Paper 182-2007

Marginal Interpretation of Subject-Specific Curves: Logistic-Normal Regression

Pedro A. Torres, University of Puerto Rico, Mayagüez, PR
Raúl E. Macchiavelli, University of Puerto Rico, Mayagüez, PR

ABSTRACT

We propose the *percentile curves* concept as conditional probabilities curves across representative percentiles of the distribution of curves induced by random effects in a logistic model with random intercepts. We extend this concept to a logistic model with random intercepts and slopes and propose a methodology to approximate the percentile curves using the Monte-Carlo technique. We apply this concept to a binary longitudinal data set. The results suggest that the percentile curves complement the analysis of longitudinal data and permit a marginal interpretation of subject-specific parameters. The first and third quartile, the median and the mean curves are the principal percentile curves to describe the behavior of the data.

INTRODUCTION

Generalized linear mixed models are widely used in medicine, agriculture, social sciences and others. In particular, the logistic-normal model for binary longitudinal is commonly applied to model binary longitudinal clustered data. Several works about binary longitudinal modeling have interpreted and compared the estimates in marginal and random effects models (McCulloch and Searle, 2001; Agresti, 2002; P. Diggle and Zeger, 2002; Lee and Nelder, 2004). These works analyze the relationship between subject-specific and population-average models. They have discussed the convenience of using either a marginal or a random effects (or conditional) models in different situations. The main difference between marginal and conditional models is based on the regression coefficients, which can describe individual responses or marginal responses due to changes in covariates (Lee and Nelder, 2004). The random-effects models are preferred by some researchers because they permit a marginal interpretation. In addition, the adequacy of using alternative models such as marginal models in binary longitudinal data have been questioned by some authors (Chaganty and Joe, 2004; Duchateau and Janssen, 2005).

In this paper we propose a methodology to interpret subject-specific curves in a logistic model with random effects applied to a longitudinal study. Our proposal is to interpret the subject-specific curves across time as percentiles of a distribution that depends on the values of the random effects variance and the covariate parameters. We apply this result in one dataset.

LOGISTIC MODEL WITH RANDOM EFFECTS

Consider a longitudinal study with binomial response variables vector $Y_i = (y_{i1}, \ldots, y_{in_i})$ measured at recorded times $t_i = (t_{i1}, \ldots, t_{in_i})$ along with possibly time-dependent covariates matrix $X_i = (X_{i1}, \ldots, X_{in_i})$, where in each observation the covariate is a $p-$dimensional vector x'_{ij}, for $i = 1, 2, ..., N$ subjects. The logistic model with random effects to analyze these data is given by:

$$\begin{aligned} Y_{ij} \mid u_i &\sim \text{Binomial}(m_{ij}, \pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x'_{ij}\beta + z'_{ij}u_i \end{aligned} \tag{1}$$

where m_{ij} is the number of trials ($j = 1, 2, \ldots, n_i$), β is p-dimensional fixed parameters vector, z'_{ij} is a $q-$dimensional vector associated to the random effects design matrix, and u_i is a $q-$dimensional vector of the random effects normally distributed, $u_i \sim N(\vec{0}, \Sigma)$.

In the simplest case, we consider a logistic model with random intercepts, $u_i = u_{i0}$, then the model presented in Equation (1) is reduced to

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x'_{ij}\beta + u_{i0} \tag{2}$$

where $u_{i0} \sim N(0, \sigma^2)$.

PERCENTILE CURVES CONCEPT

Logistic Model with Random Intercepts

A logistic model with random intercepts yields estimates with a subject-specific interpretation. These models express the conditional probabilities of success (π_{ij}) for each individual (conditionally on the random intercept). Based on the model defined in Equation (2) and omitting the estimation problem with the fixed effects parameters, it is possible to determine the density function of the conditional probabilities π (the ij indices are omitted for simplicity) based on random intercept u_{i0} (Duchateau and Janssen, 2005):

$$f_\pi(\pi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left\{\text{logit}\,(\pi) - x'_{ij}\beta\right\}^2\right] \frac{1}{\pi(1-\pi)} \tag{3}$$

The density function presented in Equation (3) allows to study the influence of random intercepts on the conditional probability distribution. Given a random intercept value, u_{i0}, and the value of the covariates, they produce a point estimate of the probability for the $i-th$ subject, which corresponds to one observation of the distribution of π. The characteristics (shape and scale) of the density of π at each specific time, determine the distribution (and therefore the mean) of the conditional probabilities, such as shown in Figure 1. The marginal mean induced by the random intercepts at each time is computed integrating out the random effects. If the marginal means are connected across time, then the marginal mean curve is generated. It should be close to the curve generated by a marginal model if both marginal and random effects models are reasonable. These procedures can be carried out with any value of the distribution of π, generating several curves of interest.

