Modelling Exposure Cadmium Data

Bayesian Data Analysis
2020-2021

2nd year Master of Statistics
Hasselt University

BY:

Nsoh Tanih Dam

Data Description

The exposure to cadmium is based on two data bases: (1) food consumption data and (2) the contamination data.

The food consumption data was obtained from the Chinese national nutrition and health survey in 2002 via a 24-hour recall method on three consecutive days (two week days and one weekend, holidays excluded). Multistage, random cluster sampling was used in the survey and 30 provinces with 132 districts and counties were covered. The study included 65,915 consumers from 22,567 households with participants aging between 2 and 100 years. Demographic information of the participants, such as age, gender and body weight, and habitat (rural/urban) of the study participant was also collected in this survey.

The study focused on adults aged between 25 and 65 years and the subjects from 2 provinces (numbers 46 and 64) were selected. In total **1006 subjects** were considered in the analysis.

In order to compute the exposure of the subjects to cadmium by intake of food, another data base was used were in the past the cadmium content of food products was measured a number of times. More specifically, the cadmium contamination data was taken from the national food contamination monitoring program during the years 2001–2010. Specified food items for the Cd monitoring program included rice and its products, wheat flour and its products, crustaceans, pak-choi, pig meat, bean and its products, root and tuber vegetables, leafy vegetables, seafood, and mushrooms. In the food consumption surveys a large number of specific foods were recorded, however they were classified into a smaller number of food categories. Thus for each food item a distribution of cadmium values is available obtained from the data of that food item sampled repeatedly.

The data set contained the following covariates: houseid, subject id, age (years), weight (kg), sex (Male, Female), ur (living area with values Urban , Rural), provinceid (46, 64) and the median exposure value of cadmium.

1 Across all subjects and irrespective of the covariates, what is the 95%-ile and 99%-ile of exposure?

To perform this analyses, three chains, 50000 iterations and a burnin of 15000 were used. This led to 35000 iterations used to compute the posterior mean and posterior variance of distribution. To validate the analyses, convergence diagnosis were used such as the trace plot, auto-correlation plot, Brooks-Gelman-Rubin diagnosis plots and a cross correlation plot for the parameters. The response variable (median exposure value of cadmium) was transformed to log form to get a normal distribution. Furthermore, sensitivity analyses was done by varying the prior of the parameters.

1.1 Sensitivity Analysis For Priors

In cases were the likelihood dominates the prior, the likelihood will also dominate in the posterior and hence the posterior is invariant to the choice of prior (Lesaffre and Lawson, 2012) (Stojanovski et al., 2011). However, when the prior dominates the likelihood, the posterior changes with the prior and thus there is a need to assess the sensitivity of the model to different prior specifications. Sensitivity analysis was performed by repeating the analysis under different prior assumptions, but within the same simulation, so that a direct comparison can be made [4].

Table 1 and 2 illustrates various choices of prior in order to verify if the posterior distribution could be affected by the prior. The priors of the parameters beta0 were changed from N(0.01,1.0E-6) to N(0.01,1.0E-6) and dt(0,0.1, 100) while the priors of the variance parameter was changed from dunif(0, 100) to dunif(0, 1000) and to IG(0.001,0.001). In all these variations of the priors the mean and standard deviation of the posterior estimates were seen to be the same implying that the prior is contributing little or no information on the parameter estimates and thus varying the prior had no impact.

Table 1: Varying prior of Beta's showing Posterior Mean(posterior standard deviation)

Parameter	**dnorm(0.01,1.0E-6)**	**dnorm(0,1.0E-9)**	**dt(0,0.1, 100)**
beta0	-0.613(0.012)	-0.613(0.012)	-0.613(0.012)

Table 2: Varying prior of Sigma showing Posterior Mean(posterior standard deviation)

Parameter	**dunif(0, 100)**	**dunif(0, 100)**	**IG(0.001,0.001)**
σ^2	0.137(0.006)	0.137(0.006)	0.137(0.006)

1.2 Convergence Diagnosis

Figure 1 shows trace plots of the parameters indicating stationarity was achieved with a thick horizontal line an indication of high mixing rate and individual moves are not easily noticeable. Figure 2 shows an auto-correlation plot indicating that starting values are easily forgotten in estimating the parameters, this implies good and independent sample moves were made during sampling. Figure 3 shows a formal graphical diagnostic using the Brooks-Gelman-Rubin procedure where convergence is attained when difference in chains cannot be noticed. The figure 3 shows all parameters attained convergence after about 3000 iterations since the chains all converged to the horizontal line at 1. Figure 4 shows a cross correlation plot which indicates very little correlation between the parameters implying the parameters where sampled independently.

