A robust estimation method for a growth curve model with balanced design

Kenichi Satoh

Research Institute for Radiation Biology and Medicine

Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, JAPAN.

Hirokazu Yanagihara

Graduate School of Science, Hiroshima University

1-3-1 Kagamiyama, Higashi-Hiroshima 739-8626, JAPAN.

Ken-ichi Kamo

Division of Mathematics, Sapporo Medical University

South 1, West 17, Chuo-ku, Sapporo, Hokkaido 060-8556, JAPAN.

Abstract

Repeated measurements at several time points can be described by a growth curve model. However, there might exist a few individual curves that are located far away from the mean trend or have large variances. These individual curves can be candidates for outliers and robust estimators of the regression coefficients are desired. In this paper, an estimation method is proposed that uses weighted least square estimators, where the weights are optimized from squared residuals following the gamma distribution.

Key words and Phrases: Growth curve model; Longitudinal data; Repeated measurements; Robust estimation.

Mathematics Subject Classification: 62H12, 62J05, 62P10.

1. Introduction

Let $y_i(t)$ be the growth amount of the ith individual at time t $(i = 1, \ldots, n)$, where n

is the sample size. Let $\boldsymbol{y}_i = (y_i(t_1), \ldots, y_i(t_p))'$ be a p-dimensional response variables vector, $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{ik})'$ be a k-dimensional between-individuals design vector, $X = (\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_p))$ be a $q \times p$ within-individuals design matrix and $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_1), \ldots, \varepsilon_i(t_p))'$ be a p-dimensional error vector distributed according to $N_p(\mathbf{0}_p, \Sigma)$. Then the $GMANOVA$ model proposed by Potthoff and Roy (1964) can be expressed as $y_i(t) = \boldsymbol{a}_i' \Theta \boldsymbol{x}(t) + \varepsilon_i(t)$ where $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$ is a $k \times q$ unknown regression coefficient matrix. We assume that all individuals are observed at the same time points and the numbers of repetitions are also the same. More general time point designs have been studied by Laird and Ware (1982), Vonesh and Carter (1987) and others.

The mean is described in terms of interactions between between-individual covariates and within-individual covariates. Note that $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_k(t))' = \Theta \boldsymbol{x}(t)$ can be considered as time-varying coefficients as proposed by Hastie and Tibshirani (1993), *i.e.*, $E[y_i(t)] = \boldsymbol{a}_i' \Theta \boldsymbol{x}(t) = \sum_{j=1}^{k} \beta_j(t) a_{ij}$, and Satoh and Yanagihara (2010) derived the confidence interval as a function of time under the growth curve model.

In matrix notation, the GMANOVA model is given by

$$Y = A \Theta X + \mathcal{E}, \qquad \mathcal{E} \sim N_{n \times p}(\mathrm{O}, \Sigma \otimes I_n)$$

where $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$, $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)'$ and $\mathcal{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$. For example, if we fit a polynomial curve of degree $q - 1$ to the trend over time, the design vector is taken to be $\boldsymbol{x}(t) = (1, t, \ldots, t^{q-1})'$. Selecting the degree of the polynomial curve or the number of base functions is known as a variable selection problem and some variable selection criterion have been proposed, *c. f.*, Satoh, Kobayashi and Fujikoshi (1997). To describe a complex non-linear time trend we can also use B-spline base functions for $\boldsymbol{x}(t)$ and the relationship with polynomial curves has been discussed by Satoh, Yanagihara and Ohtaki (2003).

In this paper we consider a robust estimation of the regression coefficients. Rousseeuw et al.(2004) and Ben et al. (2006) proposed robust estimation methods for multivariate linear regression by using the Mahalanobis distance of response and covariates at the same time. Although the growth curve model can be considered as a family of multivariate linear models,

those methods can not be applied directly because they have two types of covariates; 1) ordinary covariates $\boldsymbol{a}_i$ and 2) time-dependent covariates $\boldsymbol{x}(t)$ for which the design is common to all individuals. Therefore we need to detect outliers as curves over the observation time period under the growth curve model. In the next section weighted least squares estimators are derived for fixed weights. We then discuss ways of revising the weights and deciding the cutoff point based on the residual distribution in Section 3. An example is illustrated in section 4 and we summarize our conclusions in Section 5.

