighted rank correlation coefficients when applied to the same data structure.

Methodology I

Generating Samples of Correlated Zero-Clustered Data

Two different models were used to generate zero-clustered data for the simulation study – the binomial-lognormal model and the truncated lognormal model (Lachenbruch, 2001a; Moulton & Curriero, 2002). The first model assumes that the zero-clustered data arise from combination of binary and lognormal responses, and the second that the zeros arise from a process of truncation of lognormal data. These models are described in more detail below.

Samples sizes of 25, 50, 100, 200 and 1000 were used in the simulation study, with correlations in the data specified to be 0.30, 0.60 or 0.90, representing low, medium and high correlations respectively. The proportion of zeros in the generated samples was 10%, 20% or 30% in different series of simulations. For each of these combinations of parameters, 10000 simulated datasets were generated, and the value of each of the chosen correlation coefficients was calculated for each generated sample of data.

Binomial-Lognormal Model

For these simulations, samples of zero-clustered data were generated as a mixture of binary responses and lognormal responses, with the same correlation applied to both components of the data. This gives samples of paired, correlated data which follow the binomial-lognormal model (Lachenbruch, 2001a).

The correlated binary components were generated using the algorithm described by Kang and Jung (2001), and the correlated lognormal components were generated using the methods described by Saucier (2000). The method of Kang and Jung permits the generation of pairs of binary observations - values (0,0), (0,1), (1,0) and (1,1) - with specified probabilities and correlations. For each sample size studied, a full set of such correlated binary pairs was generated, and also a full set of correlated lognormal responses. The final correlated zero-clustered binomial-lognormal dataset was then derived simply by multiplying these two sets of values together. Thus, a binary pair (0,0) and a lognormal pair (X_1,X_2) when multiplied together yield the pair (0,0), a binary pair (0,1) and a lognormal pair (X_3,X_4) when multiplied together yield the pair $(0,X_4)$, and similarly for other combinations.

Truncated Lognormal Model

In this series of simulations, zero-clustered data were generated by truncating correlated lognormal data. To do this, correlated lognormal data were first generated, using the methods described by Saucier (2000), then, to generate a sample containing a given proportion

Table 1. Mean Value of Pearson and Spearman Correlation Coefficient Estimates
[Binomial-Lognormal Model - 10000 Simulations]

Sample Size	Coefficient	True Correlation=0.3 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.6 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.9 Proportion of Zeros 0.1	0.2	0.3
25	Pearson	0.30	0.28	0.27	0.56	0.54	0.53	0.85	0.85	0.84
	Spearman	0.32	0.30	0.29	0.56	0.56	0.56	0.81	0.83	0.85
50	Pearson	0.29	0.27	0.26	0.57	0.55	0.53	0.86	0.85	0.85
	Spearman	0.33	0.31	0.29	0.57	0.56	0.57	0.82	0.84	0.85
100	Pearson	0.29	0.27	0.26	0.57	0.55	0.53	0.87	0.86	0.86
	Spearman	0.33	0.31	0.30	0.57	0.56	0.57	0.82	0.84	0.86
200	Pearson	0.29	0.27	0.25	0.57	0.55	0.54	0.88	0.87	0.86
	Spearman	0.33	0.31	0.30	0.57	0.57	0.57	0.83	0.84	0.86
1000	Pearson	0.28	0.26	0.25	0.57	0.55	0.54	0.89	0.88	0.87
	Spearman	0.33	0.31	0.30	0.57	0.57	0.57	0.83	0.84	0.86

Table 2. Mean Value of Pearson and Spearman Correlation Coefficient Estimates
[Truncated Lognormal Model - 10000 Simulations]

Sample Size	Coefficient	True Correlation=0.3 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.6 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.9 Proportion of Zeros 0.1	0.2	0.3
25	Pearson	0.33	0.33	0.33	0.59	0.59	0.59	0.86	0.86	0.85
	Spearman	0.36	0.36	0.36	0.57	0.57	0.57	0.79	0.79	0.79
50	Pearson	0.32	0.32	0.32	0.59	0.59	0.59	0.87	0.87	0.86
	Spearman	0.36	0.36	0.36	0.58	0.58	0.58	0.80	0.80	0.80
100	Pearson	0.31	0.31	0.32	0.59	0.59	0.59	0.88	0.88	0.87
	Spearman	0.37	0.37	0.37	0.58	0.58	0.58	0.81	0.80	0.80
200	Pearson	0.31	0.31	0.31	0.60	0.60	0.60	0.89	0.89	0.88
	Spearman	0.37	0.37	0.37	0.58	0.58	0.58	0.81	0.81	0.81
1000	Pearson	0.30	0.30	0.31	0.60	0.60	0.60	0.90	0.89	0.89
	Spearman	0.37	0.37	0.37	0.58	0.58	0.59	0.81	0.81	0.81

p of zero data, any lognormal value lower than exp(probit(p)) to was set to zero.