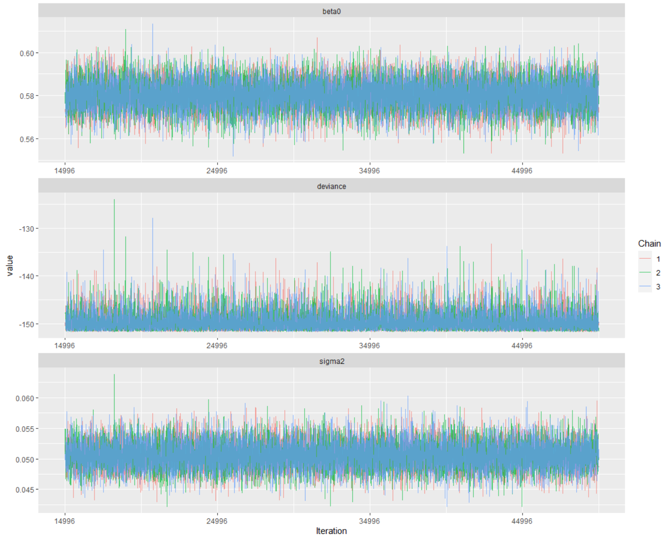

Figure 1: Trace Plots

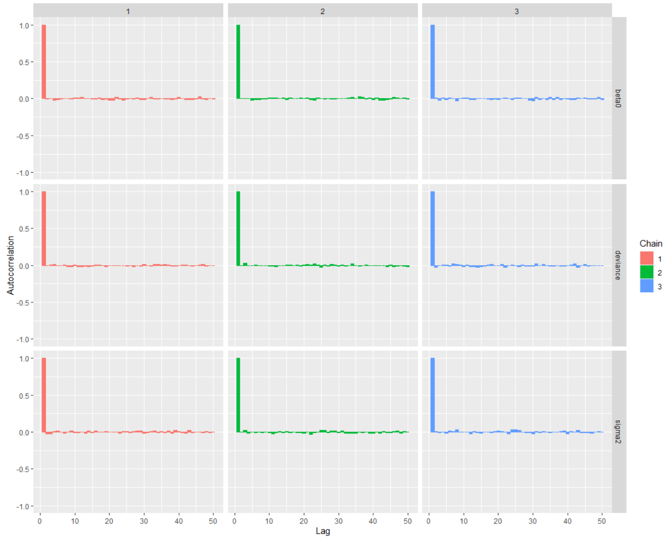

Figure 2: Auto correlation

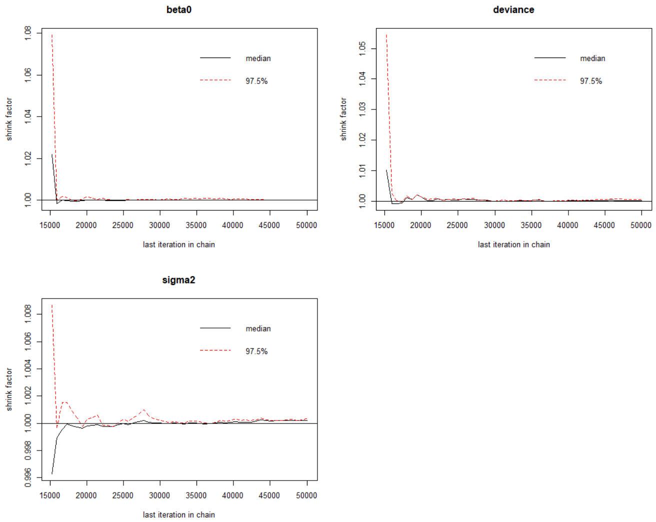

Figure 3: Brooks-Gelman-Rubin

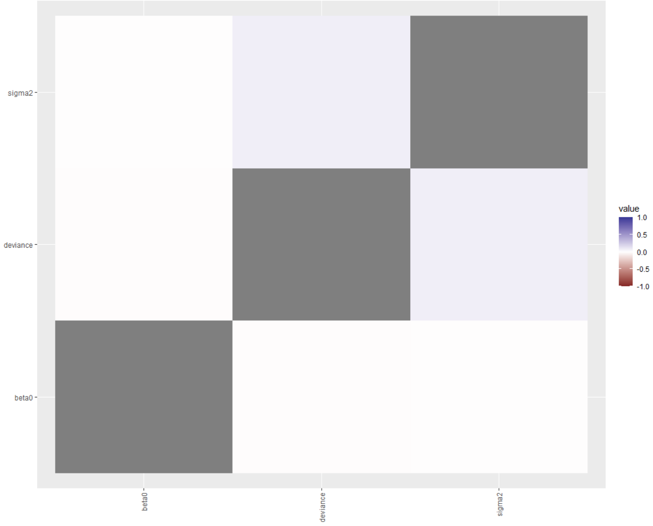

Figure 4: Cross Correlation

1.3 Posterior Estimates

Table 3 illustrates the population average cadmium exposure and the variability in the distribution. The columns 2.5%, 75%, 95%, 99% confidence intervals pertains to the plausible values given the data. In order to ensure efficient estimates are obtained from the MCMC, the Monte-Carlo (MC) error should be less than 5% of the respective standard deviations. The Monte-Carlo error is less than 5% the standard deviation for both parameters. Therefore the average cadmium exposure from the data is $0.542(e^{-0.613})$ with a variability of 0.137. Figure 5 shows the distribution of log response and the posterior mean.

Table 3: Posterior Parameter Estimates of Log Cadmium

Parameter	**mean**	**sd.dev**	**2.5%**	**75%**	**95%**	**99%**	**MCse**
beta0	-0.613	0.012	-0.636	-0.605	-0.594	-0.586	0.00008302
σ^2	0.137	0.006	0.125	0.141	0.147	0.152	0.000005

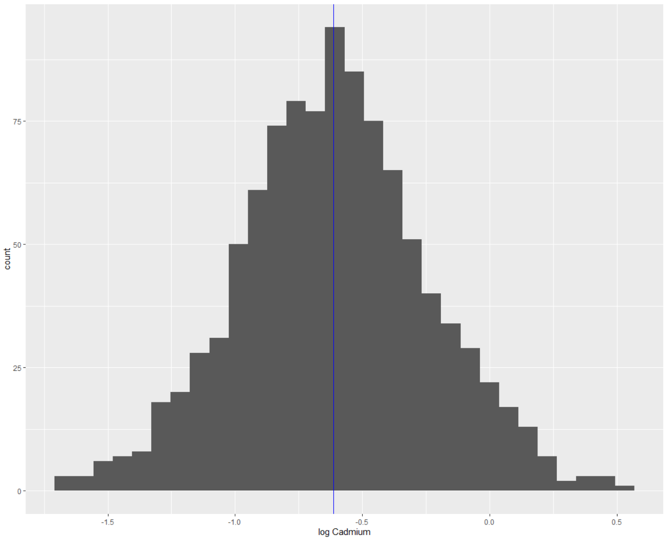

Figure 5: Distribution of Cadmium with Posterior Mean

2 What are the factors that influence the distribution of median values?

2.1 Now include the covariates and again estimate the 95%-ile and 99%-ile of the exposures and their uncertainty as a function of the covariates. Explore whether the variance of the exposures varies with the covariate values

