Austrian Journal of Statistics
April 2017, Volume 46, 3–11.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.666

Robust Regression Analysis of Longitudinal Data under Censoring

Somnath Datta
University of Florida

Abstract

We consider regression analysis of longitudinal data when the temporal correlation is modeled by an autoregressive process. Robust R estimators of regression and autoregressive parameters are obtained. Our estimators are valid under censoring caused by detection limits. Efficient computation of the estimators is discussed. Theoretical and simulation studies of the estimators are presented. We analyze a real data set on air pollution using our methodology.

Keywords: rank estimators, left-censoring, censored rank, reweighting.

1. Introduction

We consider a time series $\{X_t : t \geq 1\}$ and an associated series of covariate vectors $\{Z_t : t \geq 1\}$, in $\Re^q$, for some $q \geq 1$. We postulate a linear model of the form $X_t = \beta_0 + \beta' Z_t + \alpha_t$, where the model errors α_t, is a stationary autoregressive time series of order p, for some $p \geq 1 : \alpha_t = \phi_1 \alpha_{t-1} + \cdots + \phi_p \alpha_{t-p} + \epsilon_t$, where $\{\epsilon_t\}$ are i.i.d. from a symmetric continuous distribution, and $\alpha_s = 0$, for $s \leq 0$. We assume that the coefficients ϕ satisfy the usual invertibility condition.

In some situations, the exact values of X_t may be unavailable due to censoring. In this paper, we develop our methodology for the situation when the censoring is to the left which may occur when the values of the time series X_t fall below a detection limit D_t. Thus, the observed data consists of $X_t^c = X_t \vee D_t$ and the censoring indicators $\delta_t = I(D_t \leq X_t)$. Our method can easily be adopted to the case of right censored data by simple changes in various formulas leading to our estimators. They can also be extended to the case when an observation is doubly censored but it requires more work.

The number of papers dealing with some form of censored time series data is limited (Vasudaven et al., 1996) although Zeger and Brookmeyer (1986) argue that censoring may occur naturally in longitudinal studies when there are detection limits on the observation that are being collected in time. They took a fully parametric approach to the above problem and fitted a Gaussian error model using the maximum likelihood approach via an EM type algorithm. In this paper, we take an estimating equation approach that is a robustified form of the least squares estimating function.

There is a sizable literature on R-estimators in the regression context (Hettmansperger and McKean, 2011) with i.i.d. errors but not for auto-regressive errors. Furthermore, an added complication arises due to censoring. As use the "approximate unbiasedness" principle of re-weighting data to construct our R-type estimating equation for the set of regression and the error autoregression parameters. Since this estimating equation involve ranks of quantities that are not computable due to censoring, the re-weighting principle is used again to compute the approximate ranks to be used in the estimating equation.

The rest of the paper is organized as follows. Section 2 describes our estimation method and discusses an efficient method of computing the estimator. We also present a model based resampling procedure for making inference using our estimators. Section 3 presents results from a number of simulation studies demonstrating the performance of our estimators. We illustrate our methodology on a real dataset dealing with air pollution in Section 4. The paper ends with a discussion section (Section 5).

2. The estimators

We develop two different estimators of the regression and the error autoregression parameters. In the first approach, we ignore the fact that the models errors are dependent and estimate the regression parameters first which are then used to estimate the autoregressive parameters. In the second approach, a joint objective function of both sets of parameters is formed.

2.1. The complete data case

First we consider the situation when there is no censoring so that we have fully observed the time series $X_t, 1 \leq t \leq n$. We form an estimating equation that is partly based on ranks of certain model residuals and is therefore yields more robust estimators than the corresponding least squares estimators. Define, for any vector $b \in \Re^q$, the residuals for the linear model part $a_t(b) \doteq X_t - b^T Z_t$, for $1 \leq t \leq n$, and $a_t(b) \doteq 0$, for $t \leq 0$. Note that even though the true errors α_t are not independent, they are still ergodic and thus we could use the same estimating equation of a traditional R-estimation in this context. Thus, we could obtain a "quick and dirty" consistent estimator of β by minimizing the objective function

$$D_{1,M}(b) = \sum_{t=1}^{n} \left\{ \phi_1 \left(\frac{R(a_t(b))}{n+1} \right) - \overline{\phi}_1 \right\} a_t(b), \tag{1}$$

where $R(a_t(b))$ is the rank of $a_t(b)$ amongst $a_1(b), \cdots, a_n(b)$. Here ϕ_1 is defined on $(0,1)$ such that ϕ_1 is monotonic and $\int \phi_1^2 < \infty$, and $\overline{\phi}_1 = n^{-1} \sum_{i=1}^{n} \phi_1 \left(i/(n+1) \right)$. After obtaining an estimate $\widehat{\beta}$, the intercept parameter can be (robustly) estimated as $\widehat{\beta}_0 = \text{med}(a_1(\widehat{\beta}), \cdots, a_n(\widehat{\beta}))$.

