Analysis on CD4PCT cells

Garima Garg

Abstract

In this study we are interested in finding out the effect of baseline age and treatment in the percentage of CD4 cells. Three models have been considered. Sensitivity of the results with respect to the prior, Model checking, posterior predictive checks are used. Model comparison methods have been used to compare between these models.

KEY WORDS: Joint Posterior Distribution, Full conditional Distribution, Model Checking

1. Data Analysis

Missing data have been removed from the dataset provided. After removing the missing data, the number of observations were 436 (corresponds to n in our later analysis). After removing the missing data, some of the patients information were no longer in the dataset. No records exist for patients corresponding to newpid 36, 80 and 102 after removing the missing data. I recoded the variable treatment. If the treatment for patient i and visit j was treatment 1, then recoded treatment variable is 0 and is 1 if such treatment was treatment 2. Fig 1 shows the scatter plot between baseage and percentage of CD4 cells. Clearly, high correlation exists between baseage and percentage of CD4 cells. Fig 2. shows the box plot between recoded treatment and percentage of CD4 cells. Again, the data is highly correlated.

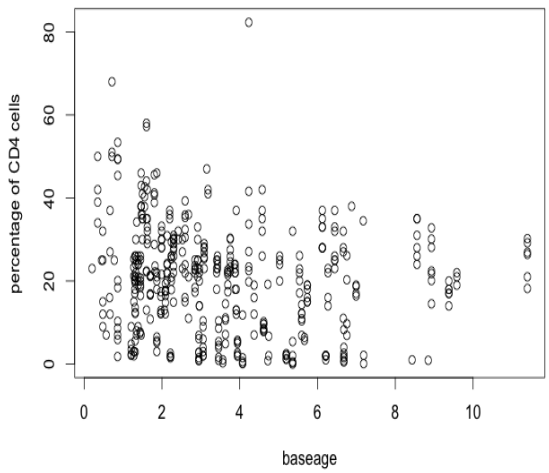

Figure 1: *Scatter plot between baseage and percentage of CD4 cells.*

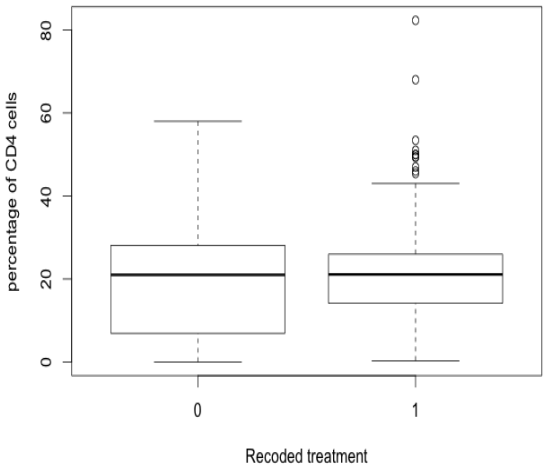

Figure 2: *Scatter plot between Recoded treatment and percentage of CD4 cells.*

2. Model I - Using non-informative prior

I fit a linear regression model, with independent and identically distributed errors, so that

$$y_{i,j} = \alpha + \beta x_{1,i,j} + \gamma x_{2,i,j} + \epsilon_{i,j} \quad \epsilon_{i,j} \sim N(0, \sigma^2)$$

where $y_{i,j}$ is the square root of the percentage of CD4 cells for the i-th child and the j-th visit. I consider $\theta = (\alpha, \beta, \gamma)$. Under the model,

$$y|\theta, \sigma^2, X \sim N(X\theta, \sigma^2 I)$$

where X is full rank matrix with rank 3.
For Model I, the non-informative priors are

$$p(\theta) \propto 1 \quad p(\sigma^2) \propto 1/\sigma^2$$

The conditional posterior distribution of θ is

$$\theta|\sigma^2, y, X \sim N(\hat{\theta}, \sigma^2 V_\theta)$$

where $\hat{\theta} = (X'X)^{-1}X'y$ and $V_\theta = (X'X)^{-1}$ Since X has full rank, $\hat{\theta}$ can be easily estimated. The marginal posterior distribution of σ^2 is

$$\sigma^2|y \sim IG((n-k)/2, (n-k)\hat{\sigma^2}/2)$$

where n is the total number of records in the data, and k is the rank of X matrix which is 3. Using the QR

Table 1: Part-I quantiles of the conditional posterior of θ.

