# Double generalized linear model for tissue culture proportion data: a Bayesian perspective

Peer-reviewed author version

CORREA VIEIRA, Afranio Marcio; Leandro, Roseli A.; DEMETRIO, Clarice & MOLENBERGHS, Geert (2011) Double generalized linear model for tissue culture proportion data: a Bayesian perspective. In: JOURNAL OF APPLIED STATISTICS, 38 (8), p. 1717-1731.

DOI: 10.1080/02664763.2010.529875
Handle: http://hdl.handle.net/1942/14380

# Double generalized linear model for tissue culture proportion data: a Bayesian perspective

**Article** *in* Journal of Applied Statistics · August 2011

DOI: 10.1080/02664763.2010.529875 · Source: RePEc



**4 authors:**

Afrânio M C Vieira
Universidade Federal de São Carlos
**36** PUBLICATIONS **277** CITATIONS
SEE PROFILE

Roseli Aparecida Leandro
University of São Paulo
**23** PUBLICATIONS **53** CITATIONS
SEE PROFILE

Clarice G. B. Demétrio
University of São Paulo
**153** PUBLICATIONS **1,958** CITATIONS
SEE PROFILE

Geert Molenberghs
Universiteit Hasselt and University of Leuven
**892** PUBLICATIONS **20,546** CITATIONS
SEE PROFILE

# Double Generalized Linear Model for Tissue Culture Proportion Data: A Bayesian Perspective[1]

**Afrânio M. C. Vieira**[2]
Departamento de Estatística, Universidade de Brasília
ICC centro, subsolo, módulo 15, CEP 70910-900
Brasília, DF, Brazil

**Roseli A. Leandro** and **Clarice G. B. Demétrio**
Departamento de Ciências Exatas, Universidade de São Paulo - ESALQ
Av. Pádua Dias 11, CP 9, CEP 13418-900
Piracicaba, SP, Brazil

**Geert Molenberghs**
I-BioStat, Universiteit Hasselt and Katholieke Universiteit Leuven
Agoralaan 1, 3590 Diepenbeek, Belgium

[1]Keywords: Bayesian data analysis; Generalized Linear Models; Tissue Culture; Markov Chain Monte Carlo; Binomial Distribution; Gibbs sampling; Random Effects
[2]Corresponding author; Email: afranio@unb.br

## Abstract

Joint generalized linear models (JGLM) and double generalized linear models (DGLM) were designed to model outcomes for which the variability can be explained using factors and/or covariates. When such factors operate, the usual normal regression models, which inherently exhibit constant variance, will under-represent variation in the data and hence may lead to erroneous inferences. For count and proportion data, such noise factors can generate a so-called overdispersion effect, and the use of binomial and Poisson models underestimates the variability and, consequently, incorrectly indicate significant effects. In this manuscript, we propose a double generalized linear model from a Bayesian perspective, focusing on the case of proportion data, where the overdispersion can be modeled using a random effect that depends on some noise factors. The posterior joint density function was sampled using Monte Carlo Markov Chain (MCMC) algorithms, allowing inferences over the model parameters. An application to a dataset on apple tissue culture is presented, for which it is shown that the Bayesian approach is quite feasible, even when limited prior information is available, thereby generating valuable insight for the researcher about its experimental results.

# 1 Introduction

Many well-known experimental designs that are applied across a diverse range of scientific domains are based on the assumption of variance homogeneity. It is a quite strong assumption when one is faced with situations where environmental or external factors influence the experimental measures. Modeling the variability from planned experiments gained momentum with Taguchi's work (Taguchi, 1985), which emphasizes the need to adequately deal with the influence of noise and control factors in industrial experimentation, as a means to reducing loss of information and hence optimizing product quality. In a conventional approach, if either environmental factors, the process factors under investigation, or a combination thereof, influences the variance of the continuous response variable, then it means that all statistical inferences from the resulting model will be based on a single dispersion measure, likely inflated by the effects not entered into the model. For proportion or count data, the effect of not taking into account such overdispersion is to produce underestimated variances if the standard, too restrictive, models, such as binomial or Poisson-based models are used. Needless to say that ultimately inference is in jeopardy then. Related to this, not taking account of this phenomenon can lead to the selection of overly complex models (Hinde & Demétrio, 1998).

The approach of Taguchi to deal with dispersion effects was criticized and a discussion started about effectiveness and alternatives to the signal-to-noise ratios (Box, 1988). One argument against signal-to-noise regards the fact that a transformation is chosen *a priori*. An alternative presents itself by way of the Box and Cox transformation family (Box & Cox, 1964), where the choice of the best variance-stabilizing transformation is data driven. However, the alternatives proposed to quantify and graph dispersion effects takes the form of exploratory tools; a joint approach was not considered. At the same time, a modeling approach was undertaken.

The concept of modeling heterogeneity through a pair of parametric non-linear predictors was formally established by Harvey (1976), with the parameters linked to the mean and variance estimated by maximum likelihood, for a normally distributed response variable.

For this case, when all factors are quantitative, alternatives exists in the form of so-called dual response surface methodology, where the mean and dispersion models are optimized simultaneously (Myers et al., 1992).

This problem was revisited later, and various regression models have been proposed to jointly model mean and dispersion (Aitkin, 1987; Wolfinger & Tobias, 1998; Smyth, 1989; Nelder & Lee, 1991). These authors base inferences, including hypothesis testing and interval estimation, on asymptotic theory (McCullagh & Nelder, 1989). Such methods work well with large sample sizes combined with modest numbers of model parameters. However, in agricultural research, many experiments exhibit a large number of parameters relative to the sample size. The asymptotic-theory-based estimators and their corresponding measures of uncertainty can then be questionable and lead to erroneous conclusions. This motivates our choice for a relative small set of data.

In this paper, we propose a *double generalized linear model* (DGLM) for proportion data using a Bayesian framework for parameter estimation. This approach allows one to incorporate the uncertainty about the unknown quantities of the model using prior information into the estimation procedure. The difficulty of obtaining the parameters' posterior marginal densities is overcome by the use of Monte Carlo Markov Chain (MCMC) algorithms (Gamerman & Lopes, 2006). The rest of the paper is organized as follows. In Section 2, the generalized linear models (GLM) framework, the extended quasi-likelihood estimation method, and the model proposed by Smyth (1989) and Nelder & Lee (1991) are briefly described and commented. A Bayesian perspective on the DGLM is presented in Section 3. In Section 4, an apple tissue culture experiment described in Ridout & Demétrio (1992) is introduced, with the results presented

and discussed in Section 4.2.

## 2 Joint Modeling of Mean and Dispersion

The methodology proposed by Smyth (1989) and Nelder & Lee (1991) for the joint modeling of mean and dispersion involves two generalized linear models (GLM; McCullagh & Nelder, 1989). For a random sample of $n$ observations $(y_i, \mathbf{x}_i)$, where $y_i$, $i = 1, \ldots, n$ is an observed value for a single response variable $Y_i$, and $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$ is a $p$-dimensional vector of explanatory variables, the three components of a GLM are (Hinde & Demétrio, 1998): (i) *independent random variables* $Y_i$, stemming from the exponential family of distributions with mean $\mu_i$ and constant scale parameter $\phi$, i.e., observations from a density of the form:

$$f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

where $a(\phi) = \phi/w$, $\phi$ is the dispersion parameter, $w$ is a prior weight, $\theta$ is the canonical parameter [it can be shown that $\mathsf{E}(Y) = b'(\theta)$ and $\mathsf{Var}(Y) = \phi b''(\theta)$]; (ii) *a linear predictor* vector $\boldsymbol{\eta}$ given by $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of $p$ unknown parameters and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ is an $n \times p$ design matrix; (iii) *a link function* $g(\cdot)$ relating the mean to the linear predictor, i.e., $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$; hence, $\mathsf{E}(Y_i) = g^{-1}(\eta_i)$.

