

## Summer school “Likelihood-based inference for hierarchical/mixed statistical models”

*Tune Landboskole, Greve, Denmark, August 7-18, 2005*

### Project

## The germination of seeds from different species of trees

*Zhanna Andrushchenko*

Department of Biometry and Engineering, SLU (Swedish University of Agricultural Sciences), Uppsala, Sweden

### Introduction

An experiment has been made\* to study the germination of seeds from different species of trees. Data consist of repeated measurements of germinated seeds at times 1, 2, 3, 4 or 8 weeks after planting. The experiment was performed in a climate chamber where external factors can be kept on a reasonably constant level.

Each line in the data set contains data from one box. At the start of the experiment, 100 seeds were planted in each box. At different points in time, the number of seeds that had germinated, was counted in each box, i.e. how many living plants there were in the box.

### Description of data

#### *Hierarchical levels:*

<b>species</b>	Species of the tree (1, 2, 3 or 4)
<b>box</b>	The seeds were planted in boxes in two rounds (because of lack of space in the climate chamber; these can be regarded as blocks (1 or 2))
<b>week</b>	The measurements were done at 1, 2, 3, 4 and 8 weeks after planting

#### *Key dependent variable*

**number of seeds** that had germinated in the box at times 1, 2, 3, 4 or 8 weeks after planting.

#### *Key independent variable*

<b>temperature</b>	Temperature at which the box was kept (10 or 25 degrees Celsius)
<b>treatment</b>	A certain treatment the seeds were exposed to before planting (1 or 2)

#### **Some special problems with the data:**

1. In some cases, the number of germinated seeds decreases with time. This is because plants that were alive at one date have died before the next count.
2. In some boxes, the highest number of live plants is larger than 100, which should be impossible. We will assume that the number of planted seeds in these boxes is equal to the highest number of live plants over the whole experiment.

#### **Purpose of the study**

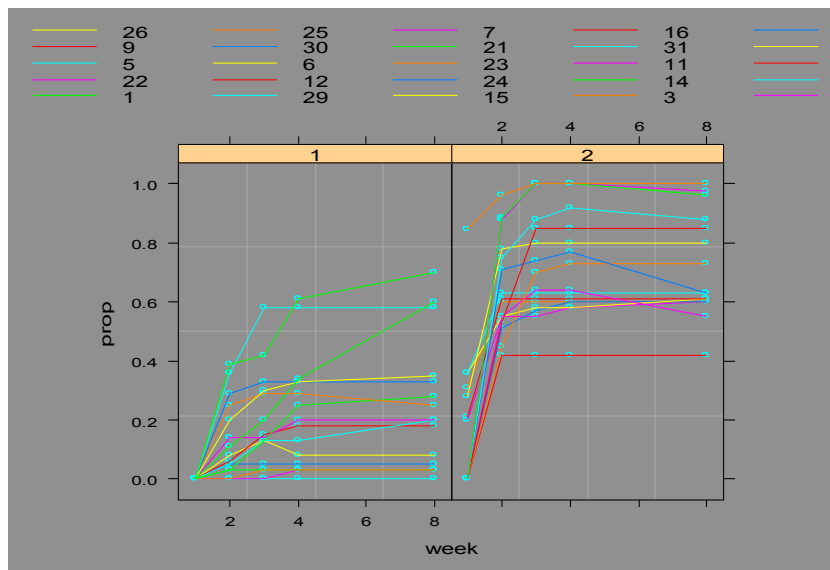
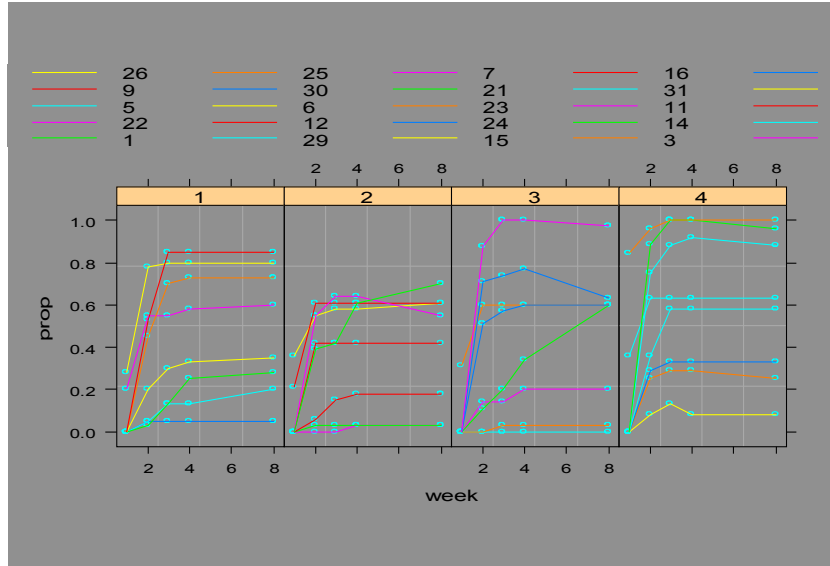
Analyze these data to investigate the effect of different factors on the germination of the seeds.