	2.5%	50%	97.5%
α	4.064643	4.371094	4.710342
β	-0.15047842	-0.08653302	-0.02049993
γ	0.03801632	0.34476879	0.6266874

factorization of X, $\hat{\theta}$ was estimated. I generated 1000 samples from the posterior distribution of σ^2. Using these samples, I generated 1000 samples of θ. Fig.3 shows the trace and density plots of θ and σ^2. Fig 4 shows the scatter plot of 1000 draws of β and γ. The mean and quartiles for θ are shown in Table1. It can be easily seen that the 95% interval of β and γ does not contain 0. So I can conclude that baseage and treatment have an effect in the percentage of CD4 cells at .05 significance level.

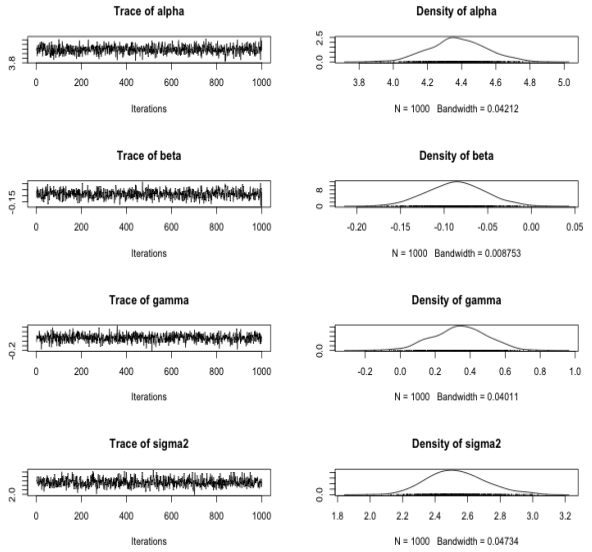

Figure 3: *Trace and density plots of θ and σ^2.*

2.1 Model I - Using non-informative prior, Sensitivity, Model checking, posterior predictive checks

For Model checking, I replicated the data and Fig5 shows the posterior predictive distribution of $y_{i,j}$ with original values of percentage of CD4 cells. Some of the outliers can be observed.

For the posterior predictive check, I used the test statistic that counts the number of times the replicated data is greater than the original data. Fig 6 shows the posterior predictive p-values for each of the $y_{i,j}$. Some of the p-values are extreme which could be due to outliers. For Residual Analysis, I plotted a graph between Residuals and fitted values. Since the pattern is random, I can

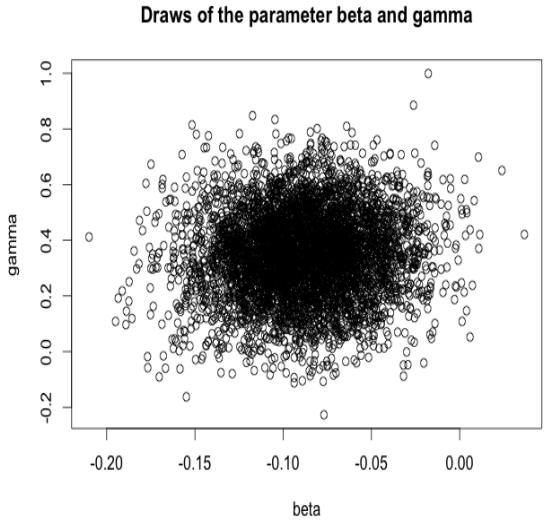

Figure 4: *Scatter plot of 1000 draws of β and γ.*

conclude that the model fits well with the data. Fig 7 shows the plot between Residuals and fitted values.

