View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by NASA Technical Reports Server

Modeling Longitudinal Data Containing Non–Normal Within Subject Errors

Alan Feiveson, Ph. D.
alan.h.feiveson@nasa.gov
Biomedical Research and Environmental Sciences Division
NASA JSC
Nancy L. Glenn, Ph. D.
nglenn@alumni.rice.edu
Department of Mathematical Sciences
Texas Southern University

August 26, 2013

Abstract

The mission of the National Aeronautics and Space Administration's (NASA) human research program is to advance safe human spaceflight. This involves conducting experiments, collecting data, and analyzing data. The data are longitudinal and result from a relatively few number of subjects; typically 10 – 20. A longitudinal study refers to an investigation where participant outcomes and possibly treatments are collected at multiple follow-up times. Standard statistical designs such as mean regression with random effects and mixed–effects regression are inadequate for such data because the population is typically not approximately normally distributed. Hence, more advanced data analysis methods are necessary.

This research focuses on four such methods for longitudinal data analysis: the recently proposed linear quantile mixed models (lqmm) by Geraci and Bottai (2013), quantile regression, multilevel mixed–effects linear regression, and robust regression. This research also provides computational algorithms for longitudinal data that scientists can directly use for human spaceflight and other longitudinal data applications, then presents statistical evidence that verifies which method is best for specific situations. This advances the study of longitudinal data in a broad range of applications including applications in the sciences, technology, engineering and mathematics fields.

1 Introduction

This research presents and improves upon computational algorithms for the analysis of longitudinal repeated measures data. Longitudinal data, which comprise repeated measurements of the same subjects over time, arise frequently in in a broad range of applications including biomedical sciences. Statistical methods for analyzing longitudinal data include the cutting edge linear quantile mixed models (lqmm) by Geraci and Bottai (2013), and standard methods such as quantile regression, multilevel mixed–effects linear regression, and robust regression. Several key algorithms for implementing these four methods using the popular statistical software packages R , S–PLUS and Stata are used for simulation studies. Major codes are written in the form of R , S–PLUS and Stata functions which can be directly used for real data applications and simulation studies. Understanding the conceptual differences among these methods is important to their proper application.

This paper is organized as follows. Section 2 describes a popular model used for longitudinal data, the mixed model. Section 3 describes estimation methods for longitudinal data. Simulation results are contained in Section 4. Indications for future research are in Section 5, and concluding remarks are in Section **??**.

2 Mixed Model

A mixed model is defined as a nonlinear statistical model that contains mixed effects (Demidenko, 2004). A mixed effects model contains both fixed effects and random effects. A fixed effect model represents observed quantities in terms of explanatory variables treated as if the quantities were nonrandom. Unlike classical statistics where observations are drawn from independent, identically distributed populations, mixed model observations are independent between levels or clusters. However, observations within clusters are dependent because they belong to the same subpopulation. Hence, there are two sources of variation, between clusters and within clusters.

To illustrate the difference between modeling data with least squares regression versus a mixed model, Figures 1 and 2 present the same data on two different graphs. The horizontal axis represents measurement time, the vertical axis represents the measurement on each subject. Measurement times range from 0 to 7. The straight line in both figures represents the overall regression line. Each additional line in Figure 2 individually connects the measurements for each of the 10 subjects. For example, the line with the highest y values connect all measurements for subject 5 for the seven time periods.

A classical statistical model is

$$y_k = \alpha + \beta x_k + \epsilon_k \tag{1}$$

where (x_k, y_k) is an observation, α is the intercept, β is the slope, ϵ_k represents independent random errors with zero mean and constant variance. A mixed model is

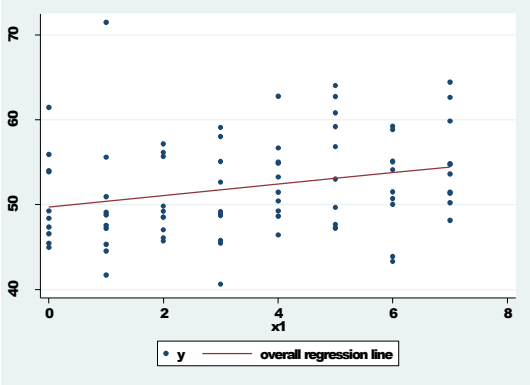

Figure 1: Graph depicts overall regression line for simulated data.