To verify factors that influences individuals median cadmium values, a Bayesian multiple linear regression is used to verify possible factors that influence or cause variability in an individuals median cadmium levels. The model with covariates is defined as follows;

$$log(Y_i) = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Sex_i + \beta_4 LivingArea_i + \epsilon_i$$

where: $log(Y_{ij})$ is the log median cadmium values of the i^{th} individual with $i = 1, ..., 1008$. Age represents the age of the individuals (continuous). Weight represents the weight of the individuals (continuous). Sex is the gender of an individual coded (1=Male, 0=Female) and living area represents the individuals' living area coded Living Area(1=urban, 0=rural). The fixed effects parameters were modelled with $N(0, 1.E-6)$. $\epsilon_{ij} \sim N(0, \tau)$ where $\tau \sim dunif(0, 100)$ and $\sigma^2 = \frac{1}{\tau}$
To perform this analyses, three chains, 50000 iterations and a burnin of 15000 was used. This led to 35000 iterations being used to compute the posterior mean and posterior variance of distribution. To validate the analyses, various convergence diagnostics were used, such as the trace plot, auto-correlation plot, Brooks-Gelman-Rubin diagnosis plots and a cross correlation plot for the parameters. The response variable (median exposure value of cadmium) was transformed to log form to get a normal distribution. Furthermore, sensitivity analyses was done by varying the prior of the parameters.