In this paper, we focus on the particular GLM with binomial distribution and logit link function. Assuming that a random variable $Y_i$, the number of successes out of $m_i$ samples, has a binomial distribution with probability of success $\pi_i$, it follows that $\theta_i = \ln[\mu_i/(m_i - \mu_i)]$, $b(\theta_i) = m_i \ln(1 + e^{\theta_i})$ and $\phi = 1$. Therefore, $\mathsf{E}(Y_i) = m_i\pi_i = \mu_i$, $\mathsf{Var}(Y_i) = m_i\pi_i(1 - \pi_i)$ and $g(\mu_i) = \ln[\mu_i/(1 - \mu_i) = \eta_i]$. Parameter estimation conventionally proceeds by maximum likelihood; in computational terms, the *iteratively re-weighted least square algorithm* (IRLS) is popular.

Note that, because the dispersion parameter $\phi = 1$, the variance function depends solely on the mean parameter. However, it is quite common in experimental situations that proportions show variability larger than that allowed by the theoretical variance of the binomial distribution, the aforementioned *overdispersion*. Hinde & Demétrio (1998a) reviewed a wide variety of avenues for overdispersion modeling, together with methods of estimation. These authors also discussed applications to agricultural experimentation data. Nelder & Pregibon (1987) proposed the *extended quasi-likelihood* (EQL) method for parameter estimation, based only on the first two moments of a distribution. The EQL method consists of maximizing the function

$$Q^+ = -\frac{1}{2}\sum_{i=1}^{n}\{\frac{d(y_i, \mu_i)}{\phi_i} + \log(2\pi\phi_i V(y_i))\},$$

where

$$d(y, \mu) = -2\int_y^\mu \frac{y - t}{\mathsf{V}(t)}dt$$

is the deviance function and $\mathsf{V}(\cdot)$ is the variance function evaluated in $y_i$. The dispersion parameter is indexed by observation, allowing for flexible modeling. For example, experimental factors and/or covariates affecting the variability of the data may be encompassed. For proportion data, the method allows for the modeling of overdispersion as a function of a linear predictor that may differ from the one describing the mean.

The joint-modeling ideas for mean and dispersion, proposed by Smyth (1989) and Nelder & Lee (1991), all share the same double structure of generalized linear models. Assuming that $\mathsf{E}(Y) = \mu$ and $\mathsf{Var}(Y) = \phi\mathsf{V}(\mu)$, and that both the mean and the dispersion parameters vary across observations in a parametric way, i.e., $\eta_i = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ and $\zeta_i = h(\phi_i) = \mathbf{z}_i^T\boldsymbol{\gamma}$ and

where $\boldsymbol{\beta}$ is a vector of mean parameters, $\boldsymbol{\gamma}$ is as vector of dispersion parameters, $g(\cdot)$ and $h(\cdot)$ are link functions for the mean and dispersion, and $\mathbf{x}_i^T$ and $\mathbf{z}_i^T$ are the row-vectors of the design matrices $\mathbf{X}$ and $\mathbf{Z}$, respectively. The matrix $\mathbf{X}$ contain covariates and/or factors affecting the mean, and the matrix $\mathbf{Z}$ contains covariates and/or factors affecting the dispersion parameter. In this model, $\phi$ represents the independent variation of the mean and V($\mu$) is the mean-dependent variation. Apart from this commonality between the modeling frameworks, they exhibit particular aspects, too.

Parameter estimation proposed by Smyth (1989) and Nelder & Lee (1991) is based on a two-step iterative algorithm: (i) holding $\boldsymbol{\gamma}$ fixed, the vector $\boldsymbol{\beta}$ is estimated; (ii) fixing the estimated value of $\boldsymbol{\beta}$, the vector $\boldsymbol{\gamma}$ is estimated. These two steps are then alternated until convergence. Although both proposals are based on different estimation methods, results are often very similar.

Nelder & Lee (1991) based estimation on extended quasi-likelihood. In their algorithm, the step where $\phi$ is assumed fixed coincides with Smyth's (1989) method, thus reducing to quasi-likelihood. When $\boldsymbol{\beta}$ is fixed, the extended quasi-likelihood function becomes a gamma likelihood function, where $d_i$ is the response variable. Lee & Nelder (1998) also considered an alternative for the estimation method based on REML with adjustment proposed by Cox & Reid (1987). Lee & Nelder (2006) extended their proposal to a larger class of *double hierarchical generalized linear models*, jointly incorporating random effects in both mean and dispersion linear predictors. This class will not be explored in this work.

At this point, it is important to emphasize key differences between the JGLM and the Bayesian DGLM explored here. The JGLM is a fixed-effects model that deals with dispersion modeled in a particular way. This involves another generalized linear model for deviance components as a response, a logarithmic link function and a linear predictor. The Bayesian perspective for the proportion data, which will be described in Section 3, proceeds by hierarchically modeling the overdispersion through a random effect, where the linear predictor is linked to the variance of the random effect. So, even though the results of both approaches may lead to the same conclusions, the interpretations are different.

# 3 The Double Generalized Linear Model from a Bayesian Perspective

## 3.1 Model for Normally Distributed Measurements

The frequentist estimation approaches of Smyth (1989) and Nelder & Lee (1991) are clearly approximate and dependent on asymptotic assumptions. In agricultural experimentation, the number of experimental units is mostly limited owing to physical space, resources, and/or ethical constraints. Situations are common where the number of parameters is relatively large compared with the number of observations. The frequentist approach can then lead to strongly biased estimates (Smyth & Verbyla, 1999). An alternative way to tackle this problem is to work with the double generalized linear model (DGLM) from a Bayesian point of view.

One proposal for a Bayesian DGLM was presented by Cepeda & Gamerman (2000), with the following structure:

$$\begin{array}{rcl} y_i & = & \mu_i + \varepsilon_i, \\ \varepsilon_i & \sim & \mathsf{N}(0, \sigma_i^2), \\ \mu_i & = & \mathbf{x}_i^T \boldsymbol{\beta}, \\ g(\sigma_i^2) & = & \mathbf{z}_i^T \boldsymbol{\gamma}, \end{array} \tag{1}$$

where $\mathbf{x}_i^T\boldsymbol{\beta}$ is the linear predictor for the mean $\mu$, $\varepsilon_i$ is the random component, the variance $\sigma_i^2$, which is linked to the linear predictor $\mathbf{Z}\boldsymbol{\gamma}$ by a non-linear link function $g(\cdot)$, and $\mathbf{x}_i^T$ and $\mathbf{z}_i^T$ are known rows of design matrices for mean and dispersion, respectively. These authors

assumed the following prior joint probability density function for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$:

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{b}_0 \\ \mathbf{g}_0 \end{pmatrix}, \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{G} \end{pmatrix}\right],$$

where the hyper-parameters $\mathbf{b}_0$, $\mathbf{g}_0$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ are assumed known. The posterior joint probability density function is given by

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{X}, \mathbf{Z}) &= \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\int_{\boldsymbol{\beta}}\int_{\boldsymbol{\gamma}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}, \boldsymbol{\gamma})\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}} \\ &\propto p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}, \boldsymbol{\gamma}). \end{aligned} \tag{2}$$

As (2) assumes an intractable analytical form for integral manipulation, the Metropolis-Hastings (MH) algorithm was employed, together with a block-wise scheme to obtain the samples of the posterior marginal density functions for each parameter (Gamerman, 1997).