For a logistic normal model, suppose that u_q is the *q-th* is percentile of the random intercepts distribution. Then the subject-specific curve with the value u_q plugged in across time is the $q-th$ percentile of the *subject-specific curves distribution* induced by the random intercepts. The subject-specific curve generated with u_q will be called the $q-th$ *percentile curve*. Particulary, if $u_{50} = 0$, the median of the random normal intercepts, the curve is the *median curve* or *typical curve*. For example, it may be of interest to consider the first and third quartile curves, denoted by $Q1$ and $Q3$, respectively. The *subject-specific curves* are generated connecting the values of the distribution of the conditional probabilities π_{ij} (Figure 1) across time. The percentiles curves can be obtained using two methods. First, percentiles can be obtained directly from the distribution of π_{ij} at each time-point. Second, to calculate the inverse logit of each percentile of the random intercepts distribution: $\text{logit}^{-1}(x'_{ij}\beta + u_q)$. In both cases, the percentiles curves across time are obtained by connecting the correspondent values of π_{ij}. The proof of this result is based on the classical inference theory and the properties of a link function in GLMM's (Casella and Berger, 2001; Demidenko, 2004; Agresti, 2002).

Logistic Model with Random Intercepts and Slopes

In a logistic model with random intercepts it is straightforward to find the density function of π_{ij}. However, in a logistic model with random intercepts and slopes, π_{ij} is a bivariate function of the random effects u_i. Because of this, an analytic expression for $f_\pi(.)$ may not be possible. Therefore, we propose a methodology to estimate the percentiles curves in these models using the Monte-Carlo technique. This method can be generalized to any GLMM.

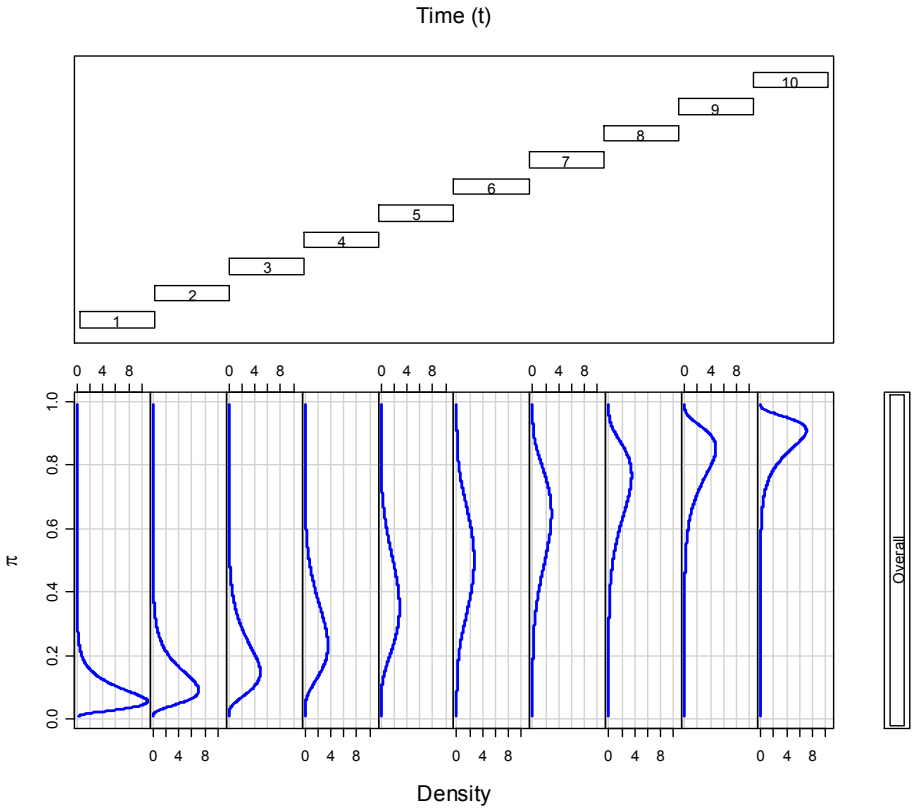

Figure 1: Densities functions of π across time for model: logit$(\pi_{ij}) = -3+0.5t+u_{i0}$, where $u_{i0} \sim N(0, 0.36)$.

