RT-MAE 9821

A ROBUST ESTIMATION OF GROWTH CURVES

by

Elisete C. Quintaneiro Aubin
and
Gabriela Stangenhaus

Palavras-Chave: Bootstrap method; Growth curves models; M-estimators; Minimum volume ellipsoid estimator (MVE); Robust estimation procedures.

Classificação AMS: 62J99, 62G35.
(AMS Classification)

- Agosto de 1998 -

A Robust Estimation of Growth Curves

Elisete C. Quintaneiro Aubin
Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil

Gabriela Stangenhaus
Statistika Consultoria, 13015-002, Campinas, SP, Brazil

Abstract

A robustified version of the parameter matrix estimators in the standard growth curve model is obtained via the Potthoff-Roy transformation. The asymptotic distribution of the robust estimators is derived and the estimation of their variance-covariance matrix is discussed.

1. Introduction

Growth curve models arise naturally in experiments where multiple measurements are recorded over a period of time on each sampling unit. Outlying observations are common in data with multiple observations on each sampling unit. In recent years, statisticians have become aware of the large influence outliers have upon maximum likelihood estimates, under the assumption of normality, and on least squares estimates, generally.

Several robust procedures have been developed to overcome the requirement of the basic assumption of multivariate normality and that are resistant to outliers. Ghosh, Grizzle and Sen (1973) developed robust and asymptotically efficient inference procedures based on rank statistics for the statistical analysis of longitudinal data that do not require the assumption of multivariate normality of the underlying distribution. Lange, Little and Taylor (1993) suggested an anlytical strategy, based on maximum likelihood for a general model with multivariate t errors, for the statistical modeling of data sets involving errors with larger-than-normal tails. Applications were presented, including repeated measures data. Pendergast and Broffitt(1985) discussed two robust estimators of the growth curve's location and scatter parameters in the general linear model considered by Potthoff and Roy (1964) based upon M-estimation. In the same line, Singer and Sen (1986) obtained two types of multivariate M-estimators of the parameter matrix in the standard growth curve model, via the Potthoff-Roy transformation. They followed Maronna (1976) to define the M-estimators. The M-estimation procedures try to downweight the outliers and depend on a tuning constant. Asymptotic distributions of these robust estimators were presented.

The object of the present article is to suggest a simple robust estimation procedure based on the median of robustified estimators of the profiles, for the growth curve models considered by Potthoff and Roy (1964).

In the following section we present the robust method with an application to the dental study discussed by Potthoff and Roy (1964). Section 3 contains the asymptotic results and discusses the use of bootstrap methods to estimate the asymptotic variance-covariance matrix of the estimates. Conclusions follow in Section 4.

2. A Median Based Method

Potthoff and Roy (1964) introduced a growth curve model based on a general statistical model which includes as special cases regression models and both univariate and multivariate analysis models. Consider the linear model

$$Y = X\beta G + \varepsilon \tag{2.1}$$

where Y is a $N \times p$ matrix of observations on each of N subjects, X is a $N \times r$ between-subjects design matrix of rank r, β is a $r \times q$ matrix of unknown parameters,

G is a $q \times p$ within subjects design matrix of rank q and ε is a $N \times p$ matrix of random errors whose rows are independent with density f, location vector $\mathbf{o}$ and positive definite scatter matrix $\sum$.

The transformation

$$Y^* = Y\Delta^{-1}G'(G\Delta^{-1}G')^{-1} \tag{2.2}$$

leads to the standard multivariate linear model

$$Y^* = X\beta + \varepsilon^* \tag{2.3}$$

where Δ is a $p \times p$ positive definite matrix.

The least squares estimator of β is

$$\tilde{\beta}_{LS} = (X'X)^{-1}X'Y^* = (X'X)^{-1}X'Y\Delta^{-1}G'(G\Delta^{-1}G')^{-1} \tag{2.4}$$

and it can easily be seen to be the componentwise average of the profile estimators

$$\tilde{\beta}_{LS}^{(i)} = Y_i'\Delta^{-1}G'(G\Delta^{-1}G')^{-1} \tag{2.5}$$

where Y_i' is the i-th row of Y.