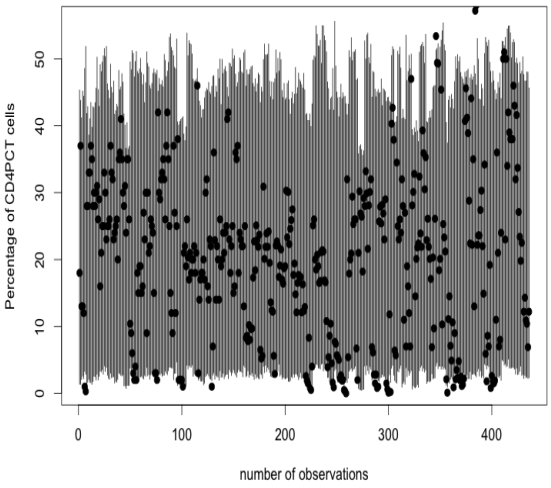

Figure 5: *Posterior predictive distribution of $y_{i,j}$ with original values of percentage of CD4 cells for Model I .*

3. Model II - Using G-prior

Under this model, the prior used is

$$\theta \sim N(0, g\sigma^2(X'X)^{-1}) \quad \epsilon_{i,j} \sim N(0, \sigma^2)$$

. θ is same for this model. The conditional posterior distribution on θ is

$$\theta|\sigma^2, y, X \sim N(\frac{g}{g+1}\hat{\theta}, \frac{g}{g+1}\sigma^2(X'X)^{-1})$$

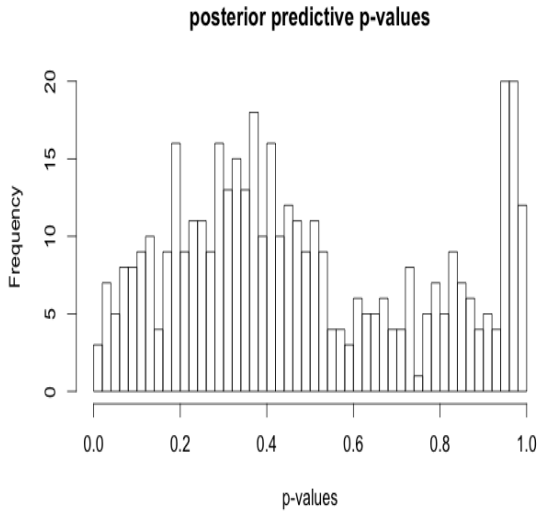

Figure 6: *posterior predictive p-values for each of the $y_{i,j}$ for Model I .*

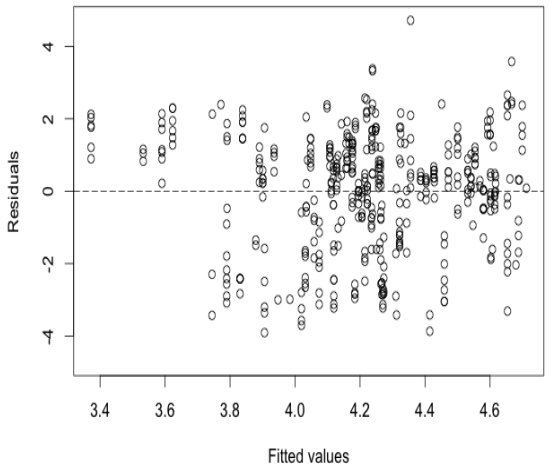

Figure 7: *Plot between Residuals and fitted values for Model I .*

The marginal posterior distribution of σ^2 is

$$\sigma^2|y, X \sim IG(\frac{n}{2}, \frac{\hat{\sigma}(n-k)}{2} + \frac{1}{2(g+1)}\hat{\theta}(X'X)\hat{\theta})$$

where $\hat{\theta} = (X'X)^{-1}X'y$, n and k have same values as before since the design matrix X is same. Using the same technique as before, QR factorization of X, I generated 1000 samples from the posterior distribution of σ^2. Using these samples, I generated 1000 samples of θ. The value of g used is 5. Fig8 shows the trace and density plots of θ, and σ^2. Table 2 shows the quartiles of marginal posterior parameters. It can be easily seen that the 95% interval of β and γ does not contain 0. So I can conclude that baseage and treatment have an effect in the percentage of CD4 cells at .05 significance level.

Table 2: Model-II quantiles of the posterior of θ.

