GENERALIZED LINEAR MIXED MODELS: AN APPLICATION

Stephen D. Kachman

Walter W. Stroup

GENERALIZED LINEAR MIXED MODELS: AN APPLICATION

Stephen D. Kachman and Walter W. Stroup
Department of Biometry University of Nebraska–Lincoln

Abstract

The purpose of this paper is to present a specific application of the generalized linear mixed model. Often of interest to animal-breeders is the estimation of genetic parameters associated with certain traits. When the trait is measured in terms of a normally distributed response variable, standard variance-component estimation and mixed-model procedures can be used. Increasingly, breeders are interested in categorical traits (degree of calving difficulty, number born, etc.). An application of the generalized linear mixed to an animal breeding study of the number of lambs born alive will be presented. We will show how the model is determined, how the estimation equations are formed, and the resulting inference.

Key Words: Generalized Linear Model, Categorical, Mixed Model, Variance Components.

1. Introduction

Researchers are often faced with analyzing data for which the assumptions of independence and/or normality are not reasonable. When one of the assumptions is violated then, mixed models or generalized linear models may be used. Mixed models can model lack of independence with the use of random effects and generalized linear models can model a large class of distributions using link functions and variance functions. Difficulties arise when both the assumptions are not reasonable. Mixed models assume that the response variable is normally distributed and generalized linear models assume the data are independently distributed. When both of the assumptions are not reasonable, generalized linear mixed models (GLMM) may be used. Generalized linear mixed models include random effects from mixed models with link functions and variance functions from generalized linear mixed models. GLMM estimators and tests are given in (Breslow and Clayton, 1993; Stroup and Kachman, 1994).

Generalized linear mixed models are made up of several components. The flexibility in selecting the particular components to use allows for a variety of models to be analyzed. Random effects along with their covariances allow the modeling of a variety of experimental designs. In addition, the inclusion of random effects allows the modeling of variation arising from other sources such as Mendelian segregation of alleles. Inverse link functions allow for the modeling of non-additive effects. For

example, a logistic link function to model probabilities or a logistic growth curve to model growth. Letting the residual variance vary as a function of the mean allows modeling the heterogeneity arising from non-normal distributions. In addition, more complicated covariance structures allow the modeling of multivariate data including multinomial data.

In Stroup and Kachman (1994) an overview of GLMM was given. The objective of this paper is to illustrate the process of using a GLMM to analyze a set of data. The paper will give a brief description of the features of a GLMM. Next a general introduction to threshold models as a special case of GLMM will be given. Finally a data set consisting of number of lambs born will be analyzed.

2. Features of a Generalized Linear Mixed Model

In this section the components of a GLMM will be briefly discussed. A more detailed overview can be found in Stroup and Kachman (1994). Three components of a GLMM are the linear predictor ($\boldsymbol{\eta}$), inverse link function ($\boldsymbol{\mu} = \boldsymbol{h}(\boldsymbol{\eta})$), and conditional covariance matrix ($\boldsymbol{R}$).

The linear predictor, $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$, is used to model systematic effects arising from the independent variables where $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{u}$ is a vector of random effects, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known incidence matrices. Random effects are assumed to be normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{G}$. It is important to note that the linear predictor does not include a random residual effect. The covariance matrix $\boldsymbol{G}$ is a function of set of variance components $\boldsymbol{\sigma}$.

The inverse link function, $\boldsymbol{\mu} = \boldsymbol{h}(\boldsymbol{\eta})$, is used to model the effect of the linear predictor on the conditional mean of the dependent variables given the random effects. For many models the value of the i^{th} conditional mean will depend only on the i^{th} linear predictor. Sometimes there will not be a one to one mapping. For example with ordinal data there is typically a single linear predictor for each individual and m observations per individual, where m is the number of ordered categories.

The conditional covariance matrix, $\boldsymbol{R}$, is used to model variability in the dependent variables around the conditional mean given the random effects. For many models the covariance matrix is a diagonal matrix. That is the dependent variables are conditionally independent and therefore their covariance is zero. The dependent variables may be correlated due to spatial variability or measuring multiple dependent variables on each individual. Furthermore, The covariance matrix may be singular as happens with multinomial data.