Methodology to compute Percentile Curves

We propose an empirical approximation to percentiles curves using the estimates of the distribution of random intercepts and slopes.

Procedure

- Generate a number K, sufficiently large, of samples u_i from a bivariate normal distribution with the estimates of $\vec{\beta}$ mean vector, and covariance matrix Σ using the Monte Carlo Method. The estimates of $\vec{\beta}$ and Σ obtained in the fitted model are used in this step.
- For each $k = 1, \ldots, K$ random effects vector generated, compute the conditional probabilities in each time using the inverse logit transformation:

$$\hat{\pi}_{ijk} = \frac{\exp[\hat{\beta}_0 + u_{i0k} + (\hat{\beta}_1 + u_{i1k})t_{ij}]}{1 + \exp[\hat{\beta}_1 + u_{i0k} + (\hat{\beta}_1 + u_{i1k})t_{ij}]}$$

- Calculate the percentiles of the empirical distribution of $\hat{\pi}$. For each time-point it is necessary to sort the $\hat{\pi}_{ijk}$ values calculated previously, and then select the respective percentiles. To find the $s-th$ percentile, take the correspondent $[\frac{Ns}{100}]$ ordered value of $\hat{\pi}_{ijk}$.
- To estimate the marginal mean induced by the random effects at a specific time, average the subject-specific probabilities estimated with the K random effects vectors u_i generated in the first step, such as is presented by some authors (Molenberghs and Verbeke, 2005):

$$\hat{\pi}_{ij} = \frac{1}{N} \sum_{r=1}^{K} \hat{\pi}_{ijk}$$

Using this methodology is possible to obtain the percentiles of π at each time-point. The percentile curves are generated connecting the respective percentiles points across time.

This methodology may be expanded for more than two random effects when the model takes the general form of any GLMM:

$$h[E(Y_{ij} \mid u_i)] = x'_{ij}\beta + z'_{ij}u_{ij}$$

where $h(.)$ is a link function and $Y_{ij} \mid u_i$ has a distribution function which belongs to the exponential family.

APPLICATION

In this section we apply the percentile curves concept to analyze binary longitudinal data. The percentile curves summarize the results of the subject-specific curves for each environment and also facilitate the comparison of the disease progress between the environments in terms of the average and the variability of the infection indices.

Dataset: White Rot in Garlic

White rot is produced by a soil-born fungus, *Scletorium cepivorum* Berk, with very serious effects such as infected fields that can not be used for garlic or onions for years. The fungus infects areas of a field and kills garlic plants. The disease progress is faster if the initial density of sclerotia (reproductive structure of this fungus which can stay inactive in the field for several years) is high. The data were obtained from a study about the epidemiology of white rot in garlic in five different environments *CE1, CE2, CE3, JM1*and *JM2* [1]. In each environment, the experimental field was divided in ten blocks. In each block five to six sampling points (*sites*) were randomly selected. Measurements of incidence over a total of fifty plants at each point (number of dead plants/total number of plants) were taken every t days, where t varies according to the environment. The number of monitored plants represented approximately 10% of the plantation. The soil at each site had been previously sampled to determine sclerotium density before planting (initial inoculum density). The objective is to characterize the progress curves of the infection.

A first exploratory analysis suggests the progress curves of infection have different trends between environments and these also depend on the initial density inoculum within each environment, such as shown Figure 2.

