Non-linear mixed models for disease incidence and severity: Modeling plant diseases in tropical crops

Raúl E. Macchiavelli
College of Agricultural Sciences, University of Puerto Rico, Mayagüez, Puerto Rico
Box 9000, Mayagüez, PR 00681-9000, Puerto Rico
raul.macchiavelli@upr.edu

Abstract

Progress curves are used in plant disease epidemiology to describe temporal changes in the proportion of diseased plants (disease incidence) or in the proportion of diseased plant material (disease severity). Models typically used are intrinsically nonlinear and the observations are taken longitudinally on the same unit (for example, the plant or the plot). Data are binary, proportions based on an integer denominator, or real numbers in (0, 1). Non-linear mixed models can accommodate all these features, since they implicitly incorporate correlation between longitudinal observations, and can be applied with different distributions. Since the full likelihood is specified, likelihood based inference can be applied and both unit specific and marginal inference are available. In this paper we fit non-linear mixed models to describe disease incidence progress curves of papaya ring spot virus in papaya and disease severity of black Sigatoka (caused by the fungus *Mycosphaerella fijiensis*) in banana. We compare the results of fitting nonlinear mixed models, discuss alternatives to conduct marginal inference to estimate percentile curves of interest, and interpret the results in the analyzed examples.

Keywords: binary data, logistic-normal, marginal models, subject specific models.

1. Introduction

Diseases are normally monitored over time, assessing the amount of disease present in a population of plants. This change over time is studied with the so-called “Disease Progress Curve,” which represents an interpretation of all host, pathogen and environmental effects occurring during an epidemic (Madden *et al.*, 2007).

These curves provide tools for analyzing plant disease epidemics, comparing different conditions (treatments), and predicting disease dynamics. The “amount” of disease, *Y*, can be the proportion of diseased/dead plants (incidence) or the proportion of diseased plant material (severity). The most common disease progress curves model *Y* as a function of time *t*, and are derived from differential equations. The exponential model assumes that the rate of change of the disease is proportional to the amount of disease present:

$$\frac{dY}{dt} = r_e Y, \quad Y = Y_0 \exp(r_e t)$$

The mono-molecular (negative exponential) model assumes that the rate of change of the disease is proportional to the proportion of plant material not affected by the disease:

$$\frac{dY}{dt} = r_m (1-Y), \quad Y = 1 - B\exp(-r_m t)$$

The logistic model assumes that the rate of change of the disease is proportional both to the amount of disease and to the proportion of plant material not affected by the disease:

$$\frac{dY}{dt} = r_l Y(1-Y), \quad Y = \frac{1}{1+\exp(-B - r_l t)}$$

The Gompertz model assumes that the rate of change of the disease is proportional both to the amount of disease and to the log of the amount not affected by the disease:

$$\frac{dY}{dt} = r_g Y\left[-\log Y\right], \quad Y = \exp\left[-B\exp(-r_g t)\right]$$

Typically we observe the “amount” of the disease in longitudinal random samples under one or more conditions (experimental treatments, environments, etc.). In each unit (plot, tree, plant) we observe periodically, for example, the number of diseased subunits, the percentage of damage area, or the presence of symptoms. From these observations we estimate the best fitting curve and make inferences.

The data can therefore be binary data, proportions (*y/m*) or estimated percentages. They typically have non constant variances and are not independent (observations from the same unit are dependent). There is also a possible dependence among neighboring units (contagion). Hence the models used need to consider the non-linear nature of the phenomenon, non-normal distributions, different variances, and the longitudinal data structure.

There are two main approaches to model longitudinal data: Marginal models and mixed effects models (Serroyen *et al.*, 2009). In normal data with linear models the two approaches yield the same results. In nonlinear models and/or nonnormal data the parameter interpretation is inherently different (Molenberghs and Verbeke, 2005).

The mixed effect model has the advantage that can explain variability among regression coefficients explicitly, its likelihood is known (and thus likelihood-based inference can be applied), and a marginal mean can be derived from it by integrating the random effects. On the other hand, modeling the marginal means explicitly with a regression model permits interpreting the regression coefficients in terms of population averages, and sometimes simplify the modeling of the dependency among observations.

