Generalized Linear Mixed Models
An Introduction for Tree Breeders and Pathologists

Fikret Isik, PhD
Quantitative Forest Geneticist,
Cooperative Tree Improvement Program, North Carolina State University

Statistic Session class notes
Fourth International Workshop on the Genetics of Host-Parasite Interactions in Forestry
July 31 – August 5, 2011. Eugene, Oregon, USA.

Table of Contents

Introduction 2
Linear Models 2
  Linear regression example 4
Linear Mixed Model 8
Generalized Linear Models 11
  Binary Data Example – Disease incidence probability 12
  Count Data Example – Number of trees infected 14
Generalized Linear Mixed Model 15
Overdispersion in Binomial and Poisson Regression Models 18
Example1: Binomial Counts in Randomized Blocks 19
  Analysis as a GLM 21
  Analysis as GLMM – Random block effects 23
  Analysis with Smooth Spatial Trends 26
  GLMM with ASReml 29
  Spatial R structure with ASReml 30
Example 2: Binary response variable with genetic effects 32
References 38

Introduction

Generalized Linear Mixed Models (GLMM) have attracted considerable attention over the last years. The word "Generalized" refers to non-normal distributions for the response variable, and the word "Mixed" refers to random effects in addition to the usual fixed effects of regression analysis. With the development of modern statistical packages such as SAS, R, and ASReml, a large variety of statistical analyses are available to a larger audience. However, along with being able to handle more sophisticated models comes a responsibility on the part of the user to be informed on how these advanced tools work.

The objective of this workshop is to provide an introduction to generalized linear mixed models by first discussing some of the assumptions and deficiencies of statistical linear models in general, then giving examples of uses in common situations in the natural sciences.

The first section reviews linear models and regression analysis for simple and multiple variables. Two numerical examples are solved using the SAS REG software.

The second section presents linear mixed models by adding the random effects to the linear model. A simple numerical example is presented using the SAS MIXED Procedure.

The third (last) section introduces generalized linear models. Two illustrative examples of binary and count data are presented using the SAS GLIMMIX procedure and ASReml software.

Linear Models

Linear models (regression) are often used for modeling the relationship between a single variable *y*, called the *response* or *dependent* variable, and one or more *predictor*, *independent* or *explanatory* variables, $X_1,\ldots,X_p$. When $p=1$, it is called simple regression but when $p>1$ it is called multiple regression.

Regression analysis can be used to assess the relationship between explanatory variables on the response variable. It is also a useful tool to predict future observations or merely describe the structure of the data.

To start with a simple example, suppose that *y* is the weight of trees, the predictors are the height (X_1), and the age of the trees (X_2). Typically the data will be available in the form of an array like the following

$$\begin{array}{ccc} y_1 & x_{11} & x_{12} \\ y_2 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} \end{array}$$

where y_i is the observation of the i-th tree and *n* is the number of observations.

There is an infinite number of ways to model the relationship between the response and the explanatory variables. However, to keep it simple, the relationship can be modeled through a linear function in the parameters as follows

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where β_i for $i = 0, 1, 2$ are unknown parameters and ε is the error term. Thus, the problem is reduced to the estimation of three parameters.

Notice that in a linear model the parameters enter linearly, but the predictors do not necessarily have to be linear. For instance, consider the following two functions

$$y = \beta_0 + \beta_1 \exp(X_1) + \beta_2 \log(X_2) + \varepsilon$$

$$y = \beta_0 + X_1^{\beta_1} + X_2 \exp(\beta_2) + \varepsilon$$

The first one is linear in the parameters, but the second one is not.

Using matrix representation, the regression equation for the above example can be written as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_0, \dots, \varepsilon_n)^T$, and the design matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

The estimation of $\boldsymbol{\beta}$ can be carried out using the least square approach. That is, we define $\widehat{\boldsymbol{\beta}}$ as the best estimate of $\boldsymbol{\beta}$ in the sense that minimizes the sum of the squared errors.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X\beta})^T(\mathbf{y} - \mathbf{X\beta})$$

Differentiating with respect to $\boldsymbol{\beta}$ and setting equal to zero, it can be shown that $\widehat{\boldsymbol{\beta}}$ satisfies the normal equations

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathbf{T}}\mathbf{y}$$

Thus, provided that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

So far, we have not assumed any distributional form for the errors $\boldsymbol{\varepsilon}$. The usual assumption is that the errors are normally distributed and in practice this is often, although not always, a reasonable assumption.

If we assume that the errors are independent and identically normally distributed with mean 0 and variance σ^2, that is to say $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$, then the expectations of observations are

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

and expectations of parameters are

$$\widehat{\boldsymbol{\beta}} \sim N(\,\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2\,)$$

Linear regression example

We would like explore linear dependency between height and weight. The linear model can be run using SAS GLM or REG procedures.

```
title 'Simple Linear Regression';
data A;
   input Height Weight Age @@;
 datalines;
  69.0 112.5 14  56.5  84.0 13 65.3  98.0 13
  62.8 102.5 14  63.5 102.5 14 57.3  83.0 12
  59.8  84.5 12  62.5 112.5 15 62.5  84.0 13
  59.0  99.5 12  51.3  50.5 11 64.3  90.0 14
  56.3  77.0 12  66.5 112.0 15 72.0 150.0 16
  64.8 128.0 12  67.0 133.0 15 57.5  85.0 11
  66.5 112.0 15
   ;

/* Create a unique age for each */

data B (drop =i ); set A;
   do i = 1 to 1 ;
      b = ranuni (i) ;
      output ;
   end;
run;

data B; set B ;
   Age=round(sum(year,b),.001) ;
run;

/* Simple linear regression */
```

```
ods graphics on;
ods listing style=statistical sge=on;
proc reg data=A ;
    model weight = height / clm cli;
run; quit ;
```

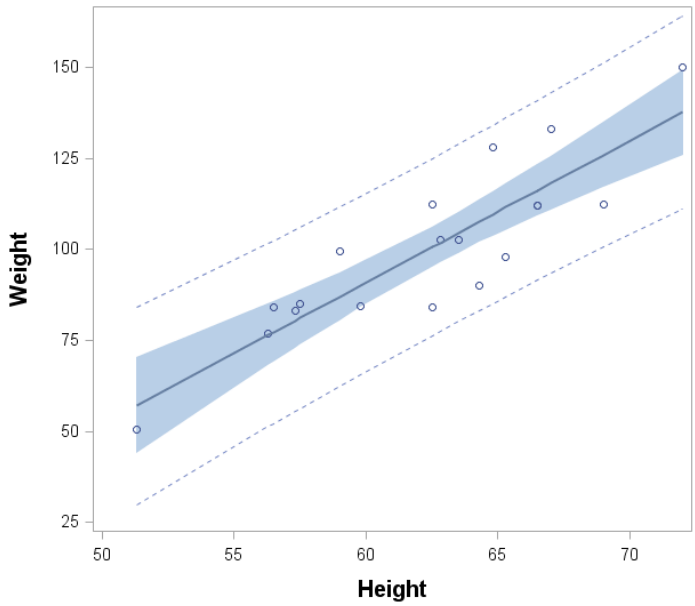


Figure1. *Scatter plot and regression line of Height on Weight*

The SAS output:

The REG Procedure
Dependent Variable: Weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

```
                          Parameter Estimates

                        Parameter    Standard
 Variable      DF        Estimate       Error     t Value     Pr > |t|

 Intercept      1      -143.02692    32.27459      -4.43       0.0004
 Height         1        3.89903     0.51609       7.55       <.0001
```

The intercept is $\hat{\beta}_0 = -143.027$ and the slope is $\widehat{\beta}_1 = 3.899$

Multiple linear regression

Now consider that we also want to include tree *age* as another predictor. That is, we want to assess the relationship between *Age* and *Weight* while keeping the effect of *Height* constant. In simple regressions, a line is fit to the data, whereas in multiple regressions, a *p-th* dimensional plane is fit, where *p* is the number of explanatory variables. To visualize the data and the fitted values, a 3D plot is necessary.

```
/* Create 3d plot */

proc g3grid data=ws.a out=a;
   grid age*height=weight  /  spline   ;
run;

goptions device=png xpixels=720 ypixels=600 noborder
gunit=pct htitle=3 htext=2 reset=all ;

ODS LISTING CLOSE;
proc g3d data=A;
   plot age*height=weight /grid  ctop=red
 cbottom=blue caxis=black  ;
run; quit;
```

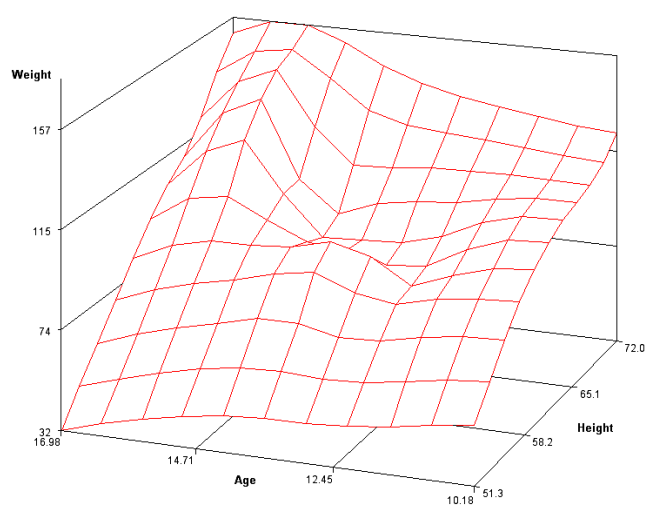


Figure2. *Regression plane of weight, age, and height*

In this case, we can use the multiple linear regression approach. So, we will need to estimate three parameters from the data. In SAS, this can be easily done by adding the following lines to the code shown above.

```
/* Multiple linear regression */

     ods html image_dpi=300;  * Image resolution ;
     ods graphics on;
     ods listing style=statistical sge=on; * Graphic type;

     proc reg data=ws.A ;
         model weight = height age;
     run; quit;
```

The SAS output is

The REG Procedure
Dependent Variable: Weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7574.69595	3787.34798	34.41	<.0001
Error	16	1761.04089	110.06506		
Corrected Total	18	9335.73684			

Root MSE	10.49119	R-Square	0.8114
Dependent Mean	100.02632	Adj R-Sq	0.7878
Coeff Var	10.48843		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-155.29024	30.87234	-5.03	0.0001
Height	1	3.49864	0.52808	6.63	<.0001
Age	1	2.68549	1.44255	1.86	0.0811

Following with the notation, the estimated parameters of the multiple linear regression are $\hat{\beta}_0 = -155.29$, $\hat{\beta}_1 = 3.498$ and $\hat{\beta}_2 = 2.685$. Thus, the relationship among variables can be expressed as: $Weight = -155.29 + 3.498\ Height + 2.685\ Age$

Linear Mixed Model

A linear mixed model is a statistical model containing both fixed effects and random effects. These models are widely used in the biological and social sciences. In matrix notation, linear mixed models can be represented as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z\gamma} + \boldsymbol{\varepsilon}$$

where:
y is the *n x 1* vector of observations,
β is a *p x 1* vector of fixed effects,
γ is a *q x 1* vector of random effects,
ε is a *n x 1* vector of random error terms,
X is the *n x p* design matrix for the fixed effects relating observations **y** to **β**,
Z is the *n x q* design matrix for the random effects relating observations **y** to **γ**.