2. Weighted Least Squares Estimator

To develop robust estimators of the unknown regression parameters, we consider minimizing the following sum of weighted squares of individual residuals,

$$l = \sum_{i=1}^{n} w_i e_i^2, \qquad e_i^2(\Theta, \Sigma) = (\boldsymbol{y}_i' - \boldsymbol{a}_i'\Theta X)\Sigma^{-1}(\boldsymbol{y}_i' - \boldsymbol{a}_i'\Theta X)',$$

where $\mathrm{Var}(\boldsymbol{y}) = \Sigma$ and w_i denotes the positive weight of the ith individual. Assuming that the covariance matrix Σ is known, a weighted least squares estimator is obtained by

$$\hat{\Theta} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k)' = (A'WA)^{-1}A'WY\Sigma^{-1}X'(X\Sigma^{-1}X')^{-1},$$

where $W = \mathrm{diag}(w_1, \ldots, w_n)$. It is an unbiased estimator of Θ for fixed weights, *i.e.*, $E[\hat{\Theta}] = \Theta$ because $E[Y] = A\Theta X$. Let $\hat{\boldsymbol{\theta}} = \mathrm{vec}(\hat{\Theta}') = (\hat{\boldsymbol{\theta}}_1', \ldots, \hat{\boldsymbol{\theta}}_k')'$; then the covariance matrix of the estimator is given by

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\theta}}) &= (A'WA)^{-1}A'W^2A(A'WA)^{-1} \otimes (X\Sigma^{-1}X')^{-1} \\ &= \begin{pmatrix} \Omega_{1,1} & \cdots & \Omega_{1,k} \\ \vdots & & \vdots \\ \Omega_{k,1} & \cdots & \Omega_{k,k} \end{pmatrix}, \end{aligned}$$

where $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\theta}}_{j'}) = \Omega_{j,j'}$ and $\mathrm{Var}(\hat{\boldsymbol{\theta}}_j) = \Omega_{j,j}$.

In an actual application, the covariance matrix of observations needs to be estimated and we use the weighted estimator defined by

$$\hat{\Sigma} = \frac{Y'HY}{\mathrm{tr}(H)} \text{ with } H = W - WA(A'WA)^{-1}A'W.$$

Note that $\hat{\Sigma}$ is an unbiased estimator of Σ for fixed weights as well as $\hat{\Theta}$, *i.e.*, $E[\hat{\Sigma}] = \Sigma$ because $HY = H\mathcal{E}$ and $\mathcal{E}\Sigma^{-1/2} \sim N_{n\times p}(\mathrm{O}, I_p \otimes I_n)$.

3. Revision of Weights

From the previous discussion, we consider fixed weights and obtained two estimators, 1) the covariance matrix $\hat{\Sigma}$ and 2) the regression coefficients $\hat{\Theta}$. Hence individual residuals can be obtained as $\hat{e}_i^2 = e_i^2(\hat{\Theta}, \hat{\Sigma})$. Here we discuss revising individual weights, according to the residuals.

Firstly, we attempt to describe the distribution of residuals. If there are no outliers, the residuals e_i^2, $i = 1, \ldots, n$ have a chi-square distribution with p degrees of freedom. On the other hand, the distribution including outliers is expected to have a heavy upper tail, and therefore the actual value for the degrees of freedom might be greater than the theoretical value p.