Models Considered

Based on the previous exploratory analysis, logistic regression models with normal random coefficients were used to fit the disease progress curves using NLMIXED procedure from SAS/STAT$^{\mathrm{TM}}$ software v9.1.3. (*see Appendix*). Models were selected using the BIC criterion. Models considered included random intercepts and fixed slopes, and random intercepts and slopes. In all cases, possible heterogeneity in the parameters due to environment and initial inoculum density (more than 15 sclerotia per 100 g of soil, or less than 15 sclerotia per 100 g of soil) was considered.

Let π_{ijkm} be the expected disease index at environment $i(i = 1, \ldots, 5)$, initial level of sclerotia $j(j = 1, 2)$, time t_k, given the effects of site $m(i)$. The models considered are specific cases of:

$$\text{logit}(\pi_{ijkm} \mid u_{m(i)}, v_{m(i)}) = \alpha_{ij} + u_{m(i)} + (\beta_{ij} + v_{m(i)})t_k$$

where $[u_{m(i)}, v_{m(i)}] \sim N(\vec{0}, \Sigma)$.

The parameters and the criteria to compare the models that were considered to analyze the data set are shown in Table 1. Model 6 is chosen, and this confirms that the disease progress curves depend on initial inoculum density and environment, but the variability of intercepts and slopes depends only on initial density.

Using the methodology presented, we obtained an approximation to the percentiles curves in each environment and level of initial inoculum density. For the final model we computed the marginal mean,

[1] *Source*: Martha Conles, Universidad Nacional de Córdoba, Argentina.

Table 1: Parameters of the Models Considered

Model	Parameters	Intercept	Slope	Var(u)	Var(v)	Cov(u, v)	BIC
1	11	α_i	β_i	σ^2_u	–	–	9851.4
2	15	α_i	β_i	σ^2_{ui}	–	–	9811.6
3	23	α_i	β_{ij}	σ^2_{uij}	–	–	9722.6
4	36	α_{ij}	β_{ij}	σ^2_{uij}	σ^2_{vij}	–	8870.0
5	22	α_{ij}	β_{ij}	σ^2_{uij}	σ^2_{vij}	–	8856.3
6	***24***	α_{ij}	β_{ij}	σ^2_{uij}	σ^2_{vij}	σ_{ujvj}	***8854.5***
7	45	α_{ij}	β_{ij}	σ^2_{uij}	σ^2_{vij}	σ_{ujvj}	8887.6

median, Q1 and Q3 curves by sampling repeatedly from an appropriate bivariate normal distribution of random effects (intercepts and slopes) such as shown Figure 3. We used SAS/IML$^{\mathrm{TM}}$ software to run the routine and GPLOT procedure from SAS/GRAPH$^{\mathrm{TM}}$ to plot the results (*see Appendix*).