Results I

First Simulation Study

Pearson and Spearman Correlations applied to the Binomial-Lognormal Model

Table 1 shows the results of the simulation study of the performance of the Pearson and Spearman correlation coefficients, when applied to zero-clustered data generated using the binomial-lognormal model.

The most obvious finding is that, under this data model, both the Pearson and Spearman coefficients on average slightly underestimate the true correlation in most simulated scenarios. The bias is relatively small, but persists across all sample sizes and for low, medium and high correlations. A second finding is that the bias of the Pearson correlation increases slightly as the proportion of zeros in the data increases. In contrast, with the Spearman estimate, the bias either remains much the same as the proportion of zeros increases, or diminishes slightly. The other interesting feature of the results is that the Spearman estimate is generally more accurate for low and medium correlations, across all sample sizes, while the Pearson estimate performs better for high correlations.

Pearson and Spearman Correlations applied to the Truncated Lognormal Model

Table 2 shows the results of the simulation study of the performance of the Pearson and Spearman correlation coefficients with correlated zero-clustered data generated using the truncated lognormal model. Under this data model, both the Pearson and Spearman coefficients tend to underestimate the true correlation for medium and high correlations, but tend to overestimate the true value when the true correlation is 0.3. When the true correlation is high, the bias of the Pearson correlation increases slightly as the proportion of zeros in the data increases, whereas with the Spearman estimate, the bias either remains much the same as the proportion of zeros increases, or diminishes slightly. Under this data model, the Pearson correlation performs better than the Spearman for most scenarios.

Methodology II

Weighted Rank Correlation Coefficients

Introduction

For the second simulation study correlation estimates were selected that were (a) based on ranks or functions of ranks, and (b) were defined in a way which allows lower weights to be attached to the zero values in the data, and higher weights to the non-zero values. These were considered likely to be properties which would result in better estimation of correlation in the presence of data containing many zeros.

Three weighted rank correlation coefficients which have these properties are described in the literature and are easily computed, but they are little known and little used in practice. They are the "top-down" correlation, the Blest-Genest-Plante correlation, and the Costa-Soares correlation. The second part of the simulation study investigated the properties of these three coefficients when applied to correlated zero-clustered data.

Top Down Correlation

Iman and Conover (1987) described a correlation estimate which they termed the "top down" correlation. This coefficient places emphasis on the higher ranked data in a sample (i.e. assigns lower weights to low-ranked zero values) by computing the correlation using Savage scores derived from the ranked data. Savage scores are defined as follows:

$$S_i = \sum_{j=1}^{n} 1/j \qquad (1)$$

where i is the rank assigned to the ith order statistic in a sample of size n. For example, with n = 3, the three Savage scores are $S_1 = 1 + 1/2 + 1/3$, $S_2 = 1/2 + 1/3$, and $S_3 = 1/3$. The top-down coefficient is calculated as:

$$r_{td} = \left(\sum_{j=1}^{n} S_{Ri} S_{Qi} - n \right) / (n - S_1) \qquad (2)$$

where S indicates the Savage score, the R_i and Q_i are the ranks of the data in the two samples, and n is the sample size. A full description of the properties of this coefficient is given by Iman and Conover (1987).

Blest-Genest-Plante correlation

Blest (2000) also defined a rank correlation coefficient which allows lower weights to be assigned to lower ranked values in a dataset. This coefficient, whilst having some desirable properties, suffers from the disadvantage that in its original form it is not symmetric (i.e. corr[X,Y] does not equal corr [Y,X]). However, the Blest estimate was later modified by Genest and Plante (2003) to a symmetrical form, and this symmetric version for the simulation study reported here. The coefficient is calculated as:

$$r_{bgp} = -((4n+5)/(n-1)) + (6/(n^3 - n) \sum_{j=1}^{n} R_i Q_i (4- (R_i + Q_i)/ n+1))) \quad (3)$$

where the R_i and Q_i are the ranks of the data in the two samples, and n is the sample size. The detailed properties of the original Blest coefficient and its symmetrical generalization are described by Genest and Plante (2003).