2. Methods

Suppose that we have *n* independent units (plants, plots, trees, etc.). Each unit *i* is monitored repeatedly t_i times, and at each time *j* an observation Y_{ij} is recorded. The nonlinear mixed model specifies the conditional distribution of Y_{ij} given u_i, the effect of the *i* th unit, and the distribution of u_i. The disease progress curves introduced in the previous section are interpreted now as the conditional expectation of Y_{ij} given u_i. Thus,

$$\begin{aligned} &E\left(Y_{ij} \mid u_i\right) = h\left(\beta, u_i, x_{ij}, t_{ij}\right), \quad && Var\left(Y_{ij} \mid u_i\right) = \sigma^2\left(E\left(Y_{ij} \mid u_i\right), \gamma\right) \\ &Y_{ij} \mid u_i \sim f\left(\beta, \gamma, u_i\right), && u_i \sim N\left(0, \Sigma_u\right) \end{aligned}$$

Here *h* is a nonlinear function which may depend on covariates and time (for example, the ones mentioned in the introduction). There may be extra parameters in the conditional variance. The conditional distribution of Y_{ij} given u_i is *f*. The distribution of the random effects is usually the normal distribution, but other distributions can be used (Nelson et al., 2006).

The main objectives of inference when applying nonlinear mixed models to describe plant diseases are to estimate disease progress curves for "typical" units (random effects =0), to study how the curves depend on covariates (for example, treatments or conditions), and to study how the curves vary in the population of interest.

The parameters β are interpreted as regression coefficients describing the relationship between the response and the covariates or time, controlling for the unit effects, as well as characteristics of a "typical" curve; i.e., a curve for a unit such that its random effects are 0.

The estimation of the parameters is typically done by maximum likelihood. In order to obtain the likelihood function, it is necessary to integrate through the random effects distribution, making the process computationally intensive (at each optimization step a numerical integration with the dimension of the random effects is necessary). Adaptive Gauss-Hermite quadrature is the method of choice for the integration step when the number of random effects is not very large, and several alternatives are used for the optimization (Pinheiro and Bates, 2000; Molenberghs and Verbeke, 2005).

The marginal distribution induced by the model can be obtained by integrating through the distribution of random effects. This can be interpreted as averaging the random effects, and can be done using Monte Carlo simulation (Molenberghs and Verbeke, 2005). The process is the following:

Suppose the model parameters were known. We first simulate random effects from the random effects distribution. Using this realization of the random effects, we then simulate observations at each time *j* from the conditional distribution. Thus we obtain one disease progress curve with correlated observations (since they all would share the same value of the random effects). By repeating this process of simulating first from the distribution of random effects and then from the conditional distribution of Y_{ij}, we generate a population of disease progress curves. At each time *j* these observations Y_{ij} are the marginal distribution of Y_{ij}. The curve obtained by joining the means of the Y_{ij} obtained in this way is the population average disease progress curve. Similarly, curves obtained by joining a given percentile of the distribution of the Y_{ij} for each time *j* can be interpreted as percentiles of the population of disease progress curves. For example, plotting the median, quartiles, and 5th and 95th percentile curves would yield a good idea of the possible curves that could be observed under conditions similar to the ones studied (Torres Saavedra, 2006).

In real applications this process must take into account the fact that the parameters β are estimated, and hence when making the simulations the approximate variance-covariance matrix of the β's must be considered. In order to attain this, we propose sampling from the (approximate) normal distribution of the estimated β parameters:

a. Obtain a sample $u_i \sim N\left(0, \hat{\Sigma}_u\right)$
b. Obtain a sample $b \sim N\left(\hat{\beta}, \hat{\Sigma}_{\hat{\beta}}\right)$
c. For each *j*, obtain a sample $y_{ij} \sim f\left(b, u_i, \hat{\gamma}\right)$
d. Repeat a-c to generate the population of marginal curves

3. Papaya Ring Spot Virus

An experiment was carried out to control certain insects (aphids), which are vectors of the virus (Robles *et al.*, 2007). There were 4 different treatments: Control 1 (bare soil), control 2 (soil with weeds present), soil covered with black plastic, soil covered with reflectant plastic (silver plastic). Each treatment was randomly assigned to 5 plots (each plot had 20 plants). The experiment was monitored every two weeks during 16 weeks (8 longitudinal observations). Each week, every plant was checked to see whether it showed symptoms. Once the plant showed symptoms, it was classified as diseased for the rest of the experiment. The variable of interest is the disease index for treatment *i*, time *j* and plot *k:*

$$DI_{ijk} = \frac{Y_{ijk}}{20} = \frac{\text{number of plants with symptoms}}{20}$$

The logistic progress curve with a random intercept was found the best model for these data:

$$E\left(Y_{ijk} \mid u_k\right) = 20 \; \frac{1}{1+\exp(-\beta_{0i}+u_k-\beta_{1i}t_j)}$$

$$Y_{ijk} \mid u_k \sim Binomial\left(20, E\left(Y_{ijk} \mid u_k\right)\big/20\right), \quad u_k \sim N\left(0,\sigma_u^2\right)$$

After considering different models (different intercepts, slopes, and variances of the random effects for each treatment, random effects for both intercept and slopes, etc.), the best fitting model had a random intercept with a constant variance, two intercepts and two slopes (one for plastic covers, one for the control). This was selected using the BIC criterion and likelihood ratio tests.