We will assume that **γ** and **ε** are uncorrelated random variables with zero means and covariance matrices **G** and **R**, respectively.

$$E[\boldsymbol{\gamma}] = \mathbf{0}, \qquad Var[\boldsymbol{\gamma}] = \mathbf{G}$$

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}, \qquad Var[\boldsymbol{\varepsilon}] = \mathbf{R}$$

$$cov(\boldsymbol{\varepsilon}, \boldsymbol{\gamma}) = \mathbf{0}$$

Thus, the expectation and variance (**V**) of the observation vector **y** are given by:

$$E[\mathbf{y}] = \mathbf{X\beta}$$

$$Var[\mathbf{y}] = \mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$$

Understanding the **V** matrix is a very important component of working with mixed models since it contains both sources of random variation and defines how these models differ from computations with Ordinary Least Squares (OLS).

If you only have random effects models (such as a randomized block design) the **G** matrix is the primary focus. On the other hand, for repeated measures or for or spatial analysis, the **R** matrix is relevant.

If we also assume the random terms are normally distributed as:

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$$

Then, the observation vector will be normally distributed $\mathbf{y} \sim N(\mathbf{X\beta}, \mathbf{V})$.

For the general linear mixed model described above, the Henderson's mixed model equations (MME) can be used to find $\widehat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, the best linear unbiased estimator (BLUE) of **β,** and the best linear unbiased predictor (BLUP) of **γ**, respectively.

$$\begin{pmatrix} \mathbf{X^T R^{-1} X} & \mathbf{X^T R^{-1} Z} \\ \mathbf{Z^T R^{-1} X} & \mathbf{Z^T R^{-1} Z + G^{-1}} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{X^T R^{-1} y} \\ \mathbf{Z^T R^{-1} y} \end{pmatrix}$$

The solutions can also be written as:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X^T V^{-1} X})^{-1} \mathbf{X^T V^{-1} y}$$

$$\hat{\boldsymbol{\gamma}} = \mathbf{G Z^T V^{-1} (y - X\widehat{\beta})}$$

If the **G** and **R** matrices are known, generalized least squares can estimate any linear combination of the fixed effects **β**. However, as usually is the case these matrices are not known, so a complex iterative algorithm for fitting linear models must be used to estimate them.

Consider the following example. Suppose that we have collected data on the growth of different trees measured in two different locations. We can assume that the trees come from a large population, which is a reasonable assumption, and therefore, we will treat them as random.

Tree (*t*)	Location (*l*)	Height (*y*)
1	1	87
2	2	84
3	2	75
4	1	90
5	2	79

The linear mixed model can be expressed in matrix notation as follows

$$\begin{bmatrix} 87 \\ 84 \\ 75 \\ 90 \\ 79 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

which has the matrix form $\mathbf{Y} = \mathbf{X\beta + Z\gamma + \varepsilon}$**.** Notice that we are treating location as fixed effects. Matrices **X** and **Z** relate phenotypic observations to location and random tree effects.

We need to make some assumptions for the variance components of the model. In this case, we will assume that trees are independent of each other, and the errors are independent. So, the structure of the variance matrices can be expressed as:

$$\mathbf{R} = \mathbf{I}_{\mathrm{n}} \sigma_{\varepsilon}^2 \qquad \mathbf{G} = \mathbf{I}_{\mathrm{n}} \sigma_{\gamma}^2 \qquad \mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$$

where $\mathbf{I}_{\mathrm{n}}$represents the *n x n* identity matrix. Also, we assume that:

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G}) \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}) \quad \mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

We can compute the solutions using R software or any other software that uses matrix algebra.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} = \begin{bmatrix} 88.50 \\ 79.33 \end{bmatrix}$$

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \end{bmatrix} = \begin{bmatrix} -1.0067 \\ 3.1320 \\ -2.9083 \\ 1.0067 \\ -0.2237 \end{bmatrix}$$

Below is a sample R code to compute the above solution.

```
# Linear Mixed Model Example y = X*B + Z*U + e
# Y, observations
Y=c(87, 84, 75, 90, 79)

# X, design matrix for the fixed effects
X=matrix(c(1,0,0,1,0,0,1,1,0,1),5,2)

# Z, design matrix for the random effects
Z=diag(5); #identity matrix of size 5x5

# G = var(U)
Su2=100; # random effect variance
G=diag(5)*Su2;

# R = var(e)
Se2=49; # error variance
R=diag(5)*Se2;

# V=var(Y)
V=Z%*%G%*%t(Z) + R

# Solutions
Vi= solve(V)
B = solve( t(X)%*%Vi%*%X )%*%t(X)%*%Vi%*%Y;
U = G%*%t(Z)%*%Vi%*%(Y - X%*%B);

# Print out solutions
print (U, quote=T, row.names=F)

# Or produce histogram of BLUP values
hist(U, col="lightblue")
```

Generalized Linear Models

Recall that linear models, as its name states, assumes that

- the relationship between the dependent variable and the fixed effects can be modeled through a linear function,
- the variance is not a function of the mean, and
- the random terms follow a normal distribution

In some situations, any or even all these assumptions may be violated.

The generalized linear model (GLM) is an extension of ordinary least squares regression. The GLM generalizes linear regression by allowing the linear model to be related to the response variable ***via a link function*** and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

In a GLM, each outcome of the dependent variables, **y**, is assumed to be generated from a particular distribution in the exponential family. The most common distributions from this family are Binomial, Poisson, and Normal.

The mean, $\boldsymbol{\mu}$, of the distribution depends on the independent variables, **X**, through the inverse link function (g^{-1}).

$$E(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X\beta}) = g^{-1}(\boldsymbol{\eta})$$

where E(**y**) is the expected value of **y**; $\boldsymbol{\eta} = \mathbf{X\beta}$ is called the *linear predictor*, a linear combination of unknown parameters, **β**, and *g* is the link function.

In this framework, the variance **V** is typically a function of the mean:

$$Var(\mathbf{Y}) = V(\boldsymbol{\mu}) = V\left(g^{-1}(\mathbf{X\beta})\right) = V\left(g^{-1}(\boldsymbol{\eta})\right)$$

As an example of generalized linear model, consider a regression analysis where the response variable follows a Bernoulli distribution (the outcome is 1 or 0 with probabilities *p* and *1-p*, respectively), and we want to estimate the probability *p* of some event of interest.

Let's focus on the linear model assumptions that may not be reasonable for a binary response variable.

- First, note that the assumption of normal distribution of the response variable is clearly violated.

- Second, the assumption of constant variance and independence between mean and variance are not satisfied. For instance, a predicted probability *p* gives a variance *p(1-p)*, which is clearly mean dependent and, therefore, non-constant.

- The mean response is not necessarily a linear function of the effects. The linearity assumption does not hold in many practical situations of interest.

- Finally, perhaps the most important, in linear models, the predictions can take any value, whereas for some response variables we would like to have predictions between given boundaries (e.g., 0 and 1 for disease).

For all the above mentioned facts, linear models are not a suitable tool for analyzing binary traits. Generalized Linear Models address the deficiencies of linear models.

Consider the situation where individual seeds are laid on damp soil in different pots. In this experiment, the pots are kept at different temperatures *T* for a number of days *D*. After an arbitrary number of days, the seeds are inspected and the outcome *y=1* is recorded if it has germinated and *y=0* otherwise. The probability of germination *p* can be modeled through a linear function of the form:

$$\eta = \beta_0 + \beta_1 D + \beta_2 T$$

Where η is the linear predictor and β_0, β_1 and β_2 are parameters to be estimated. Note that there is nothing holding the linear predictor to be between 0 and 1.