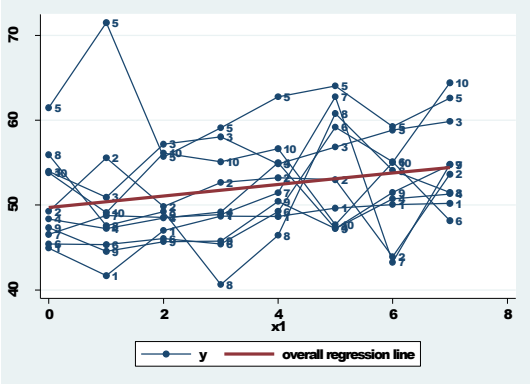

Figure 2: Graph depicts overall regression line. Additional lines individually connect measurements for each of the 10 subjects.

Method	Function	Longitudinal	Quantile	Complexity
linear quantile mixed model	*lqmm*	yes	conditional	high
quantile regression	*qreg*	no	unconditional	moderate
mixed effects linear regression	*xtmixed*	yes	no	moderate
robust regression	*mmregress*	no	no	very high

Table 1: Table contains estimation methods, named functions, whether there is an inherent longitudinal model, the type of quantile estimation, and computational complexity as measured by relative computational time. The function *lqmm* is designed for longitudinal conditional quantiles, *qreg* is for quantile regression, *xtmixed* is for multilevel means and variances, and *mmregress* is for robust central tendency.

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij} \tag{2}$$

where the observation (x_{ij}, y_{ij}) is the i^{th} subject's, j^{th} measurement, α_i is a subject specific intercept, β_i is a subject specific slope, and ϵ_{ij} is an error term.

3 Estimation Methods

Standard statistical designs such as mean regression with random effects and mixed–effects regression are inadequate to analyze longitudinal data because the errors are typically not approximately normally distributed. Additionally, standard quantile regression is inadequate since it does not include random effects. Linear quantile mixed models by Geraci and Bottai (2007) incorporates random effects in the model. It is a robust alternative to mean regression with random effects, which allows estimation of an entire family of conditional quantile functions. Thus, providing a more complete statistical analysis of the stochastic relationships among random variables (Koenker, 2000). The estimation methods and models are summarized in Table 1.

The estimation method lqmm is an *R* function designed for linear quantile mixed models. A quantile divides an ordered dataset into equally sized subsets. A quantile regression model provides estimates that approximate the conditional median and other quantiles of the response variable given certain predictor variables.

Figure 3 presents quantile regression graphs. For example, 75 percent of the measurement are below the top line, and 25 percent are above. The solid middle line represents the median. The broken middle line represents the least squares regression line, and the bottom line represents the 25^{th} percentile.

The benefits of quantile regression include the fact that it makes no distributional assumptions about the error terms in the model. It is also robust to extreme points in the response variable.