2.1 Estimating Equations

The estimating equations for the fixed and random effects are

$$\begin{pmatrix} \boldsymbol{X'H'R^{-}HX} & \boldsymbol{X'H'R^{-}HZ} \\ \boldsymbol{Z'H'R^{-}HX} & \boldsymbol{Z'H'R^{-}HZ + G^{-1}} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'H'R^{-}y^{*}} \\ \boldsymbol{Z'H'R^{-}y^{*}} \end{pmatrix} \qquad (1)$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of the fixed effects, $\widehat{\boldsymbol{u}}$ is the predicted random effects, $\boldsymbol{H} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\eta}'$ is the matrix of partial derivatives of the conditional mean with respect to the linear predictor, and $\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{\mu} + \boldsymbol{H}\boldsymbol{\eta}$. Therefore, to estimate the fixed and random effects in a GLMM, three things will be needed: 1) the inverse link function, 2) the residual covariance matrix, and 3) the partial derivatives of the inverse link function.

2.2 Variance components

Often the covariance matrix will also need to be estimated. When the response variables are normally distributed REML estimates of the variance components are obtained by solving the following set of equations

$$\widehat{\boldsymbol{u}}'\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma_i}\widehat{\boldsymbol{u}} = \mathrm{tr}\left[\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma_i}\mathrm{E}(\widehat{\boldsymbol{u}}\widehat{\boldsymbol{u}})\right] \tag{2}$$

where $\mathrm{E}(\widehat{\boldsymbol{u}}\widehat{\boldsymbol{u}}) = \boldsymbol{G} - \mathrm{Var}(\boldsymbol{u} - \widehat{\boldsymbol{u}})$ (Harville, 1977).

When the response variables are not normally distributed or the inverse link function is not linear the variance of the predicted random effects is difficult to obtain. Estimates of the covariance components can be obtained from the following estimating equations

$$\widehat{\boldsymbol{u}}'\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma_i}\widehat{\boldsymbol{u}} = \mathrm{tr}\left[\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma_i}(\boldsymbol{G} - \boldsymbol{C}^{uu})\right]$$

where $\boldsymbol{C}^{uu}$ is the asymptotic covariance matrix of $\boldsymbol{u} - \widehat{\boldsymbol{u}}$ (Harville and Mee, 1984).

3. Threshold Model

In this section a quick introduction to threshold models will be given. The details will be left for the next section when a specific example will be examined in more detail. Ordinal categorical data arises in numerous settings. The number of lambs born, calving difficulty scores, and fabric ratings are examples ordinal categorical data. Threshold models provide a means of modeling ordinal categorical data.

For normally distributed response variables it is often reasonable to model the effect of a set of independent variables as additive. That is a change in one independent variable will raise or lower the average of the dependent variable by a given amount. The challenge is then to estimate how large this amount is.

When the data is categorical, it is often not reasonable to model the effect of a set of independent variable as additive. The effect of a set of independent variables is measured by their impact on the probability of falling into a certain category. The effect of a set of independent variables can be modeled as a cumulative effect. That is unless an individual falls above a certain threshold they will be observed in lower category. The effect of a change in one independent variable will be to raise or lower the threshold by a given amount. The challenge is then to estimate how large this amount is.

As with any GLMM three features need to be specified. First, the dependent variable is needed along with its conditional mean and variance. Second, the inverse link function is needed. Third, the linear model for the linear predictor is needed.

4. Example

In this section an example involving the number of lambs born will be used to examine the features of a GLMM in more detail. Number of lambs born to 276 ewes bred in 1980 and giving birth to at least one live lamb will be used to compare the reproductive rates of Finn, Suffolk, and Targhee breeds. The ewe records included the sire and dam of the ewe, the age of the ewe, along with the breed of the ewe and number of lambs born alive. The number of lambs born was categorized as singles, twins, and triplets. Ewes giving birth to at least three live lambs were pooled into the triplet category.

