Posterior F-Value In Bayesian Analysis Of Variance Using Winbugs

Milton K. Koech,[1*] Arwings R. Otieno[2], Victor Kimeli[3], Eliud K. Koech[4].
Department of Mathematics and Computer Science
University of Eldoret,
P O Box 1125,
Eldoret, Kenya.
*Email:koechmilton@yahoo.com

ABSTRACT

Analysis of variance (ANOVA) is a standard method for describing and estimating heterogeneity among the means of a response variable across the levels of multiple categorical factors. In most experimental settings, ANOVA is used to test the presence of treatment effects. Frequentist approaches to making inferences about the variances of random cluster effects in hierarchical generalized linear models (HGLMs) have several limitations. These include reliance on asymptotic theory, questionable properties of classical likelihood ratio tests when pseudo-likelihood methods are used for estimation, and a failure to account for uncertainty in the estimation of features of prior distributions for model parameters. This paper compares and contrasts alternative approaches to making a specific type of inference about the variance components in an HGLM, focusing on the difference in the variance components. A Bayesian approach to making inferences about these types of differences is proposed that circumvents many of the problems associated with alternative frequentist approaches.Bayesian hypothesis testing literature on ANOVA is scant; the dominant treatment is still classical or frequentist. One impediment to adoption of Bayesian approach is lack of practical development, particularly a lack of ready-to-use formulas and algorithms. Markov Chain Monte Carlo (MCMC) and Gibbs sampling are used to obtain posterior point estimates from these posterior distributions. The 95% credible intervals (CI) were also obtained. Posterior F-values were obtained for the different priors and finally compared with that obtained using classical approach. The Bayesian test for ANOVA designs is useful to both researchers and students; both groups will get to appreciate the importance of Bayesian approach when applied to practical statistical problems.

Key Words: Bayesian Analysis of Variance, Variance Components, Hierarchical Generalized Linear Models, Posterior F-value, ANOVA.

1. Introduction

In many social science settings, the data available for analysis span multiple groups. In these settings it is often plausible that any statistical model that might fit to the data need to be flexible, so as to capture variation across the groups, typically accomplished by letting some or all of the parameters vary across the groups. Examples include survey data gathered over a set of locations (e.g., states, districts, countries); experimental studies deployed in multiple locations; studies of educational outcomes where the subjects are students, who are grouped in classes or schools, which are in school districts, which in turn are in states etc.

This paper considers alternative approach to making inferences about the parameters in a specific class of HGLMs.Frequentist approaches to estimation of HGLMs rely on various numerical or theoretical approaches to approximating complicated likelihood functions, especially for models involving complex random effects structures (e.g., Faraway, 2006; Molenberghs and Verbeke, 2005). In general, inferences based on these approximate likelihood-based approaches, such as residual pseudo-likelihood, penalized quasi-likelihood, and maximum likelihood based on a Laplace approximation, have the same drawback for normal outcomes in that they fail to account for the uncertainty in estimating features of prior distributions for the model parameters (Carlin and Louis, 2009). In addition, frequentist approaches to testing hypotheses about fixed effects or covariance parameters in HGLMs and making inferences about the parameters rely on asymptotic theory and asymptotic results (Zhang and Li, 2010). Molenberghs and Verbeke (2005) argue that likelihood ratio tests should not even be used to test hypotheses when models are fitted using pseudo-likelihood methods. Furthermore, the number of clusters under study may be fairly small in practice, making inferences or tests of hypotheses concerning between-cluster covariance parameters based on asymptotic theory invalid. Approximate maximum likelihood estimation methods can also lead to invalid (i.e., negative) estimates of variance components in these models. Bayesian methods for making inferences about the parameters in HGLMs can provide an attractive solution to these various problems, and this paper considers such methods.

In analysis of data of this type, the researcher is interested with the parameters that vary at each group level. These group level parameters go by different names, in different contexts, in different disciplines, and depending on the estimation method being used. Examples include “contextual effects”, “fixed effects”, “random effects”, and “varying” or “stochastic coefficients”. This between-group parameter variation is potentially of great substantive interest, since it speaks to a fundamental issue in empirical social science. Moreover, group by-group analysis is often an important preliminary step in data analysis: a useful and easily-implemented method for assessing parameter heterogeneity, but one that is often overlooked (Berger, 2006).