Costa-Soares correlation

Costa and Soares (2005) also defined a rank correlation coefficient which, like the top-down correlation and the Blest-Genest-Plante correlation, allows lower weights to be assigned to lower ranked values in a dataset, and hence in this application allows lower weights to be assigned to the zero values in the zero-clustered data. The coefficient takes the form:

$$r_{cs} = 1 - 6 \sum_{j=1}^{n} (R_i - Q_i)^2 / (n^3 - n) \quad (4)$$

where the R_i and Q_i are the ranks of the data in the two samples, and n is the sample size. The properties of this coefficient, and in particular a comparison of the properties with those of the Blest correlation, are described by Costa and Soares (2005).

Results

Second Simulation Study

Weighted Correlations with the Binomial-Lognormal Model

Table 3 shows the results of the simulation study of the performance of the weighted correlation coefficients with zero-clustered data generated using the binomial-lognormal model. These weighted correlation coefficients all slightly underestimate the true correlation in the data when the true correlation is medium or high, and overestimate the value when it is low, Their performance generally is as good as or better than that of the Spearman estimate.

Weighted Correlations with the Truncated Lognormal Model

Table 4 shows the results of the simulation study of the performance of the weighted correlation coefficients with zero-clustered data generated using the truncated lognormal model. As with the Pearson and Spearman coefficients, the general tendency of the estimates under this data model is that low correlations are overestimated and medium and high correlations are underestimated. Again under most conditions the weighted coefficients perform on average at least as well or better than the Spearman estimates.

Conclusion

The literature contains no recommendations on an appropriate choice of correlation coefficient for use with zero-clustered data, but Delucchi & Bostrom (2004) reported the results of an informal survey showing that 22 of 35 articles reported analyses of zero-clustered data that used standard normal theory methods, despite the clear non-normality of such data. Hence it seems likely that some practitioners, in the absence of any specific alternative, might choose to apply commonly-used correlation measures - such as Pearson's correlation or Spearman's rank correlation – to zero-clustered data.

Table 3 Mean Value of Some Weighted Rank Correlation Coefficients
[Binomial-Lognormal Model - 10000 Simulations]

Sample Size	Coefficient	True Correlation=0.3 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.6 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.9 Proportion of Zeros 0.1	0.2	0.3
25	Top-Down	0.31	0.28	0.27	0.55	0.54	0.53	0.83	0.82	0.83
	Blest-Genest-P	0.33	0.31	0.31	0.56	0.56	0.57	0.82	0.83	0.86
	Costa-Soares	0.33	0.30	0.29	0.56	0.55	0.56	0.82	0.83	0.84
50	Top-Down	0.31	0.29	0.27	0.56	0.55	0.54	0.84	0.84	0.84
	Blest-Genest-P	0.33	0.31	0.31	0.57	0.56	0.57	0.82	0.84	0.86
	Costa-Soares	0.33	0.31	0.29	0.57	0.56	0.56	0.82	0.83	0.85
100	Top-Down	0.31	0.29	0.28	0.57	0.55	0.54	0.85	0.85	0.85
	Blest-Genest-P	0.33	0.31	0.31	0.57	0.57	0.58	0.83	0.84	0.86
	Costa-Soares	0.33	0.31	0.29	0.57	0.56	0.56	0.83	0.84	0.85
200	Top-Down	0.32	0.29	0.28	0.58	0.56	0.55	0.86	0.85	0.85
	Blest-Genest-P	0.34	0.31	0.31	0.57	0.57	0.58	0.83	0.84	0.87
	Costa-Soares	0.33	0.31	0.29	0.57	0.56	0.56	0.83	0.84	0.85
1000	Top-Down	0.32	0.30	0.28	0.58	0.56	0.55	0.86	0.86	0.86
	Blest-Genest-P	0.34	0.32	0.31	0.58	0.57	0.58	0.83	0.85	0.87
	Costa-Soares	0.34	0.31	0.30	0.58	0.57	0.56	0.83	0.84	0.85

Table 4. Mean Value of Some Differentially Weighted Correlation Coefficients
[Truncated Lognormal Model - 10000 Simulations]