4.1 Dependent Variable

The dependent variables for ewe i is

$$\boldsymbol{y}_i = \begin{cases} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} & \text{if she gave birth to a single live lamb,} \\ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & \text{if she gave birth to two live lambs,} \\ \text{and} & \\ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} & \text{if she gave birth to at least three live lambs.} \end{cases}$$

Assuming conditional independence between ewes, the probability density function for $\boldsymbol{y}$ given the vector of random effects is

$$\Pr(\boldsymbol{y}|\boldsymbol{u}) = \prod_i \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \pi_{i3}^{y_{i3}} \tag{3}$$

where $\pi_{ij} = \Pr(y_{ij} = 1|\boldsymbol{u})$. Written another way,

$$\Pr(\boldsymbol{y}|\boldsymbol{u}) = \prod_i \exp[\sum_{j=1}^{3} y_{ij} \ln(\pi_{ij})],$$

it is clear that the conditional distribution of $\boldsymbol{y}$ is a member of the exponential family.

The conditional mean of $\boldsymbol{y}_i$ given $\boldsymbol{u}$ is

$$\mathrm{E}(\boldsymbol{y}_i|\boldsymbol{u}) = \boldsymbol{\pi}_i$$

where $\boldsymbol{\pi}_i = \{\pi_{ij}\}$. The conditional covariance matrix of $\boldsymbol{y}$ given $\boldsymbol{u}$ is

$$\boldsymbol{R} = \bigoplus_i \text{Diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i'$$

where $\oplus$ is the usual direct sum operator and $\text{Diag}(\cdot)$ is a diagonal matrix. A convenient generalized inverse of $\boldsymbol{R}$ is

$$\boldsymbol{R}^- = \bigoplus_i \text{Diag}(\boldsymbol{\pi}_i)^{-1}$$

(McCullagh and Nelder, 1989).

4.2 Linear Predictor

The linear predictor is used to model factors that have a systematic effect on the dependent variables. It will be assumed that underlying effect of each of the independent variables will be the same for the three dependent variables. That is, factors that have a large effect on pushing a ewe over the threshold for having twins will also have a large effect on pushing a ewe over the threshold for having triplets.

The linear predictor for ewe i (η_i) includes fixed effects for breed (Targhee, Suffolk, and Finnsheep), and age of the ewe (one, two, or at least three years old) along with random effects for sire of the ewe (100 sires) and dam of the ewe (244 dams). The random effects are assumed to be independently distributed as normal random variables with mean zero and variances σ_s^2 and σ_d^2 respectively. It is important to note that while each ewe has three dependent variables there is only a single linear predictor for each ewe.

4.3 Inverse Link Function

The inverse link function is used to model the effect of the linear predictor on the conditional probability of having singles, twins, or triplets. Two approaches to selecting an inverse link function are: 1) finding a process that could generate multinomial data and 2) looking at the properties the inverse link function has.

The first approach is to find a process that could generate multinomial data. Start with a hypothesized random variable X with mean η_i and variance σ^2. If the random variable falls below the first threshold then the ewe will have a single live lamb, if the random variable falls below the second threshold but above the first threshold then the ewe will have twins, otherwise the random variable will fall above the second threshold and she will have at least three live lambs. The conditional probability that a ewe i will have singles, twins, or triplets is the probability a random variable X with mean η_i falls between the corresponding thresholds. The first approach is illustrated in Figure 1.

Many choices exist for the distribution of the underlying random variable including the normal and logistic distributions. The normal distribution was selected. The normal distribution is consistent with our choice for the distribution of the random

effects. In addition the normal distribution will simplify our interpretation as will be seen in the results section. Without loss of generality the variance of our underlying random variable will be set to one. The idea of a process that could have generated the multinomial data helps in formulating a model. However, it does focus attention away from the effect of the independent variables and onto a hypothesized underlying random variable.

The second approach is to start with the properties the inverse link functions should have. Three properties the inverse link function should have are: 1) the mean should fall between zero and one, 2) singles under unfavorable conditions the probability should be close to one, and 3) as conditions improve the probability should decrease to zero where larger η_i denote more favorable conditions. These properties describe a cumulative distribution function $F(\tau_1|\eta_i)$ as a function of a location parameter η_i. For triples we expect the opposite relationship, which is as conditions improve the probability should increase to one. For twins the probability of having twins should be highest for moderate levels of η_i. Last, the probability of having singles, twins, or triples should sum to one. The inverse link functions for singles, twins, and triples are summarized in Figure 2.