From the result of Wilson and Hilferty (1931), the median value m of the chi-square distribution with f degrees of freedom can be approximated as

$$m \approx f - \frac{2}{3} + \frac{4}{27f}.$$

Conversely, it is possible to approximate the number of degrees of freedom as a function of m by solving $f^2 - (m + 2/3)f + 4/27 = 0$ with respect to f:

$$f(m) = \frac{1}{2}\left\{m + \frac{2}{3} + \sqrt{\left(m + \frac{2}{3}\right)^2 - \frac{4^2}{27}}\right\}.$$

Note that the other solution of the quadratic equation, $\{m + 2/3 - \sqrt{(m + 2/3)^2 - 4^2/27}\}/2$ is inappropriate because $g(z) = z - \sqrt{z^2 - a}$, $a > 0$ is a monotonically decreasing function for $z > 0$. Thus the number of degrees of freedom for the residual distribution including outliers can be estimated by using its median residual, which can be expressed as $\hat{f} = f(\hat{m})$ where $\hat{m}$ is the median value of $\{\hat{e}_1^2, \ldots, \hat{e}_n^2\}$.

The estimated value for the number of degrees of freedom from the median of the residuals can be regarded as a sort of robust estimator for the chi-square distribution, but $\hat{f}$ is not

always a natural number. Therefore we consider the gamma distribution with shape parameter $\hat{f}/2$ and scale parameter 2, $\mathcal{G}(\hat{f}/2, 2)$. The gamma distribution can be interpreted as a wider distribution class which includes the chi-square distribution and the number of its degrees of freedom can take positive real values. Thus, the residual distribution is described by using the gamma distribution with robust estimated degrees of freedom $\hat{f}$.

Letting c_α be the upper $\alpha \times 100$ percentage point of $\mathcal{G}(\hat{f}/2, 2)$ we take c_α as the cutoff point of weights, *i.e.*, $w_i = 0$ if $\hat{e}_i^2 > c_\alpha$ for $i = 1, \ldots, n$. In other words, we do not use observation $\boldsymbol{y}_i$ in estimating the regression coefficient if the corresponding residual is larger than the cutoff point. To give smooth and continuous weights for the other residuals smaller than the cutoff value, the following Tukey's biweight function, *c.f.* Maronna, Martin and Yohai (2006) can be applied and new weights are defined by

$$w_i = w(\hat{e}_i^2/c_\alpha) \text{with} w(z) = \begin{cases} (1-z^2)^2, & (|z| < 1) \\ 0, & (|z| \geq 1) \end{cases}.$$

Selecting candidates for outliers or the number of outliers depends on the cutoff value. If there is no prior information, α might be given by 0.01, 0.02 or 0.05.

Finally, our robust estimation method can be summarized in the following steps.

Step 0. Set α for the cutoff value and let all weights be 1.

Step 1. Calculate $\hat{\Sigma}$ and the weighted least squares estimate $\hat{\Theta}$ for the given weights.

Step 2. Obtain the residuals and their median value, then calculate the number of robust degrees of freedom $\hat{f}$.

Step 3. Approximate the residual distribution by $\mathcal{G}(\hat{f}/2, 2)$ and find the cutoff value c_α.

Step 4. Renew the weights and go to Step 1 until the estimated regression coefficients or weights converge.

4. Example

We now apply our estimation method to Table 1 considered by Potthoff and Roy (1964), consisting of measurements of the distance (mm) from the center of the pituitary to the pterygomaxillary fissure for 11 girls and 16 boys with 4 different ages (8, 10, 12, 14 years old).

Here we fit model (1) to the data by letting $\boldsymbol{a}_i = (1, 0)'$ for girls and $(1, 1)'$ for boys and $\boldsymbol{x}(t) = (1, t)'$ for $t \in \{8, 10, 12, 14\}$. The first covariate expresses the distance for girls and the second covariate expresses the sex effect, which is determined by the additional distance for boys.

