Journal of Environmental Statistics

Feburary 2013, Volume 4, Issue 1. http://www.jenvstat.org

Estimation of Regression Parameters for Binary Longitudinal Data Using GEE: Review, Extension and an Application to Environmental Data

Sudhir Paul **Xuemao Zhang** **Jianwen Xu**

Abstract

Longitudinal or clustered response data arise in many applications such as, biostatistics, epidemiology and environmental studies. The repeated responses can not in general be assumed to be independent. The generalized estimating equations (GEE) approach is a widely used method to estimate marginal regression parameters for correlated responses. The advantage of the GEE is that the estimates of the regression parameters are asymptotically unbiased, although their small sample properties are not known. In this paper we review the GEE methodology for longitudinal binary data and propose a method of correcting bias of the estimates when the sample size is potentially small. Some simulation studies are provided to illustrate the theoretical results and applications of the GEE and its bias corrected version are also discussed to a set of environmental data.

Keywords: bias reduction, generalized estimating equations, longitudinal data, marginal model.

1. Introduction

Longitudinal studies are characterized by repeated measures over a period of time from each individual. Usually the subjects are assumed to be independent while the repeated measurements taken on each subject are correlated. The complication of longitudinal data analysis is partly due to the lack of a rich class of models such as the multivariate Gaussian for the joint distribution of the correlated responses (Liang and Zeger, 1986). Liang and Zeger (1986) introduced the generalized estimating equations (GEE) approach for analyzing longitudinal data in which a working correlation matrix for the responses of each individual is used. The GEE approach requires specification of only the first two moments of a subject's responses rather than the full specification of the joint distribution. The main advantage of the GEE is that the estimators are consistent (asymptotically unbiased) even if the working correlation structure is misspecified. However, the GEE technique may produce biased estimates in the case of small to moderate sample sizes.

Under general conditions, maximum likelihood (ML) estimators are consistent. However, they are not unbiased generally. Cox and Snell (1968) provided general results for the first-order correction of bias of maximum likelihood estimators for any distribution. Cordeiro and Klein (1994) gave a general matrix formula for computing the bias of the ML estimates. In this paper we review the GEE methodology for longitudinal binary data and using Cox and Snell (1968) and Cordeiro and Klein (1994) we propose a method of correcting bias of the estimates when the sample size is potentially small. We shall evaluate the method by some simulation studies and applications of the method are discussed to a set of environmental data.

In Section 2 we review the general results for first-order correction of bias of maximum likelihood estimators by Cox and Snell (1968). The bias correction of GEE estimates and bias-corrected generalized estimating equation method are derived in Section 3. In section 4, the bias correction methods are applied to longitudinal binary and poisson data. In section 5, some simulation studies are performed. Two examples are given in section 6 and a discussion follows in Section 7.

2. The GEE estimating method in longitudinal data

Consider a longitudinal study with K subjects, each subject having d repeated measures $y_i = (y_{i1}, \ldots, y_{id})^T$ and a $d \times p$ design matrix $X_i = (x_{i1}, \ldots, x_{id})^T$ with elements $x_{ij} = (x_{ij1}, \cdots, x_{ijp})^T$. Assume that the K subjects are independent while the repeated measurements y_{ij} taken on each subject are correlated. Define $\mu_i = \mathrm{E}(y_i|X_i) = (\mu_{i1}, \ldots, \mu_{id})^T$ to be the expectation of y_i conditional on X_i and suppose $\mu_i = g(X_i\beta)$, where β is a $p \times 1$ vector of regression parameters of interest and g^{-1} is the link function. Assume that the variance of y_{ij} is given by $\phi v(\mu_{ij})$, where v is the variance function and ϕ is the overdispersion parameter. Let $R(\rho)$ be a working correlation matrix completely specified by the parameter vector ρ of length q. Then $\phi W_i = \phi A_i^{1/2} R(\rho) A_i^{1/2}$ is the corresponding working covariance matrix, where $A_i(\beta) = \mathrm{diag}\{v(\mu_{ij})\}$, $j = 1, \ldots, d$, $i = 1, \ldots, K$.