### 3.2 Model for Overdispersed Proportion Data

The ideas behind a Bayesian DGLM for normal data do not carry over to the binomial situation, because in that case there is no separate variance parameter. Hinde & Demétrio (1998b) describe a logistic-normal model with the following structure

$$\begin{aligned} Y_i|\mathbf{z}_i &\sim \text{Bin}(m_i, p_i), \\ \text{logit}(p_i) &= \mathbf{x}_i^T\boldsymbol{\beta} + \sigma\mathbf{z}_i, \\ \mathbf{z}_i &\sim \text{N}(0, 1), \end{aligned} \tag{3}$$

with the aim of accommodating the overdispersion effect through the random effect $\mathbf{z}_i$. Borgatto *et al.* (2006) proposed a hierarchical random-effects model to account for both overdispersion and zero-inflation effects, as an alternative to the model described in Vieira *et al.* (2000). A Bayesian version of (3) was also proposed by Hinde & Demétrio (1998b), taking the form

$$\begin{aligned} Y_i|\mathbf{b}_i &\sim \text{Bin}(m_i, p_i), \\ \text{logit}(p_i) &= \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{b}_i, \\ \mathbf{b}_i &\sim \text{N}(0, \sigma^2), \end{aligned} \tag{4}$$

and assuming a prior distribution for $\boldsymbol{\beta}$ and $\tau = \sigma^{-2}$ to incorporate the uncertainty associated with these parameters.

Here, we propose a generalized version of (4), to allow for covariates and/or factors affecting the dispersion parameter of the random-effect distribution. To this end, the following hierarchical double generalized linear model is assumed:

$$\begin{aligned} Y_i &\sim \text{Bin}(m_i, p_i), \\ \text{logit}(p_i) &= \mathbf{x}_i^T\boldsymbol{\beta} + \delta_i, \\ \delta_i &\sim \text{N}(a, \tau_i), \\ \boldsymbol{\tau}_i &= 1/\exp(\mathbf{z}_i^T\boldsymbol{\gamma}), \end{aligned} \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the appropriate rows of the design matrices $\mathbf{X}$ and $\mathbf{Z}$, respectively, $\delta_i$ is a random effect, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters. The normal distribution for a random effect, used to accommodate overdispersion, is a sensible choice whenever this random

variable is required to range over the entire real line. Evidently, other distributions could be entertained as well, such as, for example, a scaled $t$-distribution. We further assume that $\beta_j$ and $\gamma_k$ are independent, i.e., $p(\beta_j, \gamma_k) = p(\beta_j)p(\gamma_k)$, which is sensible given that it is difficult to establish a prior dependence structure for these parameters in common experimental situations. In this model, the link function for the mean of $Y_i$ is $\text{logit}(p_i) = \ln[p_i/(1-p_i)] = \ln[\mu_i/(m_i - \mu_i)]$. The link function for the dispersion is assumed to be $\ln \tau_i^{-1}$, to enforce positive variance; this can, of course, be modified to other monotone link functions, as appropriate. It was assumed for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ *a priori* to be normally distributed with known hyper-parameters specified by $\beta_j \sim \mathsf{N}(b, c)$, $j = 0, \ldots, r$, and $\gamma_k \sim \mathsf{N}(d, e)$, $k = 0, \ldots, s$, respectively. The normal priors with vague hyper-parameters is a way to establish non-informative uncertainty for the parameters.

The posterior joint probability density function for model (5), obtained by the Bayes' rule, can be described by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta} | \mathbf{X}, \mathbf{Z}, \mathbf{y}) \propto \mathsf{L}(\boldsymbol{\beta} | \boldsymbol{\delta}, \mathbf{y}, \mathbf{X}) p(\boldsymbol{\delta} | \boldsymbol{\gamma}, \mathbf{Z}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}). \tag{6}$$

Model (5) can be represented by a directed acyclic graph (DAG), as described in Best & Green (2005) as can be seen from Figure 1 in the Supplementary Materials. The advantage of presenting a model in DAG form is that the essence of the model structure is elucidated, making clear the functional flow of the information, thereby suppressing the distributional assumptions and deterministic relations between variables and parameters. Moreover, such a graphical model representation may suggest a conditional independence structure, convenient for efficient implementation. The Bayesian computation environment `OpenBUGS` (Thomas *et al.*, 2006) was built to sample the posterior marginal distributions of the parameters under DAGs that can be described graphically or through the BUGS language (Spiegelhalter *et al.*, 1996). Best & Green (2005) and Thomas *et al.* (2006) provide more details and information about directed acyclic graphs and the BUGS language.

## Sampling From the Posterior Marginal Densities

Assuming the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and for the random effect $\delta_i$, and using the binomial likelihood function for $Y_i$, the posterior joint density function can be written as (see the Appendix for more details):

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta} | \mathbf{X}, \mathbf{Z}, \mathbf{y}) \quad \propto \quad & \exp\left\{ \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}) - \frac{1}{2}\mathbf{1}^T\mathbf{Z}\boldsymbol{\gamma} - \sum_{i=1}^{n} \exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2 - \right. \\ & \left. - \frac{1}{2c}\sum_{j=0}^{r}(\beta_j - b)^2 - \frac{1}{2e}\sum_{k=0}^{s}(\gamma_k - d)^2 \right\} \times \\ & \times \prod_{i=1}^{n} [1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{-m_i}. \end{aligned} \tag{7}$$

From (7) it is not possible to derive analytic forms for the posterior marginal density functions for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$. Furthermore, it is not a viable alternative, neither to make use of numeric integration, because of its multi-dimensionality. Therefore, stochastic simulation of the posterior marginal densities, through the Monte Carlo Markov Chain (MCMC) methods, offers an appealing route.

So, to sample from the posterior joint density function (7), it is necessary to construct an appropriate Markov chain (Gamerman & Lopes, 2006), which can be done by using an MCMC algorithm, such as the Metropolis-Hastings algorithm, Gibbs sampling, or using a more general

MCMC algorithm such as, for example, the slice sampler (Neal, 2003). All of these algorithms are based on the full posterior marginal density function, given by

$$p(\beta_j|\boldsymbol{\beta}_{-j},\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{X},\mathbf{Z},\mathbf{y}) \propto \exp\left\{\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \frac{1}{2c}\sum_{j=0}^{r-1}(\beta_j - b)^2\right\}\prod_{i=1}^{n}[1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}+\delta_i)]^{-m_i}, \quad (8)$$

$$p(\gamma_k|\boldsymbol{\gamma}_{-k},\boldsymbol{\beta},\boldsymbol{\delta},\mathbf{X},\mathbf{Z},\mathbf{y}) \propto \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n-1}\exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i-a)^2 - \frac{1}{2e}\sum_{k=0}^{s}(\gamma_k-d)^2\right\}}{\sqrt{\exp\left(\mathbf{1}^T\mathbf{Z}\boldsymbol{\gamma}\right)}}, \quad (9)$$

$$p(\delta_i|\boldsymbol{\delta}_{-i},\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{X},\mathbf{Z},\mathbf{y}) \propto \frac{\exp\left\{\mathbf{y}^T\boldsymbol{\delta} - \frac{1}{2}\sum_{i=1}^{n-1}\exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i-a)^2\right\}}{\prod_{i=1}^{n-1}[1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}+\delta_i)]^{m_i}}, \quad (10)$$

where the vectors $\boldsymbol{\beta}_{-j}$, $\boldsymbol{\gamma}_{-k}$, and $\boldsymbol{\delta}_{-i}$ refer to the parameter vectors without the elements $\beta_j$, $\gamma_k$ and $\delta_i$, respectively.