\* Data are from the course “Repeated measurements”, given by Dr. Ulf Olsson (SLU, Spring 2005)

## Results of studies

The data under analysis are typical repeated measures data, that consists of series of measures on the same observational unit (box) over time (week 1, 2, 3, 4, 8). Hierarchical levels consist of species, boxes and time points (weeks). However, as far as there were only 4 species, they were not treated as hierarchical level, but rather as a predictor.

Simple plots of data show that the germination of seeds can be affected by different factors (temperature, treatment, etc):



The data are also clearly correlated:

### Correlations

	week1	week2	week3	week4	week8
week1	1.0000				
week2	0.5100	1.0000			
week3	0.4138	0.9730	1.0000		
week4	0.3919	0.9606	0.9902	1.0000	
week8	0.4062	0.9234	0.9587	0.9830	1.0000

As far as data consists of number of plants germinated out of planted ones, the corresponding distribution should be **Binomial** with **logit** as link function.

The number of responses  $y$  was modelled as a binomial random variable for each combination of the explanatory variables, with the number of trials parameter equal to the number of subjects  $n$  and the binomial probability equal to the probability of a response. A logistic regression for these data is a GLM with response equal to the binomial proportion  $y/n$ . The number of trials,  $n$ , was chosen as  $n=100$  except for those units where the observed number is larger than 100; for these boxes, the largest observed number was used instead of  $n$ .

Analysis was carried out using MlwiN. There are few approaches to study such a data, here analysis is limited to linear mixed models and multivariate response models. In models 1 and 2, the analysis has been done assuming one “response” variable and treating other variables as predictors. For models 3 and 4, several variables were identified as responses, and the model was fitted to each response variable.

### 1. Generalized linear mixed model (Binomial response)

The model has been chosen using Bayesian Deviance Information Criterion (DIC): the model with a random intercept and random slope for time had the lowest DIC. It has been found that intercept, treatment and time effect are all highly significant. On the other hand, effect of species, blocks, temperatures as well as interactions were found to be not significant. However, not all chains look reasonable even after 100,000 iterations; plots of residuals do not look nice too.

### 2. General linear mixed model (Normal response)

Data that follow Binomial distribution, can be transformed to follow Normal distribution:

$$x \sim \text{Bin}(n, p) \Leftrightarrow y = \text{asin}(\sqrt{x/n}) \sim N(0, 1)$$

The model has been chosen using Likelihood Ratio test (LR). As iteration methods, **IIGLS**, **RIGLS** and **MCMC** were used. It has been found again that intercept, treatment and time effect have practically the same values for all iteration methods and are all highly significant, whereas other effects are not. All chains look reasonable even after 10,000 iterations. Surprisingly for such a simple model, plot of residuals looks quite OK.

### 3. Multivariate Binomial response model

Attempt to fit multivariate model (with Binomial responses) was unsuccessful.

### 4. Multivariate Normal response model

The data were again transformed to follow Normal distribution. Using Likelihood Ratio test (LR) the model with a random intercept has been chosen. As iteration method, **IIGLS**, **RIGLS** and **MCMC** were used. It has been found that treatment is highly significant for the germination of seeds, effects of intercept and temperature are significant, whereas other effects are not. All chains look reasonable even after 50,000 iterations. Plots of residuals (via IIGLS and MCMC iteration methods) looks reasonable OK for all weeks except the week 1.