Standard working correlation matrices $R(\rho)$ (Liang and Zeger, 1986), Wang and Carey, 2003) are:

i) exchangeable correlation structure in which the diagonal elements of $R(\rho)$ are 1 and the off-diagonal elements are ρ,

ii) AR(1) correlation structure in which the diagonal elements of $R(\rho)$ are 1 and the off-diagonal elements are $\rho^{|i-j|}$, $i \neq j$,

iii) the general autocorrelation structure

$$R(\rho_1, \ldots, \rho_{d-1}) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{d-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{d-2} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \rho_{d-1} & \rho_{d-2} & \rho_{d-3} & \cdots & 1 \end{bmatrix},$$

and iv) the unstructured correlation matrix (Liang and Zeger, 1986)

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1,d-1} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2,d-2} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \rho_{1,d-1} & \rho_{2,d-2} & \rho_{3,d-3} & \cdots & 1 \end{bmatrix}. \tag{2.1}$$

Let $y^*_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}(1-\hat{\mu}_{ij})}$. Then, the method of moments estimate of (i) the common correlation coefficient ρ in the exchangeable correlation structure is

$$\hat{\rho} = \frac{\sum_{i=1}^{K} \sum_{j \neq k} y_{ij}^{*} y_{ik}^{*}}{(d-1) \sum_{i=1}^{K} \sum_{j=1}^{d} {y_{ij}^{*}}^{2}},$$

(ii) the common correlation coefficient ρ in the AR(1) correlation structure is

$$\hat{\rho} = \frac{\sum_{i=1}^{K} \sum_{j=2}^{d} y_{ij}^{*} y_{i,j-1}^{*}}{\sum_{i=1}^{K} \{ \sum_{j=2}^{d-1} {y_{ij}^{*}}^{2} + ({y_{i1}^{*}}^{2} + {y_{id}^{*}}^{2})/2 \}},$$

(iii) the correlation parameter ρ_l in $R(\rho_1, \dots, \rho_{d-1})$ is

$$\hat{\rho}_l = \frac{\sum_{i=1}^{K} \sum_{j=1}^{d-l} y_{ij}^{*} y_{i,j+l}^{*} / (d-l)}{\sum_{i=1}^{K} \sum_{j=1}^{d} {y_{ij}^{*}}^{2} / d}, \quad l = 1, \dots, d-1.$$

Finally, the estimate of the unstructured correlation matrix is given by

$$\hat{R} = \sum_{i=1}^{K} \hat{A}_i^{-1/2} S_i S_i^T \hat{A}_i^{-1/2} / K, \text{ where } S_i = y_i - \hat{\mu}_i, i = 1, \dots, K.$$

For given consistent estimates of ϕ and ρ, the estimate $\hat{\beta}$ is the solution of the GEE equations

$$\sum_{i=1}^{K} (y_i - \mu_i)^T W_i^{-1} D_i = 0, \tag{2.2}$$

where $D_i = \frac{\partial \mu_i}{\partial \beta^T}$. The estimator $\hat{\beta}_{GEE}$ obtained by solving equation (2.2) is consistent even if the covariance structure is miss-specified. However, the miss-specification of the covariance structure may result in inefficient estimates of the regression parameters (for more details, see Wang and Carey, 2003).

Under mild regularity conditions, $K^{1/2}(\hat{\beta}_{GEE} - \beta)$ is asymptotically multivariate normal with mean zero and covariance matrix

$$\lim_{K \to \infty} K \left(\sum_{i=1}^{K} D_i^T W_i^{-1} D_i \right)^{-1} \left[\sum_{i=1}^{K} D_i^T W_i^{-1} \operatorname{cov}(Y_i) W_i^{-1} D_i \right] \left(\sum_{i=1}^{K} D_i^T W_i^{-1} D_i \right)^{-1}.$$

The estimated covariance matrix obtained by this formula is called the robust covariance estimator for $\hat{\beta}_{GEE}$.

An iterative algorithm to obtain the above GEE estimate $\hat{\beta}_{GEE}$ can be described as what follows:

Step 1: Choose an initial value $\hat{\beta}$ of β.