These algorithms are implemented in the statistical computing environment `R` (R Development Core Team, 2007), using the library `BRugs` (Thomas *et al.*, 2006). The functions related with this library use the BUGS language. It is therefore necessary and sufficient to specify the model structure (5), the data set, and the initial values for each parameter, in order to start the Markov chain iterations. When the sequence generated by the Markov chain $\theta^t$, $t = 1, 2, \ldots$ reaches convergence, the sample of $\boldsymbol{\theta}^t$ can be considered a sample of $p(\theta|y)$. It can be formally shown that the convergence of the chain is to the stationary distribution. For a given set of data, it is important to construct some exploratory graphical diagnostics and to obtain further diagnostic measures to scrutinize convergence (Gelman *et al.*, 2000).

The use of a Bayesian approach via stochastic simulation methods demands that Markov chain for each parameter be examined for convergence, so as to guarantee that the samples contain the principal characteristics of the equilibrium distribution, including shape, the first few sample moments, etc. Some formal, e.g., test-based, methods, as well as informal graphically-based ones, are quality indicators for the simulated samples. Here, some graphical devices (history plot, auto-correlation function plot) and the Gelman-Rubin criterion (Gelman & Rubin, 1992) were used for Markov chain convergence diagnosis. Alternatively, the test, proposed by Raftery & Lewis (1992) and Heidelberger & Welch (1983), could be applied as well. These tests are all implemented in the `R` environment through the `coda` library (Plummer *et al.*, 2007).

To perform model selection, the Deviance Information Criteria (DIC, Spiegelhalter *et al.*, 2002) was used. This index is calculated as DIC $= p_D + \mathsf{E}_\theta[D(\theta)]$, where $p_D = \mathsf{E}_\theta[D(\theta)] - D(\mathsf{E}_\theta[p(\theta|y)])$, representing the effective number of parameters; $\mathsf{E}_\theta[D(\theta)]$ is the average of $D$ calculated over all values of $\theta$ from the sample obtained from the MCMC algorithms; and $D(\mathsf{E}_\theta[p(\theta|y)])$ is the deviance measure calculated over the average of the sampled values of $\theta$. As is well known, this measure, as well as the AIC and BIC, quantify the model fitting and, at the same time, penalize the complexity of the candidate models.

# 4 The Apple Tissue Experiment

## 4.1 Data Description

As a motivation to the modeling tools developed here, an apple tissue culture experiment, described in Ridout & Demétrio (1992), is analyzed. The treatment structure was a $2 \times 5$ factorial (2 media and 5 varieties), and the plot structure completely randomized. The experimental unit was a Petri dish, divided into a $5 \times 5$ array, hence having 25 individual compartments. In each compartment, a standard volume of some culture medium was used. A small piece of vegetation tissue, called *explant*, was added to the medium. The Petri dishes

were kept in a incubator for several weeks. During this period, new shoots could grow from the explants, which enhances the regeneration process. One aim of the researcher was to establish whether some of the five explant varieties and/or one of the two culture media have an influence on the proportion of regenerated explants. The data set is reproduced in Table 1.

The motivation to use a relatively small set of data is twofold. First, it allows focusing on the methodological contributions, without the intricacies of large and potentially complex data manipulations. Second, this type of experiment is quite common in horticulture.

Table 1 ABOUT HERE

Figure 1 shows a plot with the average of regenerated explants, for each combination of explant variety and culture method. There clearly is a lot of variability between the means of regenerated explants in culture medium $X$, when compared with that in culture medium $Y$. Explant $E$ appears to show differential behavior when compared to the other explants in culture medium $X$, suggesting a possible interaction. Ridout & Demétrio (1992) leave open the option that it could be a potential outlier. That said, we believe that the value of $2/25$ for the $E/X$ cell is scientifically plausible. Hence, the value was retained for analysis. Figure 2 shows the standard deviations for the treatment combinations. It suggests that explant $D$ probably exerts strong influence on the dispersion. It is important to note that it is a small and unbalanced data set and, in general, the asymptotic theory does not apply.

FIGURES 1 and 2 ABOUT HERE

### 4.2 Data Analysis

Model (5) was applied to the apple tissue data set and, to estimate the posterior marginal density, three Markov chains were used, with initial points dispersed across the parameter space.

Vague priors were assumed so as to incorporate the uncertainty for the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In all cases, they were assumed normally distributed, with variance 1000 for the $\boldsymbol{\beta}$ parameters and 100 for the $\boldsymbol{\gamma}$. While the value 100 may appear not sufficiently vague, choosing a larger value tends to jeopardize the convergence of the Markov chain process.

Three chains of size 200,000 were generated. The first 100,000 iterations of each chain were discarded for burn-in purposes. In an effort to minimize the within-chain autocorrelation, each 50th iteration was retained. As a result, the sample size to be used for posterior inference about the parameters is 6,000. The models with less parameters have shown faster convergence of the Markov chains, indicating a better mix throughout the parameter space. The final model took 281 seconds under a Intel Core Duo processor on a 1.66 GHz personal computer. The R code using the `BRugs` and `CODA` libraries are available at the website containing supplementary material, with name `BayesianDGLM.r`. Note that the same model specification can be used with the OpenBUGS/WinBUGS software.

Using the deviance information criteria (DIC), backward model selection was conducted. The most complex model fitted was

$$\begin{aligned} \eta &= \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}, \\ \zeta &= \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}. \end{aligned}$$

Table 2 presents each fitted model, the deviance information criteria, and their components $\mathbf{pD}$ and $\bar{\mathbf{D}}$. Models 1 to 4 describe the search for a parsimonious linear predictor of dispersion. Model 5 minimized the DIC index; it features a linear predictor for the dispersion using the dummy variable for the explant $D$; for the mean it incorporates main effects of `medium`, `explant`, and the interaction term of the culture medium $Y$ and the explant $E$. Model 6 has

a constant term only in the dispersion linear predictor. Model 7 has only a random effect $\delta$ with distribution $\mathrm{N}(0, 1)$, and Model 8 is an ordinary binomial GLM without random effects.

In Model 5, the non-significant effects in dispersion model, related with explants B, C, and E, were dropped, in view of reducing the DIC. In Model 3, these non-significant terms were kept and while a consistent picture emerges in terms of pD and the number of parameters, the DIC has deteriorated.

A strong overdispersion effect is apparent for Model 8. Models 9, 10, and 11 are nested for the linear predictor for the mean; they fit worse than Model 5.

TABLES 2 and 3 ABOUT HERE

Model checking plots are not usually presented in the Bayesian literature, where the focus is often directed towards MCMC diagnostics. However, some applied manuscripts present a type of residual analysis that mimics the frequentist approach, thereby considering some summary measure of the sampled marginal posterior for deviance residuals and predicted values (Robinson et al., 2009). Using the mean of sampled marginal posteriors, some graphical analysis were done. Figure 3 exhibits the Q-Q plot and the standardized residuals versus the predicted values, which do not indicate departure from normality or any systematic pattern that could suggest lack-of-fit. The 95% credibility interval plot for the standardized residual posterior samples, displayed in Figure 4, does indicate neither outliers nor extreme values, providing support for inference over the parameter estimates.