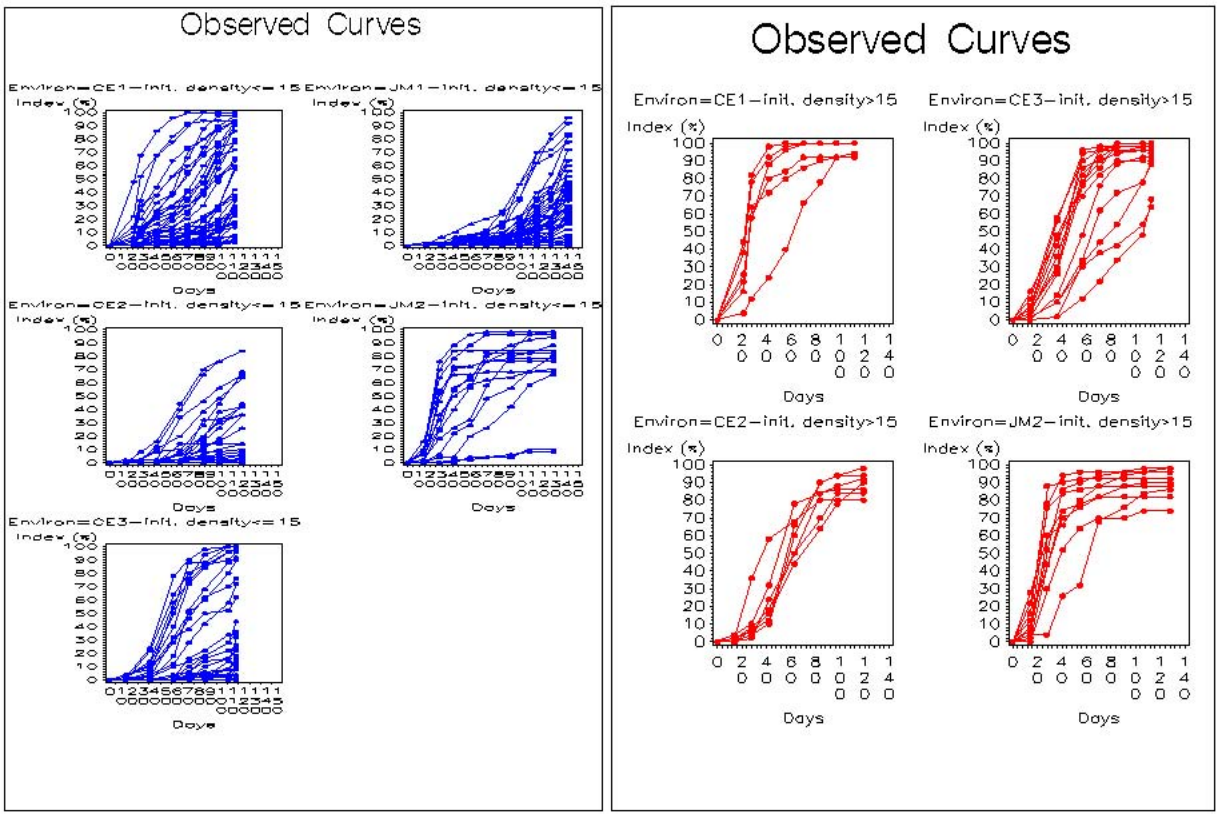

Figure 2: Observed Disease Index (%) for all environments separated by initial density inoculum .

CONCLUSIONS

The disease progress curves depend on initial inoculum density and environment, but the variability of intercepts and slopes depend only on initial density. The interpretation of disease progress curves can be done using a population-average curve (mean of the induced marginal distribution) and important subject specific curves (Q1, median and Q3 of the induced marginal distribution). Note that the percentiles curves summarize well the subject-specific curves for most subjects. The mean, the median, Q1 and Q3 curves are good candidates to represent the disease progress curves in each environment.