Step 2: For given $\hat{\beta}$, the moment estimate of the overdispersion parameter is

$$\hat{\phi} = \frac{1}{Kd}\sum_{i=1}^{K}\hat{Z}_i^T\hat{Z}_i, \text{ where } \hat{Z}_i = A_i^{-1/2}(\hat{\beta})(y_i - \hat{\mu}_i).$$

Step 3: For given $\hat{\beta}$ and $\hat{\phi}$, calculate the moment estimates $\hat{\rho}$ of α (Liang and Zeger, 1986 and Wang and Carey, 2003). For example, if the working exchangeable correlation is used, Wang and Carey (2003) estimate ρ by

$$\hat{\rho} = \frac{\sum_{i=1}^{K}\sum_{j\neq k} y_{ij}^* y_{ik}^*}{\hat{\phi}(d-1)\sum_{i=1}^{K}\sum_{j=1}^{d} {y_{ij}^*}^2}, \text{ where } y_{ij}^* = (y_{ij} - \hat{\mu}_{ij})/\sqrt{v(\hat{\mu}_{ij})}.$$

Step 4: For given estimated working correlation matrix $R(\hat{\rho})$, the estimator of β is updated according to the modified Fisher scoring formula

$$\tilde{\beta} = \hat{\beta} + \left\{\sum_{i=1}^{K}\hat{D}_i^T\hat{W}_i^{-1}\hat{D}_i\right\}^{-1}\left\{\sum_{i=1}^{K}\hat{D}_i^T\hat{W}_i^{-1}(Y_i - \hat{\mu}_i)\right\},$$

where $\hat{D}_i = \partial\mu_i/\partial\beta^T\big|_{\hat{\beta}}$ and $\hat{W}_i = A_i(\hat{\beta})R(\hat{\rho})A_i(\hat{\beta})$.

Step 5: Iterate step 2 - 4 until $|\tilde{\beta} - \hat{\beta}|$ is less than a desired convergence rate. $\hat{\beta}_{GEE} = \tilde{\beta}$ is the estimate of the regression parameter. The estimate of ρ is given by $\hat{\rho}$ and the estimate of ϕ is given by $\hat{\phi}$.

3. The bias corrected GEE estimator in longitudinal data

The left hand side of equation (2.2) which can be written as

$$U(\beta;\alpha,\phi) = \sum_{n=1}^{K}(y_n - \mu_n)^T W_n^{-1}\frac{\partial\mu_n}{\partial\beta^T} \tag{3.1}$$

is the generalized estimating function for β given α and ϕ. Let $U(\beta;\alpha,\phi) = (U_1, U_2, \ldots, U_p)$. Under general conditions, maximum likelihood (ML) estimators are consistent. However, they are not unbiased generally. Cox and Snell (1968) provided general results for the first-order correction of bias of ML estimators of parameters under any distribution. Cordeiro and Klein (1994) provided simplified form of the general results of Cox and Snell (1968). For the purpose of obtaining bias-corrected GEE estimates we treat U_i as if it were a likelihood score function for β_i, $i = 1, \ldots, p$.

Now, define $\kappa_{ij} = \mathrm{E}(\partial U_i/\partial \beta_j)$ for $i, j = 1, \ldots, p$. Further, define $\kappa_{ijl} = \mathrm{E}(\partial^2 U_i/\partial \beta_j \partial \beta_l)$, $\kappa_{ij}^{(l)} = \partial \kappa_{ij}/\partial \beta_l$ and $k_{ij,l} = E(\frac{\partial U_i}{\partial \beta_j} U_l)$ for $i, j, l = 1, \ldots, p$. Then the Fisher information matrix analogue of order p for β is $I = \{-\kappa_{ij}\}$. Now, let $I^{-1} = \{\kappa^{ij}\}$ be the inverse of I. Then, following Cordeiro and Klein (1994) the bias of $\hat{\beta}_s$ can be expressed as

$$b_s(\hat{\beta}) = \sum_{i=1}^{p} \kappa^{si} \sum_{j,l=1}^{p} \left[\kappa_{ij}^{(l)} - \frac{1}{2}\kappa_{ijl}\right] \kappa^{jl}, \quad s = 1, \ldots, p. \tag{3.2}$$

In fact, the bias of the GEE estimator of the regression coefficients could also be written in the following matrix form:

$$b_s(\hat{\beta}) = I^{-1} \cdot A \cdot Vec(I^{-1}),$$

where $Vec(I^{-1})$ denotes the vector obtained by stacking the columns of I^{-1} and

$$A_{p\times p^2} = (\{a_{ij}^{(1)}\}, \cdots, \{a_{ij}^{(l)}\}) = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1p}^{(1)} & \cdots & a_{11}^{(p)} & \cdots & a_{1p}^{(p)} \\ a_{21}^{(1)} & \cdots & a_{2p}^{(1)} & \cdots & a_{21}^{(p)} & \cdots & a_{2p}^{(p)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{p1}^{(1)} & & a_{pp}^{(1)} & & a_{p1}^{(p)} & & a_{pp}^{(p)} \end{pmatrix}$$

with $a_{ij}^{(l)} = k_{ij}^{(l)} - \frac{1}{2}k_{ijl} = \frac{1}{2}(k_{ij}^{(l)} + k_{ij,l}), i, j, l = 1, \ldots, p.$

Then the bias corrected estimate $\tilde{\beta}_s$, of β_s is given by $\tilde{\beta}_s = \hat{\beta}_s - b(\hat{\beta}_s)$.