2.1.1 Sensitivity Analysis

Sensitivity analysis was performed by changing the prior distribution of the beta and variance parameters. The priors of the beta parameters were varied by using N(0, 1.0E-3) ,N(0, 1.0E-6) and dt(0,01,100) while the prior of variance parameter was varied by using a Uniform Distribution U(0,100),U(0,1000) and IG(0.001, 0.001). Table 4 and Table 5 show that changing the prior distributions of the various parameters led to little or no change in the posterior means and standard deviation of the estimates. This implies varying the prior had no impact on the analysis (non-informative).

Table 4: Varying Priors of Beta's showing posterior means(standard deviation)

parameter	**Variable**	**dnorm(0,1.0E-3)**	**dnorm(0,1.0E-6)**	**dt(0,01,100)**
beta0	Intercept	0.342(0.074)	0.342(0.074)	0.316(0.070)
beta1	Age	-0.001(0.001)	-0.001(0.001)	-0.001(0.001)
beta2	Weight	-0.017(0.001)	-0.017(0.001)	-0.016(0.001)
beta3	Sex(male)	0.121(0.024)	0.121(0.024)	0.117(0.023)
beta4	Living Area(urban)	0.114(0.022)	0.114(0.022)	0.114(0.022)

Table 5: Varying prior of σ^2 showing Posterior Mean(posterior standard deviation)

Parameter	**dunif(0, 100)**	**dunif(0, 100)**	**IG(0.001,0.001)**
σ^2	0.109(0.005)	0.109(0.005)	0.109(0.005)

2.1.2 Convergence diagnosis for Bayesian linear regression model

Figures 6 shows trace plots of the parameters indicating stationarity was achieved with a thick horizontal line an indication of high mixing rate and individual moves are not easily noticeable which falls inline with Gelfand and Smith (1990) proposal. Figure 7 shows an auto-correlation plot indicating that starting values are easily forgotten in estimating the parameters, this implies good and independent sample moves were made during sampling. Figure 8 shows a formal graphical diagnostic using the Brooks-Gelman-Rubin procedure where convergence is attained when difference in chains cannot be noticed. The figure 8 shows all parameters attained convergence after about 3000 iterations since the chains all converged to the horizontal line at 1. Figure 9 shows a cross correlation plot which indicates very little correlation between the parameters implying the parameters where sampled independently.