A Bayesian approach to making inferences about differences in variance components in this context has several attractive features relative to the frequentist approach. The Bayesian approach would not require asymptotic theory or assumed asymptotic distributions for the test statistics computed in the frequentist approach, would account for the uncertainty in estimating features of prior distributions for model parameters, and would allow analysts to construct credible intervals for the difference between the two variance components based on draws from a posterior distribution for the two variance components (treating the fixed effects and any additional error variances allowing for possible over dispersion in the non-normal responses as nuisance parameters). This paper compares and contrasts these alternative approaches using real data.

2. Methods

Bayesian models deal with the possibility of parameter variation across groups by positioning a model for the parameters above the model for the data. The “hierarchy” then arises because the model for the parameters sits “above” the model for the data. Indeed, in this sense all Bayesian models are hierarchical, in that a prior for θ sits above the model for y, the latter indexed by the parameter θ. This notion of a statistical model as a nested hierarchy of stochastic relations permeates all hierarchical modeling, highlighting why hierarchical models are very amenable to Bayesian analysis. Generically, Bayesian hierarchical statistical models have the form:

$y_j|\theta \sim f(y_j|\theta)$ (model for the data in group j = 1, . . . , J)

$\theta|\upsilon \sim f(\theta|\upsilon)$ (between-group model or “prior” for the parameters θ)

$\upsilon \sim P(\upsilon)$ (prior for the hyper parameters, υ),

Writing the hierarchy from “bottom” to “top” i.e, the model for the parameters is above that of the data. The inferential challenge is to compute the posterior density of all the parameters, $\theta = (\theta_1, . . . ,\theta_J, \upsilon)'$ and any marginal posterior densities for specific elements of θ that are of interest. Markov chain Monte Carlo and Gibbs sampling are extremely well-suited to this task.

2.1 Multiple Regression Framework

In linear multiple regression analysis, the goal is to predict, knowing the measurements collected on N subjects, a dependent variable Y from a set of J independent variables denoted $\{X_1,...,X_j,...,X_J\}$.

We denote by X the N × (J + 1) augmented matrix collecting the data for the independent variables (this matrix is called augmented because the first column is composed only of ones), and by y the N × 1 vector of observations for the dependent variable.
The predicted values of the dependent variables are collected in a vector and are obtained as:

$y = Xb$ with $b = (X^TX)^{-1}X^Ty$. ………………………………………………….…(1)

The vector b has J components. Its first component is traditionally denoted b_0, it is called the intercept of the regression and it represents the regression component associated with the first column of the matrix X. The additional J components are called slopes and each of them provides the amount of change in Y consecutive to an increase in one unit of its corresponding column.
The regression sum of squares is obtained as

$SSregression = b^TX^Ty - \frac{1}{N}(1^Ty)^2$……………………………………………………………(2)

(with 1^T being a row vector of 1’s conformable with y).
The total sum of squares is obtained as

SStotal = $y^Ty - \frac{1}{N}(1^Ty)^2$……………………………………………………………………..(3)

The residual (or error) sum of squares is obtained as
SSerror = $y^Ty - b^TX^Ty$………………………………………………………………………. (4)

The quality of the prediction is evaluated by computing the multiple coefficient of correlation denoted $R^2_{Y.1,...,J.}$ This coefficient is equal to the squared coefficient of correlation between the dependent variable (Y) and the predicted dependent variable (b, Y).

An alternative way of computing the multiple coefficient of correlation is to divide the regression sum of squares by the total sum of squares. This shows that $R^2_{Y.1,...,J}$ can also be interpreted as the proportion of variance of the dependent variable explained by the independent variables. With this interpretation, the multiple coefficient of correlation is computed as

$$R^2_{Y.1,...,J} = \frac{\text{SSregression}}{\text{SSregression} + \text{SSerror}} = \frac{\text{SSregression}}{\text{SStotal}}$$

2.2 Significance test

In order to assess the significance of a given $R^2_{Y.1,...,J,}$ we can compute an F ratio as

$$F = \frac{R^2_{Y.1,...,J}}{1 - R^2_{Y.1,...,J}} \times \frac{N-J-1}{J} \text{………………………………………………………...……….(5)}$$

Under the usual assumptions of normality of the error and of independence of the error and the scores, this F ratio is distributed under the null hypothesis as a Fisher distribution with $\nu_1 = J$ and $\nu_2 = N - J - 1$ degrees of freedom

