Paper 260-2010

# Fitting Linear and Nonlinear Growth Curve Models Using PROC NLMIXED

Mo Zheng, University of Southern California, Los Angeles, CA

## ABSTRACT

Longitudinal data, or data that are repeated measurements on various subjects across time, are commonplace in everyday life. Multi-level mixed models are often used for analyzing longitudinal data and drawing meaningful inferences about them. This paper discusses two common mixed models, the linear growth model and the logistic growth model, and fits them to a prototypical example that involves repeated measures on forest growth. Parameter estimates and model fitting results from two analyses are compared. The nonlinear logistic growth curve is selected as the suitable model for the current data, even though evidence from model fit statistics seems to suggest otherwise. Computer implementation is via PROC NLMIXED in the SAS® 9.2 program. A plot of the data, code descriptions, and output interpretations are also presented.

KEY WORDS: mixed models, growth curves, nonlinear mixed models

## INTRODUCTION

Change is ubiquitous in our everyday life. Measuring change requests a longitudinal perspective, with repeated observations of the same items over long periods of time. A popular approach to analyze longitudinal data is latent growth analysis, which is a multi-level change model and includes both fixed and random effects. Latent growth models have a wide variety of application, and can be easily fit in SAS/STAT mixed procedures.

PROC NLMIXED is a recently developed procedure dealing with general mixed model analysis. The main feature of it is the specification of the equations with the parameters to be estimated, following some hypothetical mathematical equations. It can be viewed as generalizations of the random coefficient models fit by the popular PROC MIXED procedure. This generalization allows the random coefficients to enter the model nonlinearly, whereas in PROC MIXED they can only enter linearly. Thus Proc NLMIXED is more flexible, allowing the modeling of multiple linear and nonlinear models to be specified.

The current study shows how both the linear and nonlinear model can be fitted using the NLMIXED procedure. The general purpose of this paper is to provide a demonstration of programming features of PROC NLMIXED by fitting two different growth curve models. Another important goal is to help researcher and SAS users implement these models as a useful way to test their hypothesis of growth.

## DATA DESCRIPTION

To illustrate how to fit mixed models using NLMIXED procedures, consider a longitudinal study on the development of forest management by Marshall (2005). The study consisted of multiple waves of measurement of the forest yield and growth at 9 different installations in Pacific Northwest region. The following are variable definitions and a snapshot of the dataset.

**INST** Installation Number (1-9)

**TAGE** Total Age

**CV6PA** Merchantable volume per acre

**CUNIT** CV6PA/100

| INST | TAGE | CV6PA | CUNIT |
|---|---|---|---|
| .... | .... | .... | .... |
| 6 | 40 | 2505 | 25.05 |
| 6 | 46 | 4326 | 43.26 |
| 6 | 52 | 6035 | 60.35 |
| 7 | 25 | 29 | 0.29 |
| 7 | 31 | 219 | 2.19 |
| 7 | 37 | 677 | 6.77 |

| | | | |
|---|---|---|---|
| 7 | 44 | 1412 | 14.12 |
| 7 | 51 | 2342 | 23.42 |
| 8 | 24 | 174 | 1.74 |
| 8 | 51 | 10114 | 101.14 |
| 8 | 56 | 11719 | 117.19 |
| .... | .... | .... | |

Series plot of the growth data over last several decades show all 9 locations appear to have S-shape trajectories. That means the tree volume per acre increase gradually at first, more rapidly in the middle growth period, and slowly at the end, leveling off at a maximum value after some period of time. However, before we conclude that individual changes are best represented by a nonlinear curve, a simple questions to ask is, does there exist a linear growth trajectory?