In GLM, however, the probability is restricted to the interval (0,1) through the inverse link function:

$$p = \frac{1}{1 + e^{-\eta}} = g^{-1}(\eta)$$

It is worth noting that ***p*** is the expected value of ***y*** for a binomial distribution. Also notice the non-linear relationship between the outcome *p* and the linear predictor η is modeled by the inverse link function. In this particular case, the link function is the logistic link function or logit:

$$\eta = \log\left(\frac{p}{1-p}\right) = g(p)$$

Binary Data Example – Disease incidence probability

Here is a SAS code to reproduce data.

```
goptions reset=all cback=white htitle=15pt htext=15pt;
data pgerm;
   do T=0 to 40 by 0.5;
      do D=0 to 15 by 0.5;
         p=1/(1+exp(8 - 0.19*T - 0.37*D));
         output;
      end;
   end;
run;
```

SAS code to produce 3D plot of Figure 4

```
/* Designate a GIF file for the G3D output. */
filename anim 'c:\users\fisik\pond.gif';

/** animation. **/
goption reset dev=gifanim gsfmode=replace noborder htext=1.4
gsfname=anim xpixels=640 ypixels=480
iteration=0 delay=5
gepilog='3B'x /* add a termination char to the end of the GIF file */
disposal=background;

proc g3d data=pgerm;
   plot D*T=p / tilt=60 grid rotate=0 to 350 by 10
                xticknum=4 yticknum=5 zticknum=5
                zmin=0 zmax=1
                caxis = black ctop=red cbottom=blue;
               label T='Temperature' D='Days' p='p' ;
  run; quit;
```

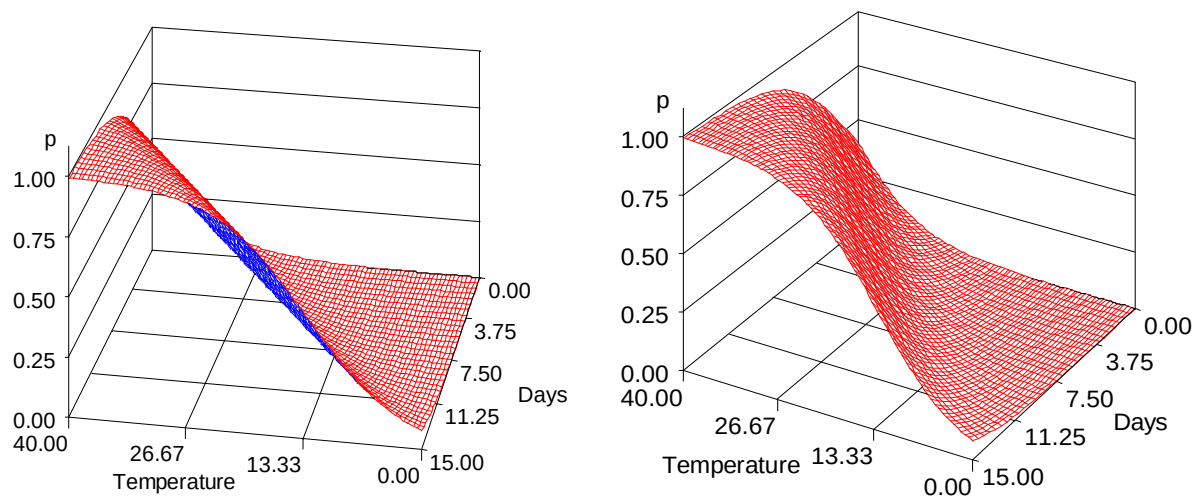


Figure 4: *Disease incidence probability as a function of Temperature and Day. The values of β_i for i = 0, 1, 2 have been chosen for illustrative purposes*

The inverse link function is, $\mathrm{p} = [1 + \exp(-\eta)]^{-1}$. Then, the linear predictor takes the form: $\eta = -8 + 0.19T + 0.37D$. Using the model we can estimate how many days are needed for the probability of disease incidence to exceed 80% at a given temperature. After some simple algebra it can be shown that at temperature 10 at least 20 days are needed to reach a probability of 0.80 germination.

The above example has no random effects so it is a generalized linear model (GLM), not a generalized mixed model (GLMM).

Count Data Example – Number of trees infected

We have previously considered an example where the outcome variable is numeric and binary. Often the outcome variable is numeric but in the form of counts. Sometimes it is a count of rare events such as the number of new cases of fusiform rust of loblolly pine occurring in a population over a certain period of time.

The Poisson probability function is appropriate for count data with no upper bound to the range, that is, *y* could be any non-negative integer. The Poisson *pmf* is described by the function:

$$\Pr(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

For such distributions, λ is the mean and the variance is also λ. This λ is often referred to as the Poisson "intensity" as it is the average event count.

As a hypothetical example, we can use data that relate the number of newly infected trees over a three-month period with its age, *A*, and height, *H*, within a population.

Our interest lies in modeling λ as a function of *age* (*A*) and *height* (*H*) for a given family and location. Using a linear model for λ could result in negative intensities, λ <0, which would make no sense. The natural link function $\boldsymbol{g(\lambda)}$ for a Poisson is the logarithm, so the model can be the following:

$$\eta = \beta_0 + \beta_1 A + \beta_2 H$$

Where η is the linear predictor and β_0, β_1, and β_2 are parameters to be estimated.

The link function and its inverse are given by:

$$\eta = \ln(\lambda) = g(\lambda)$$

$$\lambda = e^{\eta} = g^{-1}(\eta)$$

In Figure 5 the intensity λ as a function of *age* and *height* is plotted. Again, the values of β_i for *i = 0, 1, 2* have been chosen for illustrative purposes only. Then, the linear predictor takes the form: $\eta = -2 - 0.03A - 0.01H$.

Here is a SAS code to reproduce Figure 5.

```
goptions reset=all cback=white htitle=15pt htext=15pt;
data intensity;
   do A=0 to 50 by 0.5;
      do H=0 to 30 by 0.5;
         I=exp(-2 - 0.03*A - 0.01*H);
         output;
      end;
   end;
run;
```

```
proc g3d data=intensity;
title 'Intensity';
   plot A*H=I / rotate=160 tilt=80  grid
                xticknum=4 yticknum=3 zticknum=5
                zmin=0 zmax=0.15
                caxis = black ctop=blue cbottom=red;
               label A='Age (years)' H='Height (meters)' I='lambda';
  run; quit;
```

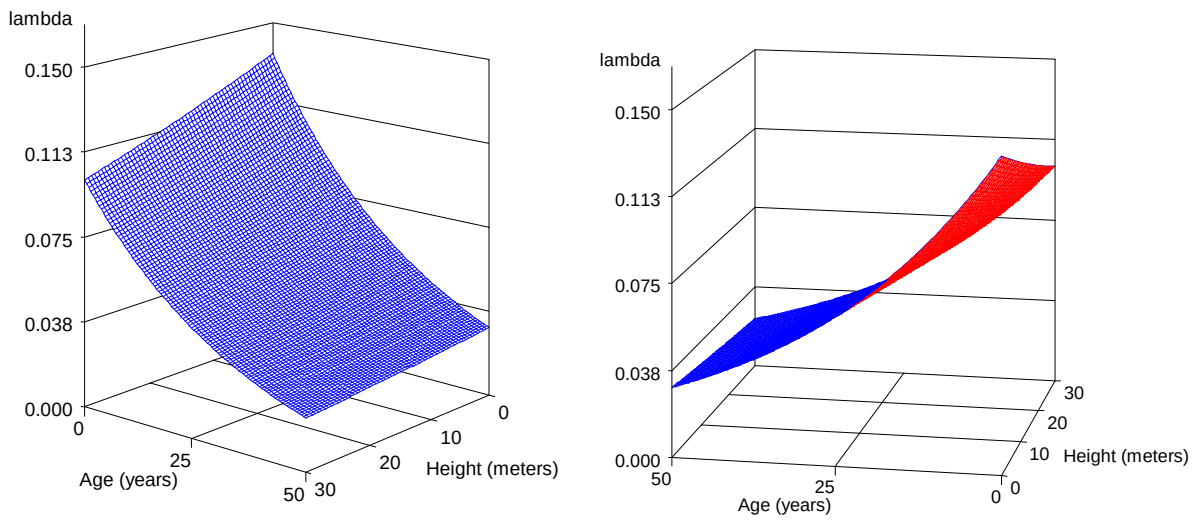


Figure 5: *Intensity as a function of Age and Height. The inverse link function is defined as* $\lambda = exp(\eta)$, *where* $\eta = -2 - 0.03A - 0.01H$.

Finally, notice that the above count data example does not include random effects, therefore, it is a generalized linear model, not a generalized linear mixed model. In the next section the generalized linear mixed model is presented.

Generalized Linear Mixed Model

A Generalized linear mixed models (GLMM) is an extension to the generalized linear model (GLM) in which the linear predictor contains random effects in addition to the fixed effects.

The expectations of the GLMM are:

$$E[\mathbf{y}|\mathbf{u}] = g^{-1}(\mathbf{X\beta} + \mathbf{Zu}) = g^{-1}(\boldsymbol{\eta})$$

Where

y represents the (*n x 1*) response vector,

X the (*n x p*) design matrix of rank *k* for the (*p x 1*) fixed effects **β** and

Z the (*n x q*) design matrix for the (*q x 1*) random effects **u**.

The random effects **u** are assumed to be normally distributed with mean **0** and variance matrix **G**, that is to say $u \sim N(\mathbf{0}, \mathbf{G})$.

$$E[\boldsymbol{u}] = \mathbf{0}, \qquad Var[\mathbf{u}] = \mathbf{G}$$

The fixed and random effects are combined to form a linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

The model for the vector of observations **y** is obtained by adding a vector of residuals, **ε**, as follows:

$$\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

The relationship between the linear predictor and the vector of observations is modeled as

$$\mathbf{y}|\mathbf{u} \sim (g^{-1}(\boldsymbol{\eta}), \mathbf{R})$$

The above notation denotes that the conditional distribution of **y** given **u** has mean $g^{-1}(\boldsymbol{\eta})$ and variance **R**. The conditional distribution of **y**|**u** is usually referred as the error distribution. Note that Instead of specifying a distribution for **y**, as in the case of a GLM, we now specify a distribution for the conditional response, **y**|**u**. This formulation is also known as the conditional model specification.

Last, the variance matrix of the observations is given by:

$$\mathrm{V}(\mathbf{y}) = \mathrm{E}[\mathrm{V}(\mathbf{y}|\mathbf{u})] + \mathrm{V}[\mathrm{E}(\mathbf{y}|\mathbf{u})] = \mathbf{A}^{\mathbf{1/2}}\mathbf{R}\mathbf{A}^{\mathbf{1/2}} + \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathbf{T}}$$

Where matrix **A** is a diagonal matrix that contains the variance functions of the model.