	2.5%	50%	97.5%
α	3.249518	3.633094	4.067342
β	-0.16027842	-0.07173302	0.0089493
γ	-0.06981632	0.28576879	0.6756874

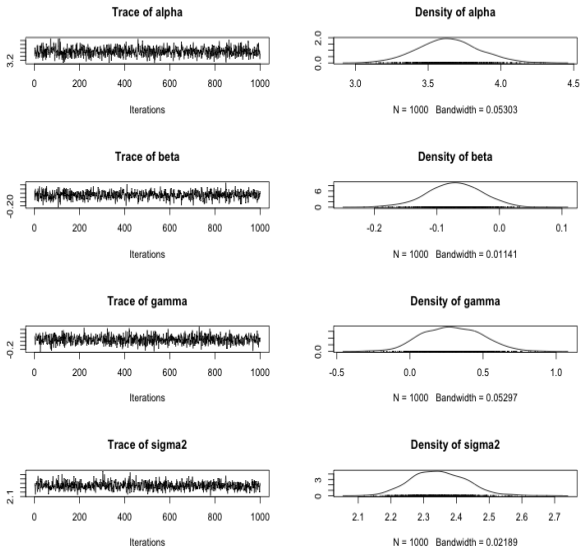

Figure 8: *Model II - Trace and density plots of θ, and σ^2.*

3.1 Model II - Using G-prior, Sensitivity, Model checking, posterior predictive checks

Under different values of g, the posterior samples of the parameters varies, which suggests that our results are sensitive with respect to the prior. Fig 9 shows the scatterplot between beta and gamma under different values of g. Fig 10 shows the scatterplot between alpha and beta under different values of g. For Model checking, I replicated the data and Fig 11 shows the posterior predictive distribution of $y_{i,j}$ with original values of percentage of CD4 cells. Very few outliers can be observed in this model as compared to Model I. For the posterior predictive check,

I used the test statistic that counts the number of times the replicated data is greater than the original data. Fig 12 shows the posterior predictive p-values for each of the $y_{i,j}$. Most of the p-values are not extreme. For Residual Analysis, I plotted a graph between Residuals and fitted values. Since the pattern is random, I can conclude that the model fits well with the data. Fig 13 shows the plot between Residuals and fitted values.

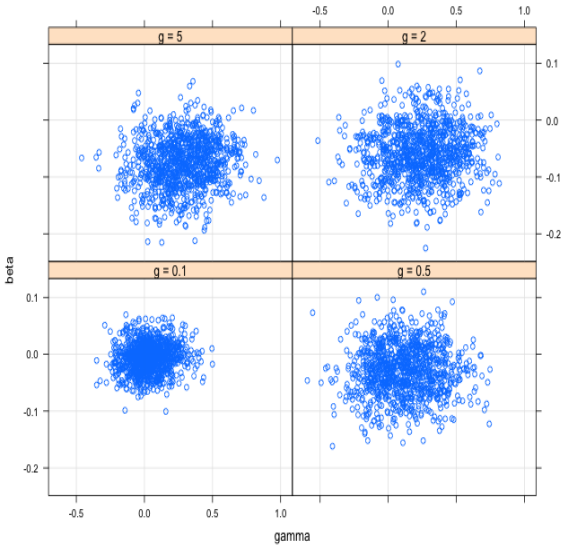

Figure 9: *Model II - scatterplot between beta and gamma under different values of g.*

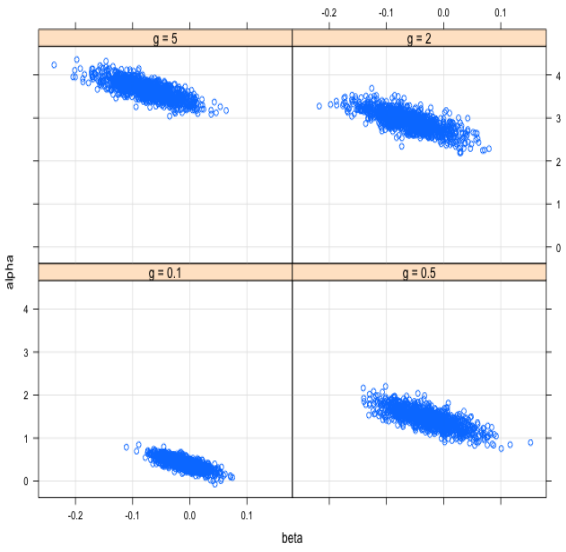

Figure 10: *Model II - scatterplot between alpha and beta under different values of g.*