As with the first approach many choices exist for the distribution selected. Looking at the properties of the inverse link function does add a level of abstraction. However, attention is focused on the effect of the independent variables as opposed to a hypothesized underlying random variable.

The inverse link function selected was

$$\boldsymbol{\pi}_i = \boldsymbol{\mu}_i = \begin{pmatrix} \Phi(\tau_1 - \eta_i) \\ \Phi(\tau_2 - \eta_i) - \Phi(\tau_1 - \eta_i) \\ 1 - \Phi(\tau_2 - \eta_i) \end{pmatrix} \tag{4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

To complete the estimation equations for the fixed random effects the partial derivatives of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\eta}$ is needed. The partial derivatives are

$$\boldsymbol{H} = \bigoplus_i \boldsymbol{H}_i$$

where

$$\boldsymbol{H}_i = \begin{pmatrix} -\phi(\tau_1 - \eta_i) \\ \phi(\tau_1 - \eta_i) - \phi(\tau_2 - \eta_i) \\ \phi(\tau_2 - \eta_i) \end{pmatrix} \tag{5}$$

an $\phi(\cdot)$ is the standard normal density function.

Estimates of the thresholds will also be needed and can be obtained from the estimating equations

$$(\boldsymbol{T}'\boldsymbol{R}^-\boldsymbol{T})\widehat{\boldsymbol{\tau}} = \boldsymbol{T}'\boldsymbol{R}^-(\boldsymbol{y} - \widehat{\boldsymbol{\mu}} + \boldsymbol{T}\widehat{\boldsymbol{\tau}}) \tag{6}$$

where $\boldsymbol{T} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\tau}'$. The partial derivatives are

$$\boldsymbol{T} = \{\boldsymbol{T}_i\}$$

where

$$\boldsymbol{T}_i = \begin{pmatrix} \phi(\tau_1 - \eta_i) & 0 \\ -\phi(\tau_1 - \eta_i) & \phi(\tau_2 - \eta_i) \\ 0 & -\phi(\tau_2 - \eta_i) \end{pmatrix}. \tag{7}$$

The estimating equations (6) are equivalent to the estimating equations in Misztal, Gianola and Foulley (1989).

4.4 Computations

SAS macros based on PROC MIXED (SAS Institute Inc., 1992) for analyzing binary data are not sufficiently general for analyzing three ordered categories. Using a FORTRAN program for analyzing mixed models as a basis, a FORTRAN program for analyzing ordinal categorical data was developed. Changes in the mixed model program included: 1) using $\boldsymbol{H}_i'\boldsymbol{R}_i^-\boldsymbol{H}_i$ in place of $1/\sigma^2$, 2) using $\boldsymbol{H}_i'\boldsymbol{R}_i^-\boldsymbol{y}_i^*$ in place of y_i/σ^2, and 3) estimating the thresholds.

Starting with a set of estimates for the fixed effects, random effects, variance components and thresholds the program iteratively updates the estimates. The fixed effects, random effects, and variance components are updated each iterate. The threshold estimates are updated after every five iterations.

Each iterate starts by building the estimating equations (1) for the fixed and random effects, and the estimating equations (6) for the thresholds. For each observation $\boldsymbol{H}_i'\boldsymbol{R}_i^-\boldsymbol{H}_i$, $\boldsymbol{H}_i'\boldsymbol{R}_i^-\boldsymbol{y}_i^*$, $\boldsymbol{T}_i'\boldsymbol{R}_i^-\boldsymbol{T}_i$, and $\boldsymbol{T}_i'\boldsymbol{R}_i^-(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i + \boldsymbol{T}_i\widehat{\boldsymbol{\tau}}_i)$ are obtained and used to update the estimating equations. After the equations have been built estimates of the fixed and random effects are obtained by solving (1). After updating the estimates of the fixed and random effects estimates of the variance components are obtained by using (2). Every fifth iteration estimates for the thresholds are obtained by solving (6).