In order to decide the cutoff point, we tentatively set $\alpha = 0.01$. After the iterative calculation described as Step 0 to Step 4 in Section 3, the regression coefficients or renewed weights converge, which is shown in Figure 1. Estimators of the covariance matrix of observation and regression coefficients are finally obtained as

$$\hat{\Sigma} = \begin{pmatrix} 3.343 & 2.549 & 3.659 & 2.729 \\ 2.549 & 3.573 & 3.551 & 2.831 \\ 3.659 & 3.551 & 5.180 & 4.175 \\ 2.729 & 2.831 & 4.175 & 4.349 \end{pmatrix}, \hat{\Theta} = \begin{pmatrix} 17.974 & 0.468 \\ 0.560 & 0.178 \end{pmatrix},$$

respectively. The median of the residuals and the robust number for the degrees of freedom are given by $\hat{m} = 4.696$ and $\hat{f} = 5.334$, respectively. Note that $\hat{f}$ is larger than $p = 4$, which is an ordinal value for the number of degrees of freedom for the case when there are no outliers. We then approximate the residual distribution by the gamma distribution and obtain the upper $\alpha \times 100$ percentage point, $c_\alpha = 15.671$. Figure 2 shows the fitted gamma distribution and smoothed experimental density function which is expected to be close to the original residual distribution. Note that the range of the horizontal axis is restricted to $[0, 20]$ and there are two residuals 55.8 and 123.7 outside this range.

The limit values of the weights are listed in Table 1. The result weights of $F10$, $M09$ and $M13$ are almost zero; this implies that these observations are not used for estimation of regression coefficients. Robust estimated growth curves are illustrated with candidates for outliers in Figure 3.

5. Discussion

We regard an individual observation vector $\boldsymbol{y}_i$ for the ith subject as one unit for detecting outliers. It might be possible for us to consider finer weights w_{ij} for $y_i(t_j)$, $j = 1, \ldots, p$. However we need to take into account the covariance structure between observed time points; therefore we summarize the residual for the ith subject as e_i^2 and consider its weight w_i.

Although we estimated the unknown parameter f of $\mathcal{G}(f/2, 2)$ by using the relation between the degrees of freedom and its median of the chi-squared distribution, there is another way using the maximum likelihood method for which the estimator was 5.840. The maximum likelihood estimator might be sensitive to outliers and become a larger value than that of our proposed estimator.

Deciding the cutoff point for weights is an important problem for constructing robust estimators and it is sometimes discussed in term of asymptotic efficiency against the maximum likelihood estimator. For example Huber (1964, 1973) derived the asymptotic covariance matrix of proposed robust estimators for the multiple linear regression model, and this makes it possible to discuss the optimal cutoff point which attains a given efficiency, e.g. 85%. We derived an asymptotic covariance matrix for our estimator only for fixed weights w_i, $i = 1, \ldots, n$, and it is a future problem to examine the assumption that weights are estimators or random variables.

Acknowledgement

This research was supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007-2010 and the Fujii Fund of Hiroshima University, 2008-2009.

References

Ben, M. G., Martinez, E. and Yohai, V. J. (2006): Robust estimation for the multivariate linear model based on a T-scale, *Journal of Multivariate Analysis* **97**, 1600-1622.

Hastie, T. and Tibshirani, R. (1993): Varying-coefficient models, *Journals of the Royal Statistical Society Series* **B 55**, 757–796.

Huber, P. J. (1964): Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **35**, 73-101.

Huber, P. J. (1973): Robust regression: asymptotics, conjectures and monte calro, *The Annals of Statistics* **1**, 799-821.

Maronna, R., Martin, D. R. and Yohai, V. J. (2006): Robust Statistics - Theory and Methods, Wiley.

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., Aelst, S. V., Driessen, K. V. and Gulló J. A. (2004): Robust Multivariate Regression, *Technometrics* **46**, 293-305.

Satoh, K., Kobayashi, M. and Fujikoshi, Y. (1997): Variable selection for the growth curve model, *Journal of Multivariate Analysis* **60**, 277–292.