Table 3 presents the posterior summary. Using the median as a point estimate for the model parameters, it can be seen that explant $D$ increases the variance of the random effect $\boldsymbol{\delta}$ by $\exp(-6.624 + 7.533) = 2.48$ units, while the other type of explants reduce the variance of the said random effect to almost zero. We conclude that explant $D$ is responsible for the overdispersion in the proportion of regenerated explants. Turning to the factorial effects on the mean of the regenerated explants, the larger influence owes to the interaction between culture medium $Y$ and explant $E$. Whenever this combination occurs, the chance of explant regeneration increases $\exp(2.185) = 8.89$-fold, when compared to the other explants and culture medium $X$.

FIGURES 3 and 4 ABOUT HERE

# 5 Concluding Remarks

The objective of this work was to propose a Bayesian double generalized linear model for overdispersed proportion data, thereby providing an alternative to the frequentist approach of Smyth (1989) and Nelder & Lee (1991). Our model includes a normally distributed random effect in the linear predictor of the generalized linear model, where the variance of the random effect is linked non-linearly to another linear predictor. The model was successfully applied to an agricultural data set of a type frequently encountered. Evidently, it can be applied to related situations too. Experiments with the aim of identifying factors affecting the variability in industrial processes, such as Taguchi experiments, can be analyzed using this approach as well.

The Bayesian approach adopted here and based on stochastic simulation is a very convenient and flexible mode for fitting our class of models. It conveniently extends to other such situations, without limitation to the exponential family of distributions. The use of directed acyclic graphs makes easier the presentation of the model and suggests the hierarchical construction of the probabilistic model, based on conditioning the random values on their parent random parameters. The use of MCMC algorithms efficiently deals with the problem of calculating high-dimensional integrals, thus allowing to generate samples of the posterior marginal

densities of the associated parameters. Using such samples, summary statistics are calculated that render feasible inferences about the fitted model. The price to pay is the method's computational intensity, needing high-quality computational resources. Furthermore, the additional step of diagnosing the convergence of the Markov chain guarantee the coherence of the inferences and has to be taken seriously. The statistical computing environment `R`, jointly with the libraries `BRugs` and `coda` have shown quite flexible and efficient to the estimation process and data analysis.

The proposed modeling framework can be extended in various ways, including the incorporation of the time dimension when measurements are taken longitudinally, as well as particular implementations or count and time-to-event data. Also, the assumption of normal prior distributions can be modified or relaxed. Here, they have shown to be reasonably insensitive to variations of the hyper-parameters for the mean linear predictor, even though they are sensitive to the choice of vague priors for the parameters in the linear predictor of the random effect variance; this calls for careful illicitation, preferably with the help of substantive researchers. Finally, identifiability, parameterization and tests with other prior densities need further research.

## Acknowledgments

The first and third authors were partially supported by The Brazilian National Council for Scientific and Technological Development (CNPq). Part of this work was developed during the sandwich scholarship program of the first author, funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), processo BEX 4344/07-3. We thank Dr. Rosângela Loschi (UFMG) for the useful discussion of ideas. The fourth author gratefully acknowledges grant #P6/03 of the Belgian Government (Belgian Science Policy).

## 6 References

1. Aitkin, M. (1987) Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, **36**, 332–339.

2. Best, N. and Green, P. (2005) Structure and uncertainty: graphical models for understanding complex data. *Significance*, 177–181.

3. Borgatto, A.F., Demétrio, C.G.B., and Leandro, R.A. (2006) Modelos para proporões com superdispersão e excesso de zeros — um procedimento Bayesiano. Revista de Matemática e Estatística, **24** 121–131.

4. Box, G. (1988) Signal-to-noise ratios, performance criteria and transformations. Technometrics, 30:1, p.1–40.

5. Box, G.E.P., Cox, D.R. (1964) An analysis of transformations. Journal of the Royal Statistical Society, B, 26, p. 211-52.

6. Cepeda, E. and Gamerman, D. (2000) Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, **14**, 207–221.

7. Cox, D.R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society B*, **49**, 1–39.

8. Gamerman, D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57–68.

9. Gamerman, D. and Lopes, H. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall.

10. Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginals. *Journal of the American Statistical Association*, **87**, 523–532.

11. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2000) *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC.

12. Givens, G.H. and Hoeting, J.A. (2005) *Computational Statistics*. Hoboken: John Wiley & Sons.

13. Harvey, A.C. (1976) Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, **44**, 461–465.

14. Heidelberger,P. and Welch, P.D. (1983) Simulation run length control in presence of a initial transient. *Operations Research*, **31**, 1109–1144.

15. Hinde, J. and Demétrio, C.G.B. (1998a) Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, **27**, 151–170.

16. Hinde, J. and Demétrio, C.G.B. (1998b) *Overdispersion: Models and Estimation.* In: Simpósio Nacional de Probabilidade e Estatística, Caxambú: Associaão Brasileira de Estatística **13**.

17. Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society, Series B*, **49,** 127–162.

18. Lee, Y. and Nelder, J.A. (1998) Generalized linear models for the analysis of quality-improvement experiments. *The Canadian Journal of Statistics*, **26**, 95–105.

19. Lee, Y. and Nelder, J.A. (2006) Double hierarchical generalized linear models. *Applied Statistics*, **55**, 139–185.

20. Lee, Y., Nelder, J.A., and Pawitan Y. (2006) *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman & Hall/CRC.

21. McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models ( 2nd ed.)* London: Chapman & Hall.

22. Myers, R.H., Khuri, A.I., Vining, G. (1992) Response surface alternatives to the Taguchi robust parameter design approach. *The American Statistician*, 46, 2, 131-139.

23. Neal, R.M. (2003) Slice sampling. *The Annals of Statistics*, **31**, 705–767.

24. Nelder, J.A. and Lee, Y. (1991) Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, **7**, 107–120.

25. Nelder, J.A. and Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

26. Paulino, C.D., Turkman, M.A.A., and Murteira, B. (2003) *Estatística Bayesiana*. Lisboa: Fundaão Calouste Gulbenkian.

27. Plummer, M., Best, N., Cowles, K., and Vines, K. (2007) *coda: Output Analysis and Diagnostics for MCMC*. R package version 0.12-1.

28. Raftery, A.E. and Lewis, S.M. (1992) One long run with diagnostics: implementations strategies for Monte Carlo Markov Chains. *Statistical Science*, **7**, 493–497.

29. R Development Core Team. (2007) *R: A Language and Environment for Statistical Computing*, version 2.6.0. Vienna: R Foundation for Statistical Computing.

30. Ridout, M. and Demétrio, C.G.B. (1992) Generalized linear models for positive count data. *Revista de Matemática e Estatística*, **10**, 139–148.

31. Robinson, T.J., Anderson-Cook, C.M., Hamada, M.S. (2009) Bayesian analysis of split-plot experiment with nonnormal response for evaluating nonstandard performance criteria. Technometrics, 51:1, p. 56–65.

32. Smyth, K. (1989) Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society Series B*, **51**, 47–60.

33. Smyth, G.K. and Verbyla, A.P. (1999) Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, **10**, 696–709.

34. Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996) *BUGS 0.5: Bayesian Inference Using Gibbs Sampling* — Manual (version ii). Cambridge: Medical Research Council Biostatistics Unit.

35. Spiegelhalter, D., Best, N.G., Carlin, B.P., and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

36. Taguchi, G. (1985) Quality engineering in Japan. *Communication in Statistics: Theory and Methods*, **14**, 2785–2801.

37. Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006) Making BUGS open. *R News*, **6**, 12–17.