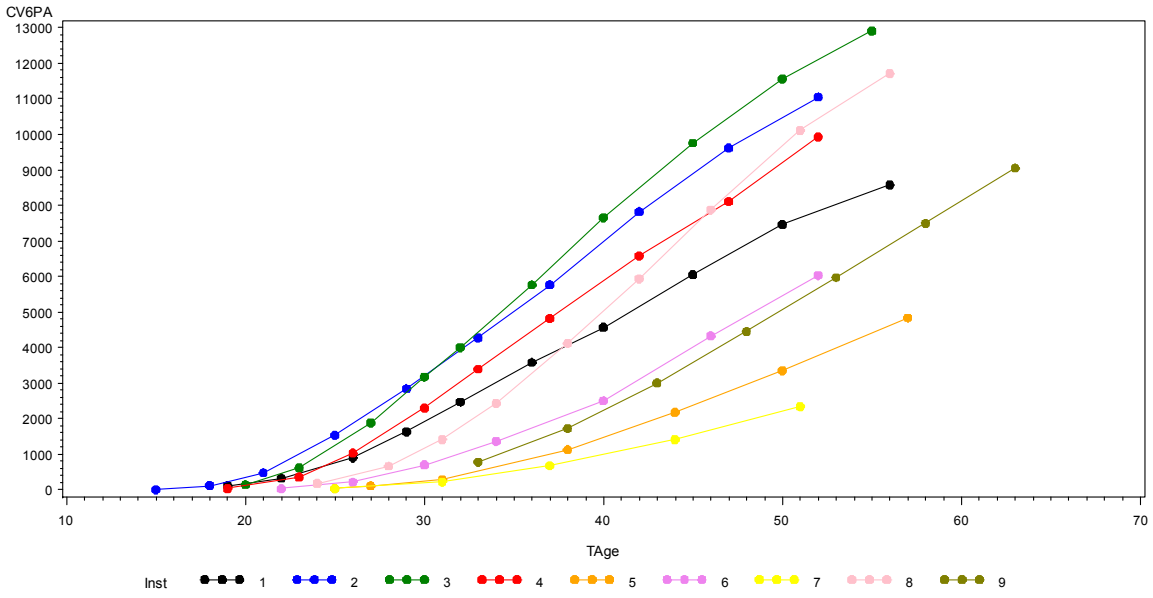


**Figure 1. Serial plot of the forest yield and growth at 9 different installations**

## ANALYSIS I: LINEAR MODEL

In linear growth models, we define an individual $i$ on occasion $j$ is a linear function of time:

$$y_{ij} = b_{0i} + b_{1i}*Time_{ij} + e_{ij}$$
$$b_{0i} = g_0 + u_{0i} \quad [1]$$
$$b_{1i} = g_1 + u_{1i}$$

Here $Time_{ij}$ is the corresponding time, $B_0$ and $B_1$ are random intercepts and slopes, and $e_{ij}$ are the residual errors assumed to be normally distributed and independent of deviation of $B_0$ and $B_1$. $G_0$ and $G_1$ represent the fixed components in the linear model, while $U_{0i}$ and $U_{1i}$ are random components in the model. The equation indicates that the trajectory for the outcome variable is a function of the intercept and slop, which themselves are random variables. $U_{0i}$ and $U_{1i}$ are normally distributed with zero means but nonzero variances ($S^2_{U0}$, $S^2_{U1}$) and covariance (COV). Below is the SAS code to fit the above linear growth curve using NLMIXED procedure.

```
PROC NLMIXED DATA=DataA;
   PARMS  g0=-50 g1=1 s2e=20 s2u0=500 cov=-20 s2u1=1;
          b0 = g0 + u0;          /*Random effect intercept*/
          b1 = g1 + u1;          /*Random effect slope*/
          y  = b0 + b1*TAge;     /*Linear equation*/
          p  = g0 + g1*TAge;     /*Predicted value from estimates of fixed effects*/
   MODEL Cunit ~ NORMAL(y, s2e);
   RANDOM u0 u1 ~ NORMAL([0,0],[s2u0,cov,s2u1]) SUBJECT=Inst;
   PREDICT p OUT=DataB;
   TITLE 'Model 1: Linear growth curve model using NLMIXED';
RUN;
```

The PARMS statement identifies the unknown parameters and their starting values. There are two fixed effects parameters (*G0* and *G1*) and three variance components (*S2U0, S2U1,* and *COV*). The next three statements specify the linear mixed mode. The MODEL statement defines the dependent variable and its conditional distribution given the random effects. Here a normal conditional distribution is specified with mean *B0* + *B1**TAGE and variances *S2E*. The RANDOM statement defines the double random effect to be *U0* and *U1*, and specifies that they follow a bivariate normal distribution. The SUBJECT argument defines a variable indicating when the random effect obtains new realizations; in this case, it changes according to the values of the TAGE variable. The PREDICT statement enables us to construct predictions for every observation in the input data set and output a new dataset, allowing us to plot the growth trajectory of predicted outcome values against time.

The main output from this analysis is as follows.

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 494.8 |
| AIC (smaller is better) | 506.8 |
| AICC (smaller is better) | 508.0 |
| BIC (smaller is better) | 507.9 |