4. Application to binary data

For the vector of binary responses y_i, the variance function is given by $v(\mu) = \mu(1-\mu)$ and we consider the logit and probit link functions. The inverse of the probit link is given by $g(x) = \Phi(x)$, namely the cumulative distribution function of the standard normal distribution, thus the 1st and 2nd derivatives are $\dot{g}(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), \ddot{g}(x) = \frac{-x}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. The inverse of the logit link is given by $g(x) = \frac{\exp(x)}{1+\exp(x)}$, then the 1st and 2nd derivatives are $\dot{g}(x) = \frac{\exp(x)}{(1+\exp(x))^2}, \ddot{g}(x) = \frac{\exp(x)(\exp(x)-1)}{(1+\exp(x))^3}$.

Let I_p be a p-dimensional identity matrix, 1_p be a p-dimensional column vector with all elements being 1, then the quantities required for the calculation of the bias $b_s(\beta)$ are given by $I = \{-k_{ij}\} = \{-E(\frac{\partial U_i}{\partial \beta_j})\} = \{E(U_i U_j)\} = E(UU^T) = \sum_{i=1}^{n} D_i^T W_i^{-1} D_i$ and

$$A^T = \begin{pmatrix} \{a_{ij}^{(1)}\} \\ \vdots \\ \{a_{ij}^{(l)}\} \end{pmatrix} = \frac{1}{2}(\begin{pmatrix} \{k_{ij}^{(1)}\} \\ \vdots \\ \{k_{ij}^{(p)}\} \end{pmatrix} + \begin{pmatrix} \{k_{ij,1}\} \\ \vdots \\ \{k_{ij,p}\} \end{pmatrix}) = \frac{1}{2}\sum_{i=1}^{n}(Q_i - P_i),$$

where

$$P_i = (I_p \otimes X_i^T) \cdot \{diag(Vec(X_i)) \cdot [I_p \otimes (\dot{\Delta}_i W_i^{-1} \Delta_i)] \\ + [I_p \otimes (\Delta_i W_i^{-1} \dot{\Delta}_i)] \cdot diag(Vec(X_i))\} \cdot (1_p \otimes X_i)$$

$$Q_i = \{I_p \otimes \{(1_p^T \otimes X_i^T) diag(Vec(X_i))[I_p \otimes \dot{\Delta}_i W_i^{-1} \Delta_i]\}\} \cdot \begin{pmatrix} I_p \otimes \tilde{X}_{i1} \\ \vdots \\ I_p \otimes \tilde{X}_{ip} \end{pmatrix}$$

$$\Delta_i = diag(\dot{g}(x_{i1}^T\beta), \cdots, \dot{g}(x_{id}^T\beta)), \dot{\Delta}_i = diag(\ddot{g}(x_{i1}^T\beta), \cdots, \ddot{g}(x_{id}^T\beta))$$

and $\tilde{X}_{ij}$ denotes the j-th column of $X_i, j = 1, \dots, p$. In particular, for the probit link, $\Delta_i = diag(\phi(x_{i1}^T\beta), \cdots, \phi(x_{id}^T\beta))$, while $\Delta_i = diag(\frac{\exp(x_{i1}^T\beta)}{(1+\exp(x_{i1}^T\beta))^2}, \cdots, \frac{\exp(x_{id}^T\beta)}{(1+\exp(x_{id}^T\beta))^2})$ for the logit link.