The class of generalized linear mixed models contains several important types of statistical models. For example,

• Linear models: no random effects, identity link function, and normal distribution
• Generalized linear models: no random effects present
• Linear mixed models: random effects, identity link function, and normal distribution

The generalized linear mixed models have been developed to address the deficiencies of linear mixed models.

There are many cases when the implied assumptions are not appropriate.

For instance, the linear mixed model assumes that the relationship between the mean of the dependent variable **y** and the fixed and random effects can be modeled through a linear function. This assumption is questionable, for example, in modeling disease incidence.

Another assumption of the linear mixed model is that the variance is not a function of the mean, and the random effects follow a normal distribution. The assumption of constant variance is violated when analyzing a zero/one trait, such as diseased (1) or not diseased (0). In this case, the response variable is Binomial. So, for a predicted disease incidence, the variance is $\mu(1-\mu)$, which is a function of the mean.

The assumption of normality is not valid for a binary trait. The outcome is a random variable that can only take two values, zero or one. In contrast, the normal distribution is a bell shaped curve that can take any real number.

Finally, the predictions from linear mixed models can take any value. Whereas predictions for a binary variable is bounded (0,1) or for a count variable, it cannot take negative values.

Probability, Odds and Odds Ratio

In order to understand the output and correctly interpret results, we need to be familiar with the probability, odds and odds ratio. Here is an example data and the P, ODDS and OR.

	Outcome		
	No	**Yes**	Sum
Sample1	30	70	100
Sample2	20	180	200
Sum	50	250	**300**

Sample1
Probability of a **Yes outcome** = 70/100 **(0.70)**
Probability of a **No outcome** = 30/100 **(0.30)**
The **odds** of Outcome in Sample1 Odds = p / (1-p) = 0.70 /0.30 = **2.3**
We expect 2.3 times as many occurrences as non-occurrences in Sample1.

Sample2
Probability of a **Yes outcome** = 180/200 **(0.90)**
Probability of a **No outcome** = 20/200 **(0.10)**
The **odds** of Outcome in Sample2 Odds = p / (1-p) = 0.90 /0.10 = **9**
We expect 9 times as many occurrences as non-occurrences in Sample2.

The odds ratio of Sample2 to Sample1 is OR = 9/2.3 = 3.86
The odds of having outcome (Yes) in Sample2 is almost 4 times those in Sample1.

Over-dispersion in Binomial and Poisson Regression Models

(Modified from http://rfd.uoregon.edu/files/rfd/StatisticalResources/gnmd08_overdisp.txt)

Over-dispersion results when the data appear more dispersed than is expected under some reference model. It may occur with count data analyzed with binomial or Poisson regression models, since the variance of both distributions is a function of the mean. That is,

$$Var[Y] = f(E[Y]) * \varphi$$

With both distributions the scale parameter phi is assigned a value of 1.

To understand what over-dispersion implies, first review the linear regression model (computed with ordinary least squares). Under the normal distribution data are never over-dispersed because the mean and variance are not related. The expectations of a linear model (**y = X'β + e**) are

$$\mathbf{e} \sim NID\ (0, \sigma^2_e)$$

The variance of the residuals (sigma^2) is assumed constant for all linear combinations of the covariates. This variance is estimated from the data and can assume any value greater than zero no matter what the mean value is. Thus, the response values are assumed to have constant variance:

$$Var(\mathbf{y}) = \sigma^2_e * 1$$

The normal errors and identity link function (linear regression) have the variance function Var(μ) =1. This variance is constant for all y_i.

For a generalized linear model:

$$g(\mu) = \mathbf{X'\beta}$$

where **μ**=E(y) and ***g*** is the link function. The variance of **y** is:

$$Var(y) = \varphi * V(\boldsymbol{\mu})$$

That is, the variance of an observation equals some constant **φ (the scale parameter)** times a function of the mean of y.

- For a binary variable y, the variance is multiplicative function of its mean: Var(y) = **μ** (1-**μ**)
- Under the Poisson distribution the variance of is the mean itself: Var(y) = E(y)= **μ**.

In either case, the observed counts have variances that are functions which depend on the value of the mean. That is, the variance of y depends on the expectation of y, which is estimated from the data.

When either model is fit under the assumption that the data were generated from a binomial distribution or by a Poisson process, the scale parameter, phi, is automatically set equal 1. That is why we see 1.00 for error variance in the ASReml output or in SAS GENMOD procedure output when we fit Poisson or Binomial distributions. The value 1 is not error variance, but it is a scale parameter and should not be used as variance component to calculate heritability for binomial distribution.

For binomial and Poisson regression models, the covariance matrix (and hence the standard errors of the parameter estimates) is estimated under the assumption that the chosen model is appropriate. More variation in the data may be present than is expected by the distributional assumption. This is called **over-dispersion** (also known as heterogeneity) which typically occurs when the observations are correlated or are collected from "clusters".

To identify possible over-dispersion in the data for a given model, divide the deviance by its degrees of freedom: this is called the **dispersion parameter**. If the deviance is reasonably "close" to the degrees of freedom (i.e., the scale parameter=1) then evidence of over-dispersion is lacking.

Dispersion parameter (or scaled deviance)=Deviance/DF

A scale parameter that is greater than 1 does not necessarily imply over-dispersion is present. This can also indicate other problems, such as an incorrectly specified model (omitted variables, interactions, or non-linear terms), an incorrectly specified functional form (an additive rather than a multiplicative model may be appropriate), as well as influential or outlying observations.

If you believe you have correctly specified the model and the scale estimate is greater than 1, then conclude your data are over-dispersed. You should be able to identify possible reasons why your data are over-dispersed. If you do not correct for over-dispersion, the estimates of the standard errors are too small which leads to biased inferences (i.e. you will observe smaller p-values than you should and thus make more Type I errors). As a result, confidence intervals will also be incorrect.

When you have the "correct" model, outliers are not a problem, and the scaled deviance is large, there are various choices for SAS procedures (GENMOD, GLIMMIX, NLMIXED) or for ASReml to correct for over-dispersion.

Example1: Binomial Counts in Randomized Blocks

(Modified from SAS GLIMMIX Help system)

In the context of spatial prediction in generalized linear models, Gotway and Stroup (1997) analyze data from an agronomic field trial. Researchers studied sixteen varieties (entries) of wheat for their resistance to infestation with the Hessian fly. They arranged the varieties in a randomized complete block design on an 8 × 8 grid. Each 4 × 4 quadrant of that arrangement

constitutes a block. The outcome of interest was the number of damaged plants (Y_{ij}) out of the total number of plants growing on the unit (n_{ij}). In other words, determine if there are significant differences between entries (Plant varieties). The two subscripts identify the block ($i = 1,...,4$) and the entry ($j = 1,...,16$).

The following SAS statements create the data set. The variables *lat* and *lng* denote the coordinate of an experimental unit on the 8 × 8 grid.

```
data HessianFly;
label Y = 'No. of damaged plants'
      n = 'No. of plants';
input block entry lat lng n Y @@;
datalines;
1 14 1 1 8 2     1 16 1 2 9 1
1 7 1 3 13 9     1 6 1 4 9 9
1 13 2 1 9 2     1 15 2 2 14 7
1 8 2 3 8 6      1 5 2 4 11 8
1 11 3 1 12 7    1 12 3 2 11 8
1 2 3 3 10 8     1 3 3 4 12 5
1 10 4 1 9 7     1 9 4 2 15 8
1 4 4 3 19 6     1 1 4 4 8 7
2 15 5 1 15 6    2 3 5 2 11 9
2 10 5 3 12 5    2 2 5 4 9 9
2 11 6 1 20 10   2 7 6 2 10 8
2 14 6 3 12 4    2 6 6 4 10 7
2 5 7 1 8 8      2 13 7 2 6 0
2 12 7 3 9 2     2 16 7 4 9 0
2 9 8 1 14 9     2 1 8 2 13 12
2 8 8 3 12 3     2 4 8 4 14 7
3 7 1 5 7 7      3 13 1 6 7 0
3 8 1 7 13 3     3 14 1 8 9 0
3 4 2 5 15 11    3 10 2 6 9 7
3 3 2 7 15 11    3 9 2 8 13 5
3 6 3 5 16 9     3 1 3 6 8 8
3 15 3 7 7 0     3 12 3 8 12 8
3 11 4 5 8 1     3 16 4 6 15 1
3 5 4 7 12 7     3 2 4 8 16 12
4 9 5 5 15 8     4 4 5 6 10 6
4 12 5 7 13 5    4 1 5 8 15 9
4 15 6 5 17 6    4 6 6 6 8 2
4 14 6 7 12 5    4 7 6 8 15 8
4 13 7 5 13 2    4 8 7 6 13 9
4 3 7 7 9 9      4 10 7 8 6 6
4 2 8 5 12 8     4 11 8 6 9 7
4 5 8 7 11 10    4 16 8 8 15 7
;
```

If infestations are independent among experimental units, and all plants within a unit have the same propensity of infestation, then the Y_{ij} are binomial random variables.

