Residuals Plots for Repeated Measures Data [1]

TAESUNG PARK [2]

SUMMARY

In the analysis of repeated measurements, multivariate regression models that account for the correlations among the observations from the same subject are widely used. Like the usual univariate regression models, these multivariate regression models also need some model diagnostic procedures. In this paper, we propose a simple graphical method to detect outliers and to investigate the goodness of model fit in repeated measures data. The graphical method is based on the quantile-quantile(Q-Q) plots of the χ^2 distribution and the standard normal distribution. We also propose diagnostic measures to detect influential observations. The proposed method is illustrated using two examples.

Key words: multivariate linear models, outliers, residuals, Q-Q plot

1 Introduction

Like univariate regression models, the multivariate regression models for repeated measures data also need some model diagnostic procedures for checking model adequacy, and for detecting outliers and influential observations. Though the multivariate regression models have been widely used to analyze repeated measures data, not many studies have been performed for model diagnostics. Recently, the idea of using residuals in repeated measures data has been introduced. Weiss and Lazaro (1992) proposed parallel plots of residuals to check model fits and to identify outlying observations. However, their plots are not easily applicable to cases where the sample size is large or the number of repeated observations is large. Dawson, et. al. (1997) also proposed two graphical techniques useful in detecting correlation structure in repeated measures data.

The main objective of this paper is to propose a simple graphical method and diagnostic measures to detect outliers in repeated measures data. We focus on repeated measures data where responses are normally distributed. The graphical method uses the quantile-quantile(Q-Q) plot of the χ^2 distribution and the standard normal distribution. The Q-Q

[1] The work was supported by the Brain Korean 21 Project.

[2] Department of Statistics, Seoul National University, San 56-1 Shin Lim-Dong, Kwanak-Gu, Seoul, Korea 151-742

plots can handle repeated measures data where the sample size is large and the number of repeated observations is also large.

This paper is organized as follows. In Section 2, the models of repeated measures data are described. In Section 3, residuals are defined. Three plots are proposed which are useful in detecting outlying observations. They include the Q-Q plots based on the χ^2 distribution and the standard normal distribution, and the normalized residual plot.

2 Models

Consider repeated measures obtained from n subjects at t different time points. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{it_i})^T$ denote the $t_i \times 1$ vector of responses and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{it_i}^T)^T$ the $t_i \times p$ matrix of covariates, where $\boldsymbol{x}_{ij}$ is $1 \times p$ covariate vector for $j = 1, \ldots, t_i$ $(\leq t)$.

The $\boldsymbol{y}_i$ are assumed to follow the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{e}_i, \qquad (1)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{it_i})^T$ is the error vector. The error vectors are assumed to be independent and normally distributed with the mean vector $\mathbf{0}$ and the common covariance matrix $\boldsymbol{\Sigma}$.

Commonly used structures for $\boldsymbol{\Sigma}$ are simple, first-order autoregressive (AR-1), and compound symmetry(CS). A simple structure uses the identity matrix that assumes independence among observations. AR-1 assumes that first-order autoregressive correlation exists among the repeated measurements. CS assumes that the correlation is the same among the repeated measurements. This model can also handle the unstructured covariance that treats all elements as parameters. As special cases, mixed models with random effects can also be expressed by (1) with appropriate covariance structures (Jennrich and Schluchter, 1986). Model parameters can be obtained using the MIXED procedure of the SAS statistical package via maximum likelihood or restricted maximum likelihood estimation.

Note that (1) can be represented using the following concatenation response vector $\boldsymbol{Y}$ and the design matrix $\boldsymbol{X}$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}, \qquad (2)$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}, \text{ and } \boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_n \end{pmatrix}.$$

Let the MLE of β be $\hat{\beta}$ and the MLE of Σ_i be $\hat{\Sigma}_i$ which are usually obtained by an iterative algorithm (Laird, Lange, and Stram, 1987). Then

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i^T \hat{\Sigma}_i^{-1} x_i\right)^{-1} \left(\sum_{i=1}^{n} x_i^T \hat{\Sigma}_i^{-1} y_i\right).$$

Alternatively, using (2) $\hat{\beta}$ can be written as

$$\hat{\beta} = \left(X^T \hat{V}^{-1} X\right)^{-1} \left(X^T \hat{V}^{-1} Y\right),$$

where $\hat{V} = diag(\hat{\Sigma}_1, \cdots, \hat{\Sigma}_n)$ is a $t^* \times t^*$ block diagonal matrix with $t^* = \sum_{i=1}^n t_i$.