Having estimated the regression parameters, a similar objective function can now be formed to estimate the autoregressive part of the error time series:

$$D_{2,M}(h) = \sum_{t=1}^{n} \phi_2 \left(\frac{R(e_t(h))}{n+1} \right) \left\{ e_t(h) - \overline{e}(h) \right\}, \tag{2}$$

where $e_t(h) \doteq \{a_t(\widehat{\beta}) - \widehat{\beta}_0\} - \sum_{j=1}^{p} h_j \{a_{t-j}(\widehat{\beta}) - \widehat{\beta}_0\}$, $1 \leq t \leq n$, $R(e_t(h))$ is the rank of $e_t(h)$ amongst $e_1(h), \cdots, e_n(h)$, $\overline{e}(h) = n^{-1} \sum_{i=1}^{n} e_i(h)$; ϕ_2 is defined on $(0,1)$ such that ϕ_2 is monotonic and $\int \phi_2^2 < \infty$. In the rest of the paper, we refer to these estimators as "partial R estimators".

Finally, we consider a second approach where estimators are obtained by minimizing a joint objective function. For $b_0 \in \Re$, $b \in \Re^q$ and $h \in \Re^p$, define the model residuals (as a function of b_0, b and h), by $e_t(b_0, b, h) \doteq a_t(b_0, b) - \sum_{j=1}^{p} h_j a_{t-j}(b_0, b)$, where $a_t(b_0, b) = X_t - b_0 + b' Z_t$, $t \geq 1$, and $e_t(b_0, b, h) \doteq 0$, for $t \leq 0$. We then form a joint objective function

$$D_J(b_0, b, h) = \sum_{t=1}^{n} \phi_3 \left(\frac{R(e_t(b_0, b, h))}{n+1} \right) \Big\{ e_t(b_0, b, h) - \overline{e}(b_0, b, h) \Big\}, \tag{3}$$

where R and ϕ_3 are as before. In accordance with the earlier name, the resulting estimators obtained by minimizing D_J will be called the "full R-estimators".

2.2. Modification for censored data

Next, we will describe how to modify this estimating function in presence of left-censoring. Both the estimating function and the ranks have to be computed on the basis of observed data. However, in order to avoid any selection bias, the contribution of such a term has to be re-weighted by the corresponding inverse selection probability. These probabilities will have to be estimated from appropriate models fitted to the censoring distribution.

The censored data version of the objective function corresponding to (1) is of the form

$$D_{1,M}(b) = \sum_{t=1}^{n} \frac{\delta_t}{W_t} \Big\{ \phi_1 \left(\frac{R^c(a_t(b))}{n+1} \right) - \overline{\phi}_{1,c} \Big\} a_t(b), \tag{4}$$

with $\overline{\phi}_{1,c} = \left\{ \sum_{t=1}^{n} \frac{\delta_t}{W_t} \phi_1 \left(\frac{R^c(a_t(b))}{n+1} \right) / \sum_{t=1}^{n} \frac{\delta_t}{W_t} \right\}$, where the presence of δ_t indicates that the corresponding $a_t(b)$ is computable from the available data and W_t is the corresponding selection weight that is described later. The quantity R^c denotes a modified "rank" that accounts for censoring. To motivate the definition of rank for censored data, it will be useful to first consider the following sum representation for $R(a_t(b))$ in the full (uncensored) data situation:

$$R(a_t(b)) = \sum_{j=1}^{n} I[a_j(b) \leq a_t(b)].$$

Using the same re-weighting principle as before, we can define an "estimated rank" that is computable from left censored data by

$$R^c(a_t(b)) = \frac{\sum_{j=1}^{n} \frac{\delta_j}{W_j} I[a_j(b) \leq a_t(b)]}{\frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{W_j}},$$

for any t with $\delta_t = 1$.