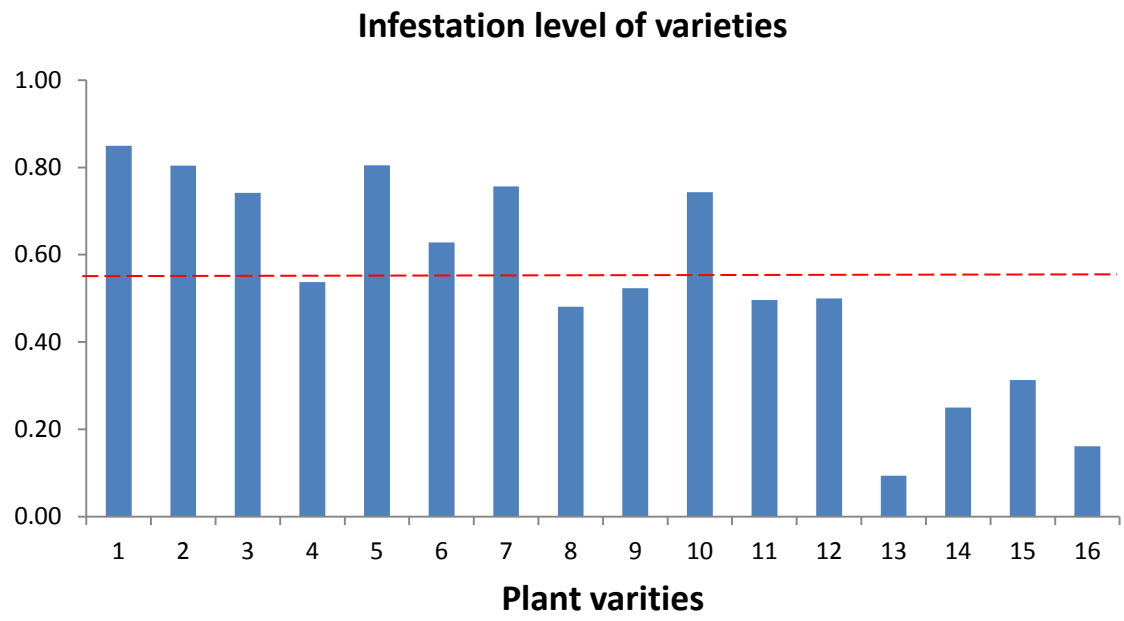


Figure 1. *Data visualization and summary is an important step before any statistical analysis. The chart shows large differences between varieties for infestation. The horizontal dashed line shows the overall mean (0.54) incidence.*

Analysis as a GLM

Let's consider first a standard generalized linear model for independent binomial counts. The SAS statements would be as follows:

```
proc glimmix data=HessianFly;
class block entry;
model y/n = block entry / solution;
run;
```

It may worth noting that the GLIMMIX procedure supports two kinds of syntax for the response variable. This example uses the *events/trials* syntax. The variable *y* represents the number of successes (events) out of n Bernoulli trials.

When the ***events/trials*** syntax is used, the GLIMMIX procedure automatically selects the binomial distribution as the response distribution. Once the distribution is determined, the procedure selects the ***link function*** for the model. The default link for binomial data is the ***logit link***. The above statement is equivalent of the following statements.

```
proc glimmix data=HessianFly ;
class block entry;
model y/n = block entry/solution dist=binomial link=logit ;
run;
```

Part of the output is given below:

The “Model Information” table describes the model and methods used in fitting the statistical model.

Model Information

Data Set	WORK.HESSIANFLY
Response Variable (Events)	Y
Response Variable (Trials)	n
Response Distribution	Binomial
Link Function	Logit
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Maximum Likelihood
Degrees of Freedom Method	Residual

The GLIMMIX procedure recognizes that this is a model for uncorrelated data (variance matrix is diagonal) and that parameters can be estimated by maximum likelihood.

The “Class Level Information” table lists the levels of the variables specified in the CLASS statement and the ordering of the levels.

Class Level Information

Class	Levels	Values
block	4	1 2 3 4
entry	16	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Number of Observations Read	64
Number of Observations Used	64
Number of Events	396
Number of Trials	736

The “Dimensions” table lists the size of relevant matrices.

Dimensions

Columns in X	21
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	64

Because of the absence of random effects in this model, there are no columns in the Z matrix. The 21 columns in the X matrix comprise the intercept, 4 columns for the block effect and 16 columns for the entry effect.

The “Fit Statistics” table lists information about the fitted model.

Fit Statistics

-2 Log Likelihood	265.69
AIC (smaller is better)	303.69
AICC (smaller is better)	320.97
BIC (smaller is better)	344.71
CAIC (smaller is better)	363.71
HQIC (smaller is better)	319.85
Pearson Chi-Square	106.74
Pearson Chi-Square / DF	2.37

The -2 Log Likelihood values are useful for comparing nested models, and the information criteria AIC, AICC, BIC, CAIC, and HQIC are useful for comparing non-nested models. On average, the ratio between the Pearson Chi-square statistic and its degrees of freedom should equal one in GLMs. Values larger than one are indicative of *over-dispersion*. With a ratio of 2.37, these data appear to exhibit more dispersion than expected under a binomial model with block and varietal effects.

The “Type III Tests of Fixed Effect” table displays significance tests for the two fixed effects in the model.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	3	45	1.42	0.2503
entry	15	45	6.96	<.0001

These tests are Wald-type tests, not likelihood ratio tests. The entry effect is clearly significant in this model with a p-value of < 0.0001, indicating that the 16 wheat varieties are not equally susceptible to damage by the Hessian fly.

Analysis as GLMM – Random block effects

There are several possible reasons for the *over-dispersion* noted in the “Fit Statistics” table (Pearson ratio = 2.37). The data may not follow a binomial distribution, one or more important effects may have not been accounted for in the model, or the data are positively correlated.

If important fixed effects have been omitted, then you might need to consider adding them to the model. Since this is a designed experiment, it is reasonable not to expect further effects apart from the block and entry effects that represent the treatment and error control design structure. The reasons for the over-dispersion must lie elsewhere. If over-dispersion stems from

correlations among the observations, then the model should be appropriately adjusted. The correlation can have multiple sources. First, it may not be the case that the plants within an experimental unit responded independently.

If the probability of infestation of a particular plant is altered by the infestation of a neighboring plant within the same unit, the infestation counts are not binomial, and a different probability model should be used. A second possible source of correlations is the lack of independence of experimental units. Even if treatments were assigned to units at random, they may not respond independently. Shared spatial soil effects, for example, may be the underlying factor. The following analyses take these spatial effects into account.

First, assume that the environmental effects operate at the scale of the blocks. By making the block effects random, the marginal responses will be correlated due to the fact that observations within a block share the same random effects. Observations from different blocks will remain uncorrelated, in the spirit of separate randomizations among the blocks.

The next set of SAS statements fits a **generalized linear mixed model (GLMM**) with random block effects:

```
proc glimmix data=HessianFly;
class block entry;
model y/n = entry / solution;
random block;
run;
```

Treating the block effects as random changes the estimates compared to a model with fixed block effects.

Selected tables of output are given below.

Model Information

Data Set	WORK.HESSIANFLY
Response Variable (Events)	Y
Response Variable (Trials)	n
Response Distribution	Binomial
Link Function	Logit
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

In the presence of random effects and a conditional binomial distribution, PROC GLIMMIX does not use maximum likelihood for estimation. Instead, the GLIMMIX procedure applies a restricted (residual) pseudo-likelihood algorithm.

The “Dimensions” table has changed from the previous model. The “Dimensions” table indicates that there is a single **G**-side parameter, the variance of the random block effect.

Dimensions

G-side Cov. Parameters	1
Columns in X	17
Columns in Z	4
Subjects (Blocks in V)	1
Max Obs per Subject	64

Note that although the block effect has four levels, only a single variance component is estimated. The Z matrix has four columns, however, corresponding to the four levels of the block effect. Because no SUBJECT= option is used in the RANDOM statement, the GLIMMIX procedure treats these data as having arisen from a single subject with 64 observations.

The "Optimization Information" table indicates that a Quasi-Newton method is used to solve the optimization problem. This is the default method for GLMM models.

Optimization Information

Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled
Starting From	Data

The "Fit Statistics" table shows information about the fit of the GLMM.

Fit Statistics

-2 Res Log Pseudo-Likelihood	182.21
Generalized Chi-Square	107.96
Gener. Chi-Square / DF	2.25

The generalized chi-square statistic measures the residual sum of squares in the final model and the ratio with its degrees of freedom is a measure of variability of the observation about the mean model. The over-dispersion parameter (2.25) is still larger than 1.

The variance of the random block effects in the following table is rather small. The random block model does not provide a suitable adjustment for dispersion.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error
block	0.01116	0.03116

Because the block variance component is small, the Type III test for the variety effect in is affected only very little compared to the standard GLM.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Entry	15	45	6.90	<.0001

In the experimental designs, researchers had row-column (longitude and latitude) values for each data to account for spatial scale micro-site variation.

Analysis with Smooth Spatial Trends

If the environmental effects operate on ***a spatial scale smaller than the block size***, the random block model does not provide a suitable adjustment. From the coarse layout of the experimental area, it is not surprising that random block effects alone do not account for the over-dispersion in the data. We need to add a multiplicative over-dispersion component in PROC GLIMMIX.

```
random _residual_;
```

Such over-dispersion components do not affect the parameter estimates, only their standard errors. A genuine random effect, on the other hand, affects both the parameter estimates and their standard errors.

```
Ods graphics on/noborder ;
 proc glimmix data=HessianFly PLOTS=All;
   class entry ;
   model y/n = entry / solution ddfm=contain;
   random _residual_ / subject=intercept type=sp(exp)(lng lat);
   lsmeans entry / plots=mean(sliceby=entry join);
 run;
```

The keyword _RESIDUAL_ in the RANDOM statement instructs the GLIMMIX procedure to model the **R** matrix. Here, **R** is to be modeled as an *exponential covariance structure matrix*. The SUBJECT=INTERCEPT option means that all observations are considered correlated. The block effects can be kept in the model. But correlated **R** structure sucks out all the variation and does not leave anything (zero variance) for Block effect to explain. We dropped the Block effects from the final model. PLOTS=All option produces model diagnostic plots.