5. Simulation studies

In this section, we are to assess and compare the performances of bias corrected estimators under discrete binary responses with different sample size by simulation studies. The binary response is generated using *mvtBinaryEP* package (Emrich and Piedmonte, 1991) in R software from a logistic regression model with the mean structure $\mu_{it} = \frac{\exp(0.5+0.6x_{i1}+0.6x_{i2})}{1+\exp(0.5+0.6x_{i1}+0.6x_{i2})}$, where the AR(1) intra-individual correlation structure with correlation parameter $\rho = 0.5$ is considered. The dispersion parameter is 1 and the covariates are generated at random from uniform distribution $U[0, 1]$. Relative large sample size with 300, 200 and 100 subjects and 5 repeated measurements are considered. In addition, the simulation is repeated with smaller sample sizes of 50, 30 and 10 subjects and 5 repeated measurements to check the performance of the bias corrected GEE estimator. The simulation results are shown in Table 1, where B-C GEE denotes the bias corrected GEE estimators,

Table 1: The GEE and bias-corrected GEE estimators with binary response.

	n=300		n=200		n=100	
	GEE(SD)	B-C GEE(SD)	GEE(SD)	B-C GEE(SD)	GEE(SD)	B-C GEE(SD)
β_0	0.4021(0.0187)	0.4013(0.0187)	0.3602(0.0256)	0.3560(0.0254)	0.2233(0.0422)	0.2224(0.0420)
β_1	0.5711(0.0285)	0.5692(0.0284)	0.8765(0.0366)	0.8710(0.0365)	0.9435(0.0570)	0.9348(0.0563)
β_2	0.4370(0.0257)	0.4356(0.0257)	0.6871(0.0395)	0.6833(0.0392)	0.3231(0.0695)	0.3201(0.0689)
	n=50		n=30		n=10	
	GEE(SD)	B-C GEE(SD)	GEE(SD)	B-C GEE(SD)	GEE(SD)	B-C GEE(SD)
β_0	0.2760(0.0915)	0.2749(0.0892)	0.6828(0.2074)	0.6662(0.2044)	-0.1604(0.3767)	-0.1410(0.3566)
β_1	0.8141(0.1870)	0.7974(0.1811)	0.7461(0.3334)	0.7096(0.3194)	1.7232(0.7646)	1.5818(0.7370)
β_2	0.6126(0.0778)	0.5988(0.0771)	1.0338(0.3660)	0.9982(0.3430)	-0.0874(0.4943)	-0.0841(0.4428)

SD is the robust standard error.

It could be observed from the simulation results that for large sample size, the bias is very small, while for relative small sample size, the bias would be significant and the corresponding bias-corrected GEE estimator for regression coefficient would have smaller variance than the GEE estimator. The simulation results for other cases with different working models and link functions are similar and are not reported here to save space.

6. Examples

Example 1: We consider the subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution that was analyzed by Fitzmaurice and Laird (1993). The data set contains complete records on 537 children from Steubenville, Ohio, each of whom was examined annually at ages 7, 8, 9 and 10. The repeated binary response is the wheezing status (1=yes, 0=no) of a child at each occasion. The purpose of the study is to model the probability of the wheezing status as a function of the child's age, his/her mother's maternal smoking habit (a binary variable MS with 1 if the mother smoked regularly and 0 otherwise) and their interactions. We consider the same marginal model used by Fitzmaurice and Laird (1993) with a logit link

$$\text{logit}(\mu) = \beta_0 + \beta_1\text{Age} + \beta_2\text{MS} + \beta_3\text{Age*MS}, \tag{6.1}$$

where 'age' is the age in years since the child's 9th birthday.

The GEE estimates and the robust standard errors of the regression parameters β_0, β_1, β_2 and β_3 are $-1.9005(0.1191)$, $-0.1412(0.0582)$, $0.3138(0.1878)$ and $0.0708(0.0883)$, respectively. The bias-corrected GEE estimates and the robust standard errors are $-1.8942(0.1185)$, $-0.1404(0.0579)$, $0.3160(0.1868)$ and $0.0706(0.0878)$, respectively. We can see that the differences between GEE and the bias-corrected GEE estimates are very small. This is because of the large sample size 537.

Example 2: In order to check what happens if the sample size is small, we consider a sub-sample of the data set of size 50. The data are given in Table 2.

For this data set the GEE estimates and the robust standard errors of the regression parameters β_0, β_1, β_2 and β_3 are $-2.3598(0.4900)$, $-0.1205(0.1853)$, $0.9674(0.6496)$ and $0.2854(0.2681)$, respectively. The corresponding bias corrected GEE estimates and the robust standard error are $-2.2404(0.4441)$, $-0.1079(0.1683)$, $0.9133(0.6049)$ and $0.2602(0.2503)$, respectively.

We see that there is significant difference between the GEE and the bias-corrected GEE estimates. Also, the standard errors of the bias-corrected GEE estimates are smaller, indicating that bias corrected estimates might have higher precision. This property was observed for many other small sub-samples investigated.