In the censored data case, a robust estimator of the intercept term can be obtained as

$$\widehat{\beta}_0 = \widehat{F}_a^{-1} \left(\frac{1}{2} \right), \tag{5}$$

where F_α is an estimator of the distribution function of the stationary distribution of the a_t based on the same re-weighting principle

$$\widehat{F}_\alpha(t) = \left(\sum_{t=1}^{n} \frac{\delta_t}{W_t} I[\, a_t(\widehat{b}\,) \leq t] \right) \Big/ \left(\sum_{t=1}^{n} \frac{\delta_t}{W_t} \right).$$

Next note that, in order to calculate the residual function $e_t(h)$ corresponding to time t, we need to have complete (i.e., uncensored) observations on X_j, $t-p \leq j \leq t$. Thus, the modified objective function for estimating ϕ will be of the form

$$D^c_{2,M}(h) = \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \phi_2 \left(\frac{R^c(e_t(h))}{n+1} \right) \Big\{ e_t(h) - \overline{e}(h) \Big\}, \tag{6}$$

where the Ws are as before,

$$\overline{e}(h) = \left\{ \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) e_t(h) \right\} \Big/ \left\{ \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \right\}$$

and

$$R^c(e_t(h)) = \frac{\sum_{j=1}^{n} \left(\prod_{k=j-p}^{j} \frac{\delta_k}{W_k} \right) I[e_j(h) \le e_t(h)]}{\frac{1}{n} \sum_{j=1}^{n} \left(\prod_{k=j-p}^{j} \frac{\delta_k}{W_k} \right)}, \tag{7}$$

for any t with $\prod_{j=t-p}^{t} \delta_j = 1$.
In the same way, the joint objective function can be modified to account for censored data as

$$D_J^c(b_0, b, h) = \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \phi_3 \left(\frac{R^c(e_t(b_0, b, h))}{n+1} \right) \Big\{ e_t(b_0, b, h) - \overline{e}(b_0, b, h) \Big\}, \tag{8}$$

where W_j are the same as before and R^c is similarly defined as in (7).

2.3. Computation of the estimator

The estimating function can be optimized using a general purpose optimizer such as "optim" or "optimize" in R. For the $p = q = 1$ case, we can perform a grid search algorithm which we describe below.

Note that $D_{1,M}(b)$ is a linear function in b in regions where the ranks $R^c(a_t(b))$ do not change for t's with $\delta_t = 1$. For b to be such a change point, there will exist pairs of integers t and i such that $\delta_t = \delta_i = 1$ and $a_t(b) = a_i(b)$. Thus $X_t - bZ_t = X_i - bZ_i$ implying $b = (X_i - X_t)/(Z_i - Z_t)$, provided $Z_i \neq Z_t$. Let $\mathcal{B} = \{b_j : j = 1, \cdots, M\}$ be the sorted values of $\{(X_i - X_t)/(Z_i - Z_t) : 1 \le i \neq t \le n,\ \delta_i \delta_t = 1,\ Z_i \neq Z_t\}$. Then $D_{1,M}(b)$ is piecewise linear and continuous on $\mathcal{B}$. Hence $\widehat{\beta}$ can be obtained by maximizing $D_{1,M}$ on the grid of points $\mathcal{B}$. If $D_{1,M}(b)$ is constant on $[b_j, b_{j+1}]$, we will take $\widehat{\beta}$ to be the midpoint $(b_j + b_{j+1})/2$.
In the same way, $D_{2,M}^c$ can be maximized over the grid of points

$$\mathcal{H} = \left\{ \left(a_t(\widehat{\beta}) - a_i(\widehat{\beta}) \right) / \left(a_{t-1}(\widehat{\beta}) - a_{i-1}(\widehat{\beta}) \right) : 1 \le i \neq t \le n,\ \delta_i \delta_{i-1} \delta_t \delta_{t-1} = 1,\ a_{t-1}(\widehat{\beta}) \neq a_{i-1}(\widehat{\beta}) \right\}.$$