Satoh, K., Yanagihara, H. and Ohtaki, M. (2003): Bridging the gap between B-spline and polynomial regression model, *Communications in Statistics - Computation and Simulation* **32**, 179-190.

Satoh, K. and Yanagihara, H. (2010): Estimation of varying coefficients for a growth curve model, *American Journal of Mathematical and Management Sciences*, in press.

Vonesh, E. F. and Carter, R. L. (1987): Efficients inference for random-coefficient growth curve models with unbalanced data, *Biometrics* **43**, 617–628.

Wilson, B. and Hilferty, M. (1931): The distribution of chi-square, *Proceedings of the National Academy of Sciences* **17**, 684-688.

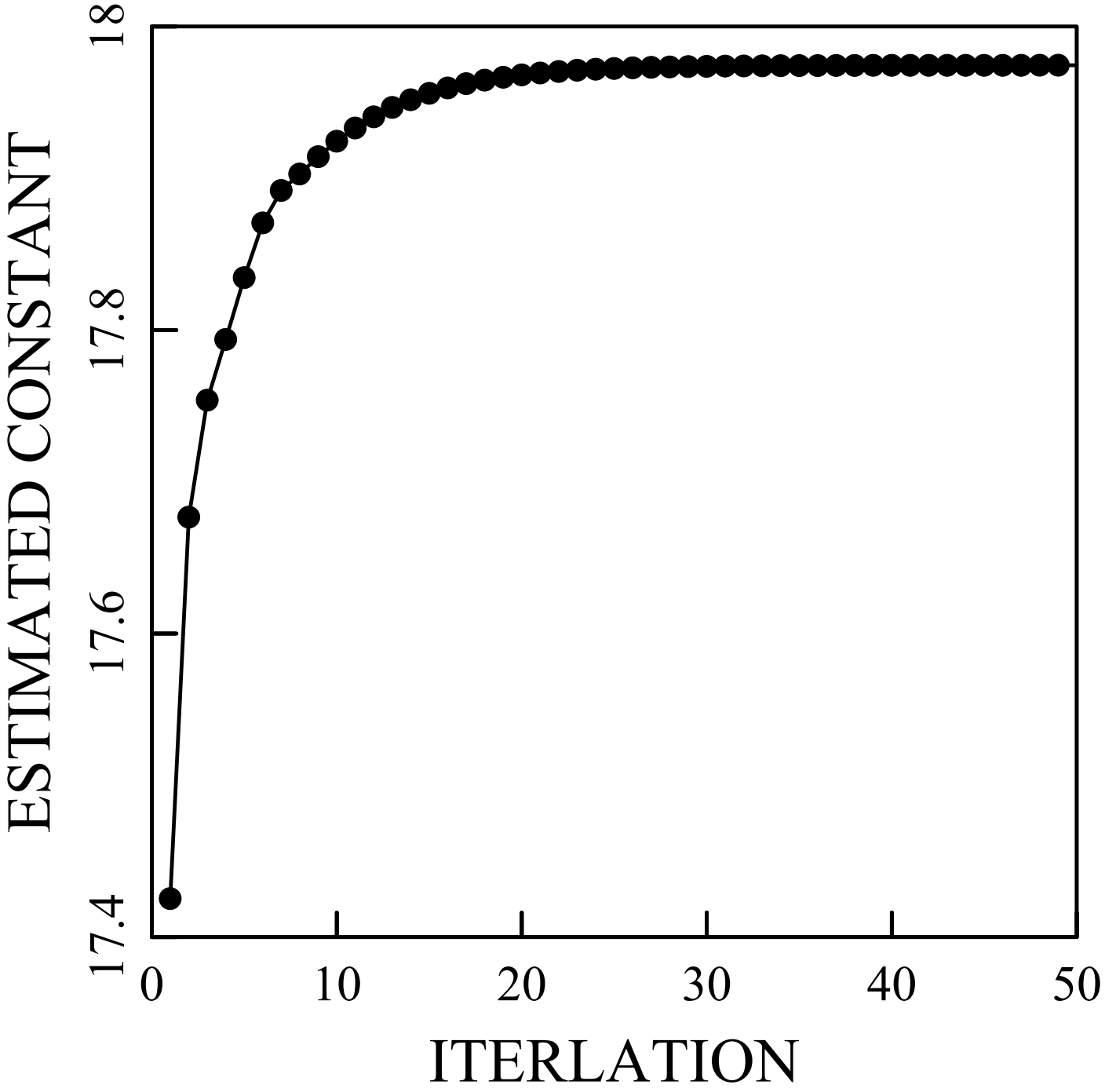
TABLE 1. Growth curve data for orthodontic measurements. Optimized weights w were obtained by a robust estimation method and F10, M09 and M13 were regarded as outliers because those weights were close to zero.