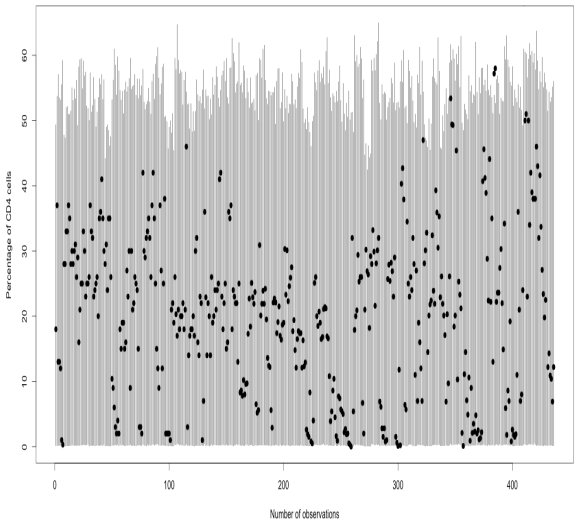

Figure 11: *Model II- posterior predictive distribution of $y_{i,j}$ with original values of percentage of CD4 cells.*

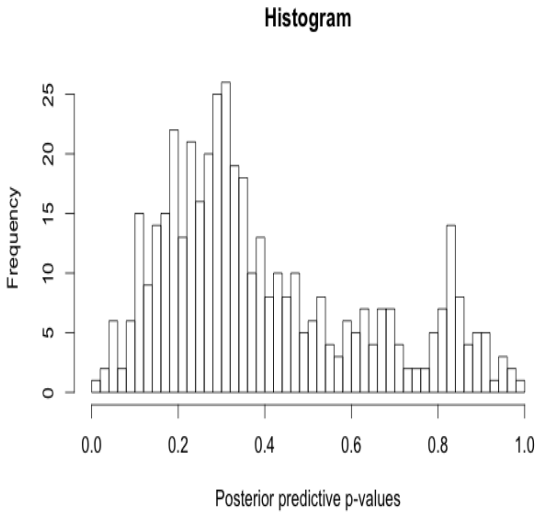

Figure 12: *Model II-posterior predictive p-values for each of the $y_{i,j}$*

4. Model- III Using random mixed effects

In this model, I considered random effects and the model looks as:

$$y_{i,j} = \alpha_i + \beta_i x_{1,i,j} + \gamma_i x_{2,i,j} + \epsilon_{i,j} \quad \epsilon_{i,j} \sim N(0, \sigma^2)$$

where $y_{i,j}$ is the square root of the percentage of CD4 cells for the i-th child and the j-th visit. Under this model, the prior used is

$$(\alpha_i, \beta_i, \gamma_i)' \sim MVN_3((\alpha, \beta, \gamma)', diag(\tau_\alpha^2, \tau_\beta^2, \tau_\gamma^2))$$

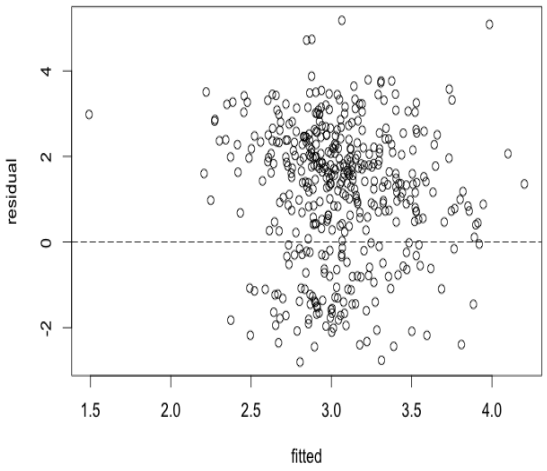

Figure 13: *Model II- Plot between Residuals and fitted values*

and

$$p(\alpha,\beta,\gamma) \propto 1 \quad p(\sigma^2) \propto 1/\sigma^2 \quad p(\tau^2_{(.)}) = Uniform(0,A)$$