3 Analysis of variance framework

For an ANOVA, the goal is to compare the means of several groups and to assess if these means are statistically different. For the sake of simplicity, we assume that each experimental group comprises the same number of observations denoted I (i.e., we are analyzing a “balanced design”). So, if we have J experimental groups with a total of K observations per group, we have a total of J × K = N observations denoted $Y_{i,j}$. The first step is to compute the J experimental means denoted α_j and the grand mean denoted μ. The ANOVA evaluates the difference between the means by comparing the dispersion of the experimental means to the grand mean (i.e., the dispersion between means) with the dispersion of the experimental scores to the means (i.e., the dispersion within the groups). Specifically, the dispersion between the means is evaluated by computing the sum of squares between means, denoted $SS_{Between}$ and computed as:

$$SS_{Between} = k \times \sum_j^J (\alpha_j - \mu)^2$$

The dispersion within the groups is evaluated by computing the sum of squares within groups, denoted SS_{Within} and computed as:

$$SS_{Within} = \sum_j^J \sum_k^K (y_{ij} - \alpha_j)^2$$

If the dispersion of the means around the grand mean is due only to random fluctuations, then the $SS_{Between}$ and the SS_{Within} should be commensurable. Specifically, the null hypothesis of no effect can be evaluated with an F-ratio computed as

$$F = \frac{SS_{Between}}{SS_{Within}} \times \frac{N-J}{J-1}$$

3.1 Bayesian model

The Bayesian approach to fitting the HGLM uses a Gibbs sampler based on the adaptive rejection sampling methodology (Gilks and Wild, 1992), as implemented in the BUGS (Bayesian Inference using Gibbs Sampling) software, to simulate draws from the posterior distribution for the parameters in the general model defined in (7). Diffuse noninformative priors for the fixed effects and the variance parameters were specified for the simulations, to let the data provide the most information about the posterior distributions of the parameters. This approach enables inferences based on simulated draws from the marginal posterior distributions of the two fixed effect parameters, the three variance parameters. This paper focuses on the marginal posterior distribution of the difference in the random effect variances. Specifically, the following prior distributions for these parameters were used. The following hierarchical model therefore operationalizes the above possibility and is fitted to the data to demonstrate Bayesian variance components comparison.

$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ $\quad$ i=1,2,…,n_j
j=1,2,….,J

$V(\varepsilon_{ij}) = \sigma^2$

$y_{ij} \mid \alpha_j, \sigma^2 \sim \text{Normal}(\alpha_j, \sigma^2)$……...…………………………...……....…………………….….(6)

$$\alpha_j \mid \mu_0, \omega^2 \sim \text{Normal}(\mu_0, \omega_0^{\,2}) \cdots\cdots (7)$$

Equation (6) is a normal model for the data, with parameters α_j and σ^2, while equation (7) is a model for how α_j (means), vary across the groups.The parameter μ_0 is the mean of the distribution of the group means, and this group-level distribution has variance,$\omega_0^{\,2}$, also known as the between variance;,σ^2 is known as the within variance for groups J. The parameters in the group-level model, μ_0 and $\omega_0^{\,2}$ are known as hyperparameters. Prior densities for these Parameters, along with a prior for the σ^2 “within variance”, are necessary to complete the specification of this model (A.Gelman, 2005). We used inverse Gamma priors for variance parameters and normal priors for means.

Here, we presented a one-way ANOVA. In random-effects models, a set of effects (group means) are constrained to come from some distribution, which is most often a normal.

A full specification of the normal, one-way Bayesian hierarchical ANOVA model is given below:

$$y_{ij}|\alpha_j, \sigma^2 \sim \text{Normal}(\alpha_j,\sigma^2) \cdots\cdots (8)$$

$$\alpha_j \mid \mu_o, \omega_o^{\,2} \sim \text{Normal}(\mu_{o,} \omega_o^{\,2}) \cdots\cdots (9)$$

$$\mu_0 \sim \text{Normal}(b_0, B_0) \cdots\cdots (10)$$

$$\sigma^2 \sim \text{inverse-Gamma}(v_0/2, \sigma^2\ v_0/2) \cdots\cdots (11)$$

$$\omega^2 \sim \text{inverse-Gamma}(k_0/2, k_0\omega_0^{\,2}/2) \cdots\cdots (12)$$

A model with unit-wise heteroskedasticity results when we let the “within-unit” variance parameter σ^2 vary over units (i.e.instead of σ^2 we would have the parameters $(\sigma_1^{\,2},\sigma_2^{\,2},\ldots\ldots\sigma_J^{\,2})$.The hyperparameters of the normal prior for μ_0 (the mean b_0 and the variance B_0) and the hyperparameters of the priors for the model parameters are in the vector,$\theta = (\alpha_1, \ldots \alpha_J,\mu_0,\sigma^2,\omega_0^{\,2})$.