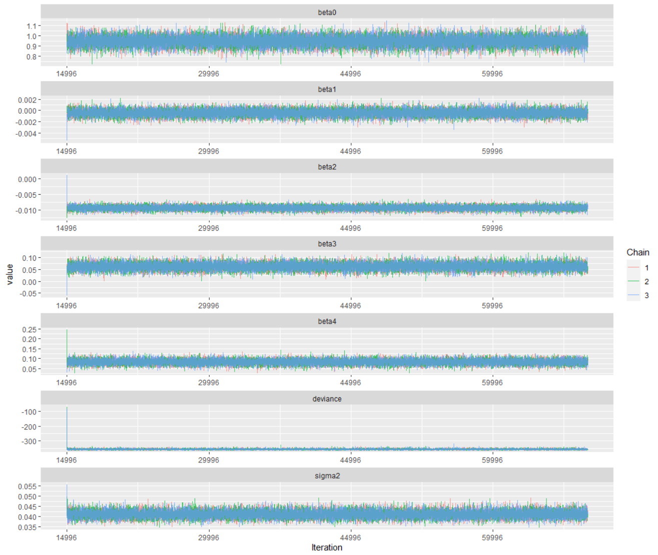

Figure 6: Trace Plots

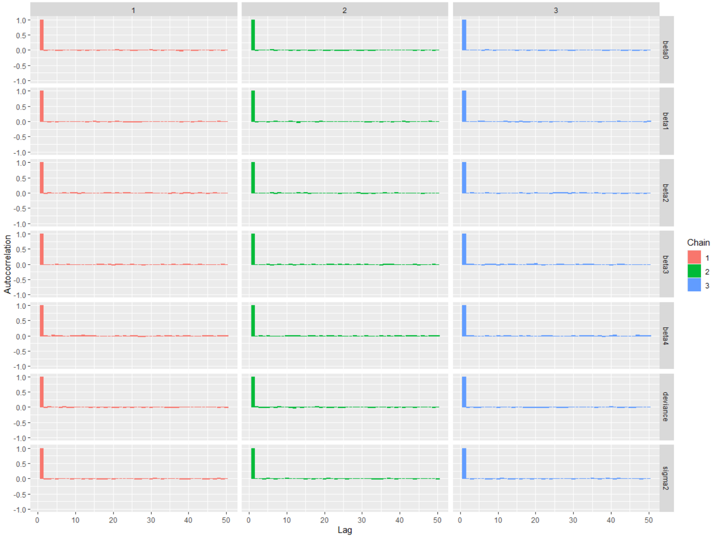

Figure 7: Auto correlation

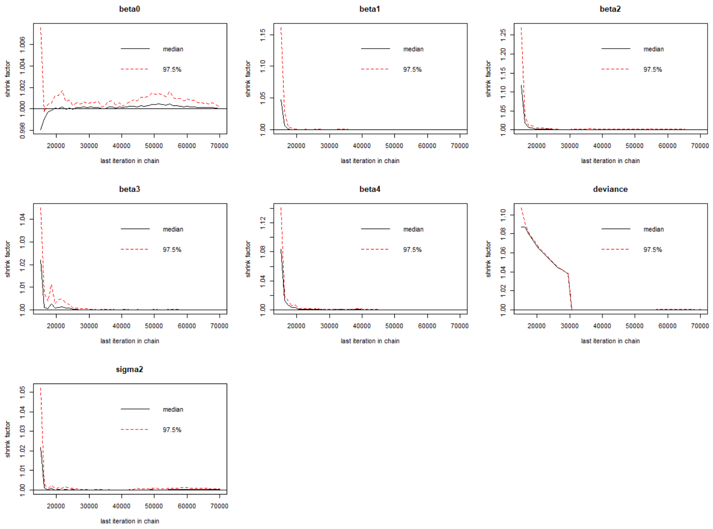

Figure 8: Brooks-Gelman-Rubin

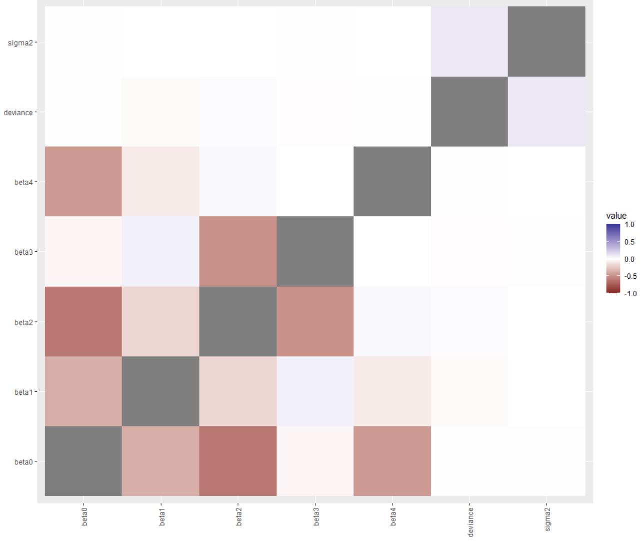

Figure 9: Cross Correlation

2.1.3 Posterior Estimates

Table 6: Posterior Parameter Estimates of Bayesian Linear Regression Model

Parameter	**Variable**	**mean**	**sd.dev**	**2.5%**	**75%**	**95%**	**99%**	**MCse**
beta0	Intercept	0.342	0.074	0.197	0.392	0.463	0.515	0.0003718
beta1	Age	-0.001	0.001	-0.003	-0.001	0.000	0.001	0.00000546
beta2	Weight	-0.017	0.001	-0.019	-0.016	-0.015	-0.014	0.000005998
beta3	Sex(male)	0.121	0.024	0.074	0.137	0.160	0.177	0.0001352
beta4	Living Area(urban)	0.114	0.022	0.070	0.128	0.150	0.165	0.0001253
σ^2		0.109	0.005	0.100	0.112	0.118	0.121	0.0195