The joint posterior distribution of all the parameters in the model is

$$p(\alpha_1,..,\alpha_I,\beta_1,..,\beta_I,\gamma_1,..,\gamma_I,\alpha,\beta,\gamma,\sigma^2,\tau^2_\alpha,\tau^2_\beta,\tau^2_\gamma|y,X)$$

$$\propto (\frac{1}{\sigma^2})[\prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{1}{\sigma}exp(-\frac{(y_{i,j}-(\alpha_i+\beta_i x_{1,i,j}+\gamma_i x_{2,i,j}))^2}{2\sigma^2})]$$

$$\times[\prod_{i=i}^{I}\frac{1}{\tau_\alpha}exp(-\frac{(\alpha_i-\alpha)^2}{2\tau^2_\alpha})][\prod_{i=i}^{I}\frac{1}{\tau_\beta}exp(-\frac{(\beta_i-\beta)^2}{2\tau^2_\beta})]$$

$$\times[\prod_{i=i}^{I}\frac{1}{\tau_\gamma}exp(-\frac{(\gamma_i-\gamma)^2}{2\tau^2_\gamma})]$$

where I is the total number of patients and n_i is the number of visits corresponding to the patient i. Now I reduced the model to the multivariate model, and for this I considered $\theta_i = (\alpha_i,\beta_i,\gamma_i)$. $\breve{\theta} = (\alpha,\beta,\gamma)$ and $\Sigma_\theta = diag(\tau^2_\alpha,\tau^2_\beta,\tau^2_\gamma)$. Now $\theta = (\theta_1,\theta_2....\theta_I)'$. Thus the prior distribution now follows $\theta_i \sim MVN(\breve{\theta},\Sigma_\theta)$. Under this setting , the joint posterior distribution is

$$p(\theta,\breve{\theta},\sigma^2,\tau^2_\alpha,\tau^2_\beta,\tau^2_\gamma|y,X)$$

$$\propto (\frac{1}{\sigma^2})[\frac{1}{(\sigma^2)^{n/2}}exp(-\frac{(y-X\theta)'(y-X\theta)}{2\sigma^2})]$$

$$\times[\prod_{i=1}^{I}\frac{1}{|\Sigma_\theta|^{1/2}}MVN_3(-\frac{1}{2}(\theta_i-\breve{\theta})'\Sigma_\theta^{-1}(\theta_i-\breve{\theta}))]$$

where n is the total number of observations in the data provided. Here the design matrix X is different from Model I and Model II.

$$\boldsymbol{X} = \begin{bmatrix} \underset{n_1\times 3}{\boldsymbol{X}_1} & \underset{n_1\times 3}{\boldsymbol{0}} & \cdots & \underset{n_1\times 3}{\boldsymbol{0}} \\ \underset{n_2\times 3}{\boldsymbol{0}} & \underset{n_2\times 3}{\boldsymbol{X}_2} & \cdots & \underset{n_2\times 3}{\boldsymbol{0}} \\ \vdots & \vdots & \ddots & \vdots \\ \underset{n_I\times 3}{\boldsymbol{0}} & \underset{n_I\times 3}{\boldsymbol{0}} & \cdots & \underset{n_I\times 3}{\boldsymbol{X}_I} \end{bmatrix}$$

where $\boldsymbol{X}_1$ corresponds to the matrix related with Patient 1. The full conditional distribution of θ follows a multivariate normal distribution

$$(\theta|\breve{\theta},\sigma^2,\Sigma_\theta,y,X) \sim MVN(\theta^*,\Sigma^*)$$