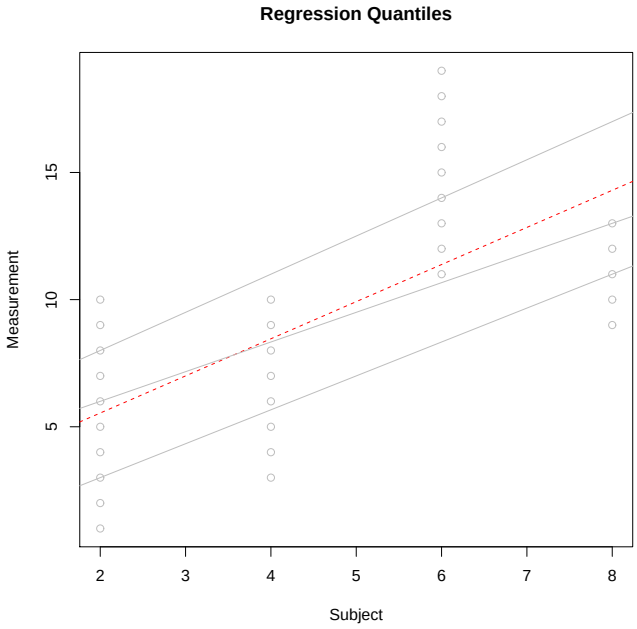

Figure 3: Graph depicts quantile regression lines. The top line represents the 75^{th} percentile. The solid middle line represents the median. The broken middle line represents the least squares regression line, and the bottom line represents the 25^{th} percentile.

The remaining three methods are in Stata. The function qreg performs quantile regression, xtmixed is designed for mixed effects linear regression, and mmregress is designed for robust regression.

A main objective of this research is to determine which of the four methods in Table 1 is the best method for modeling a dataset that involves multiple measurements on one subject over a period of time.

4 Results

Initial results, conducted through a simulation study, determine the efficiency of the models. Efficiency of the models is judged by bias and estimation accuracy. Error magnitude is judged by the standard deviation, mean or median absolute error, and quartiles / distributions of errors or absolute errors. Data were generated based on the following random slope model:

$$y_{ij} = 50 + u_i + (\beta_1 + \beta_i)x_j + \epsilon_{ij}, \tag{3}$$

where
subject $i = 1, \ldots, n_k$, $k = 1, 2$, sample sizes $n_1 = 15$, $n_2 = 30$, and measurement $j = 1, \ldots, 4$. The variable β_1 equals the values 1, 1.5, 2, 2.5, 3, 3.5, and 4. The parameter β_i represents the random slope of the i^{th} subject. The $x_j = j - 1$. The variable u_i represents the random intercept of the i^{th} subject. Probability distributions considered for u_i are:

(i) normal distribution $N(0, \sigma_u^2)$

(ii) asymmetric Laplace distribution $ALD(0.5, \sigma_u)$

where
$\sigma_u = 2, 7, 12$ The variable ϵ_{ij} represents the residual error term associated with the i^{th} subject's j^{th} measurement. Three different probability distributions were considered for the error term:

(i) asymmetric Laplace, $ALD(0.5, \sigma_e)$

(ii) chi–squared with 3 degrees of freedom, $\chi^2(3)$

(iii) t–distribution with 3 degrees of freedom, $t(3)$

The first distribution is symmetric, the second is skewed, and the third is symmetric with heavier tails than the normal distribution. These distributions were chosen to help determine the effects of the shape of the residual's distribution on estimation accuracy.

The absolute error, which is the absolute value of the difference between the value of the coefficient and the estimate of the coefficient $|\beta_1 - \hat{\beta}_1|$, is computed for each method. A log rank test for equality of functions was conducted to simultaneously compare the four cumulative distribution functions of the absolute errors.

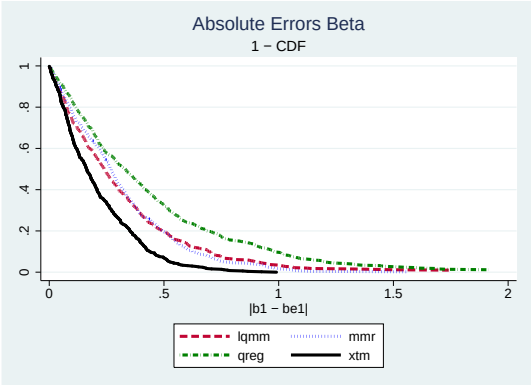

Figure 4: Graph depicts complementary cumulative distribution of absolute errors.

The p–value is approximately zero, indicating a difference in absolute error among the methods. Figure 4 indicates that method with the smallest absolute error is xtmixed. This is followed by lqmm, mmregress, and qreg. Further studies will investigate the variance. A Kolomogrov–Smirnov test could have been used for comparing any two methods, but not all four methods at once.