5. Results

An initial analysis was run which included separate effects for the nine Breed–Age combinations. Tests using Wald statistics for main effects and the two-way interaction are presented in Table 1. There were significant differences among the three breeds and among three ages ($P < .001$). However, the interaction between breed and age was not significant ($P = .222$). The model used for the remaining analyses did not include the interaction between breed and age. Estimates of the thresholds, variance components are in Table 2. Using a Wald statistic breed and age effects were highly significant ($P< .001$).

Recall that ewes with a linear predictor below the lower threshold have at least a 50% chance of having singles and that ewes above the upper threshold have at least a 50% chance of having triplets. The Finnsheep breed was the only breed estimated to have over a majority of multiple births for all three age groups. For older ewes the

Finnsheep are expected to have three or more live lambs the majority of the time. The Targhee and Suffolk breeds had the lowest proportion of multiples. The Suffolk breed tended to have a larger proportion of multiples births compared to the Targhee (P< .1). Age of the ewe also played a large role, with the proportion of multiple live lambs increasing for older ewes.

The effects of the independent variables on the observed scale are obtained using the inverse link function. The probability that a one year old Suffolk ewe will give birth to a single live lamb is

$$\mathrm{E}\left[\Phi(\tau_1 - age_1 - breed_s - sire - dam)\right] = \Phi\left(\frac{\tau_1 - age_1 - breed_s}{\sqrt{1 + \sigma_s^2 + \sigma_d^2}}\right)$$

where age_1 is the age effect for a one year old ewe, $breed_s$ is the breed effect for a Suffolk ewe, $sire$ is the random effect of a Suffolk sire, and dam is the random effect of a Suffolk dam. An estimate is then obtained by replacing the unknown parameters with their estimates yielding

$$\Phi\left(\frac{-.29 + .92 + .31}{\sqrt{1 + .014 + .036}}\right)$$

or 82%. Estimates for probability of having singles, twin, and triplets for the nine breed×age groups are given in Table 2 along with the sample proportions.

6. Summary

Generalized linear mixed models provide a framework that brings together linear mixed models and generalized linear models. The addition of a link function or an inverse link function to linear mixed models allows the researcher to model the mean of a response variable separately from residual variance of a response variable. The addition of random effects to generalized linear models allows the researcher to model experimental designs such as split plots.

Besides modeling underlying systematic effects, a researcher needs to select an appropriate distribution for the response variable and an inverse link function. After taking partial derivatives of the inverse link function estimating equations for the fixed and random effects can be readily obtained. Because of the similarity between the estimating equations for a generalized linear mixed model and the mixed model equations, mixed model software can be modified to analyze generalized linear mixed models. Currently software with the flexibility of PROC MIXED (SAS Institute Inc., 1992) is not generally available for the analysis of generalized linear mixed models.

Many questions still need to be addressed. When there are additional parameters to be estimated how should uncertainty in their value be incorporated in the analysis? How well do these tests and estimators perform in small samples? How well do different designs perform? How sensitive are results to model miss-specification?

7. References

Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* 88, 9–25.

Harville, David A. (1977), Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association* 72, 320–338.

Harville, David A. and Mee, Robert W. (1984), A Mixed-Model Procedure for Analyzing Ordered Categorical Data, *Biometrics* 40, 393–408.

McCullagh, P. and Nelder, J. A. (1989), in *Generalized Linear Models* (second), Monographs on Statistics and Applied Probability, Chapman & Hall, New York.

Misztal, I., Gianola, D. and Foulley, J. L. (1989), Computing Aspects of a Nonlinear Method of Sire Evaluation for Categorical Data, *J. Dairy Sci.* 72, 1557–1568.

SAS Institute Inc. (1992), The Mixed Procedure, in *SAS Technical Report P-229 SAS/STAT Software: Release 6.07*, Cary, NC, 287–368.

Stroup, Walter W. and Kachman, Stephen D. (1994), Generalized Linear Mixed Models: An Overview, *Sixth Annual Kansas State University Conference on Applied Statisics in Agriculture*, In press.