Table 6 illustrates the posterior parameter estimates obtained by fitting a Bayesian linear regression to the log cadmium values with age, weight, sex(male) and living area(urban) as covariates. The intercept refers to the expected log cadmium of a female who stays in a rural area with constant age and weight. In order to determine which covariates influence the response, the confidence interval should not contain zero. The table 6 shows the variables Age is not significant. However the variable Weight is seen to have a significant diminishing effect on the levels of cadmium. The columns 2.5%, 75%, 95%, 99% confidence intervals pertains to the plausible values given the data. In order to ensure efficient estimates are obtained from the MCMC,the Monte-Carlo (MC) error should be less than 5% of the respective standard deviations. The Monte-Carlo error is less than 5% the standard deviation for both parameters.

2.1.4 Model Diagnosis

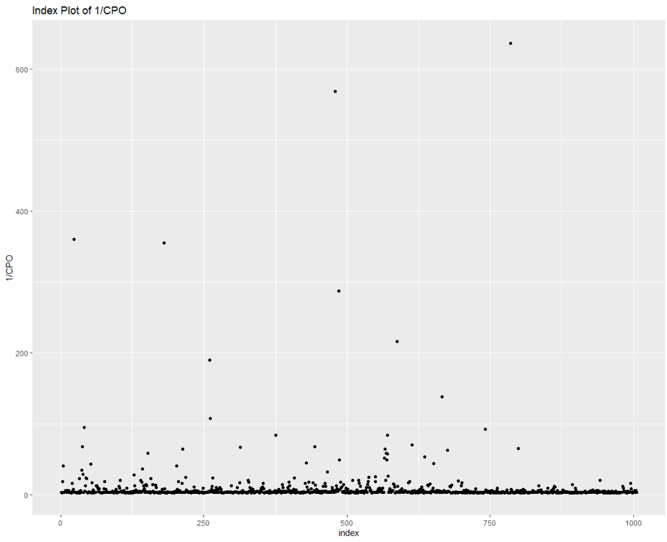

Figure 10: Outlying Observations

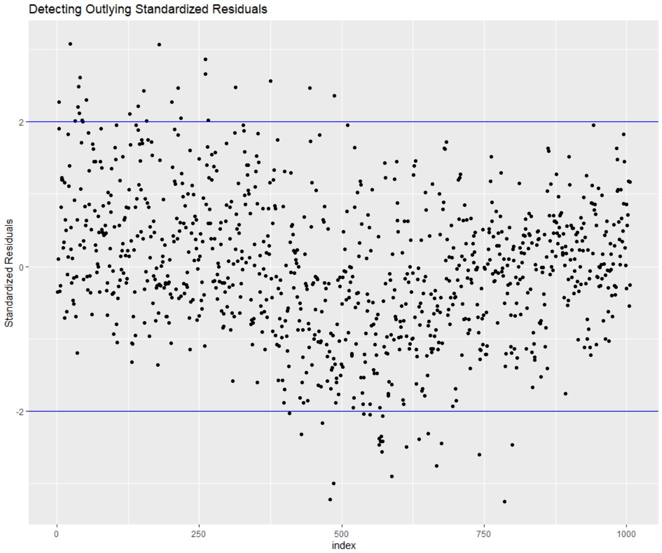

Figure 11: Bayesian Residual Analysis

Table 7: Comparing Posterior Parameter Estimates without and with outlier observations

		Without Outlying Values			With Outlying Values		
Parameter	***Variable***	***Mean***	***Std.dev***	***MCse***	***Mean***	***Std.dev***	***MCse***
beta0	Intercept	0.365	0.068	0.0003285	0.342	0.074	0.0003718
beta1	Age	-0.001	0.001	0.000004778	-0.001	0.001	0.00000546
beta2	Weight	-0.017	0.001	0.000005248	-0.017	0.001	0.000005998
beta3	Sex	0.118	0.021	0.0001186	0.121	0.024	0.0001352
beta4	Livinv Area	0.097	0.020	0.0001045	0.114	0.022	0.0001253
σ^2		0.082	0.004	0.02013	0.109	0.005	0.0195