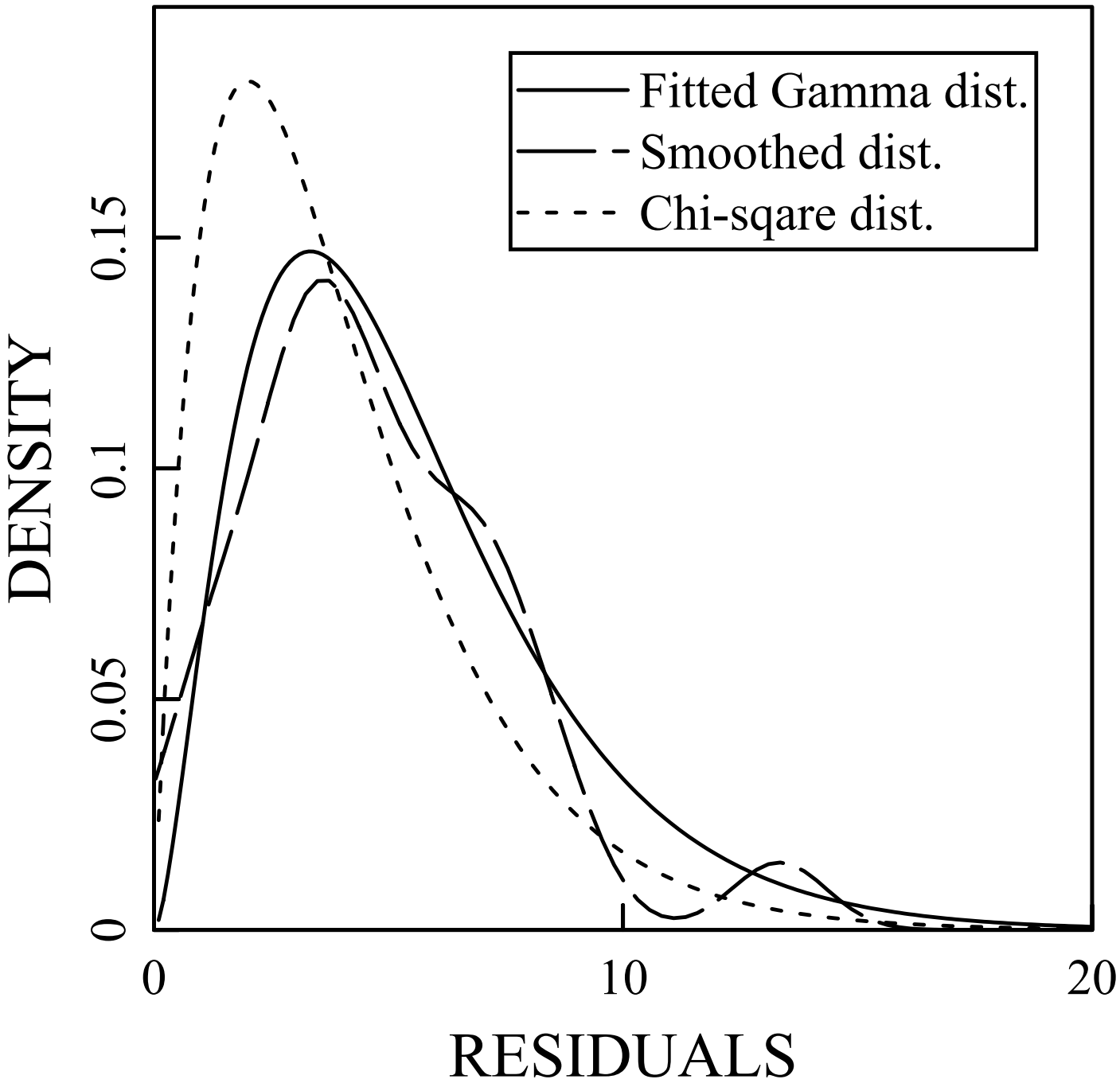
Female						Male					
ID	8	10	12	14	w	ID	8	10	12	14	w
F01	21	20	21.5	23	0.92	M01	26	25	29	31	0.68
F02	21	21.5	24	25.5	0.92	M02	21.5	22.5	23	26.5	0.81
F03	20.5	24	24.5	26	0.78	M03	23	22.5	24	27.5	0.61
F04	23.5	24.5	25	26.5	0.92	M04	25.5	27.5	26.5	27	0.53
F05	21.5	23	22.5	23.5	0.98	M05	20	23.5	22.5	26	0.65
F06	20	21	21	22.5	0.98	M06	24.5	25.5	27	28.5	1.00
F07	21.5	22.5	23	25	0.99	M07	22	22	24.5	26.5	0.90
F08	23	23	23.5	24	0.95	M08	24	21.5	24.5	25.5	0.47
F09	20	21	22	21.5	0.79	M09	23	20.5	31	26	0.00
F10	16.5	19	19	19.5	0.08	M10	27.5	28	31	31.5	0.83
F11	24.5	25	28	28	0.85	M11	23	23	23.5	25	0.90
						M12	21.5	23.5	24	28	0.71
						M13	17	24.5	26	29.5	0.00
						M14	22.5	25.5	25.5	26	0.93
						M15	23	24.5	26	30	0.62
						M16	22	21.5	23.5	25	0.88