38. Vieira, A.M.C., Hinde, J.P., and Demétrio, C.G.B. (2000) Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27**, 373–389.

39. Wolfinger, R.D. and Tobias, R.D. (1998) Joint estimation of location, dispersion and random effects in robust design. *Technometrics*, **40**, 62–70.

## Appendix

The posterior density function for model (5) can be built using (6). The prior density functions for $\beta_j$ and $\gamma_k$ are, respectively

$$p(\beta_j) = \sqrt{\frac{c^{-1}}{2\pi}} \exp\left[-\frac{c^{-1}}{2}(\beta_j - b)^2\right] \propto \exp\left[-\frac{c^{-1}}{2}(\beta_j - b)^2\right], \qquad j = 0, \ldots, r \tag{11}$$

and

$$p(\gamma_k) = \sqrt{\frac{e^{-1}}{2\pi}} \exp\left[-\frac{e^{-1}}{2}(\gamma_k - d)^2\right] \propto \exp\left[-\frac{e^{-1}}{2}(\gamma_k - d)^2\right], \qquad k = 0, \ldots, s. \tag{12}$$

Assuming that $Y_i \sim \text{Bin}(m_i, p_i)$ and that $\ln[p_i/(1-p_i)] = \mathbf{x}_i^T\boldsymbol{\beta} + \delta_i$, the likelihood function is

$$\begin{aligned} p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \delta_i) &= \binom{m_i}{y_i} \left[\frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)}\right]^{y_i} \left[1 - \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)}\right]^{m_i - y_i} \\ &= \binom{m_i}{y_i} \frac{[\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{y_i}}{[1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{m_i}}, \end{aligned}$$

and

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \delta_i) \propto [\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{y_i} [1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{-m_i}. \tag{13}$$

The conditional density of $\delta_i$, given the parameter vector $\boldsymbol{\gamma}$, is

$$\begin{aligned} p(\delta_i|\boldsymbol{\gamma}) &= \sqrt{\frac{1}{2\pi \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}} \exp\left\{-\frac{1}{2}[\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})]^{-1}(\delta_i - a)^2\right\} \\ &= (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{z}_i^T\boldsymbol{\gamma} - \frac{1}{2}\exp(-\boldsymbol{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2\right\} \end{aligned}$$

and, therefore,

$$p(\delta_i|\boldsymbol{\gamma}) \propto \exp\left\{-\frac{1}{2}[\boldsymbol{z}_i^T\boldsymbol{\gamma} + \exp(-\boldsymbol{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2]\right\}. \tag{14}$$

Applying (11)–(14) to (6), considering the vector of observations $\mathbf{y}$, it can be shown that the posterior joint probability density function is

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}|\mathbf{X}, \mathbf{Z}, \mathbf{y}) \propto & \prod_{i=1}^{n} \frac{[\exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{y_i}}{[1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{m_i}} \exp\left\{-\frac{1}{2}[\mathbf{z}_i^T\boldsymbol{\gamma} + \exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2]\right\} \times \\ & \times \prod_{j=0}^{r} \exp\left[-\frac{c^{-1}}{2}(\beta_j - b)^2\right] \prod_{k=0}^{s} \exp\left[-\frac{e^{-1}}{2}(\gamma_k - d)^2\right] \\ \propto & \exp\left\{\sum_{i=1}^{n} \mathbf{x}_i^T\boldsymbol{\beta} y_i + \sum_{i=1}^{n} y_i\delta_i - \frac{1}{2}\sum_{i=1}^{n} \mathbf{z}_i^T\boldsymbol{\gamma} - \frac{1}{2}\sum_{i=1}^{n} \exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2 - \right. \\ & \left. - \frac{1}{2c}\sum_{j=0}^{r}(\beta_j - b)^2 - \frac{1}{2e}\sum_{k-0}^{s}(\gamma_k - d)^2\right\} \prod_{i=1}^{n}[1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{-m_i} \end{aligned}$$

and, finally,

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}|\mathbf{X}, \mathbf{Z}, \mathbf{y}) \propto \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}) - \frac{1}{2}\mathbf{1}^T\mathbf{Z}\boldsymbol{\gamma} - \sum_{i=1}^{n} \exp(-\mathbf{z}_i^T\boldsymbol{\gamma})(\delta_i - a)^2 -\right.$$

$$-\frac{1}{2c}\sum_{j=0}^{r}(\beta_j - b)^2 - \frac{1}{2e}\sum_{k=0}^{s}(\gamma_k - d)^2\Bigg\} \times$$
$$\times \prod_{i=1}^{n}[1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \delta_i)]^{-m_i}. \tag{15}$$

**Table 1:** Number of explants ($y_i$) of apple trees that regenerated, considering 16 Petri dishes. SOURCE: Hinde & Demétrio (1998b).

| | | Explant Variety | | | | |
|---|---|---|---|---|---|---|
| | | A | B | C | D | E |
| Culture Medium | X | 8, 10 | 9, 10 | 7 | 20, 12 | 2 |
| | Y | 9, 11 | 18, 12 | 13 | 20, 5 | 13 |

**Table 2:** Quality-of-fit and model complexity measures for fitted Bayesian binomial DGLM models.

| **Fitted Models** | | **DIC** | **pD** | $\mathsf{E}_\theta[D(\theta)]$ |
|---|---|---|---|---|
| Model 1: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}$<br>$\zeta = \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}$ | 86,31 | 13,15 | 73,16 |
| Model 2: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}$<br>$\zeta = \texttt{medium} + \texttt{explant}$ | 86,10 | 13,44 | 72,66 |
| Model 3: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}$<br>$\zeta = \texttt{explant}$ | 84,69 | 12,79 | 71,9 |
| Model 4: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium} \times \texttt{explant}$<br>$\zeta = \gamma_0 + \texttt{explant D}$ | 84,49 | 18,81 | 71,67 |
| Model 5: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium Y} \times \texttt{explant E}$<br>$\zeta = \gamma_0 + \texttt{explant D}$ | 82,56 | 11,00 | 71,56 |
| Model 6: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium Y} \times \texttt{explant E}$<br>$\zeta = \gamma_0$ | 87,16 | 14,52 | 72,64 |
| Model 7: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium Y} \times \texttt{explant E}$<br>$\zeta = 0 \Rightarrow \boldsymbol{\delta} \sim \mathsf{N}(\mathbf{0}, \mathbf{1})$ | 86,16 | 14,77 | 71,39 |
| Model 8: | $\eta = \texttt{medium} + \texttt{explant} + \texttt{medium Y} \times \texttt{explant E}$ | 107,10 | 7,03 | 100,10 |
| Model 9: | $\eta = \texttt{medium} + \texttt{explant}$<br>$\zeta = \gamma_0 + \texttt{explant D}$ | 85,77 | 11,19 | 74,58 |
| Model 10: | $\eta = \texttt{explant}$<br>$\zeta = \gamma_0 + \texttt{explant D}$ | 88,66 | 14,37 | 74,29 |
| Model 11: | $\eta = \texttt{medium}$<br>$\zeta = \gamma_0 + \texttt{explant D}$ | 85,18 | 8,82 | 76,30 |

**Table 3:** Model 5: posterior summary for the parameters.

| Parameters | Mean | Standard Deviation | Standard Error | IC(2.5%) | Median | IC(97.5%) |
|---|---|---|---|---|---|---|
| $\beta_0$ | −0.8034 | 0.2784 | 0.004268 | -1.34200 | -0.80550 | -0.2595 |
| $\beta_Y$ | 0.5925 | 0.2900 | 0.003788 | 0.02434 | 0.59190 | 1.1540 |
| $\beta_B$ | 0.4657 | 0.3246 | 0.004614 | -0.17570 | 0.46720 | 1.0840 |
| $\beta_C$ | 0.0795 | 0.4101 | 0.004761 | -0.75690 | 0.08712 | 0.8642 |
| $\beta_D$ | 0.8618 | 1.1030 | 0.019380 | -1.21000 | 0.84280 | 3.2070 |
| $\beta_E$ | −1.8840 | 0.9127 | 0.011330 | -3.92100 | -1.80400 | -0.3263 |
| $\beta_{Y,E}$ | 2.1850 | 1.0190 | 0.013750 | 0.41940 | 2.09300 | 4.3770 |
| $\gamma_0$ | −6.6240 | 3.9760 | 0.170800 | -15.79000 | -5.84400 | -1.1140 |
| $\gamma_D$ | 7.5330 | 4.0870 | 0.171400 | 1.46700 | 6.75600 | 16.8100 |

Summary statistics calculated for 6000 observations.