2.4. Computation of the weights

The weights W_j are estimates of the conditional (given X_j and Z_j) cumulative distribution function of the D_j, i.e., $W_j = \widehat{Pr}\{D_j \le X_j | X_j, Z_j\}$. The simplest way to estimate these will be to consider the corresponding (forward in time) hazard of $C_j = -D_j$, given X_j, Z_j,

$$\begin{aligned} &\lim_{dc \downarrow 0} \frac{\lambda_c(c \le C_j < c + dc | C_j \wedge (-X_j) \ge c, X_j, Z_j)}{dc} \\ &= \lim_{dc \downarrow 0} \frac{\lambda_c(c \le C_j < c + dc | C_j \wedge (-X_j) \ge c, Z_j)}{dc} \ \overset{=}{\cdot} \ \lambda_c(c | Z_j), \end{aligned}$$

where the equality is an independent censoring assumption that we impose throughout this paper. We now need a regression model on these hazard rates on the C. A flexible model is given by Aalen's linear hazards model that admits a closed form estimates of these quantities; see, e.g., Aalen (1989) or Datta and Satten (2002). A special case of these models, where we assume that $\lambda_c(c|Z_j)$ is free of the covariate Z_j, also yields the simplest choice of W_j obtained by the Kaplan-Meier estimator of the survival function based on the (right censored) C_j evaluated at $(-X_j)^-$.

2.5. Bootstrap inference

While it is possible to develop a large sample theory for our estimators by combining elements from Hettmansperger and McKean (2011), Datta and Satten (2002), and Datta and Beck (2014), we prefer to use model based resampling to perform statistical inference since it avoids the use of tuning parameter that is necessary for smoothing based asymptotic variance estimation.

Having fitted the regression model to the original data, we compute the model residuals $\widehat{\epsilon}_t = \widehat{\alpha}_t - \widehat{\phi}_1\widehat{\alpha}_{t-1} - \cdots - \widehat{\phi}_p\widehat{\alpha}_{t-p}$, with $\widehat{\alpha}_t = X_t - \widehat{\beta}_0 - \widehat{\beta}'Z_t$. Next, we resample the centered residuals to obtain ϵ_t^* which are used to compute $\alpha_t^* = \widehat{\phi}_1\alpha_{t-1}^* + \cdots + \widehat{\phi}_p\alpha_{t-p}^* + \epsilon_t^*$, and finally the bootstrapped complete data $X_t^* = \widehat{\beta}_0 + \widehat{\beta}'Z_t^* + \alpha_t^*$. The corresponding censoring times are independently generated from the fitted censoring hazard rate function based on the original data $D_t^* \sim \widehat{\lambda}_C$. Finally, we let $X_{t^c}^* = X_t^* \vee D_t^*$ and $\delta_t^* = I(D_t^* \leq X_t^*)$.

The rest of the bootstrap procedure is standard leading to either a percentile based confidence interval or a large sample normality based confidence interval where the asymptotic variance is replaced by the empirical variance of independent replicates of bootstrapped estimates of the parameter of interest.

3. Simulation studies

We consider a single continuous covariate Z that is generated from a $N(0, .64)$ distribution; we simulate the errors from an AR(1) model $\alpha_t = 0.5\alpha_{t-1} + \epsilon_t$. A number of distributions for the ϵ were investigated. The regression parameters used for the simulation were $\beta_0 = 2$ and $\beta_1 = 1$. The censoring times were generated as $D = 1/(E + 3) + m$, where E has a standard exponential distribution and $m \in \Re$ is chosen to control the censoring rate. Three choices of the censoring rates were used. The bias and variance of the estimators were empirically estimated based on $M = 1000$ Monte Carlo samples each.

Table 1 reports the results of the simulation. Some general trends are observed from this table. The joint estimators of the slope and autoregressive parameters have better performance that the corresponding partial estimators. The joint estimator of intercept parameter, on the other hand, exhibits substantial bias which worsens with the amount of censoring; however it is corrected by the modified estimator in all cases. The standard deviation, of the estimators of slope and autoregression parameters increases, albeit slightly, with the censoring level.

Next we compute the empirical coverage of the bootstrap based confidence intervals using the percentile methods, as well as the standardized statistic using bootstrap based variance estimate. The coverage appears to be very good when there is no censoring and is adequate even with 30% censoring (Table 2). Overall, confidence intervals using standardized statistics have better coverage as expected.

4. An application

We illustrate our methodology using monthly data on the chemical composition of atmospheric deposition of dry NH_4 collected by the Environmental Measurements Laboratory between 1977 and early 1980 at a number of sites in the United States (Toonkel 1981); the same data set was used by Zeger and Brookmeyer (1986) to illustrate their method. Since there are lower detection limits of the assays, the data is left-censored. Altogether, there were 43 data points out of which 6 were left-censored. In addition, there were three observations that were missing; in order to accommodate them into our framework, we treat them as left censored by an artificially set high value (larger than all the observed values in the data set). The data were log-transformed as in Zeger and Brookmeyer (1986). A plot of the log-transformed data is shown in Figure 1; where the incomplete observations are denoted by the symbol "+".