Model Information

Data Set	WORK.HESSIANFLY
Response Variable (Events)	Y
Response Variable (Trials)	n
Response Distribution	Binomial
Link Function	Logit
Variance Function	Default
Variance Matrix Blocked By	Intercept
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

Dimensions

R-side Cov. Parameters	2
Columns in X	17
Columns in Z per Subject	0
Subjects (Blocks in V)	1
Max Obs per Subject	64

Fit Statistics

-2 Res Log Pseudo-Likelihood	158.85
Generalized Chi-Square	121.51
Gener. Chi-Square / DF	2.53

Covariance Parameter Estimates

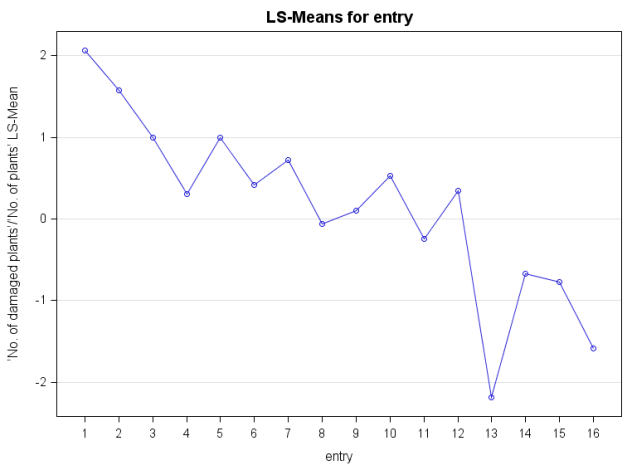
Cov Parm	Subject	Estimate	Standard Error
SP(EXP)	Intercept	0.9052	0.4404
Residual		2.5315	0.6974

The sill of the spatial process, the variance of the underlying residual effect, is estimated as 2.5315. The one third of practical range of a spatial process is 0.9052.

Type III Tests of Fixed Effects

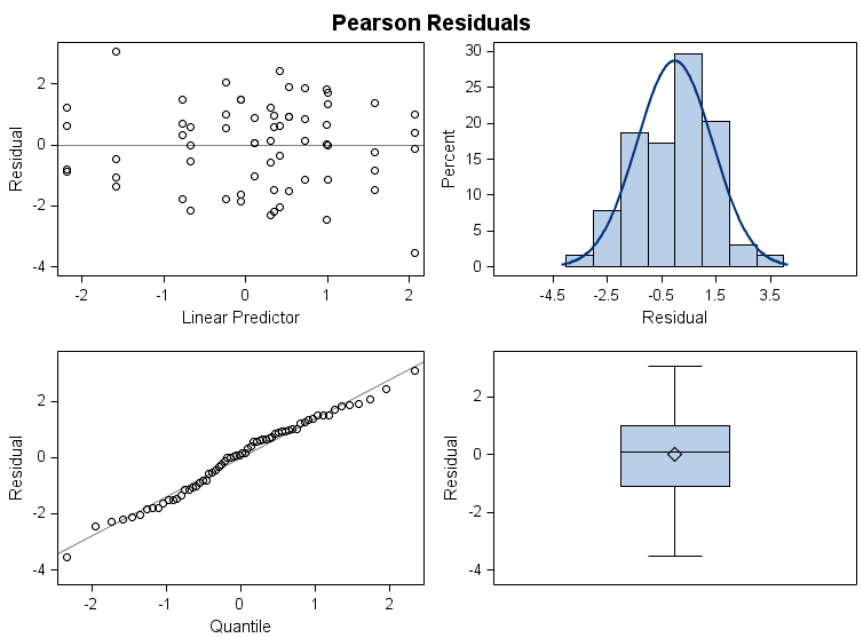
Effect	Num DF	Den DF	F Value	Pr > F
entry	15	48	3.60	0.0004

The F value (3.6) for the entry effect has been sharply reduced compared to the previous analyses. The smooth spatial variation accounts for some of the variation among the varieties



The following plot compares the LS-means of varieties. Varieties with negative LS-means have less infestation with Hessian fly.

The `PLOTS=All` statement produces diagnostic plots



Conclusions

In this example three models were considered for the analysis of a randomized block design with binomial outcomes.

If data are correlated, a standard generalized linear model often will indicate over-dispersion relative to the binomial distribution. Two courses of action are considered in this example to address this over-dispersion.

First, the inclusion of ***G-side random effects*** models the correlation <u>indirectly</u>; it is induced through the sharing of random effects among responses from the same block.

Second, ***the R-side spatial covariance structure*** models covariation directly.

In generalized linear (mixed) models, these two modeling approaches can lead to different inferences, because the models have different interpretation. The random block effects are modeled on the linked (logit) scale, and the spatial effects were modeled on the mean scale.

Only in a linear mixed model are the two scales identical.

GLMM with ASReml

ASReml is widely used in genetic data analysis designed experiments. We analyzed Hessian Fly data using the following ASReml code. The block effect is fitted as random again.

```
Title: Hessianfly.
#Damaged,Plants,block,Entry,lat,lng
#2,8,1,14,1,1
#1,9,1,16,1,2
 y
 N
 block  *
 entry  *
 lat  *
 lng  *
 yRatio !=y !/N

Hessianfly.csv  !SKIP 1

y !bin !TOTAL=N ~ mu entry !r block
```

Partial output from the primary output file (HessianFly.ASR) file is given below.

```
Distribution and link: Binomial; Logit  Mu=P=1/(1+exp(-XB))
                                          V=Mu(1-Mu)/N
 Warning: The LogL value is unsuitable for comparing GLM models

 Notice:      1 singularities detected in design matrix.
   1 LogL=-46.8426     S2=  1.0000         48 df    Dev/DF=   2.671
   2 LogL=-46.8446     S2=  1.0000         48 df    Dev/DF=   2.671
   3 LogL=-46.8446     S2=  1.0000         48 df    Dev/DF=   2.671
   4 LogL=-46.8446     S2=  1.0000         48 df    Dev/DF=   2.671
   5 LogL=-46.8446     S2=  1.0000         48 df    Dev/DF=   2.671

 Final parameter values                        1.0000
 Deviance from GLM fit              48      128.23
 Variance heterogeneity factor [Deviance/DF]     2.67
```

The heterogeneity factor [Deviance / DF] gives some indication as how well the discrete distribution fits the data. A value greater than 1 suggests the data are over-dispersed, that is the data values are more variable than expected under the chosen distribution.

```
          - - - Results from analysis of y - - -

 Source        Model  terms     Gamma     Component    Comp/SE   % C
 Variance         64     48   1.00000       1.00000       0.00   0 F

                           Wald F statistics
     Source of Variation  NumDF    DenDF    F-inc        P-inc
   7 mu                     1      48.0     4.64        0.036
   4 entry                 15      48.0     6.87        <.001

Finished: 26 Jul 2011 13:38:06.968   LogL Converged
```

The F-value for the Entry (plant varieties) is large and significant.

We can adjust for heterogeneity (over-dispersion) by using !DISPERSION qualifier in ASReml. The dispersion parameter is estimated from the residuals if not supplied by the analyst. Here is the model to account for over-dispersion.

```
y !bin !TOTAL=N !dispersion ~ mu entry
```

Output

```
                           Wald F statistics
     Source of Variation   NumDF     DenDF    F-inc     P-inc
   8 mu                        1      48.0     2.05     0.159
   4 entry                    15      48.0     3.03     0.002
```

After adjusting for heterogeneity in the variance we see a much smaller F test for Entry. It is still significant. The predictions for plant varieties do not change but their standard errors change.

Spatial R structure with ASReml

ASReml is powerful to model the **R** side of the GLMM models, especially fitting spatial and power models. In the following example, we used two dimensional autoregressive order 1 (AR1 x Ar1) correlation structure for **R** to account for heterogeneity in data as a demonstration.

```
Title: Hessianfly.
#Damaged,Plants,block,Entry,lat,lng
#2,8,1,14,1,1
```

```
#1,9,1,16,1,2
 y
 N
 block  *
 entry  *
 lat  *
 lng  *
 yRatio !=y !/N
Hessianfly.csv  !SKIP 1 !DOPART 2

!PART 2
! Generalized Linear Mixed Model with spatial R
y !bin !TOTAL=N ~ mu entry mv
1 2 0
8 row AR1 0.1
8 column AR1 0.1
```

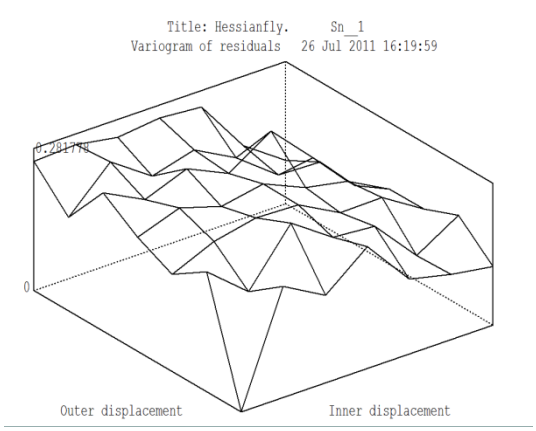
Partial output from the primary output file (HessianFly.ASR) file

```
Deviance from GLM fit                48      130.21
Variance heterogeneity factor [Deviance/DF]     2.71


          - - - Results from analysis of y - - -

 Source        Model  terms     Gamma      Component     Comp/SE    % C
 Residual    AR=AutoR      8 -0.183224      -0.183224       -1.62    0 U
 Residual    AR=AutoR      8  0.709363E-01  0.0709363    0.70    0 U

                            Wald F statistics
     Source of Variation    NumDF      DenDF     F-inc      P-inc
   8 mu                         1        7.1      5.12      0.058
   4 entry                     15       43.0      7.42      <.001
```



The AR1 structure for residuals was not successful as shown by a small correlation (0.0709). The variance heterogeneity factor is still 2.71. Sample variogram of residuals based on AR1 x AR1 is given below. There are no apparent trends in the row or column directions, but the zigzag surface suggests heterogeneity in the data.