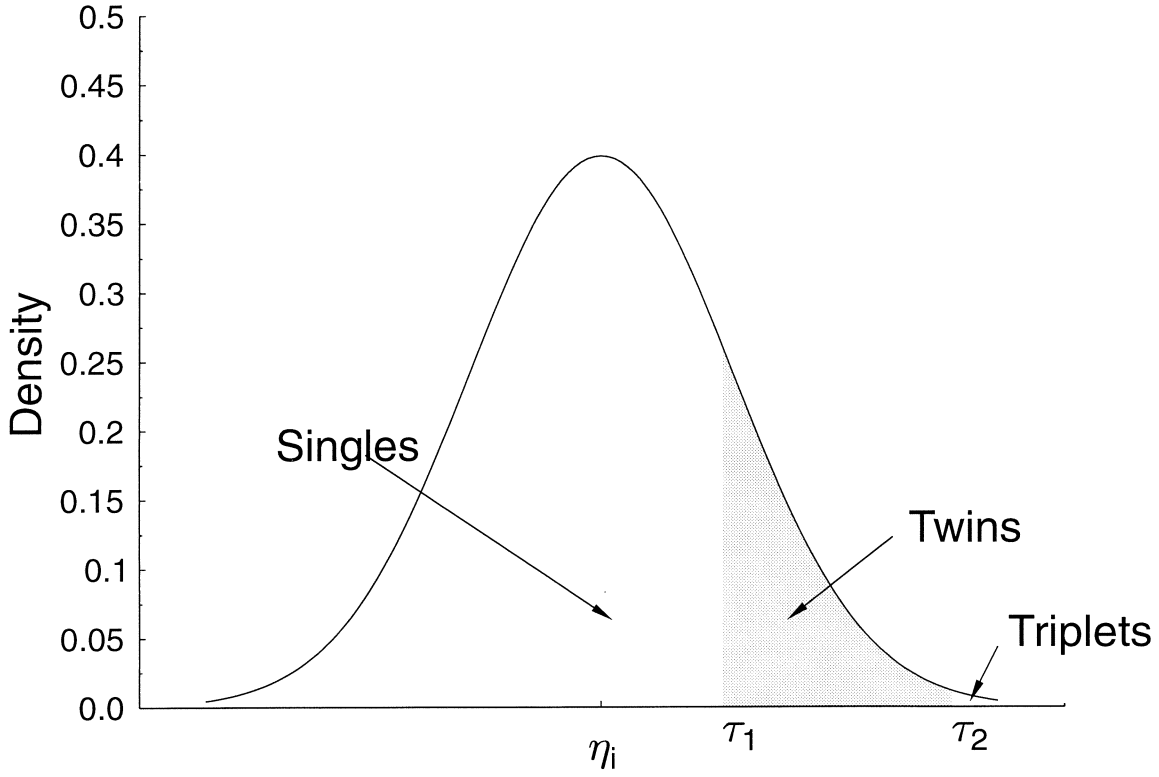

Figure 1: Probability that ewe i has twins. The probability is calculated by finding the probability that random variable X with mean η_i falls between the first threshold τ_1 and the second threshold τ_2.