Sample Size	Coefficient	True Correlation=0.3 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.6 Proportion of Zeros 0.1	0.2	0.3	True Correlation=0.9 Proportion of Zeros 0.1	0.2	0.3
25	Top-Down	0.33	0.33	0.33	0.57	0.57	0.57	0.83	0.83	0.83
	Blest-Genest-P	0.36	0.36	0.37	0.58	0.58	0.59	0.81	0.81	0.81
	Costa-Soares	0.36	0.36	0.36	0.58	0.58	0.58	0.81	0.81	0.81
25	Top-Down	0.33	0.33	0.33	0.58	0.58	0.58	0.84	0.84	0.84
	Blest-Genest-P	0.36	0.37	0.37	0.59	0.59	0.59	0.82	0.82	0.82
	Costa-Soares	0.36	0.36	0.36	0.58	0.58	0.58	0.82	0.82	0.81
100	Top-Down	0.34	0.34	0.34	0.59	0.59	0.59	0.86	0.86	0.86
	Blest-Genest-P	0.37	0.37	0.37	0.59	0.59	0.60	0.82	0.82	0.83
	Costa-Soares	0.37	0.37	0.37	0.59	0.59	0.59	0.82	0.82	0.82
200	Top-Down	0.34	0.34	0.34	0.60	0.60	0.60	0.86	0.86	0.86
	Blest-Genest-P	0.37	0.37	0.37	0.59	0.59	0.60	0.82	0.83	0.83
	Costa-Soares	0.37	0.37	0.37	0.59	0.59	0.59	0.82	0.82	0.82
1000	Top-Down	0.34	0.34	0.34	0.60	0.60	0.60	0.87	0.87	0.87
	Blest-Genest-P	0.37	0.37	0.38	0.60	0.60	0.60	0.83	0.83	0.83
	Costa-Soares	0.37	0.37	0.37	0.59	0.59	0.59	0.83	0.82	0.82

The first part of the simulation study reported here was designed to examine the performance of these common correlation coefficients when applied to this type of data. The second part of the study investigated the properties of three little-known weighted rank correlation coefficients. This summary suggests that, overall, the Pearson estimate is in fact, for most practical purposes, an adequate choice from the coefficients studied, and that among rank correlation coefficients, those allowing differential weighting of zero values generally perform better than the much more widely known Spearman coefficient.

.

References

Berk, K.N. (2002) Repeated measures with zeros. *Statistical Methods in Medical Research* 11:303-316.

Blest, D.C. (2000). Rank correlation—an alternative measure. *Australian and New Zealand Journal of Statistics* 42:101—111.

Buntin, M.B., Zaslavsky, A.M. (2004) Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures. *Journal of Health Economics* 23: 525–542.

Chang, B-H., Pocock, S. (2000) Analyzing data with clumping at zero: an example demonstration. *Journal of Clinical Epidemiology* 53:1036–1043.

Costa, J., Soares, C. (2005) A weighted rank measure of correlation. *Australian and New Zealand Journal of Statistics* 47(4):515–529.

Delucchi, K.L., Bostrom, A. (2004) Methods for Analysis of Skewed Data Distributions in Psychiatric Clinical Studies: Working With Many Zero Values. *American Journal of Psychiatry* 161:1159–1168.

Genest, C., Plante, J-F. (2003) On Blest's measure of rank correlation. *The Canadian Journal of Statistics* 31(1): 1–18.

Iman, R.L., Conover, W.J. (1987) A Measure of Top-Down Correlation. *Technometrics* 29(3): 351-357.

Kang, S-H., Jung, S-H. (2001) Generating Correlated Binary Variables with Complete Specification of the Joint Distribution. *Biometrical Journal* 43(3): 263–269.

Lachenbruch, P.A. (1976) Analysis of data with clumping at zero. *Biometrische Zeitschrift* 18: 851-856.

Lachenbruch, P.A. (2001a) Comparisons of two-part models with competitors. *Statistics In Medicine* 20:1215–1234.

Lachenbruch, P.A. (2001b) Power and sample size requirements for two-part models. *Statistics In Medicine* 20:1235–1238.

Lachenbruch, P.A. (2002) Analysis of data with excess zeros. *Statistical Methods in Medical Research* 11: 297.

Moulton, L.H., Curriero, F.C. (2002) Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* 11: 317-325.

Saucier, R. (2000) Computer Generation of Statistical Distributions. U.S. Army Research Laboratory Technical Report 2168, March 2000.

Schafer, J.L., Olsen, M.K. (1999) Modeling and imputation of semicontinuous survey variables In *Proceedings of Federal Committee on Statistical Methodology (FCSM) Research Conference*, Nov. 1999. 15.

Tu,W., Zhou, X.H. (1999) A Wald Test Comparing Medical Costs Based on Log-Normal Distributions With Zero-Valued Costs. *Statistics In Medicine* 18: 2749-2761.

Tu, W. (2002) Zero-inflated Data. *In: Encyclopedia Of Environmetric, Eds: El-Shaarawi AH, Piegorsch WW John Wiley & Sons, Chichester.* Volume 4: 2387-2391.

Van Riel, P.L.C.M., van Gestel, A.M. (2000) Clinical outcome measures in rheumatoid arthritis. *Annals of Rheumatic Disease* 59 (supplement): 28-31