## Conclusions

- Analysis based on different models, has shown that the **treatment** the seeds were exposed to before planting has highly significant effect on the germination of seeds; temperature can be significant as well.
- Among the model under consideration, the simple model 3 (General linear mixed model (with Normal response)) seems to fit data properly.
- It should be noted that data are highly correlated, but “within individuals” correlation was not modeled by MlwiN.

# 1. Generalized linear mixed model (Binomial response)

$$\text{prop}_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_{0j}\text{cons} + \beta_{1j}\text{week}_{ij} + \beta_2\text{trt}_j$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

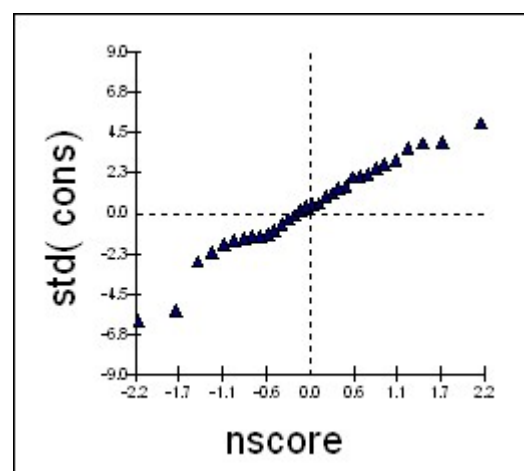
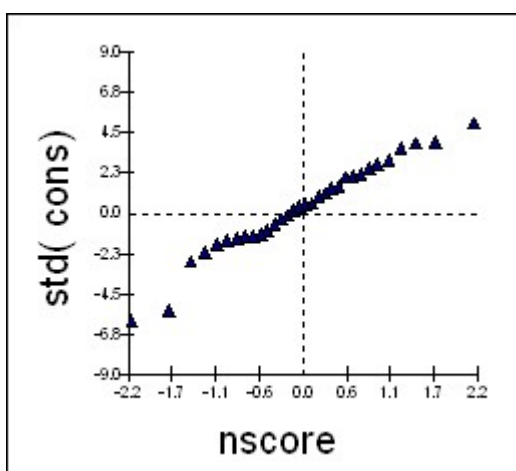
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}$$

$$\text{var}(\text{prop}_{ij} | \pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/n_{ij}$$

*Deviance(MCMC) = 2514.633(160 of 160 cases in use)*

	<i>MCMC (100,000)</i>	<i>P</i>
$\beta$ for cons	-5.333936	0.000
$\beta$ for week	0.3830456	0.000
$\beta$ for trt	2.228406	0.000
<b><math>\sigma^2</math> for cons</b>	<b>1.466892</b>	
$\sigma$ for cons,week	-0.3464975	
<b><math>\sigma^2</math> for week</b>	<b>0.370059</b>	
DIC	2575.79	



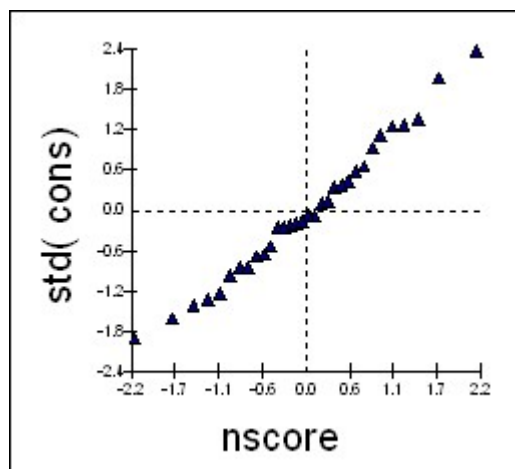
## 2. General linear mixed model (Normal response)

transformation of the data:

$$x \sim \text{Bin}(, ) \Leftrightarrow y = \text{asin}(\sqrt{x}) \sim N(, )$$

$$\begin{aligned} \text{tr\_resp}_{ij} &\sim N(XB, \Omega) \\ \text{tr\_resp}_{ij} &= \beta_{0ij} \text{cons} + \beta_1 \text{week}_{ij} + \beta_2 \text{trt}_j \\ \beta_{0ij} &= \beta_0 + \mu_{0j} + e_{0ij} \\ \begin{bmatrix} \mu_{0j} \end{bmatrix} &\sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} e_{0ij} \end{bmatrix} &\sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{bmatrix} \\ \text{Deviance(MCMC)} &= 39.443(160 \text{ of } 160 \text{ cases in use}) \end{aligned}$$