7. Discussion

In this paper we obtain a bias corrected GEE estimates of the regression parameters in longitudinal data. The proposed approach is based on correcting the GEE following a method by Cox and Snell (1968) which was later simplified by Cordeiro and Klein (1994). The bias corrected GEE estimates shows superior performance in terms of bias and efficiency compared to the GEE estimates for small samples. Some simulation results and an example provided confirms these findings.

The standard errors of the estimates in the two examples are calculated by the sandwich formula in the GEE estimation. When the sample size is small, the sandwich estimator usually underestimates the true variance of the estimate of the regression parameters. In this case, the two bias-corrected covariance estimators proposed by Kauermann and Carroll (2001) and Mancl and DeRouen (2001) can be used to correct the bias of the covariance estimator.

References

[1] Cordeiro, G. M. and Klein, R. (1994). *Bias correction in ARMA models.* Statist. Probab. Lett. 19, 169-176.

[2] Cox, D. R. and Snell, E. J. (1968). *A general definition of residuals (with discussion).* J. Roy. Statist. Soc. B 30, 248-275.

[3] Emrich, L. J. and Piedmonte, M. R. (1991). *A method for generating high-dimensional multivariate binary variates.* The American Statistician, 45, 302-304.

[4] Fitzmaurice, G. M. and Laird, N. M. (1993). *A likelihood-based method for analysing longitudinal binary responses*, Biometrika 80, 141-151.

[5] Kauermann, G. and Carroll, R. J. (2001). *A note on the efficiency of sandwich covariance matrix estimation.* J. Amer. Statistical Assoc., 96, 1387-1396.

[6] Liang, K.-Y. and Zeger S. L. (1986). *Longitudinal data analysis using generalized linear models.* Biometrika, 73, 45-51.

[7] Mancl, L. A. and DeRouen, T. A. (2001). *A covariance estimator for GEE with improved small-sample properties.* Biometrics, 57, 126-134.

[8] Wang, Y.-G. and Carey, V. (2003). *Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance.*Biometrika,90, 29-41.

Affiliation:

Sudhir Paul

Department of Mathematics and Statistics,University of Windsor, Winsor, ON N9B 3P4 Canada

E-mail: smjp@uwindsor.ca

Xuemao Zhang

Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 US

Jianwen Xu

Department of Statistics and Actuarial Science, Chongqing University, Chongqing 401331, China

Centre for Applications in Natural Resource Mathematics (CARM), School of Mathematics and Physics,

The University of Queensland, QLD 4072, Australia.

Journal of Environmental Statistics http://www.jenvstat.org

Volume 4, Issue 1 Submitted: 2012-07-20

Feburary 2013 Accepted: 2012-09-29

Table 2: A sample of the subset of data from the six cities study.

ID	Y_1	Y_2	Y_3	Y_4	MS	ID	Y_1	Y_2	Y_3	Y_4	MS
4	0	0	0	0	0	279	0	1	0	0	0
15	0	0	0	0	0	280	0	1	0	0	0
25	0	0	0	0	0	290	0	1	1	0	0
57	0	0	0	0	0	303	1	0	0	0	0
67	0	0	0	0	0	347	1	1	1	1	0
70	0	0	0	0	0	352	0	0	0	0	1
76	0	0	0	0	0	359	0	0	0	0	1
78	0	0	0	0	0	361	0	0	0	0	1
86	0	0	0	0	0	378	0	0	0	0	1
106	0	0	0	0	0	399	0	0	0	0	1
110	0	0	0	0	0	405	0	0	0	0	1
111	0	0	0	0	0	409	0	0	0	0	1
125	0	0	0	0	0	414	0	0	0	0	1
155	0	0	0	0	0	423	0	0	0	0	1
180	0	0	0	0	0	446	0	0	0	0	1
183	0	0	0	0	0	448	0	0	0	0	1
185	0	0	0	0	0	452	0	0	0	0	1
199	0	0	0	0	0	470	0	0	0	1	1
200	0	0	0	0	0	474	0	0	0	1	1
218	0	0	0	0	0	495	0	1	0	0	1
228	0	0	0	0	0	498	0	1	1	0	1
229	0	0	0	0	0	502	0	1	1	0	1
236	0	0	0	0	0	503	0	1	1	1	1
238	0	0	0	1	0	509	1	0	0	0	1
277	0	1	0	0	1	535	1	1	1	1	1