Example 2: Binary response variable with genetic effects

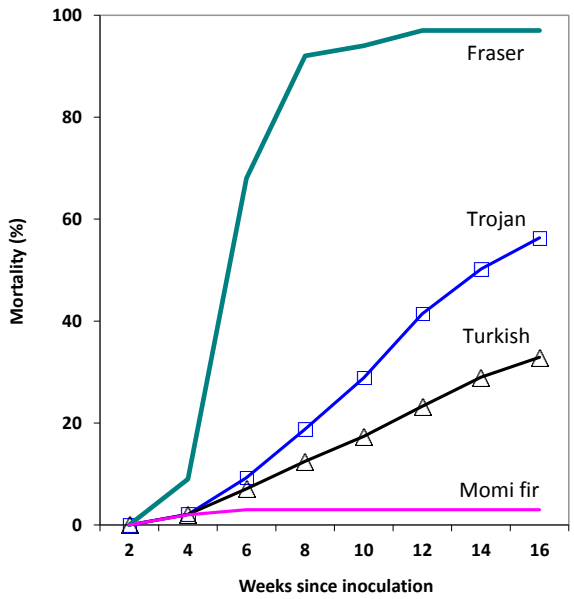
Variation in resistance to *Phytophthora* root rot in Turkish and Trojan fir

Fraser fir, a major Christmas tree is susceptible to *Phytophthora* root rot. Researchers (John Frampton and Fikret Isik) from North Carolina State University have been exploring native firs of Turkey as alternative to control disease. They embarked on a cone collection trip in northwestern Turkey in 2005. They visited four provenances (geographic regions) of Turkish fir and two provenances of Trojan fir. At each provenance, they collected cones from 20 trees. Seedlings were grown in a greenhouse and inoculated with *Phytophthora cinnamomi*, a soil borne pathogen. Subsequently, survival or mortality of each seedling was assessed biweekly. Fraser fir was used as control. The objective of the research was to examine the genetic variation between and within seed sources of Turkish firs and estimate heritability values for disease susceptibility.

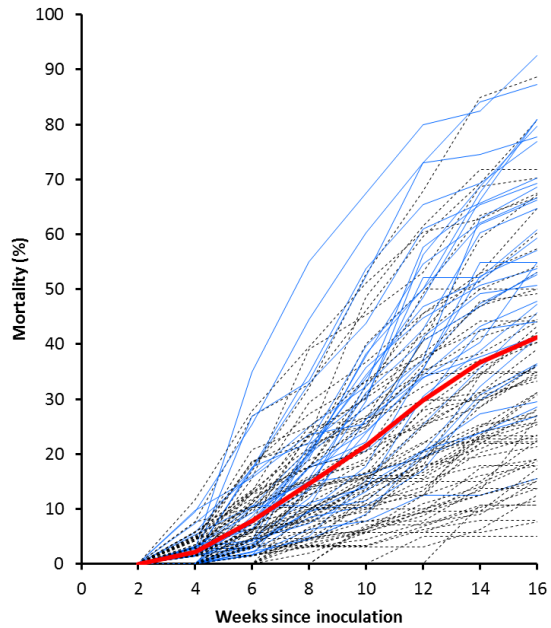
The first 10 lines of data are presented here

```
Sort,Rep,Tray,Species,Prov,Family,Tree,Wk2,Wk4,Wk6,Wk8,Wk10,Wk12,Wk14,Wk16
 1,1,1,Turkish,SAF,120, 1, 0,0,0,0,0,0,0,0
 2,1,1,Turkish,SAF,120, 2, 0,0,0,0,0,0,0,0
 3,1,1,Turkish,SAF,120, 3, 0,0,0,0,0,0,0,0
 4,1,1,Turkish,SAF,120, 4, 0,0,0,0,0,0,0,0
 5,1,1,Turkish,SAF,120, 1, 0,0,0,0,0,0,0,0
 6,1,1,Turkish,SAF,120, 6, 0,0,0,0,0,0,0,0
 7,1,1,Turkish,SAF,120, 7, 0,0,0,0,0,0,0,0
 8,1,1,Turkish,SAF,120, 8, 0,0,0,0,0,0,0,0
 9,1,1,Turkish,SAF,120, 9, 0,0,0,0,0,0,0,0
10,1,1,Turkish,SAF,120,10, 0,0,0,0,0,0,0,0
```

The means were plotted against time (weeks) to examine the trends (linear, quadratic etc.) in disease incidence but also visually depict the interactions of species and provenances with time.



Mortality means (%) of fir species over time. The mean mortality of Fraser fir was 97%, whereas the mean mortality of Trojan (56.5%) and Turkish (35%) were considerably smaller.



Plot of 105 family profiles and the average trend of mortality (thick red line) over time. The blue lines are Trojan fir families. Black lines are families from Turkish fir.

The probability of infection ($\hat{\pi}$) of a single seedling was modeled with the generalized linear mixed model using a logit (canonical) link function to partition phenotypic variance into genetic and environmental components.

$$\eta_{ijkl} = \log [\pi/(1-\pi)] = \mu + R_i + P_j + RP_{ij} + F(P)_{k(j)} + RF(P)_{ik(j)} + e_{ijkl}$$

Where;
η_{ijk} is the link function [g($\boldsymbol{\mu}$)],
$\boldsymbol{\mu}$ is the conditional mean,
π is the proportion of diseased seedlings,
R_i is the fixed effect of the *i*th replication,
P_j is the fixed effect of the *j*th provenance,
RP_{ij} is the fixed interaction effect between i*th* replication and j*th* provenance,
$F(P)_{j(k)}$ is the random effect of k*th* family nested within provenance with N(0, $\mathbf{I}\sigma^2_{f(p)}$),
$RF(P)_{ij(k)}$ is the random interaction of the kth family and ith replication with N(0, $\sigma^2_{rf(p)}$), and
e_{ijkl} is the random residual with N(0, $\mathbf{I}\sigma^2_e$).

The model was run using the ASReml software (Gilmour et al. 2009).

```
Title: Pc inoculation
 Sort   *
 Rep   *          # Rep is numeric
 Tray   *         # Tray is numeric
```

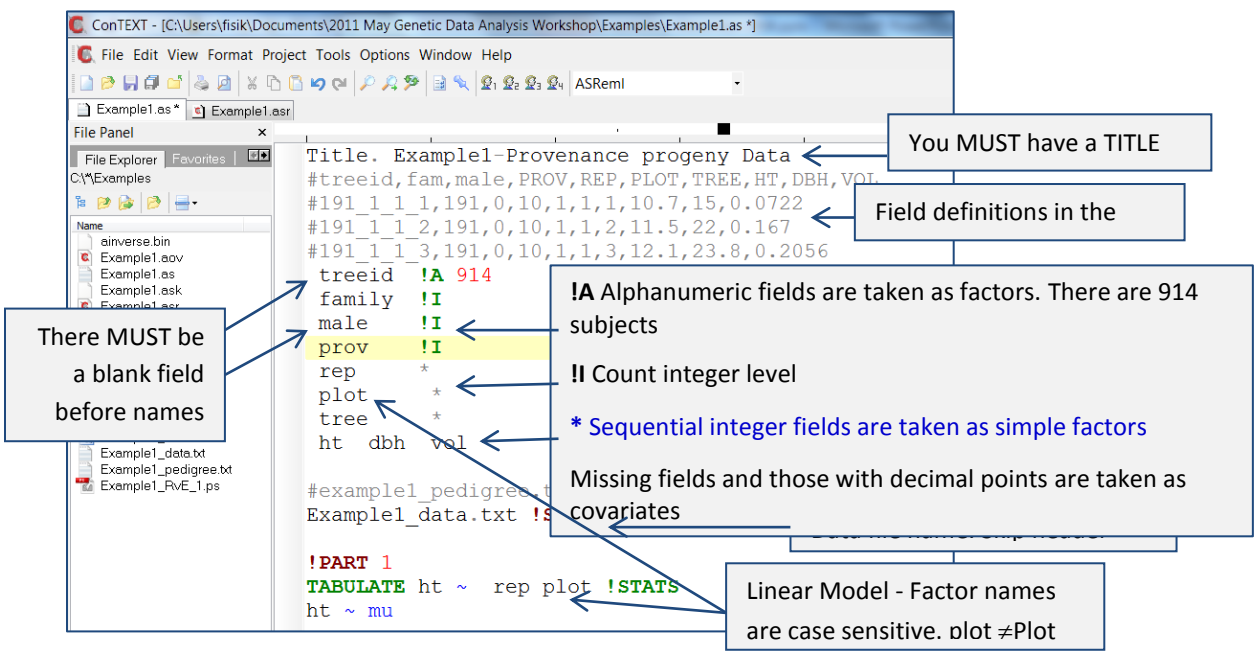
```
 Species !A     # Specify is alpha numeric, 1 level
 Prov   !A      # Provenance is alpha numeric, 3 levels
 Family  !I      # Family Id is non-integer numeric
 Tree   *
 Wk_2      Wk_4     Wk_6      Wk_8
 Wk_10     Wk_12    Wk_14     Wk_16

Tfir_ped.csv !SKIP 1 !MAKE !ALPHA !GROUPS 4 # Pedigree file
Tfir_dat.csv !SKIP 1 !DOPART 1  !CONTINUE  # Data file

!PART 1
Wk_16  !bin !logit ~ mu Rep*Prov  ,         # Specify fixed model
        !r Prov.Family Rep.Prov.Family  # Random effects
```

WK_16 is the response variable.

Here is a simple ASReml command file written using a text editor (ConTEXT) with more description.



Notice that we define the distribution using the **!BIN** qualifier (binomial) and underlying link function using !LOGIT qualifier in ASReml. The logit is the default link function. The variance on the underlying scale is $\pi^2/3$ = 3.28987=~ 3.298 (underlying logistic distribution) for the logit link (Gilmour et al. 2006).