	<i>IGLS</i>	<i>RIGLS</i>	<i>MCMC (10,000)</i>	<i>P</i>
$\beta$ for cons	-0.4681394	-0.4681389	-0.4733575	0.000
$\beta$ for week	0.06209975	0.06209977	0.06184402	0.000
$\beta$ for trt	0.5653316	0.5653313	0.5680597	0.000
$\sigma^2$ for cons	<b>0.02502852</b>	<b>0.02755995</b>	<b>0.0290016</b>	
$\sigma^2$ for week	<b>0.07317872</b>	<b>0.07375494</b>	<b>0.07551546</b>	
-2*log L	67.588	67.656		
DIC			62.60	



### 3. Multivariate Binomial response model

$$\begin{aligned} \text{resp}_{1j} &\sim \text{Binomial}(\text{denomin}_{1j}, \pi_{1j}) \\ \text{resp}_{2j} &\sim \text{Binomial}(\text{denomin}_{2j}, \pi_{2j}) \\ \text{logit}(\pi_{1j}) &= \beta_0 \text{cons.r\_w1}_{ij} \\ \text{logit}(\pi_{2j}) &= \beta_1 \text{cons.r\_w2}_{ij} \end{aligned}$$

does not work !

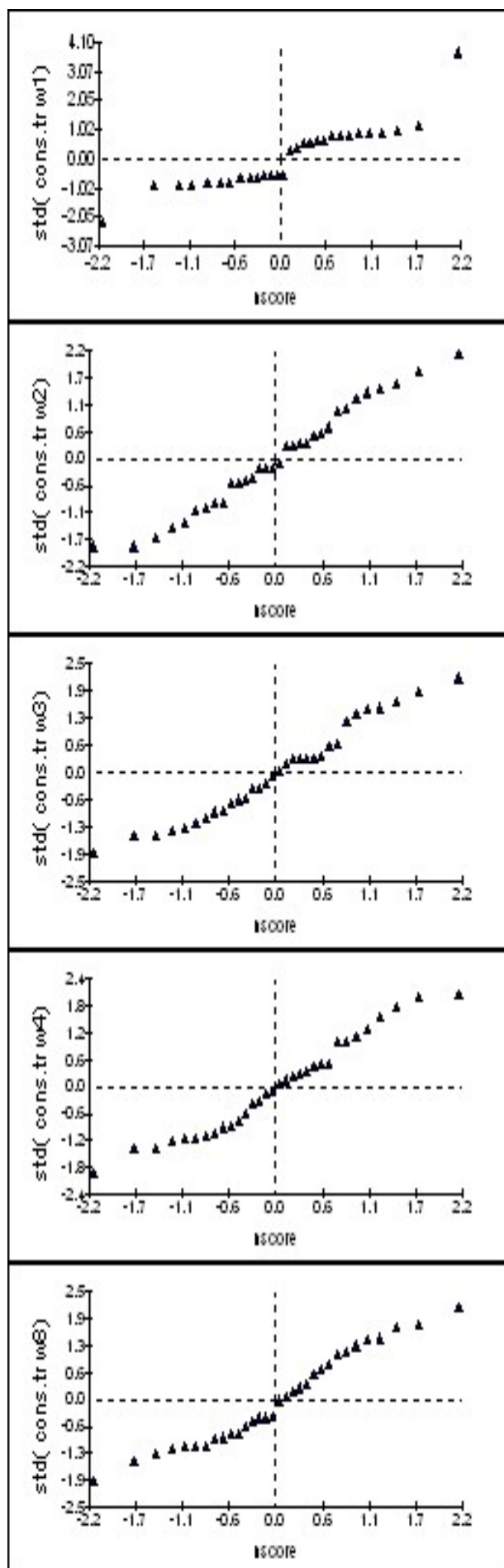
### 4. Multivariate Normal response model