The estimated marginal curves were obtained used the algorithm indicated in Section 2, and are shown in Figure 1.

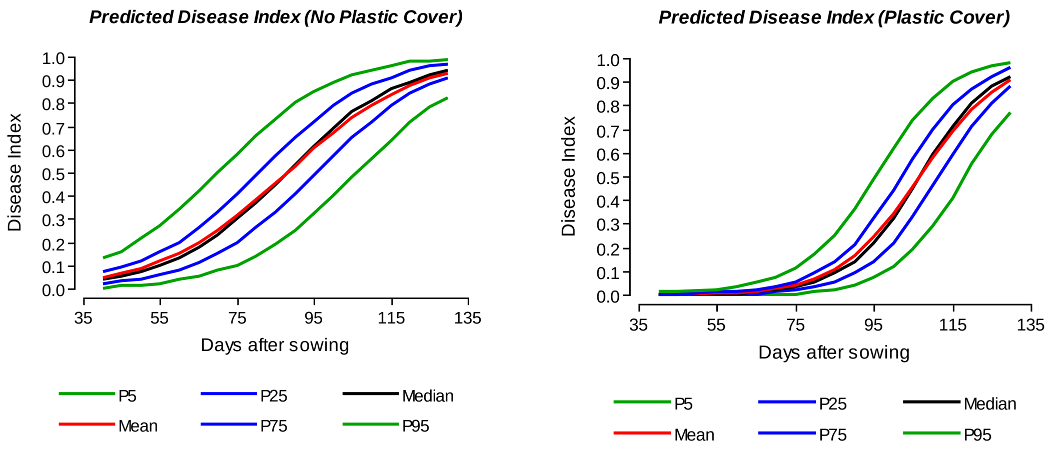

Figure 1. Predicted marginal disease progress curves under two treatments to control papaya ring spot virus.

From these curves it is clear that the plastic cover delays the progress of the disease, and, although there is variability among the curves from the same treatment, this is clearly noticed in all curves (in fact, there are significant differences between the intercepts of plastic covers and controls, and between the slopes of plastic covers and controls). Other interesting findings (not detailed here) were that no difference was found between reflectant and black cover, and that the use of plastic cover (either of them) delayed the disease progress in such a way that 50% of the plants (the inflexion point in the logistic curve) became diseased 19 days later on average in the treatments with plastic covers than in the controls. The difference between the mean and the median curves is small. In this logistic-normal case with random intercept, the median curve is also the conditional curve when u_k=0, and in this situation the slope of the mean curve is always smaller than the median curve (Agresti, 2002).

4. Black Sigatoka

Black Sigatoka, a disease caused by the fungus *Mycosphaerella fijiensis*, was studied in an experimental banana plantation in Isabela, Puerto Rico (García Saavedra *et al*., 2012). A total of 36 plants under chemical control were monitored from planting through their second ratoon (i.e., three harvest seasons). Here we analyze the data for the first ratoon (the second season, which grows from one secondary stem after the banana bunch from the main stem is harvested). This stage is characterized by a large variability between plants in the growth rate, yield, and disease severity. Since plants are treated chemically, there is little variability in the times at which disease symptoms appear. Each leaf *k* in plant *i* and time *j* is given a grade b_{ijk} using the Stover-Gauhl scale (0: no symptom, 1: less than 1% of the leaf affected, 2: 1%-5% of leaf area affected, 3: 6%-15% of leaf area affected, 4: 16%-33% of leaf area affected, 5: 34%-50% of leaf area affected, 6: more than 50% of leaf area affected). All leaves from the same plant are then averaged and expressed in a 0-1 (or 0-100%) scale (Severity Index, *SI)*:

$$SI_{ij} = \frac{\sum_{b=0}^{6} n_{ijb} b}{6\sum_{b=0}^{6} n_{ijb}}$$

where n_{ijb} is the number of leaves in plant *i* at time *j* having grade *b* (*b*=0 to 6).