Parameter Estimates

| Parameter | Estimate | Standard Error | DF | t Value | Pr > \|t\| | Alpha | Lower | Upper | Gradient |
|---|---|---|---|---|---|---|---|---|---|
| g0 | -62.8550 | 7.7088 | 7 | -8.15 | <.0001 | 0.05 | -81.0835 | -44.6265 | 7.339E-6 |
| g1 | 2.6514 | 0.3282 | 7 | 8.08 | <.0001 | 0.05 | 1.8753 | 3.4274 | 0.000137 |
| s2u0 | 500.13 | 271.14 | 7 | 1.84 | 0.1076 | 0.05 | -141.01 | 1141.26 | 0.000431 |
| cov | -19.4519 | 10.8137 | 7 | -1.80 | 0.1151 | 0.05 | -45.0222 | 6.1184 | -8.39E-7 |
| s2u1 | 0.9474 | 0.4800 | 7 | 1.97 | 0.0890 | 0.05 | -0.1876 | 2.0824 | 0.000213 |
| s2e | 18.4339 | 3.5057 | 7 | 5.26 | 0.0012 | 0.05 | 10.1442 | 26.7236 | -9.84E-7 |

The ”Fit Statistics” table lists the final maximized value of the log likelihood as well as the information criteria of Akaike (AIC) and Bayesian (BIC). These statistics can be used to compare different nonlinear mixed models. The “Parameter Estimates” table lists the maximum likelihood estimates of the five parameters and their approximate standard errors, computed using the final Hessian matrix. Approximate *t* values and Wald-type confidence limits are also provided, with degrees of freedom equal to the number of subjects minus the number of random effects.

From the above results, we have the following linear fitted model:

$$\mathbf{CUNIT = -62.86 + 2.65*TAGE} \qquad [2]$$

Here you would expect the average yearly increase of merchantable volume is 2.65 units per acre. However, when TAGE is near zero, the initial value of CUNIT is -63, a negative volume number – that is not possible!

## ANALYSIS II: NONLINEAR LOGISTIC MODEL

If a linear change trajectory with a negative intercept at the initial point does not make sense, it may be appropriate to hypothesize the true change trajectory that gave rise to this sample is nonlinear. As we have noticed earlier, those series plots are S-shaped curves, including (1) a lower asymptote ("floor") from which all installations rise, (2) an upper asymptote ("ceiling") where tree volume of all locations will reach a maximum at some point, (3) a smooth curve joining the two asymptotes. These three features define a so called "logistic" trajectory. We therefore adopt the following logistic nonlinear mixed model (see SAS Institute Inc, 2008 for more references),

$$y_{ij} = \frac{b_1 + u_{1i}}{1 + exp[-(Time_{ij} - b_2)/b_3]} + e_{ij} \qquad [3]$$

Here $Time_{ij}$ is the corresponding time; *B1, B2, B3* are the fixed-effects parameters; $U_{1i}$ is the random-effect parameter assumed to be iid N(0,$\sigma^2$u), and $e_{ij}$ is the residual error assumed to be iid N(0,$\sigma^2_e$) and independent of the $U_{1i}$. This model has a logistic form, and the random-effect parameters $U_{1i}$ enter the model linearly.

The NLMIXED procedure to fit this nonlinear mixed model is as follows:

```
PROC NLMIXED DATA=DataA;
  PARMS b1=100 b2=30 b3=10 s2u1=1 s2e=1;
  y = (b1 + u1)/ (1.0 + EXP(-(TAge-b2)/b3));
  p =  b1/ (1.0 + EXP(-(TAge-b2)/b3));
  MODEL Cunit ~ NORMAL(y,s2e);
  RANDOM u1 ~ NORMAL(0,s2u1) SUBJECT=Inst;
  PREDICT p OUT=DataC;
  TITLE 'Model 2: Logistic growth curve model using NLMIXED';
RUN;
```

In the PARMS statement there are three fixed-effects parameters (*B1, B2, B3*) and two variance components (*S2U*, *S2E*). *B1* determines the maximum or the asymptote value that *Y* can approach; *B2* is related to the intercept but can only determine the intercept jointly with B3. B3 also determines the rapidity with which the trajectory approaches the upper asymptote or the ceiling.

The next four statements specify the logistic mixed model, where the Y is predicted dependent variable and p is the fixed component in Y. The dependent variable CUNIT is normally distributed with mean Y and variance S2E. The RANDOM statement defines the single random effect to be U1, and specifies that it follows a normal distribution with mean 0 and variance S2U1.

The main output from the logistic analysis is as follows.