$$\begin{aligned} \text{resp}_{1j} &\sim N(XB, \Omega) \\ \text{resp}_{2j} &\sim N(XB, \Omega) \\ \text{resp}_{3j} &\sim N(XB, \Omega) \\ \text{resp}_{4j} &\sim N(XB, \Omega) \\ \text{resp}_{5j} &\sim N(XB, \Omega) \\ \text{resp}_{1j} &= \beta_{0j} \text{cons.tr\_w1}_{ij} + \beta_5 \text{species.tr\_w1}_{ij} + \beta_{10} \text{temp.tr\_w1}_{ij} + \beta_{15} \text{treatm.tr\_w1}_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \text{resp}_{2j} &= \beta_{1j} \text{cons.tr\_w2}_{ij} + \beta_6 \text{species.tr\_w2}_{ij} + \beta_{11} \text{temp.tr\_w2}_{ij} + \beta_{16} \text{treatm.tr\_w2}_{ij} \\ \beta_{1j} &= \beta_1 + u_{1j} \\ \text{resp}_{3j} &= \beta_{2j} \text{cons.tr\_w3}_{ij} + \beta_7 \text{species.tr\_w3}_{ij} + \beta_{12} \text{temp.tr\_w3}_{ij} + \beta_{17} \text{treatm.tr\_w3}_{ij} \\ \beta_{2j} &= \beta_2 + u_{2j} \\ \text{resp}_{4j} &= \beta_{3j} \text{cons.tr\_w4}_{ij} + \beta_8 \text{species.tr\_w4}_{ij} + \beta_{13} \text{temp.tr\_w4}_{ij} + \beta_{18} \text{treatm.tr\_w4}_{ij} \\ \beta_{3j} &= \beta_3 + u_{3j} \\ \text{resp}_{5j} &= \beta_{4j} \text{cons.tr\_w8}_{ij} + \beta_9 \text{species.tr\_w8}_{ij} + \beta_{14} \text{temp.tr\_w8}_{ij} + \beta_{19} \text{treatm.tr\_w8}_{ij} \\ \beta_{4j} &= \beta_4 + u_{4j} \end{aligned}$$

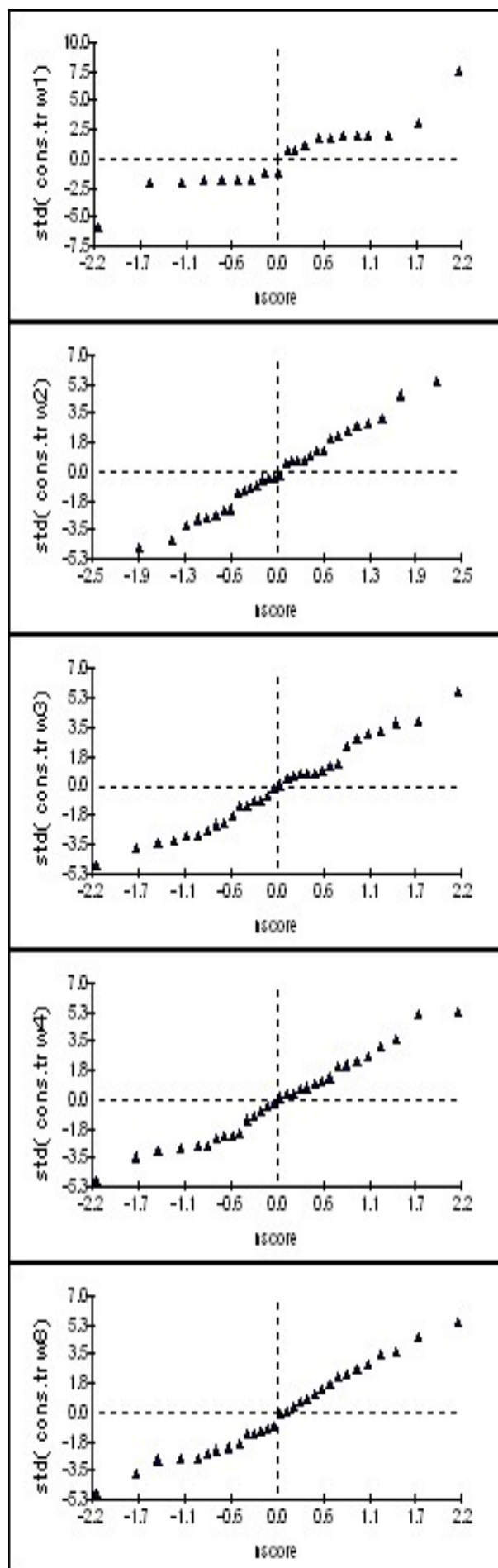
$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & & & & \\ \sigma_{u01} & \sigma_{u1}^2 & & & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 & & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma_{u3}^2 & \\ \sigma_{u04} & \sigma_{u14} & \sigma_{u24} & \sigma_{u34} & \sigma_{u4}^2 \end{bmatrix}$$