Based on the shape of the severity progress curves, S-shaped curves (logistic and Gompertz) were considered. Different combination of random effects on intercept and/or slope were considered, and using BIC the Gompertz model with random slope was selected:

$$E(SI_{ij} \mid \text{plant}_i) = \mu_{ij} = \exp\left[-\beta_0 \exp(-(\beta_1 + v_i)t_j)\right]$$
$$SI_{ij} \mid v_i \sim N\left(\mu_{ij}, \sigma^2\right)$$
$$v_i \sim N\left(0, \sigma_v^2\right)$$

This model could also be fitted using a beta distribution (which is, in fact, more realistic for continuous or “quasi-continuous” data in [0,1]). But the observed data showed less variability for data points near 0.5, and more variability for severities closer to 0 or 1, which is opposite to what is expected in a beta distribution. The normal distribution, on the other hand, seems to provide a reasonable approximation. Figure 2 shows he plant-specific fitted curves (left panel) and the induced marginal curves (right panel). We can see that the mean and median marginal curves are almost indistinguishable.

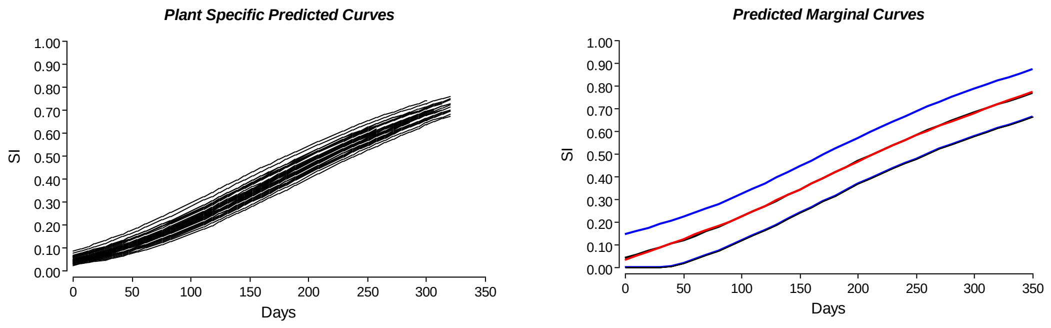

Figure 2. Plant specific and marginal severity progress curves for Black Sigatoka in banana under chemical control. Marginal curves shown are P25 and P75 (blue), mean (red), and median (black).

5. Conclusions

Nonlinear mixed models are a very natural approach to study disease progress curves in plants. We can use them to understand how individual plants or plots vary in their disease incidence or severity, and also to study the marginal curves induced by the model (Serroyen et al., 2009). Full likelihood methods for many distributions for both the observed data (conditional on the random effects) and the random effects are well known and computationally feasible. The examples presented here are not the only possibilities for applying these models in plant pathology. We have not considered curve comparisons, specific predictions, parameter estimation, and smoothing. All these tools can be applied in the context of nonlinear mixed models and would yield results very useful to plant pathologist and crop managers to understand, monitor, and control diseases.

References

Agresti, A. (2002), Categorical Data Analysis, 2nd ed., Wiley, New York.

García Saavedra, Y., Macchiavelli, R., J. Chavarría Carvajal (2012), Using nonlinear mixed models with beta distribution to fit disease progress curves to model black Sigatoka epidemics in banana in Puerto Rico, *Abstracts, XXVIth International Biometric Conference*, Kobe, Japan.

Madden, L. V., Hughes, G., and van den Bosch, F. (2007), *The Study of Plant Disease Epidemics*, APS Press, St. Paul, MN.

Molenberghs, G. and G. Verbeke (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Nelson, K., S. Lipsitz, G. Fitzmaurice, J. Ibrahim, M. Parzen, R. Strawderman (2006), Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects, *J. Comput. and Graphical Stat.*1, 39-57

Pinheiro, J. and D. Bates (2000), *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

Robles, W., Pantoja, A., Abreu, E., Pena, J., Ortiz, J., Lugo, M.D., Cortes, M., Macchiavelli, R. (2007), Effects of cultural practices on the incidence of aphids and virosis on *Carica papaya* L., *Manejo Integrado de Plagas y Agroecología*, 77, 38-42.

Serroyen, J., G. Molenberghs, G. Verbeke, and M. Davidian (2009), Nonlinear models for Longitudinal Data, *The American Statistician*, 63, 378-388.

Torres Saavedra, P. (2006), Percentile Curves in Binary Longitudinal Data, M.S. Thesis, Dept. of Mathematical Sciences, University of Puerto Rico, Mayaguez.