5 Future Research

Future research will also answer the following questions:

1. What are effects of small sample sizes on estimation accuracy for each method?
2. What are alternative methods for determining the variance–covariance matrix of the lqmm parameter estimates? The variance is a measure of the dispersion or spread of estimates for the parameter. The covariance indicates how two parameters vary together.
3. What are benefits of determining conditional quantiles for quantiles other than the median? A quantile divides an ordered dataset into equally sized subsets. A conditional quantile function is a model in which the quantiles of the conditional distribution of the dependent variable are expressed as functions of observed independent variables.
4. What are effects of not including random effects in a quantile regression model? Random effect vary across subjects while fixed effects are constant.
5. How do analysis results of the four methods differ when the model differs from Equation (3).
6. How do the methods compare when using real data?

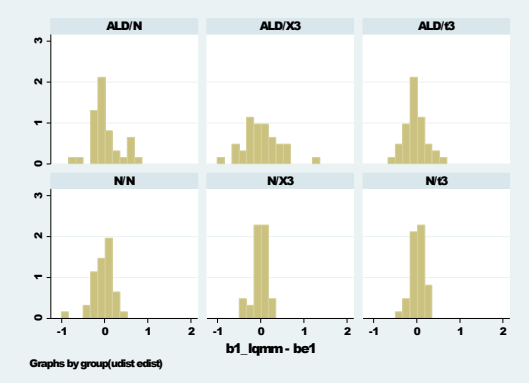

Figure 5: Graph depicts the error for the regression parameter for various combinations of the u_i distribution and the ϵ_{ij} using lqmm for estimation.

6 Summary

This research has three main objectives:

1. Present models for analyzing longitudinal data where within subject errors are not normally distributed.
2. Determine which method is best in specific situations.
3. Provide implementation algorithms in Stata, R, and Splus.

These objective were carried out through simulation studies presented in Section 4. Preliminary results indicate that the method with the smallest absolute error for the coefficients is xtmixed. This is followed by lqmm, mmregress, and qreg. Future research will focus on the variance-covariance matrix of the coefficients as well other research questions posed in Section 5.

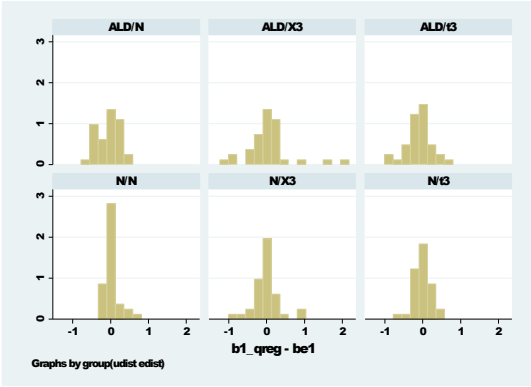

Figure 6: Graph depicts the error for the regression parameter for various combinations of the u_i distribution and the ϵ_{ij} using quantile regression for estimation.

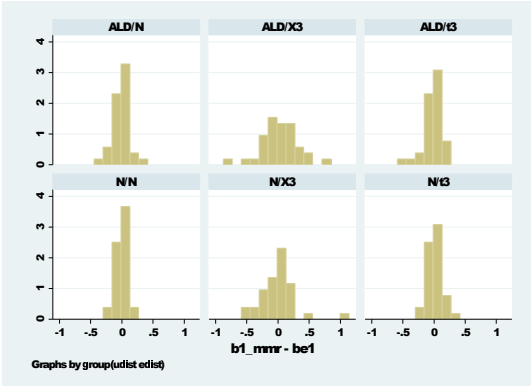

Figure 7: Graph depicts the error for the regression parameter for various combinations of the u_i distribution and the ϵ_{ij} using mmregress for estimation.

References

Geraci, Marco and Bottai, Matteo (2013), Linear Quantile Mixed Models, R package version 1.03, http://CRAN.R-project.org/package=lqmm.

Geraci, Marco and Bottai, Matteo (2007), Quantile regression for longitudinal data using the asymmetric Laplace distribution, *Biostatistics* **8**(1), 140–154.

Koenker, Roger (2000), Quantile Regression, International Encyclopedia of the Social Sciences.