Conditional predictive ordinates (CPO) is a good measure used to examine extremeness or to detect outliers in the data. An index plot of 1/CPO as proposed by Ntzoufras(2009) with values greater than 40 can be used to detect outliers. Also a Bayesian residual analysis can be used to detect outliers by making a plot of studentized residuals and individuals above 2 or below -2 can be considered outlying. Therefore Figures 10 and 11 can be used to demonstrate the presence of outliers in the data. Furthermore, in order to verify if these outlying observations are influential, a different Bayesian multiple linear regression model is fitted to data without these outlying observations (47 outlying observations were noticed) and the results are compared in Table 7. The comparison showed little or no change in the parameter estimates thus implying that the 47 observations were termed outlying but were not influential. However, a difference was observed in the DIC values with the model without outlying values showing a smaller value of 329.3 and the model with outlying showing a value of 631.84. This implies the model without outlying values has a better predictive ability.

3 Now take the clustering into households also into account. Compute now the 95%-ile and 99%-ile of the exposures and their uncertainty by averaging over the random effect distribution. Note justify the choice of the random effects distribution.

Clustering was taken into account and a Bayesian hierarchical model was fitted with random intercepts. The model with covariates is defined as follows;

$$log(Y_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Weight_{ij} + \beta_3 Sex_{ij} + \beta_4 LivingArea_{ij} + b_{0i} + \epsilon_{ij}$$

where: $log(Y_{ij})$ is the log median cadmium value of the i^{th} individual in the j^{th} household, with $i = 1, ..., 1008$ and $j = 1, ..., 503$. b_{0i} represents the random intercept for the i^{th} subject. Age represents the age of the individuals (continuous). Weight represents the weight of the individuals (continuous). Sex is the gender of an individual coded (1=Male, 0=Female) and living area represents individuals living area coded Living Area(1=urban, 0=rural). The fixed effects parameters were modelled with $N(0, 1.E - 4)$. The random intercept $b_{0i} \sim N(0, \tau_{b0})$ where $\tau_{b0} \sim dunif(0, 100)$ and $\sigma^2_{b0} = \frac{1}{\tau_{b0}}$. $\epsilon_{ij} \sim N(0, \tau)$ where $\tau \sim IG(0.001, 0.001)$ and $\sigma^2 = \frac{1}{\tau}$
To perform this analyses, three chains, 50000 iterations and a burnin of 15000 is used. This led to 35000 iterations used to compute the posterior mean and posterior variance of the distribution. To validate the analyses, convergence diagnostic tools were used, such as the trace plots, auto-correlation plots, Brooks-Gelman-Rubin diagnosis plots and a cross correlation plot for the parameters. The response variable (median exposure value of cadmium) was transformed to log form to get a normal distribution. Furthermore, sensitivity analyses was done by varying the prior of the parameters.

3.0.1 Sensitivity analysis

Sensitivity analysis was performed by changing the prior distribution of the beta and variance parameters. The priors of the beta parameters were varied by using N(0.01, 1.0E-4) ,N(0, 1.0E-9) and dt(0,01,100). The prior of variance parameters was varied by using a IG(0.001, 0.001) for σ^2_{b0} and dunif(0,100) for σ^2 and on the other hand a dunif(0,100) for both σ^2_{b0} and σ^2. Changing these priors led to little changes in the beta parameters and no changes for the variance parameters implying that changing the priors had no impact on the analysis and thus non-informative priors were utilised.

Table 8: Varying Prior of Beta's

Parameter	**dnorm(0.01,1.0E-4)**	**dnorm(0,1.0E-9)**	**dt(0,0.1,100)**
beta0	0.267(0.065)	0.267(0.065)	0.292(0.072)
beta1	-0.002(0.001)	-0.002(0.001)	-0.002(0.001)
beta2	-0.015(0.001)	-0.015(0.001)	-0.015(0.001)
beta3	0.105(0.016)	0.105(0.016)	0.105(0.016)
beta4	0.098(0.029)	0.098(0.029)	0.098(0.029)

Table 9: Varying Prior distribution for variance parameters

Parameter	**dgamma(0.001,0.001)**	**dunif(0,100)**
σ^2_{b0}	0.069(0.006)	0.069(0.006)
	dunif(0,100)	**dunif(0,100)**
σ_2	0.041(0.003)	0.041(0.003)

3.1 Convergence diagnosis for Hierarchical Bayesian linear regression model

Figure 12 shows trace plots of the parameters indicating stationarity was achieved with a thick horizontal line an indication of high mixing rate and individual moves are not easily noticeable which falls inline with Gelfand and Smith (1990) proposal. Figure 13 shows an auto-correlation plot indicating that starting values are easily forgotten in estimating the parameters, this implies good and independent sample moves were made during sampling. Figure 14 shows a formal graphical diagnosis using the Brooks-Gelman-Rubin procedure where convergence is attained when difference in chains cannot be noticed. The figure 14 shows all parameters attained convergence after about 3000 iterations since the chains all converged to the horizontal line at 1. Figure 15 shows a cross correlation plot which indicates correlation between the parameters. The parameters correlation between σ and deviance was showed to be pretty high and this is as a result of similarities in the parameter.