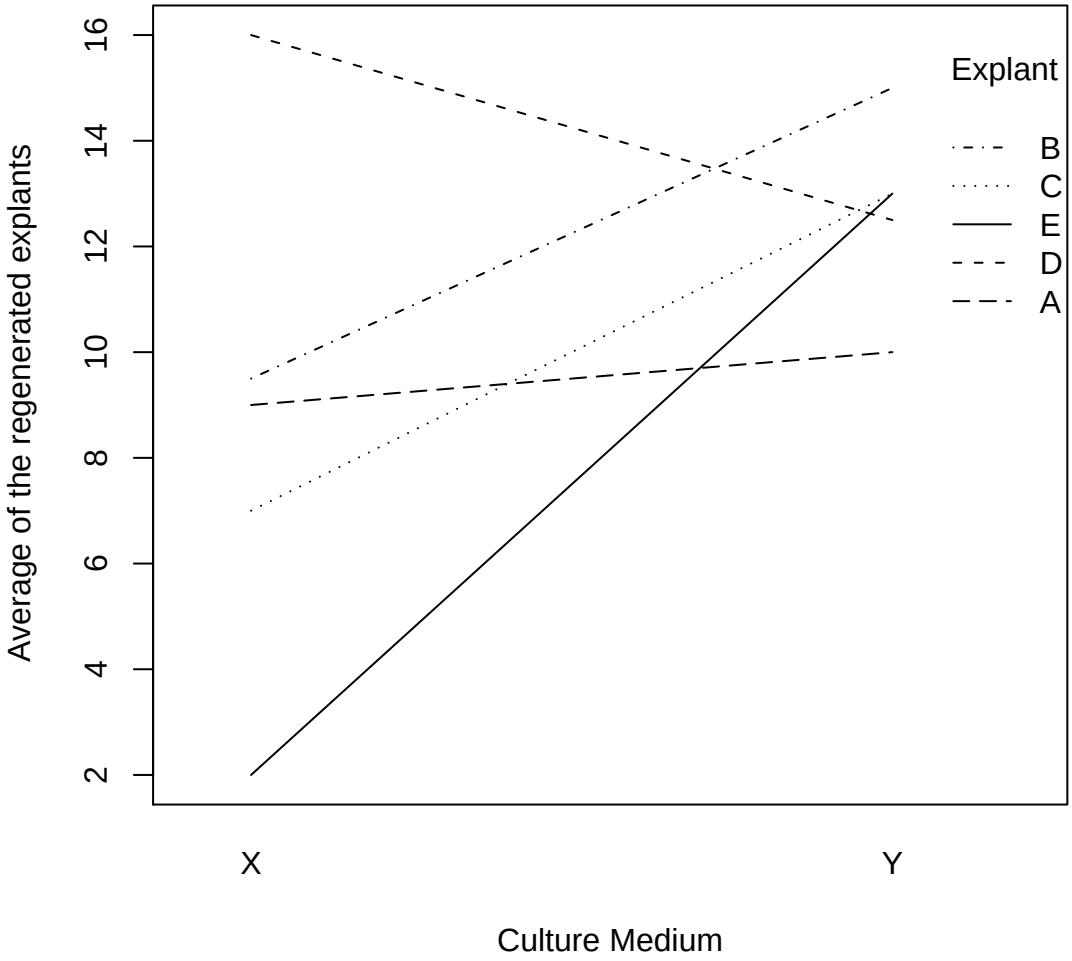


**Figure 1:** Average number of regenerated explants as function of the culture media and the explant varieties.

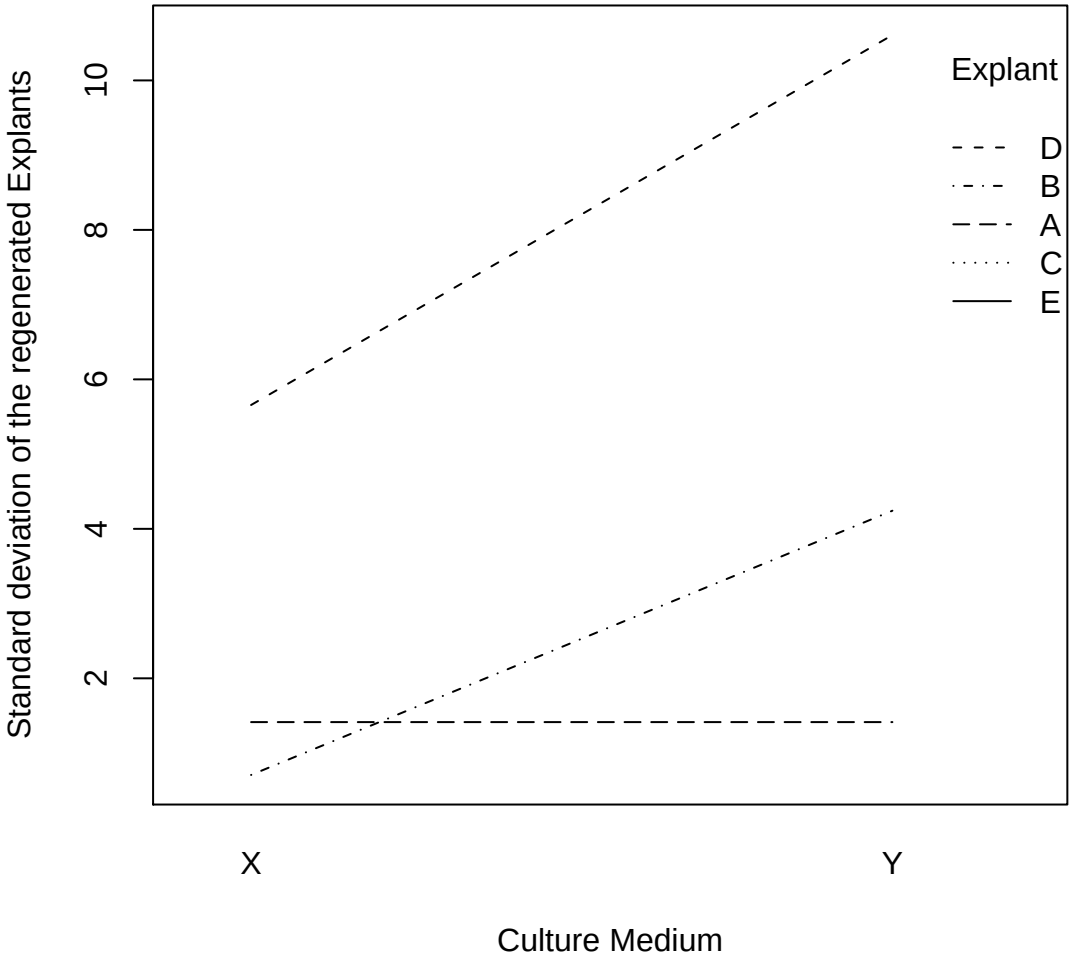


**Figure 2:** Standard deviation of the number of regenerated explants as function of the culture media and the explant varieties.