The maximum likelihood estimators of β and $\sum$, when the rows of ε follow a p-variate normal distribution with mean $\mathbf{o}$ and dispersion matrix $\sum$, were derived by Khatri (1966):

$$\tilde{\beta}_{ML} = (X'X)^{-1}X'YS^{-1}G'(GS^{-1}G')^{-1} \tag{2.6}$$
$$\widetilde{\sum}_{ML} = N^{-1}(Y - X\tilde{\beta}_{ML}G)'(Y - X\tilde{\beta}_{ML}G) \tag{2.7}$$

where $S = Y'(I - X(X'X)^{-1}X')Y$.

The presence of outliers in the data set inflates S and has a strong influence in the estimate (2.6) through the averaging of the profile estimates. This estimation procedure based on the profile estimates (2.5) can be robustified by letting Δ be a robust estimate of the dispersion matrix of $Y_i', i = 1, 2, ..., N$, denoted by C, and the estimator of β given by the componentwise median of the profiles estimates. So, the robustified estimator can be written as

$$\widehat{\beta}^{(i)} = Y_i'C^{-1}G'(GC^{-1}G')^{-1}, i = 1, ..., N, \tag{2.8}$$

$$\widehat{\beta} = \text{ Median } \{\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}, ..., \widehat{\beta}^{(N)}\} \tag{2.9}$$

and $$\widehat{\Sigma} = N^{-1}(Y - X\widehat{\beta}G)'(Y - X\widehat{\beta}G) \tag{2.10}$$

Robust estimation of a dispersion matrix was studied by Rousseeuw and Leroy (1987). We use here the Minimum Volume Ellipsoid estimator (MVE) with maximal breakdown point. It can be noticed that robustified methods can also be obtained by using alternative estimators to the least squares, as the L_1-method.

A classical example of growth curve given by Potthoff and Roy (1964), and reanalyzed by Pendergast and Broffitt (1985) and Singer and Sen(1986) consists of a data set collected by investigators at the University of North Carolina Dental School. The distance, in milimeters, from the center of the pituitary to the pteriomaxillar fissure of each of 11 girls and 16 boys at the ages of 8, 10, 12 and 14 years was measured. The distance is modelled as a simple linear function of time

$$Y_1(t_\ell) = \alpha_1 + \beta_1 t_\ell + \varepsilon \text{ (girls)},$$
$$Y_2(t_\ell) = \alpha_2 + \beta_2 t_\ell + \varepsilon \text{ (boys)}, \ \ \ell = 1, 2, 3, 4.$$

The following MVE estimate was computed using the program MINVOL: Version 1995, kindly provided by Prof. Peter Rousseeuw, whose algorithm is described in Rousseeuw and Leroy (1987),

$$C = \begin{bmatrix} 5.240 & 3.928 & 5.414 & 5.132 \\ & 3.668 & 4.828 & 4.681 \\ & & 6.744 & 6.483 \\ & & & 7.760 \end{bmatrix}.$$

Expressions (2.8) and (2.9) provide the estimates

$$\widehat{\alpha}_1 = 18.091 \quad , \quad \widehat{\beta}_1 = 0.427$$
$$\widehat{\alpha}_2 = 17.338 \quad , \quad \widehat{\beta}_2 = 0.638$$

Robust distances based on MVE, obtained with MINVOL, pointed out two outlying observations, namely, the ninth and thirteenth boy, also noted by Pendergast and Broffitt (1985). Nevertheless, the M-methods were not able to sufficiently downweight these observations. This was observed by the closer agreement among the least-squares and M-estimates reported for the boys growth curve:

$$\begin{aligned}
\widetilde{\alpha}_{2,LS} &= 18.697 \quad , \quad \widetilde{\beta}_{2,LS} = .784 \text{ (least squares)} \\
\widehat{\alpha}_{2,PB} &= 19.274 \quad , \quad \widetilde{\beta}_{2,PB} = .743 \text{ (Pendergast-Broffitt)} \\
\widehat{\alpha}_{2,SS} &= 18.930 \quad , \quad \widehat{\beta}_{2,SS} = .745 \text{ (Singer-Sen)}
\end{aligned}$$