The hierarchical structure of the model implies that the prior density for θ can be factored as follows:

$$\begin{aligned} f(\theta) &= f(\alpha_1, \ldots, \alpha_J,\mu_o, \sigma^2,\omega_0^{\,2}) \\ &= f((\alpha_1, \ldots, \alpha_J \mid \mu_0,\omega_0^{\,2})\ f(\mu_0)f(\sigma^2)f(\omega_0^{\,2}) \\ &= \prod_{i=1}^{n} f(\alpha_j \mid \mu_0, \omega_0^{\,2})\ f(\mu_0)f(\sigma^2)f(\omega_0^{\,2}) \end{aligned}$$

3.2 Data

This followed Box and Tiao (1973) and data from an experiment that was set up to investigate to what extent yield of dyestuff differs between batches was used. The experiment featured six batches with five observations each.

Table1: Data from a balanced experiment with five samples each with six randomly chosen bathes of raw material

Batch	Yield (in grams)				
1	1545	1440	1440	1520	1580
2	1540	1555	1490	1560	1495
3	1595	1550	1605	1510	1560
4	1445	1440	1595	1465	1545
5	1595	1630	1515	1635	1625
6	1520	1455	1450	1480	1445

The data in table 1, above arose from a balanced experiment in which the total product yield was determined for 5 samples from each of 6 randomly chosen batches of raw material. In order toillustrate the behavior of the various parameters when the null hypothesis is true, the difference between the batch mean and the overall mean was subtracted from the batch data.The objective was to determine the relative importance of between batch variation versus variation due to sampling and analytic errors. We assume that the batches and samples vary independently, and contribute additively to the total error variance.

First, a classical one-way ANOVA is carried out to compute the F statistic and the corresponding p value for the data set. We used the following model for the yield

3.3 Frequentist (Classical) approach

The parameters in the model will be estimated using the maximum-likelihood (ML) estimation, and implemented in the procedure in the R software (R, 2010).

4. Results

4.1 Descriptive Statistics

Table 2: Coefficient values obtained using the classical approach

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1527.50	9.04	168.985	< 2e-16 ***
Batch1	-22.50	20.21	-1.113	0.27666
Batch2	0.50	20.21	0.025	0.98047
Batch3	36.50	20.21	1.806	0.08351
Batch4	-29.50	20.21	-1.459	0.15739
Batch5	72.50	20.21	3.587	0.00149
s-within	42.00			
s-between	49.5			

Residual standard error: 49.51 on 24 degrees of freedom. Multiple R-Squared: 0.4893, Adjusted R-squared: 0.3829 .F-statistic: 4.598 on 5 and 24 DF, p-value: 0.004398

Most of the coefficients are non-significant, suggesting that the batch means do not differ significantly from the grand mean. The coefficients for batch6 is –sum(the rest) = -57.5.

Table 1 presents descriptive statistics for the interviewers in each of the groups defined by the three binary interviewer-level factors. These descriptive statistics include the number of interviewers in each group (out of 38 total), the mean, standard deviation (SD) and range for the number of cases (sample sizes) assigned to each interviewer, and the range of observed means on the parity variable.

Table 4: ANOVA Table for classical approach

ANOVA

Source	DF	Sum of squares	Mean square	F
Between	5	56,357.5	11,271.5	
Within	24	58,830	2,451.25	F = $\frac{MS_B}{MS_W}$ = 4.598
Total	29	115,187.5	4,215.98	

At a level of α= 0.05, the classical approach gave an F value (calculated) of 4.598 which was then compared with table values.