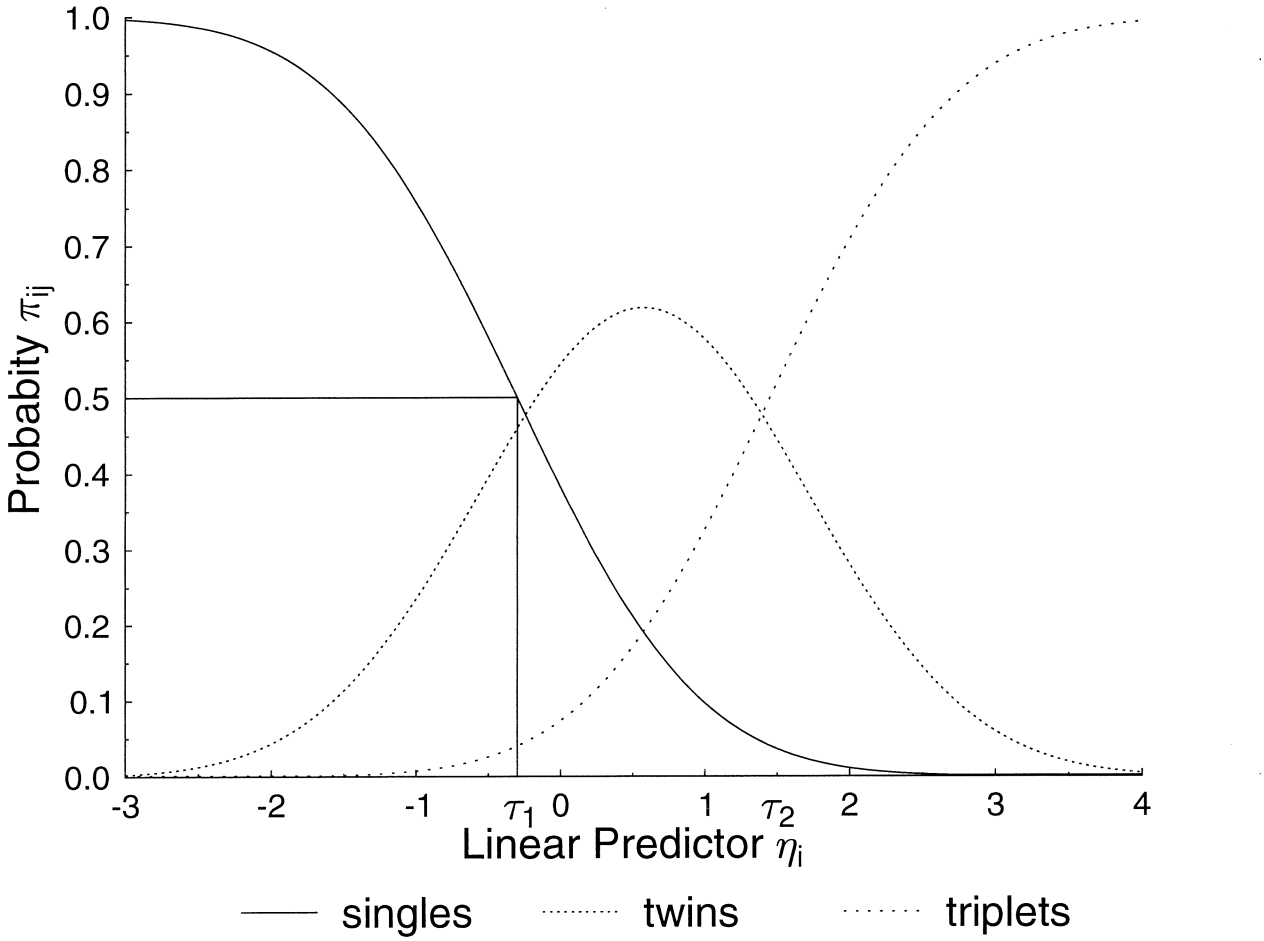

Figure 2: Inverse link functions for ewe i where η_i is the linear predictor for ewe i and π_{ij} is the probability the ewe i has j live lambs given the random effects. Thresholds τ_1 and τ_2 denote the 50% probability cutoffs for singles and triples respectively.

Table 1: Test of fixed effects for model including an interaction between age and breed.

Source	df	χ^2	p-value
Breed	2	80.26	$< .001$
Age	2	41.89	$< .001$
Breed$\times$Age	4	5.71	.222

Table 2: Estimates of thresholds, variance components, and fixed effects for number of lambs born alive.

Thresholds			
Lower	$-.29$	Upper	1.48

Variance Components			
Sire	.014	Dam	.036

Breed effects		Age effects	
Targhee	$-.71$	1 year	$-.92$
Suffolk	$-.31$	2 year	.19
Finnsheep	1.01	$\geq$3 year	.73

			Estimator					
			Sample			GLMM		
			Number Born			Number Born		
Breed	Age	n	1	2	≥ 3	1	2	≥ 3
Targhee								
	1	7	100	0	0	91	9	0
	2	23	65	35	0	59	39	3
	≥ 3	40	28	68	5	38	54	8
Suffolk								
	1	8	75	25	0	82	18	0
	2	13	54	46	0	43	51	6
	≥ 3	28	21	61	18	24	61	15
Finnsheep								
	1	65	32	62	6	36	56	9
	2	33	9	36	54	7	54	39
	≥ 3	59	7	36	58	2	38	60

Table 3: Estimated and sample percentages of one, two, or at least three live births for Targhee, Suffolk, and Finnsheep ewes at one, two, or at least three years old.