where

$$\Sigma^* = (\Sigma_\theta^{-1} + X'X\sigma^{-2})^{-1}$$

$$\theta^* = \Sigma^*(\Sigma_\theta^{-1}\tilde{\theta} + X'\sigma^{-2}y)$$

where $\tilde{\theta} = ((\alpha,\beta,\gamma),(\alpha,\beta,\gamma),....I\ \ times)'$. The full conditional distribution of $\breve{\theta}$ $((\alpha,\beta,\gamma)'|\theta,\sigma^2,\Sigma_\theta,y,X)$

$$\sim MVN_3((\bar{\alpha},\bar{\beta},\bar{\gamma})', diag(\tau^2_\alpha/I,\tau^2_\beta/I,\tau^2_\gamma/I))$$

where $\bar{\alpha},\bar{\beta},\bar{\gamma}$ is the mean of $\alpha_i,\beta_i,\gamma_i$ respectively. The full conditional distribution of τ^2_α is

$$(\tau^2_\alpha|\theta,\breve{\theta},\sigma^2,y,X) \sim IG(I/2-1,\sum_{i=1}^{I}\frac{(\alpha_i-\alpha)^2}{2})$$

Similarly distributions for τ^2_β and τ^2_γ can be computed. The full conditional distribution of σ^2 is

$$\sigma^2|\theta,\breve{\theta},\Sigma_\theta,y,X \sim IG(n/2,\frac{(y-X\theta)'(y-X\theta)}{2})$$

I used the Gibbs sampling algorithm, but first I need to initialize the values of σ^2 which I set equal to 2.3 from earlier models, $\breve{\theta}$=(3.8, -0.12, 0.4) also from previous models and $(\tau^2_\alpha,\tau^2_\beta,\tau^2_\gamma)$=(141,99,180). Using this values, I simulate θ and rest of the parameters in the model. Fig 14 shows the trace plots and density plots of $\alpha_1,\beta_1,\gamma_1$ for patient 1. Fig 15 shows the scatterplot between β_1 and γ_1. Table3 shows the quantiles of θ_4 for patient 4 when $(\tau^2_\alpha,\tau^2_\beta,\tau^2_\gamma)$=(1,4,6). The findings for all the patients were same, the 95% posterior interval contains 0, so I can conclude that baseline age and treatment have no effect in the percentage of CD4 cells for Patients, though the data is quite correlated which is unusual.

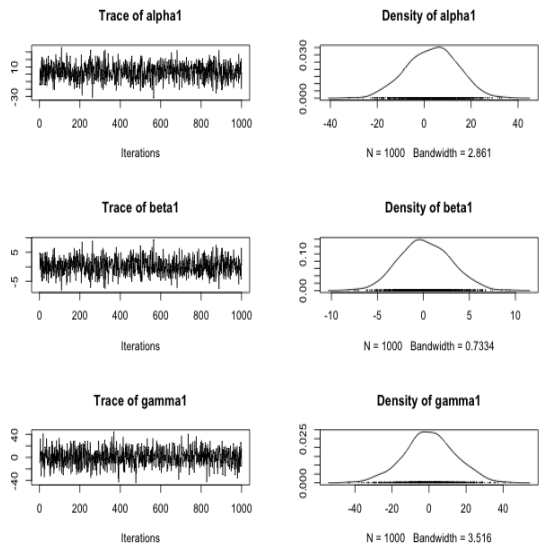

Figure 14: *Model III- Trace plots and density plots of* $\alpha_1, \beta_1, \gamma_1$ *for patient 1*

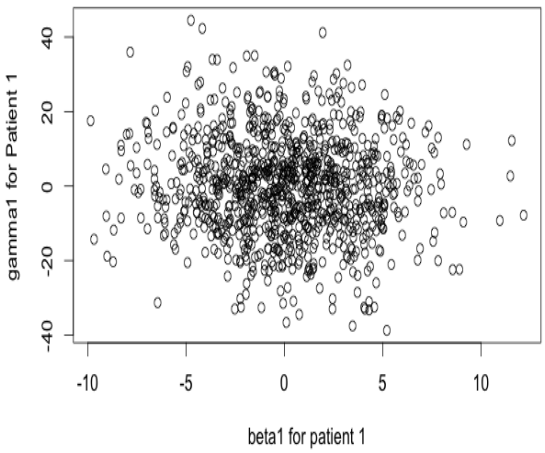

Figure 15: *Model III- Scatterplot between* β_1 *and* γ_1 *for patient 1*

Table 3: Model-III quantiles of the posterior of θ_4 for Patient 4.