Table 5: Table of posterior point estimates

	Mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
mu[1]	1514.0	20.47	0.1963	1471.0	1515.0	1552.0	5000	100002
mu[2]	1528.0	19.31	0.1142	1489.0	1528.0	1566.0	5000	100002
mu[3]	1550.0	22.19	0.2991	1510.0	1550.0	1595.0	5000	100002
mu[4]	1509.0	21.28	0.241	1466.0	1510.0	1549.0	5000	100002
mu[5]	1572.0	29.16	0.5545	1516.0	1575.0	1625.0	5000	100002
mu[6]	1492.0	25.92	0.4349	1443.0	1491.0	1541.0	5000	100002
s-with	49.74	9.24	0.1301	39.35	52.47	75.03	5000	100002
s-btw	41.65	27.15	0.4727	0.3101	37.34	102.4	5000	100002
sigma2.with	2474.54	4151.0	33.04	0.09619	1394.0	10490.0	5000	100002
sigma2.btw	1734.72	1069.0	15.35	1548.0	1753.0	5630.0	5000	100002
theta	1528.0	21.98	0.116	1483.0	1528.0	1572.0	5000	100002
F	4.56	8.355	0.06948	1.122E-4	2.589	21.19	5000	100002

The results in table 5,above gives posterior numerical summaries from the model after 100,002 iterations and additional discarded 5,000 burn-in iterations using Normal prior for the mean and Inverse-Gamma prior for the variance parameters. It gives the posterior means for the batches, posterior between and within variances. It also gives 95% credible set analog to confidence interval in frequentist approach. This gives a grand posterior mean of 1528.0, posterior within variance of 2474.54 and posterior between variance of 1734.72. These results closely agree with those obtained using frequentist approach. Posterior F-value was 4.56 which is similar to that obtained using Classical approach.