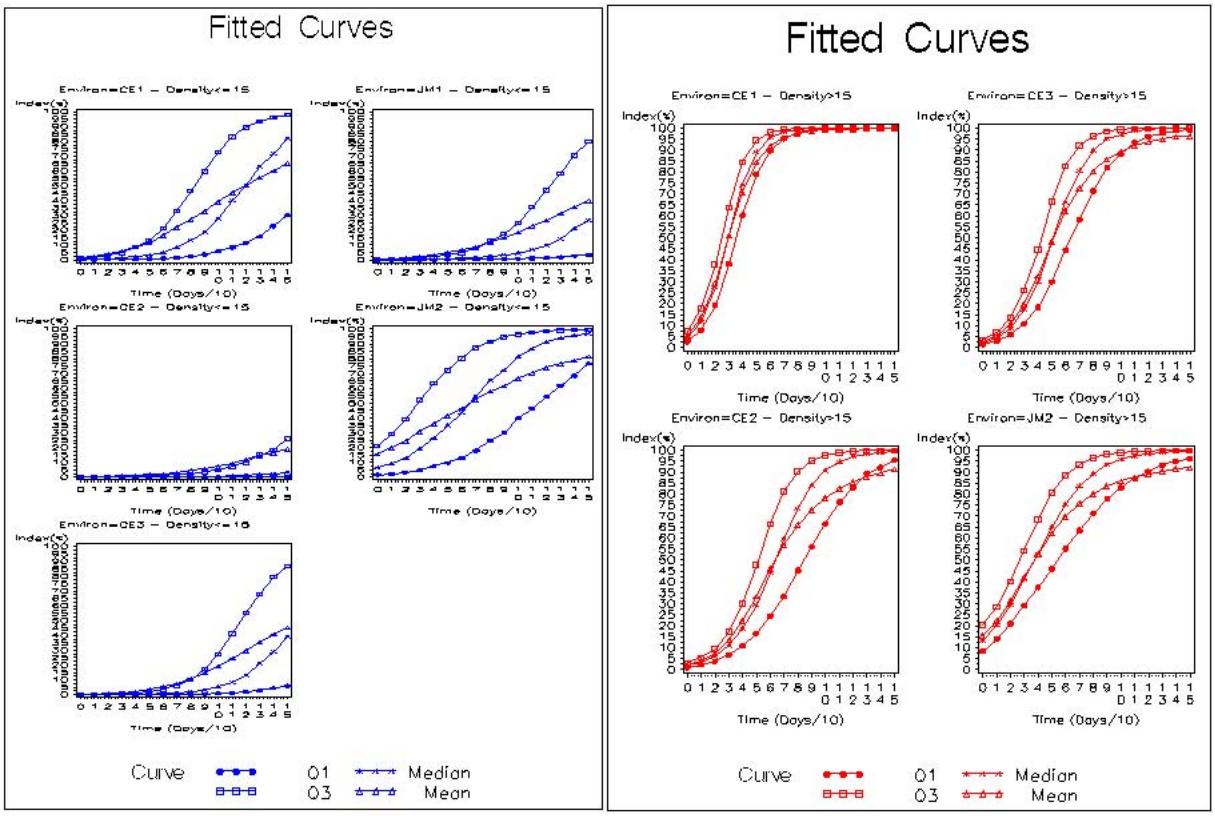

Figure 3: Induced mean, Median, Q1 and Q3 disease index (%) for all environments separated by initial density inoculum .

This methodology can be easily implemented in software such as SAS$^{\mathrm{TM}}$ and R, and can be extended to others GLMMs.

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley &Sons, Inc., second edition.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Thompson Learning, second edition.

Chaganty, N. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Society*, 66(4):851–860.

Demidenko, E. (2004). *Mixed Models. Theory and Applications*. John Wiley & Sons, first edition.

Duchateau, L. and Janssen, P. (2005). Understanding heterogeneity in generalized mixed and frailty models. *The American Statistician*, 59(2):143–6.

Lee, Y. and Nelder, J. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2):219–238.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons, first edition.

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

P. Diggle, P. Heagerty, K. L. and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, second edition.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Pedro A. Torres
Department of Mathematics, University of Puerto Rico
PO Box 5421
Mayagüez, PR 00681
Work Phone: (787) 832-4040
E-mail: pedro.torres@math.uprm.edu

Raúl E. Macchiavelli
Department of Agronomy and Soils, University of Puerto Rico
PO Box 9030
Mayagüez, PR 00681
Work Phone: (787) 832-4040
E-mail: rmacchia@uprm.edu

APPENDIX: SAS™ CODE TO ESTIMATE THE PERCENTILE CURVES

SAS™ Code to fit the model 6 in PROC NLMIXED

™ Model 6: "Random intercepts and slopes for each environment, random intercepts and slopes variances equal for all environments, correlated random effects and different distributions for each group of initial density inoculum ".

```
PROC NLMIXED data=garlic;
parms bu1a=-6 bu1b=-5
      bu2a=-9 bu2b=-6
      bu3a=-8 bu3b=-4
              bu4b=-7
      bu5a=-2 bu5b=-2
      r1a=0.5 r1b=0.4
      r2a=0.5 r2b=0.2
      r3a=0.6 r3b=0.2
              r4b=0.1
       r5a=0.4 r5b=0.1
      sua=1 sub=3
      lsva=-2 lsvb=-1
      corra=.5 corrb=.5;
if densescler>15 then
    mub=bu1a*(amb='CE1')+bu2a*(amb='CE2')+bu3a*(amb='CE3') +bu5a*(amb='JM2');
else
    mub=bu1b*(amb='CE1')+bu2b*(amb='CE2')+bu3b*(amb='CE3')+bu4b*(amb='JM1')+bu5b*(amb='JM2');
if densescler>15 then
   mur=r1a*(amb='CE1')+r2a*(amb='CE2')+r3a*(amb='CE3') +r5a*(amb='JM2'); else
    mur=r1b*(amb='CE1')+r2b*(amb='CE2')+r3b*(amb='CE3')+r4b*(amb='JM1')+r5b*(amb='JM2');
if densescler>15 then su=sua;
   else su=sub; if densescler>15 then sv=exp(lsva);
   else sv=exp(lsvb); if densescler>15 then suv=corra*sua*exp(lsva);
   else suv=corrb*sub*exp(lsvb);
eta = -u - r * time ;
expeta = exp(eta);
p = 1 /(1+expeta);
model enfermas   binomial(50,p);
random u r   normal([mub, mur],[su*su,suv,sv*sv]) subject=cuadrositio;
title "Model 6";
ods output fitstatistics=fit6 ;
ods output dimensions=dim6;
run;
```