-2\*loglikelihood(IGLS Deviance) = -301.332(160 of 160 cases in use)

### IGLS



### MCMC (50,000)



	<i>IIGLS</i>	<i>RIGLS</i>	<i>MCMC</i> (5,000)	<i>MCMC</i> (50,000)	<i>P</i>	<i>P</i>
$\beta$ for cons (week 1)	-0.67269	-0.67263	-0.67525	-0.67330	0.000	0.167
$\beta$ for cons (week 2)	-0.68147	-0.68195	-0.67953	-0.68234	0.000	
$\beta$ for cons (week 3)	-0.55883	-0.55957	-0.55693	-0.56037	0.003	
$\beta$ for cons (week 4)	-0.45661	-0.45741	-0.45634	-0.45797	0.019	
$\beta$ for cons (week 8)	-0.40033	-0.40112	-0.40061	-0.40166	0.037	
$\beta$ for species (week 1)	0.023014	0.023008	0.023528	0.022927	0.476	0.377
$\beta$ for species (week 2)	0.079109	0.079145	0.079246	0.079393	0.003	
$\beta$ for species (week 3)	0.085247	0.085303	0.085071	0.085556	0.021	
$\beta$ for species (week 4)	0.074016	0.074077	0.073904	0.074325	0.055	
$\beta$ for species (week 8)	0.054944	0.055003	0.054562	0.055186	0.149	
$\beta$ for temp (week 1)	0.018928	0.018927	0.019003	0.018952	0.000	0.006
$\beta$ for temp (week 2)	0.008571	0.008577	0.008537	0.008578	0.030	
$\beta$ for temp (week 3)	0.003292	0.003301	0.003228	0.003296	0.549	
$\beta$ for temp (week 4)	0.003713	0.003722	0.003687	0.003704	0.519	
$\beta$ for temp (week 8)	0.008751	0.008760	0.008764	0.008745	0.123	
$\beta$ for trt (week 1)	0.28391	0.28389	0.28440	0.28392	0.000	0.003
$\beta$ for trt (week 2)	0.64181	0.64199	0.64019	0.64179	0.000	
$\beta$ for trt (week 3)	0.67918	0.67946	0.67812	0.67960	0.000	
$\beta$ for trt (week 4)	0.64455	0.64485	0.64389	0.64499	0.000	
$\beta$ for trt (week 8)	0.57673	0.57702	0.57671	0.57723	0.000	
$\sigma^2$ for week 1	<b>0.041790</b>	<b>0.047761</b>	<b>0.058264</b>	<b>0.060230</b>		
$\sigma$ for week 1,2	-0.000237	-0.000266	-0.000274	-0.000264		
$\sigma$ for week 1,3	0.027962	0.031829	0.039181	0.040497		
$\sigma$ for week 1,4	-0.003640	-0.004153	-0.005121	-0.005162		
$\sigma$ for week 1,8	0.035943	0.040887	0.050421	0.052073		
$\sigma^2$ for week 2	<b>0.054295</b>	<b>0.061769</b>	<b>0.076275</b>	<b>0.078678</b>		
$\sigma$ for week 2,3	-0.005063	-0.005779	-0.007032	-0.007200		
$\sigma$ for week 2,4	0.036967	0.042045	0.051805	0.053532		
$\sigma$ for week 2,8	0.055408	0.063022	0.077760	0.080257		
$\sigma^2$ for week 3	<b>0.059574</b>	<b>0.067761</b>	<b>0.083537</b>	<b>0.086261</b>		
$\sigma$ for week 3,4	-0.003451	-0.003937	-0.004780	-0.004893		
$\sigma$ for week 3,8	0.033836	0.038470	0.047462	0.049012		
$\sigma^2$ for week 4	<b>0.052115</b>	<b>0.059265</b>	<b>0.073218</b>	<b>0.075512</b>		
$\sigma$ for week 4,4	0.057227	0.065086	0.080352	0.082894		
$\sigma^2$ for week 8	<b>0.057899</b>	<b>0.065861</b>	<b>0.081374</b>	<b>0.083886</b>		
-2*log L	-301.332	-299.995	-	-		
DIC	-	-	-235.51	-235.84		