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 517.2 |
| AIC (smaller is better) | 527.2 |
| AICC (smaller is better) | 528.1 |
| BIC (smaller is better) | 528.2 |

Parameter Estimates

| Parameter | Estimate | Standard Error | DF | t Value | Pr > \|t\| | Alpha | Lower | Upper | Gradient |
|---|---|---|---|---|---|---|---|---|---|
| b1 | 100.83 | 15.9830 | 8 | 6.31 | 0.0002 | 0.05 | 63.9754 | 137.69 | -8.86E-8 |
| b2 | 42.6062 | 1.1624 | 8 | 36.65 | <.0001 | 0.05 | 39.9258 | 45.2867 | 1.258E-6 |
| b3 | 7.5580 | 0.5309 | 8 | 14.24 | <.0001 | 0.05 | 6.3338 | 8.7822 | -5.48E-6 |
| s2u1 | 1980.61 | 974.22 | 8 | 2.03 | 0.0765 | 0.05 | -265.95 | 4227.16 | 1.024E-9 |
| s2e | 36.6171 | 6.4246 | 8 | 5.70 | 0.0005 | 0.05 | 21.8021 | 51.4322 | 6.236E-7 |

For the logistic fitted model, the predicted volume values for an individual forest location is thus

```
CUNIT = 100.83 / (1.0 + EXP(-(TAGE-42.61)/7.56));
```
[4]

## CONCLUSION

Plotting predicted trajectory of linear and logistic models side by side allows direct comparison of the two fitted models. Figure 2 shows that the linear trajectory does not flatten out asymptotically as the curve approaches an upper or lower limit, while the logistical model does. This difference explains why the predicted value of the linear model is negative even at the initial time point. The difference also pointed out a very important feature of logistic growth curves in biology and many other sciences when changes often include a plateau, such as when body shape and size typically level off with age. Based on these observations, we conclude that the nonlinear logistic trajectory is the better fitting model for the present study.

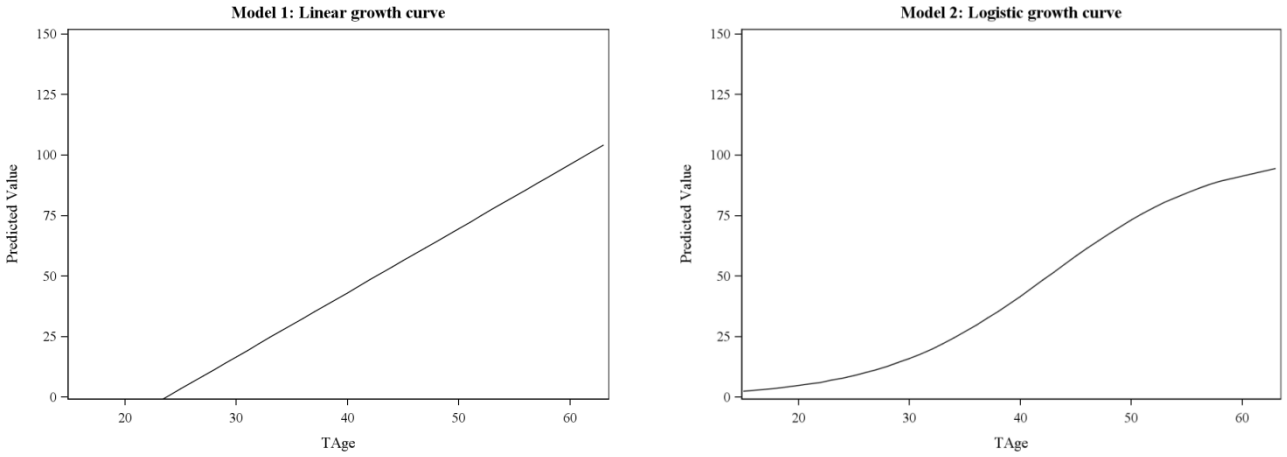


**Figure 2. Plot of predicted linear and nonlinear trajectory**

It should be pointed out that our decision to pick the logistic model over linear model is based on the substantial consideration, not on the comparison of model fit statistics. If we are negligent and ignore the common sense, we might reach an opposite conclusion. Indeed, the linear model has lower goodness of fit values –2LL, AIC, and BIC. Inspection of these output "Fit statistics" may suggest that the linear model fits data better than the logistic model. However, the linear model does not make sense for the current data – it is not possible for the forest volume to have a negative initial value and to grow infinitely. Therefore we should be cautious when we choose among a group of well-fitting growth models in longitudinal research. Blind numeric comparison of model fit indexes will rarely help us to pick the correct model. "Substance is paramount", as Singer and Willett (2003) have warned. The best way to select an appropriate growth model is to work with a theoretical framework and blend it with strong empirical evidence.

## REFERENCES


- Ferrer, E., Hamagami, F. & McArdle, J. J. (2004). Modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. Structural Equation Modeling, 11(3), 452-483.
- Marshall, D (2005). http://www.growthmodel.org/mixedmodels/mixedmodels.htm
- SAS Institute Inc. (2008). SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.
- Singer, J. D.,& Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press.

## ACKNOWLEDGMENTS

I would like to thank Dr. John McArdle for his support and advisory during my studies at USC, from whom I learned SAS and longitudinal analysis methods.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mo Zheng
University of Southern California
Email: mzheng@usc.edu