SAS™ macro to generate the percentile curves

```
%macro curves(nrep);
proc iml;
/*Estimates matrix (b,r,su, sv, corr)*/
/*You can use the output of PROC NLMIXED*/
pars=
-3.0122 1.0177 0.7654 0.2561 -0.4734,
-5.8952 0.496 2.0471 0.1536 0.2682,
-3.9562 0.6224 0.7654 0.2561 -0.4734,
-7.7227 0.2885 2.0471 0.1536 0.2682,
-3.796 0.7418 0.7654 0.2561 -0.4734,
-7.8876 0.5005 2.0471 0.1536 0.2682,
```

```
-7.0021 0.4016 2.0471 0.1536 0.2682,
-1.8618 0.4992 0.7654 0.2561 -0.4734,
-2.6818 0.4134 2.0471 0.1536 0.2682;
mattrib pars colname=(b r su sv corr);
out=J(16*9,6,0);
cont=0;
/*For each environment*/
%do i=1 %to 9;
      par=pars[&i,{'b' 'r'}];
      sigma=J(2,2,0);
      sigma[1,1] = pars[&i,{'su'}]**2;
      sigma[2,2] = pars[&i,{'sv'}]**2;
     %let cov = pars[&i,{'corr'}]*pars[&i,{'su'}]*pars[&i,{'sv'}];
     sigma[1,2]=&cov;
     sigma[2,1]=&cov;
     l=half(sigma); /*Cholesky descomposition matrix*/
     /*For each time (days/10)*/
     %do t=0 %to 15 %by 1;
         z=J(&nrep,2,0);
        /*Normal random effects (z0,z1)*/
         %do k=1 %to &nrep;
           z[&k,1]=rannor(-1);
           z[&k,2]=rannor(-1);
         %end;
        u=t(l)*t(z);
        x=1 &t;
        xb=par*t(x);
        zu=x*u;
        lin=xb+zu;
        prob=exp(lin)/(1+exp(lin));
        tprob=t(prob);
        call sort(tprob,1);
        nq1=int(0.25*&nrep);nq2=int(0.5*&nrep);nq3=int(0.75*&nrep);
        call sort(tprob,1);
        q1=tprob[nq1];q2=tprob[nq2];q3=tprob[nq3];
        mm=sum(tprob)/&nrep;
        cont=cont+1;
        out[cont,1]=&i;out[cont,2]=&t;out[cont,3]=q1;out[cont,4]=q2;out[cont,5]=q3;out[cont,6]=mm;
     %end;
%end;
create pest from out[colname=({amb time q1 q2 q3 mm})];
append from out;
quit;
%mend curves;
```

Plotting the percentile curves

```
data curves;
  length Environ $30;
  set curves;
  if amb=1 then Environ="CE1 - Density>15";
  if amb=2 then Environ="CE1 - Density<=15";
  if amb=3 then Environ="CE2 - Density>15";
  if amb=4 then Environ="CE2 - Density<=15";
  if amb=5 then Environ="CE3 - Density>15";
  if amb=6 then Environ="CE3 - Density<=15";
  if amb=7 then Environ="JM1 - Density<=15";
```

```
  if amb=8 then Environ="JM2 - Density>15";
  if amb=9 then Environ="JM2 - Density<=15";
run;

symbol1 color=green interpol=join
        width=1 value=dot height=1;
symbol2 color=blue interpol=join
        width=1 value=start height=1;
symbol3 color=red interpol=join
        width=1 value=square height=1;
symbol3 color=brown interpol=join
        width=1 value=triangle height=1;
title "Fitted Curves";
proc gplot data=curves;
  plot index*time=curve;
  by amb;
run;quit;
```