One of the main research question was to determine if the amount of deposit is increasing

with time. To that end we fit a regression model taking time as a covariate of the form $X_t = \beta_0 + \beta_1 t + \alpha_t$, where α_t was modeled by an AR(1) process. The resulting parameter estimates are given in Table 1; we include the parametric estimates by Zeger and Brookmeyer (ZB) for comparison.

While the two sets of parametric estimates are similar, the robust estimate of the intercept term is slightly smaller. More importantly, all confidence intervals for the slope term include 0 indicating that there is no significant change of the deposit levels with time. In a sense, the fact that the different analyses yielded the same scientific conclusion is reassuring.

5. Discussion

We introduce a robust and relatively model free technique of analyzing temporally correlated data that are subject to left-censoring. Although, our formulas are given here for left censored data, it is a mater of triviality to change them for right censored. With additional effort, it may be possible to extend the basic regression technique to other form of incomplete data. Another technical extension will be to consider other form of temporal correlation structures for the longitudinal responses.

This paper presents a number of novel components which may be useful for other incomplete data problems. In particular, the concept of an approximate or estimated "rank" may be applied to extend other rank based inference for censored data settings.

References

Aalen O.O. (1989). "A Linear Regression Model for the Analysis of Lifetimes." *Statistics in Medicine*, **8**, 907–925.

Datta S. and Beck J.D. (2014). "Robust Estimation of Marginal Regression Parameters in Clustered Data." *Statistical Modelling*, **14**, 489–501.

Datta S. and Satten G.A. (2002). "Estimation of Integrated Transition Hazards and Stage Occupation Probabilities for Non-Markov Systems under Stage Dependent Censoring." *Biometrics*, **58**, 792–802.

Hettmansperger T.P. and McKean J.W. (2011). *Robust Nonparametric Statistical Methods, 2nd ed.*. New York: Chapman & Hall.

Toonkel L.E. (1981). Appendix to Environmental Measurements Laboratory Environmental Report, New York: U.S. Department of Energy (available from the National Technical Information Service, U.S. Department of Commerce, Springfield, VA).

Vasudaven M. and Nair M.G. and Sithole M.M. (1996). "On Estimation for Censored Autoregressive Data." *Statistics & Probability Letters*, **31**, 97–105.

Zeger S.L. and Brookmeyer R. (1986). "Regression Analsis with Censored Autocorrelated Data." *Journal of the American Statistical Association*, **81**, 722–729.

Table 1: Performance of various estimators as measured by empirical bias and standard deviation in a simulation experiment.

			Partial R-estimators			Joint R-estimators			Modified
Sample size	Censoring %	Parameter	β_0	β_1	ϕ_1	β_0	β_1	ϕ_1	β_0
50	40%	Bias	0.002	-0.050	-0.086	-0.289	-0.032	-0.067	0.002
		SD	0.154	0.111	0.146	5.654	0.094	0.148	0.153
	30%	Bias	-0.015	-0.004	-0.070	-0.012	-0.001	-0.053	-0.015
		SD	0.157	0.107	0.138	1.429	0.087	0.139	0.157
	0%	Bias	-0.015	0.003	-0.067	-0.016	0.004	-0.051	-0.015
		SD	0.158	0.107	0.136	1.411	0.086	0.138	0.157
200	40%	Bias	0.009	-0.057	-0.040	0.042	-0.036	-0.034	0.009
		SD	0.072	0.057	0.069	1.084	0.048	0.069	0.072
	30%	Bias	-0.009	-0.011	-0.022	0.052	-0.007	-0.018	-0.009
		SD	0.073	0.053	0.063	1.075	0.043	0.064	0.073
	0%	Bias	-0.010	-0.004	-0.019	0.055	-0.002	-0.015	-0.010
		SD	0.073	0.053	0.063	1.072	0.043	0.063	0.073
500	40%	Bias	0.020	-0.050	-0.027	0.035	-0.031	-0.023	0.020
		SD	0.049	0.036	0.045	1.042	0.028	0.045	0.048
	30%	Bias	0.002	-0.005	-0.010	0.043	-0.003	-0.008	0.001
		SD	0.049	0.033	0.040	1.046	0.025	0.039	0.049
	0%	Bias	0.001	0.001	-0.008	0.044	0.001	-0.006	0.000
		SD	0.049	0.033	0.040	1.042	0.025	0.039	0.049