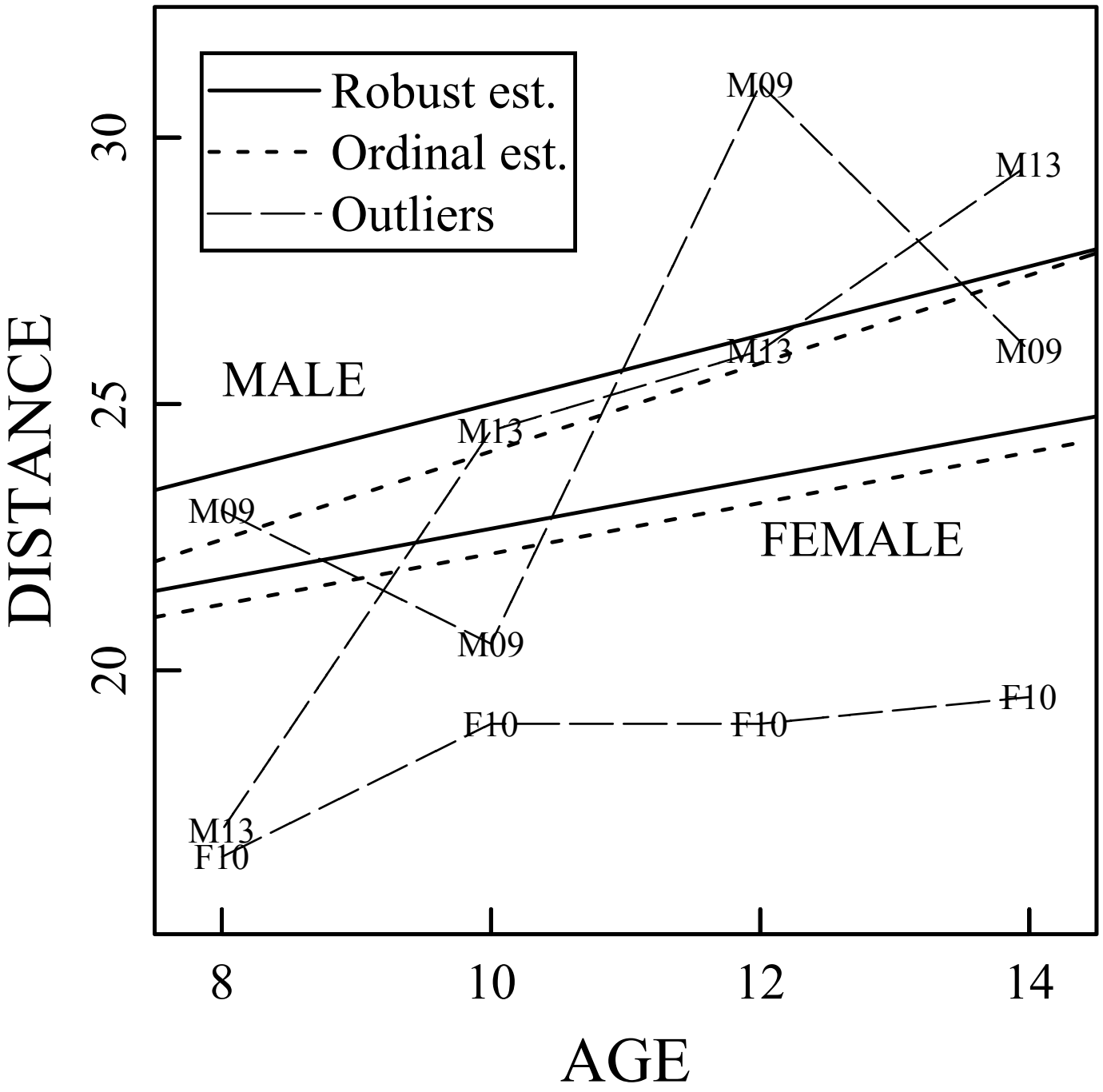
FIGURE 1. Convergence of the estimated regression coefficients, $(1,1)$ element of $\hat{\Theta}$. The proposed estimation method required iterative calculation on the weighted least squares estimator and the revision of weights based on the residuals, one after the other. The estimated constant seems to converge after about twenty repetitions.

FIGURE 2. Approximation of the residual distribution. The residual distribution is expressed by a kernel smoothing density function and theoretically it should be distributed as a chi-square distribution with $p = 4$ degrees of freedom. The gamma distribution estimated by the robust degrees of freedom $\hat{f}$ is closer to the smoothed empirical distribution than the chi-square distribution.

FIGURE 3. The growth curve model derived from the proposed robust estimation method. Solid lines show the fitted curves and the dotted lines considered by Potthoff and Roy (1964) are superimposed. Three candidates for outliers are shown by dashed lines for which the weights are almost zero.


ESTIMATED CONSTANT
18
17.8
17.6
17.4
0
10
20
30
40
50
ITERLATION


Fitted Gamma dist.
Smoothed dist.
Chi-sqare dist.
DENSITY
0
0.05
0.1
0.15
0
10
20
RESIDUALS


Robust est.
Ordinal est.
Outliers
MALE
FEMALE
M09
M13
M09
M13
M13
M09
M09
F10
F10
F10
M13
F10
DISTANCE
20
25
30
8
10
12
14
AGE