3. Asymptotic Results

Let $\widehat{\beta}^{(j)} = (\widehat{\beta}_1^{(j)}, ..., \widehat{\beta}_p^{(j)}), j = 1, 2, ..., N_k$, be the estimator of the j-th profile in-group k. They are independent vectors with marginal distribution function F_i and density function $f_i, i = 1, 2, ..., p, N_1 + N_2 + ... N_r = N$. Let F_{i_1,i_2} be the joint distribution of $\widehat{\beta}_{i_1}^{(j)}$ and $\widehat{\beta}_{i_2}^{(j)}, j = 1, 2...., N_k, i_1 \neq i_2 = 1, ..., p$ and $\sigma_{i_1 i_2} = F_{i_1,i_2}(\theta_{i_1}, \theta_{i_2}) - 1/4$, where θ_i is the median of f_i.

If $F_i, i = 1, 2, ..., p$, are continuous and twice differentiable in a neighborhood of θ_i and $\delta_i = f_i(\theta_i) > 0, i = 1, 2, ..., p$, then the asymptotic distribution of $(\sqrt{N}(\widehat{\beta}^{(1)} - \theta_1, ..., \widehat{\beta}^{(p)} - \theta_p))$ is p-variate normal with mean vector $\mathbf{o}$ and variance-covariance matrix

$$\sum{}_R = \begin{bmatrix} \frac{1}{4\delta_1^2} & \frac{\sigma_{12}}{\delta_1\delta_2} & \cdots & \frac{\sigma_{1p}}{\delta_2\delta_p} \\ \\ \frac{\sigma_{21}}{\delta_1\delta_2} & \frac{1}{4\delta_2^2} & \cdots & \frac{\sigma_{2p}}{\delta_2\delta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\sigma_{p1}}{\delta_p\delta_1} & \frac{\sigma_{p2}}{\delta_p\delta_2} & \cdots & \frac{1}{4\delta_p^2} \end{bmatrix} . \tag{3.1}$$

The derivation of the asymptotic distribution follows Babu and Rao (1988) closely.

The $(i,j)th$ element of $\sum_R$ can be consistently estimated using consistent estimates of σ_{ij} and $\delta_i, \widehat{\sigma}_{ij}$ and $\widehat{\delta}_i$, respectively, leading to $\widehat{\sigma}_{ij}/\widehat{\delta}_i\widehat{\delta}_j$. A direct estimate of $\sigma_{ij}/\delta_i\delta_j$ can be obtained by the bootstrap method

$$\widehat{\sigma}_{ij}/\widehat{\delta}_i\widehat{\delta}_j = E^*[n(\theta_i^* - \widehat{\theta}_i)(\theta_j^* - \widehat{\theta}_j)] \tag{3.2}$$

where E^* is the expectation under the bootstrap distribution function. The estimator (3.2) is also consistent (see Babu (1986) and Babu and Rao (1988)). For a bootstrap scheme one can resample whole cases or resample residuals.

Tests of significance based on the medians can be devised as in Babu and Rao (1988) to compare the growth curves of the r distinct groups.

4. Conclusion

A set of procedures to analyse growth curve models can be obtained by robustifying traditional methods of growth curve models estimation via the use of a resistant estimator of the variance-covariance matrix and the median of the profile estimates. The profiles, on the other hand, can also be fitted by a robust method, as the L_1 method. We believe that this procedure can be extended to non-linear growth curves.

5. References

Babu, G.J. (1986), "A Note on Bootstrapping the variance of sample quantile", *Annals of the Institute of Statistical Mathematics*, **38**, Part A, 439-443.

Babu, G.J. and Rao, C.R. (1988), "Joint Asymptotic Distribution of Marginal Quantiles and Quantile Functions in Samples from Multivariate Population", *Journal of Multivariate Analysis*, **27**, 15-23.

Ghosh,M., Grizzle, J.E. and Sen, P.K. (1973), "Nonparametric Methods in Longitudinal Studies", *Journal of the American Statistical Association*, **68**, 29-36.

Khatri, C.G. (1966), "A Note on a MANOVA Model Applied to Problems in Growth Curve", *Annals of the Institute of Statistical Mathematics*, **18**, 75-86.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989), "Robust Statistical Modeling Using the t Distribution", **84**, 881-896.

Maronna, R.A. (1976), "Robust M-estimation of Multivariate Location and Scatter", *Annals of Statistics*, **4**, 51-67.