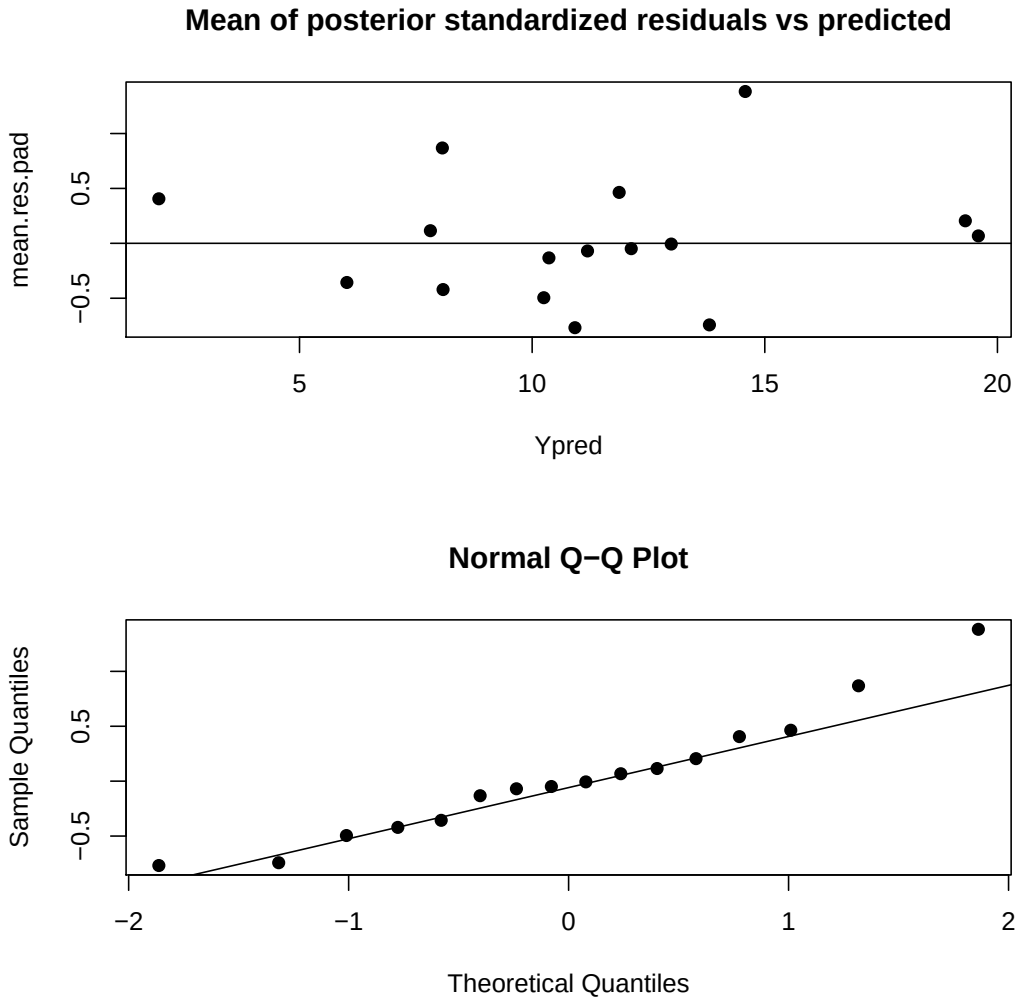


**Figure 3:** Residual analysis for Model 5. The top plot is the mean for samples of marginal posteriors for the standardized residuals versus the mean of posterior marginal samples of predicted values; the bottom plot is the normal probability plot for the mean marginal posterior of the standardized residuals.

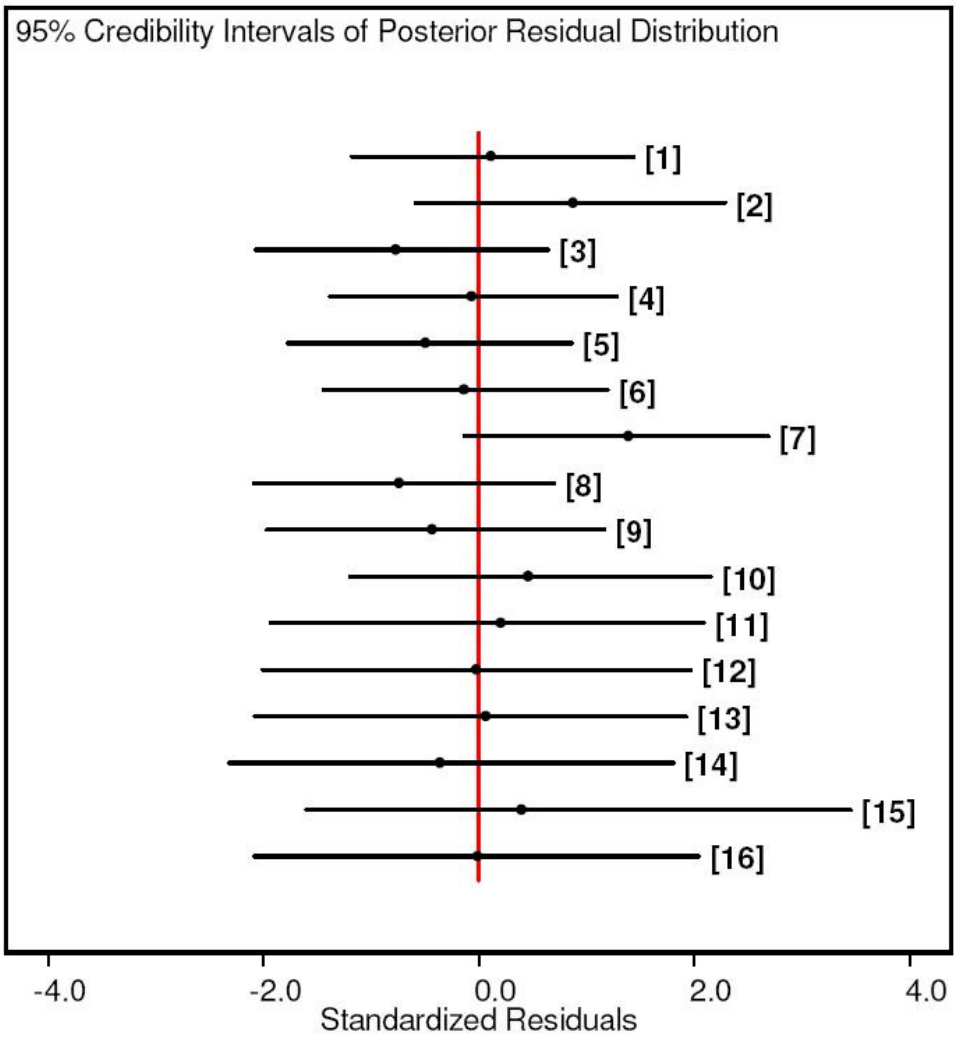


**Figure 4:** 95% credibility interval plots for marginal posterior samples of standardized residuals.