	2.5%	50%	97.5%
α	3.231661	3.856718	4.431001
β	-0.3820915	0.3134229	1.0494312
γ	-0.7494458	0.6382201	2.0184368

4.1 Model III - Using Random effects Model, Sensitivity, Model checking, posterior predictive checks

Fig 16 shows the posterior predictive distribution of $y_{i,j}$ and the points corresponds to original data. There is just one or two outliers, so we can see that this model replicates data much closer to the original data when compared to other models. We can easily see that this model is better in terms of predicting the data.
The results are sensitive to the value of A. If I use large values of A, then posterior samples of A are more spread out, but if i use small values of A, then posterior samples of A are concentrated in a smaller area. Fig 17 shows the posterior predictive p-values for each of the $y_{i,j}$. For Residual Analysis, I plotted a graph between Residuals and fitted values. Since the pattern is random, I can conclude that the model fits well with the data. Fig 19 shows the plot between Residuals and fitted values.

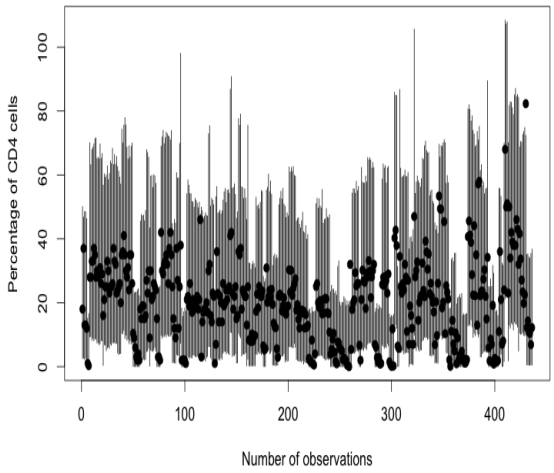

Figure 16: *Model III -Posterior predictive distribution of* $y_{i,j}$ *and the points corresponds to original data*

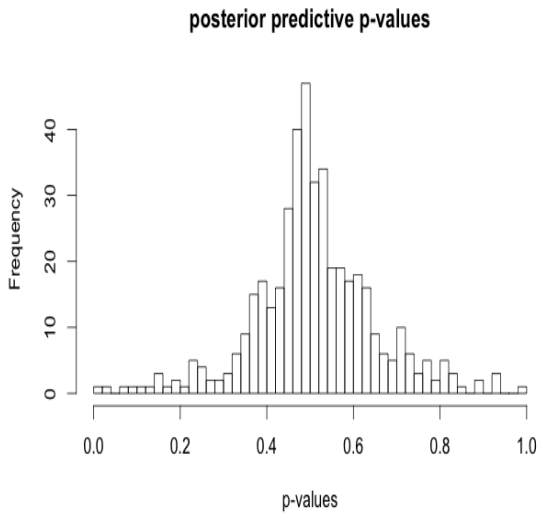

Figure 17: *Model III- posterior predictive p-values for each of the* $y_{i,j}$

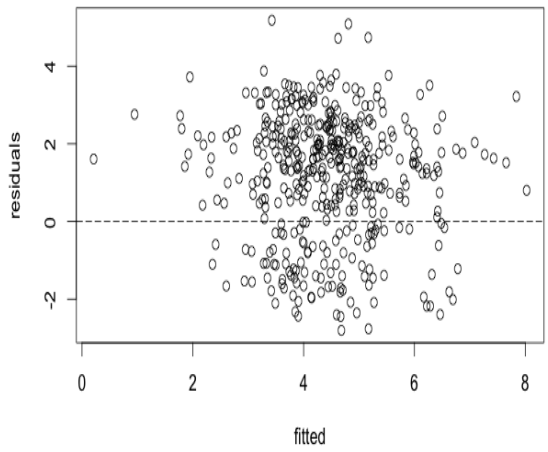

Figure 18: *Model III- plot between Residuals and fitted values*