3 Residual Plots and Diagnostic Measures

For the ith subject, the residual vector is defined as the observed vector subtracted by the predicted vector. That is, r_i is given by

$$\begin{aligned} r_i = & \quad y_i - \hat{y}_i \\ = & \quad y_i - x_i\hat{\beta}. \end{aligned} \tag{3}$$

Let R be a concatenation vector of r_is. Then,

$$R = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} y_1 - x_1\hat{\beta} \\ \vdots \\ y_n - x_n\hat{\beta} \end{pmatrix} = Y - X\hat{\beta}.$$

Weiss and Lazaro (1992) introduced two types of residuals with and without adjusting random effects. We do not distinguish these two residuals in (3), since random effects model can be easily incorporated in Model (1).

When the model fits the data well, the residual vectors are normally distributed with the mean zero vector. The next theorem summarizes the distribution of residual vectors.

THEOREM 1. *When the model fits the data well, R is normally distributed with the zero mean vector and the following covariance matrix W:*

$$W\text{“} = \text{“}Var(R)\text{“} = \text{“}HVH^T,$$

where $V = diag(\Sigma_1, \cdots, \Sigma_n)$ is a $t^* \times t^*$ block diagonal matrix and $H = I - X(X^TV^{-1}X)^{-1}X^TV^{-1}$.

The proof of Theorem is given in Appendix. Let W_i be the variance matrix of r_i which is a $t \times t$ block diagonal submatrix of W. From Theorem 1, we then have the following result.

THEOREM 2. *When the model fits the data well, $q_i = r_i^T W_i^{-1} r_i$, $i = 1, \cdots, n$, are distributed with the χ^2-distribution with t_i degrees of freedom.*
The proof of Theorem is given in Appendix.

When we have balanced and complete observations, say the number of responses from the same subject is equal to t for all i, we can construct a Q-Q plot easily with observed q_is using the χ^2-distribution. Let $q_{(1)} \leq \cdots \leq q_{(n)}$ be ordered values of q_is. Then, $q_{(i)}$ is in fact the empirical $100 \times i/n$ percentile. From the χ^2-distribution with t degrees of freedom, we can obtain the corresponding quantiles $\psi_{(1)} \leq \cdots \leq \psi_{(n)}$. Then the Q-Q plot is the graph of $(q_{(i)}, \psi_{(i)})$ from which we can investigate model fits and identify outliers.

When we have unbalanced or incomplete observations, the number of responses from the same subject t_i may differ from subject to subject. Then, the degrees of freedom of q_i also differ and it is not plausible to construct a Q-Q plot based on the χ^2−distribution. In that case, we suggest using the Q-Q plot based on the standard normal distribution.

THEOREM 3. *When the model fits the data well, $q_i^N = (q_i - t_i)/\sqrt{2t_i}$, $i = 1, \cdots, n$, are distributed with the standard normal distribution.*
The q_i^N is a summary measure for the residual vector from the i subject. We call q_i^N the *normalized residual* for the ith subject. We can construct a Q-Q plot similarly with observed q_i^Ns using the standard normal distribution. This plot can be easily obtained from the standard softwares.

REFERENCES

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15-18.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

Dawson, K. S., Gennings, C., and Carter, W. H. (1997). Two graphical techniques useful in detecting correlation structure in repeated measures data. *American Statistician* **51**, 275-283.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Laird, N. M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Laird, N. M., Lange, N., and Stram, D. O.(1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 82, 97-105.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

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.

Weiss, R. E. and Lazaro, C. G. (1992). Residual Plots for Repeated Measures. *Statistics in Medicine* **11**, 115-124.