4.2 Posterior densities

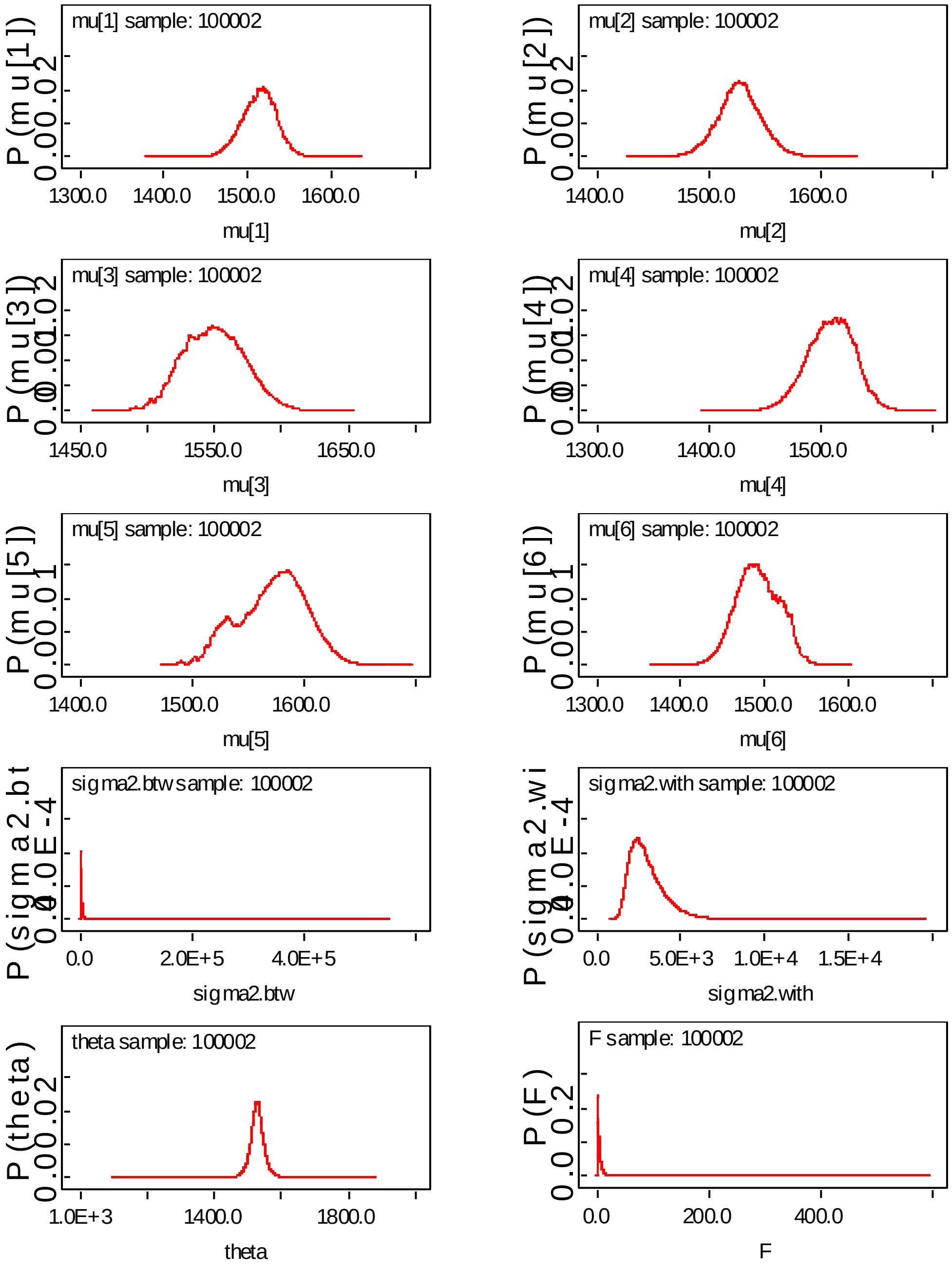

Figure 1: MCMC Posterior densities for the parameters

5.0 Discussion of results

This paper has demonstrated Bayesian approach to finding the posterior F-value and fitting HGLMs with heterogeneous random effect variance parameters and making inferences about differences in those variance parameters. Analysis of real experimental data have shown how the Bayesian approach do a better job than frequentist approach by accommodating uncertainty in the estimation of parameters in these models, and lead to more appropriate inferences when the number of clusters under study is fairly small. Specifically, inferences when following the Bayesian approach to analyzing this problem can be based on 95% credible sets for the difference in the two variance components, defined by the differences in simulated draws of the two variance components from the joint posterior distribution for a given model. This approach provides a more natural form of inference for this problem than the more problematic likelihood ratio testing in the frequentist setting, which relies on asymptotic theory and should not be applied when using pseudo-likelihood estimation approaches. After a burn-in of 5,000 draws (the first 5,000 draws from each Markov chain are discarded as not representative of the stationary distribution of the chain i.e the posterior distribution of the parameters in the model) and a

further 100,002 iterations for each chain, the MCMC produced the summary statistics for the samples as shown in Table 5. As a Bayesian point estimate, typically the posterior means or the posterior medians (or sometimes also the mode), were reported in these table, while the posterior standard deviation was used as a standard error of the parameter estimate. The range between the 2.5th and 97.5th percentiles represents a 95% Bayesian confidence interval and is called a credible interval.
Numerical summaries of the model using the priors appear in Table 5, for the posterior grand mean μ, the "between" variance (ω^2) and the "within", variance (σ^2). The left column summarizes the results of the WinBugs run, showing the mean of the MCMC output for each of the parameters, the standard deviation, and an estimate of the 95% HDR of the marginal posterior density of each parameter.

Assessing the trace plots indicates that the parameter traces look like straight hairy colorful caterpillars, with the two chains fluctuating rapidly around their equilibrium, and that there are no obvious upward or downward trends. Besides, the autocorrelation plots show little correlations, and kernel density plots show bell-like posterior distributions, and the Gelman-Rubin statistic show that the ratio of between to within variability is close to 1. All plots assume us that the model is converged.

These posterior point estimates give results similar to those obtained when using the classical or frequentist approach.

REFERENCES

A. Gelman (2005). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis.

Box, G. E. P. and Tiao, G. C. (1973), Imperical statistical analysis, 112-127.

Carlin, B.P., and Louis, T.A. (2009). *Bayesian Methods for Data Analysis*. Chapman and Hall / CRC Press.

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and*

Nonparametric Regression Models. Chapman and Hall / CRC Press.

Gilks, W.R., Thomas, A., Spiegelhalter, D.J.(1994),A language and program for complex Bayesian modeling. Statistician 43, 169 -178.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, Berlin.

R Development Core Team. (2010), R: A language and Environment for Statistical Computing

Zhang, D. and Lin, X. (2010). Variance component testing in generalized linear mixed models for longitudinal / clustered data and other related topics. *Random Effect and Latent Variable Model Selection*. Springer Lecture Notes in Statistics, Volume 192.