The results are

```
Univariate analysis of Wk_16

Using     3662 records of  3675 read
```

```
Univariate analysis of Wk_16
 Summary of 3662 records retained of 3675 read

  Model term        Size #miss #zero   MinNon0    Mean       MaxNon0  StndDevn
   1 Sort           4999     0     0      1  2244.2174       4999
   2 Rep               4     0     0      1     2.4913          4
   3 Tray              9     0     0      1     4.6182          9
   4 Species           2     0     0      1     1.0000          1
   5 Prov              4     0     0      1     2.5800          4
   6 Family   !P      70     0     0      5         36         70
   7 Tree             18     0     0      1     7.0866         18
   8 Wk_2                    0  3662  0.000      0.000      0.000      0.000
   9 Wk_4                    0  3578  1.000     0.2294E-01  1.000     0.1497
  10 Wk_6                    0  3394  1.000     0.7318E-01  1.000     0.2605
  11 Wk_8                    0  3176  1.000     0.1327      1.000     0.3393
  12 Wk_10                   0  2988  1.000     0.1841      1.000     0.3876
  13 Wk_12                   0  2751  1.000     0.2488      1.000     0.4324
  14 Wk_14                   0  2531  1.000     0.3088      1.000     0.4621
  15 Wk_16     Variate       0  2379  1.000     0.3504      1.000     0.4771
  16 mu               1
  17 Rep.Family        280  2 Rep       :   4   6 Family         :   70

 Forming    355 equations:   5 dense.
 Initial updates will be shrunk by factor    0.010
 Notice: Algebraic ANOVA Denominator DF calculation is not available
         Numerical derivatives will be used.
 Distribution and link: Binomial; Logit  Mu=P=1/(1+exp(-XB))
                                          V=Mu(1-Mu)/N
 Warning: The LogL value is unsuitable for comparing GLM models
 Notice:      9 singularities detected in design matrix.
   1 LogL=-4728.06     S2=  1.0000       3646 df    Dev/DF=   1.126
   2 LogL=-4728.23     S2=  1.0000       3646 df    Dev/DF=   1.126
   3 LogL=-4729.80     S2=  1.0000       3646 df    Dev/DF=   1.124
   4 LogL=-4735.13     S2=  1.0000       3646 df    Dev/DF=   1.119
   5 LogL=-4744.98     S2=  1.0000       3646 df    Dev/DF=   1.114
   6 LogL=-4759.29     S2=  1.0000       3646 df    Dev/DF=   1.109
   7 LogL=-4766.54     S2=  1.0000       3646 df    Dev/DF=   1.106
   8 LogL=-4768.15     S2=  1.0000       3646 df    Dev/DF=   1.106
   9 LogL=-4768.30     S2=  1.0000       3646 df    Dev/DF=   1.106
  10 LogL=-4768.31     S2=  1.0000       3646 df    Dev/DF=   1.106
  11 LogL=-4768.31     S2=  1.0000       3646 df    Dev/DF=   1.106
 Final parameter values                    0.45778    0.12291     1.0000

 Deviance from GLM fit            3646     4031.00
 Variance heterogeneity factor [Deviance/DF]     1.11

Source        Model  terms     Gamma    Component  Comp/SE   % C
Family           70     69  0.457785    0.457785     4.14   0 P
Rep.Family      280    277  0.122911    0.122911     2.41   0 P
Variance       3662   3646   1.00000     1.00000     0.00   0 F


 Analysis of Variance    NumDF     DenDF    F_inc      Prob
  16 mu                      1       3.0     7.26     0.074
   2 Rep                     3     174.5     6.24     <.001
```

```
            Estimate       Standard Error    T-value     T-prev
   2 Rep
            2  -0.486743       0.125249         -3.89
            3  -0.365991       0.124154         -2.95      0.95
            4  -0.127976       0.122878         -1.04      1.90
  16 mu
            1  -0.572161       0.308671         -1.85
   5 Prov                  6 effects fitted (       2 are zero)
  17 Prov.Family         540 effects fitted (     473 are zero)
  18 Rep.Prov.Family    2160 effects fitted (    1892 are zero)

Finished: 18 Dec 2008 14:39:08.860   LogL Converged
```

We are interested in the variance components (Component) in this study to understand the effect of genetics and environment on the disease incidence. Heritability will tell us about the effects of genetics (family differences) on the incidence compared to phenotypic variance.

The family effect and other random effects are on logistic scale with a variance of $3.28987=\pi^2/3$ (Gilmour et al. 1985). Because we have wind-pollinated families assumed to be half-siblings, variance explained by family effect is 1/4 of additive genetic variance (Falconer and Mackay 1996). The total additive genetic variance would be 4* Var(Prov.Family).

We are mostly interested in selection of families and thus the heritability of interest would be family mean heritability.

$$h_f^2 = \frac{\sigma^2_{f(p)}}{\left(\sigma^2_{f(p)} + \frac{\sigma^2_{rf(p)}}{r} + \frac{\sigma^2_e}{n}\right)}$$

Where $\sigma^2_{f(p)}$ is the aggregate family variance component across provenances, $\sigma^2_{rf(p)}$ is the replication by family interaction variance, σ^2_e is the fixed error variance, *r* is the number of replications and *n* is the number of seedlings per family. The error variance was set to 3.29 in calculation of phenotypic variances as suggested by Gilmour *et al*. (1985). Standard errors of heritabilities were estimated using Delta method (Lynch and Walsh 1998).

The denominator in the above formula is the phenotypic variance of family means, *r* is the number of replications (4), and *n* is the number of seedlings per family (on average it is 52 seedlings). Using the numbers from the output file given above column named 'Component', the heritability would be = 0.457785 / (0.457785 + 0.1229/4 + 3.29/52) = 0.83

We can use ASReml to calculate all sorts of functions of variance components.

```
Title: Pc inoculation
 Sort  *
 Rep  *          # Rep is numeric
 Tray  *         # Tray is numeric
```

```
 Species !A     # Specify is alpha numeric, 1 level
 Prov  !A       # Provenance is alpha numeric, 3 levels
 Family  !I       # Family Id is non-integer numeric
 Tree  *
 Wk_2     Wk_4    Wk_6     Wk_8
 Wk_10    Wk_12   Wk_14    Wk_16

Tfir_ped.csv !SKIP 1 !MAKE !ALPHA !GROUPS 4 # Pedigree file
Tfir_dat.csv !SKIP 1 !DOPART 1  !CONTINUE  # Data file

!PART 1
Wk_16  !bin !logit ~ mu Rep*Prov  ,      # Specify fixed model
        !r Prov.Family Rep.Prov.Family  # Random effects
!PIN !DEFINE
P Total 1+2+3*3.29                  # 4 Total Variance
P Pheno 1+2+3*3.29                  # 5 Phenotypic Variance
P Pheno_FamMean 1+2*0.25+3*0.0598   # 6 Family Mean Phen Variance
P ErrorVar 3*3.29                   # 7 Error Variance
P AddVar 1*4                        # 8 Additive Genetic Variance
H P_Fam 1 4                         # % Variance of Family(Prov)
H P_RepFam 2 4                      # % Variance of Rep*Family
H P_Error 7 4                       # % Variance of Error
H H2I 8 5                         # Individual Tree herit
H H2F 1 6                         # Family Mean Heritability
```

Output

```
   4 Total  1       3.871      0.1166
   5 Pheno  1       3.871      0.1166
   6 Pheno_Fam  1  0.5483      0.1099
   7 ErrorVar  3    3.290       0.000
   8 AddVar  1      1.831      0.4421
     P_Fam        = Family     1/Total  1   4= 0.1183    0.0254
     P_RepFam     = Rep.Fami   2/Total  1   4= 0.0318    0.0129
     P_Error      = ErrorVar   7/Total  1   4= 0.8500    0.0256
     H2I          = AddVar     8/Pheno  1   5= 0.4731    0.1016
     H2F          = Family     1/Pheno_Fa   6= 0.8349    0.0403
Notice: The parameter estimates are followed by their approximate
standard errors.
```

Additive genetic variance is 1.83 ± 0.442. Family effect explained about 12% of total variance (0.118) observed in the study. Family mean heritability is 0.83 which is high, suggesting that if we select families with low disease incidence and use them for plantations, we will be able to control the disease successfully.

Acknowledgement

Alfredo Farjat, PhD student with the Cooperative Tree Improvement Program in the Department of Forestry and Environmental Resources at NC State University contributed the document, especially the theory parts.

References

[1] *SAS System for Mixed Models*, July 1996, SAS Publishing. Ramon C. Littell, George A. Milliken, Walter W. Stroup and Russell Wolfinger.

[2] *An Introduction to Generalized Linear Mixed Models***.** Stephen D. Kachman, Department of Biometry, University of Nebraska–Lincoln.

[3] GLIMMIX Procedure: http://support.sas.com/rnd/app/papers/glimmix.pdf

[4] *Repeated Measures Modeling With PROC MIXED*. E. Barry Moser, Louisiana State University, Baton Rouge, LA. SUGI 29 Proceedings, Paper 188-29.

[5] PROC *MIXED: Underlying Ideas with Examples*. David A. Dickey, NC State University, Raleigh, NC. SAS Global Forum 2008, Statistics and Data Analysis, Paper 374-2008.

[6] *Ideas and Examples in Generalized Linear Mixed Models*. David A. Dickey, N. Carolina State U., Raleigh, NC. SAS Global Forum 2010, Statistics and Data Analysis, Paper 263-2010.

[7] *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*. Oliver Schabenberger, SAS Institute Inc., Cary, NC. SUGI 30, Paper 196-30.

[8] *Practical Regression and Anova using R*. Julian Faraway. http://www.r-project.org/

[9] Falconer, D.S., and Mackay, T.F.C. 1996. Introduction to Quantitative Genetics. Fourth Edition, Longman Group Ltd., Essex, England, 464 p.

[10] Gilmour, A.R., Anderson, R.D., and Rae, A.L. 1985. The analysis of binomial data by a generalized linear mixed model. Biometrika, 72: 593–599. doi:10.1093/biomet/72.3.593.

[11] Gilmour, AR, Gogel BJ, Cullis BR, Thomson R. 2009. ASREML User Guide, Release 3.0. VSN International Ltd, Hemel Hempstead, HP1, 1ES, UK. 267 p.

[12] Littell RC, Henry PR, CB Ammerman (1998) Statistical analysis of repeated measures data using SAS procedures. J. Anim. Sci. 76: 1216-1231.

[13] Lynch, M., and Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, Mass.