Table 2: Empirical coverages of bootstrap confidence intervals in simulated data

Censoring Rate		Parameter					
		Percentile Method			Normal Approximation		
	Nominal Coverage	β_0	β_1	ϕ_1	β_0	β_1	ϕ_1
0%	80%	0.772	0.766	0.794	0.800	0.772	0.802
	85%	0.826	0.810	0.826	0.850	0.814	0.852
	90%	0.880	0.882	0.888	0.896	0.880	0.902
	95%	0.926	0.940	0.938	0.940	0.952	0.950
	99%	0.964	0.990	0.984	0.974	0.988	0.984
30%	80%	0.732	0.730	0.800	0.742	0.732	0.818
	85%	0.776	0.778	0.848	0.802	0.792	0.870
	90%	0.836	0.832	0.898	0.846	0.840	0.904
	95%	0.882	0.924	0.938	0.912	0.922	0.942
	99%	0.964	0.978	0.984	0.980	0.984	0.988
40%	80%	0.714	0.426	0.662	0.756	0.644	0.748
	85%	0.766	0.472	0.720	0.806	0.702	0.808
	90%	0.828	0.560	0.780	0.854	0.760	0.874
	95%	0.892	0.678	0.860	0.928	0.842	0.936
	99%	0.954	0.846	0.958	0.974	0.946	0.974

Table 3: Parameter estimates for the Dry Deposition data

Parameter	β_0	β_1	ϕ_1	CI for ϕ_1
Our method	4.646	0.017	0.315	$(-0.018,\ 0.042)$ BS percentile method $(-0.011,\ 0.045)$ Using normal approximation
ZB	5.020	0.015	0.380	$(-0.042,\ 0.066)$

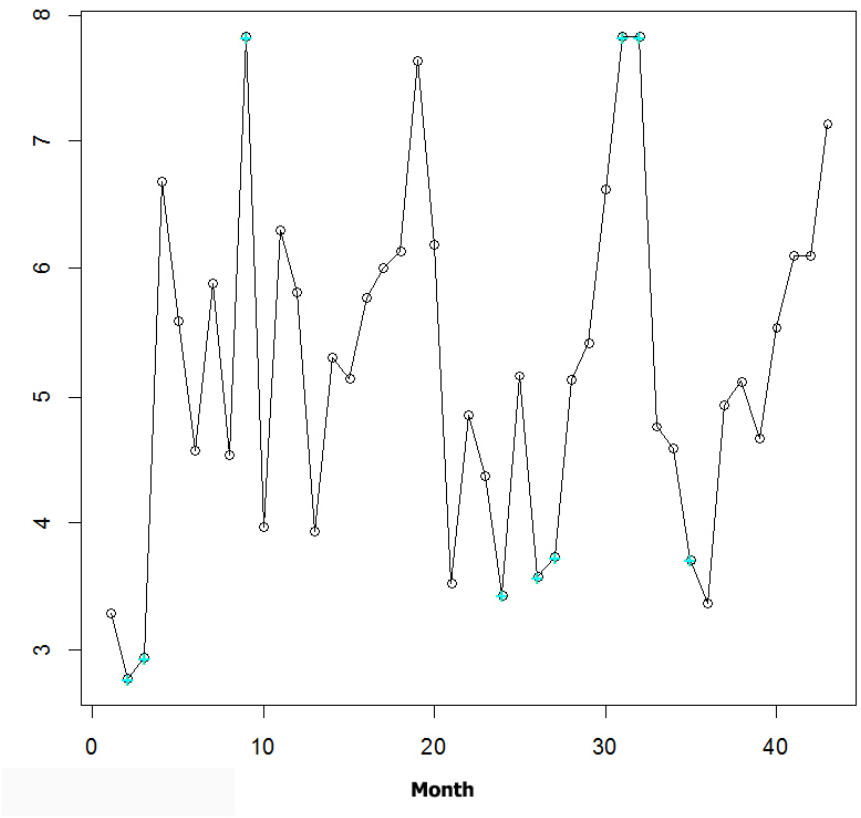

Figure 1: Log-transformed data of monthly deposition of dry NH_4; the incomplete data values are indicated by "+".

Affiliation:

Somnath Datta
Department of Biostatistics
University of Florida
Gainesville, FL 32610
E-mail: somnath.datta@ufl.edu
URL: http://www.somnathdatta.org/uf

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02