Pendergast, J.F. and Broffitt, J.D.(1985), "Robust Estimation in Growth Curve Models". *Communications in Statistics - Theory and Methods*, **14(8)**, 1919-1939.

Potthoff, R.F. and Roy, S.N.(1964), "A generalized multivariate analysis of variance model useful especially for growth curve problems", *Biometrika*, **51**, 313-326.

Rousseeuw, P.J. and Leroy, A.M.(1987), *Robust Regression and Outlier Detection*, Wiley-Interscience, New York.

Singer, J.M and Sen, P.K. (1986), "M-Methods in Growth Curve Analysis", *Journal of Statistical Planning and Inference*, **13**, 251-261.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

9801 - GUIOL, H. Some properties of K-step exclusion processes. 1998. 21p. (RT-MAE-9801)

9802 - ARTES, R., JØRSENSEN, B. Longitudinal data estimating equations for dispersion models. 1998. 18p. (RT-MAE-9802)

9803 - BORGES, W.S.; HO, L.L. On capability index for non-nornal processes. 1998. 13p. (RT-MAE-9803)

9804 - BRANCO, M.D. , BOLFARINE, H., IGLESIAS, P., ARELLANO-VALLE, R. Bayesian analysis of the calibration problem under elliptical distributions. 1998. 15p. (RT-MAE-9804)

9805 - IGLESIAS, P., MATÚS, F., PEREIRA, C.A.B., TANAKA, N.I. On finite sequences conditionally uniform given minima and maxima. 1998. 13p. (RT-MAE-9805)

9806 - IGLESIAS, P., PEREIRA, C.A.B., TANAKA, N.I. Characterizations of multivariate spherical distribution in l_∞ norm. 1998. 19p. (RT-MAE-9806)

9807 - GIAMPAOLI, V., SINGER, J.M. A note on the comparison of two normal populations with restricted means. 1998. 12p. (RT-MAE-9807)

9808 - BUENO, V.C. A note on conditional reliability importance of components. 1998. 07p. (RT-MAE-9808)

9809 - GIMENEZ, P., BOLFARINE, H. Consistent estimation in comparative calibration models. 1998. 19p. (RT-MAE-9809)

9810 - ARTES, R., PAULA, G.A., RANVAUD, R. Estudo da direção tomada por pombos através de equações de estimação para dados circulares longitudinais. 1998. 16p. (RT-MAE-9810)

9811 - LIMA, A.C.P., SEM, P.K. Time-Dependent Coefficients in a Multi-Event Model for Survival Analysis. 1998. 21p. (RT-MAE-9811)

9812 - KOLEV, N., MINKOVA, L. Over-and underdispersed models for ruin probabilities. 1998. 34p. (RT-MAE-9812)

9813 - KOLEV, N. Negative Binomial Cross-Classifications. 1998. 16p. (RT-MAE-9813)

9814- BAKEVA, V., GEORGIEVA, M., KOLEV, N. Two Characterizations of the Geometric Distribution Related to an Unreliable $Geo/G_D/1$ System. 1998. 11p. (RT-MAE-9814)

9815 - KOLEV, N., MINKOVA, L. Some basic Inflated-Parameter Discrete Distributions. 1998. 19p. (RT-MAE-9815)

9816 - ARTES, R., PAULA, G.A., RANVAUD, R. Strategy for the analysis of circular longitudinal data based on generalized estimating equations. 1998. 15p. (RT-MAE-9816)

9817 - PAULA, G.A., ARTES, R. A Score-Type Test to Assess Overdispersion in Linear Logistic Models. 1998. 17p. (RT-MAE-9817)

9818 - LOSCHI, R.H., IGLESIAS, P.L., WECHSLER, S. Unpredictability and Probability Updating. 1998. 11p. (RT-MAE-9818)

9819 - KOLEV, N., MINKOVA, L. New Over-/Under- Dispersed Class of Inflated-Parameter Discrete Probability Distributions. 1998. 10p. (RT-MAE-9819)

9820 - KOLEV, N., MINKOVA, L. Inflated-Parameter Family of Generalized Power Series Distributions. 1998. 11p. (RT-MAE-9820)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05315-970 - São Paulo, Brasil