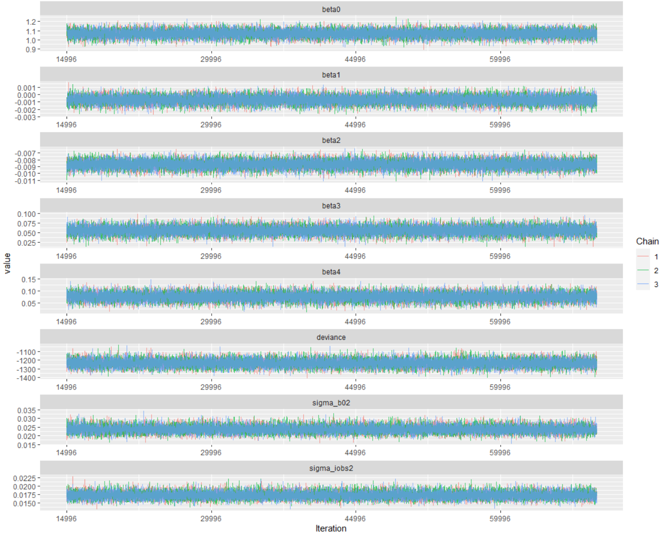

Figure 12: Trace Plots

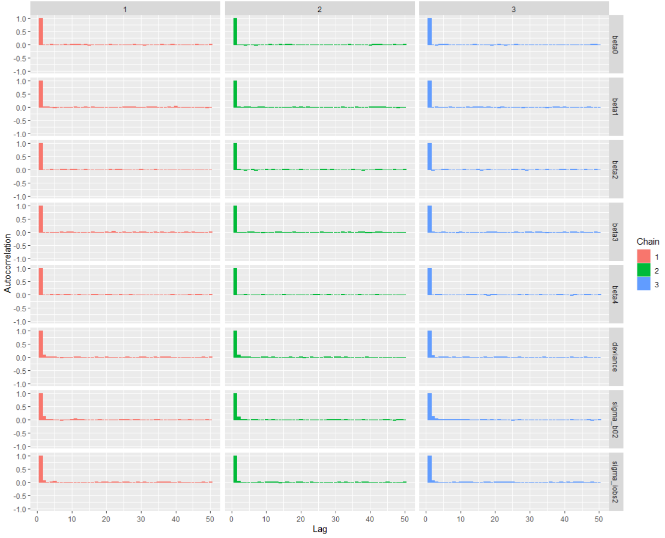

Figure 13: Auto correlation

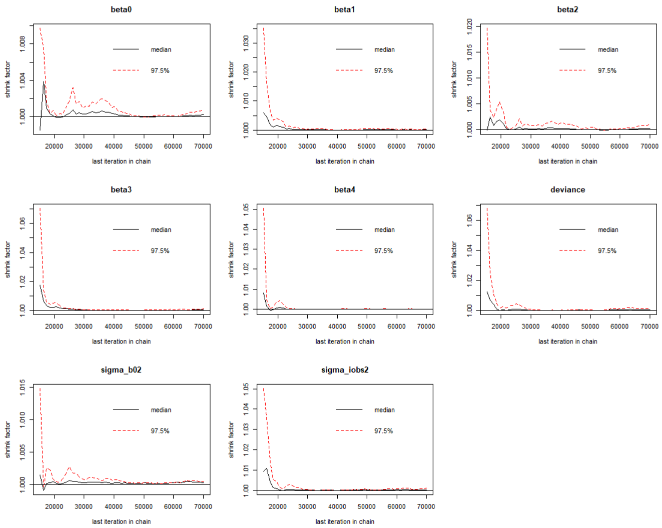

Figure 14: Brooks-Gelman-Rubin

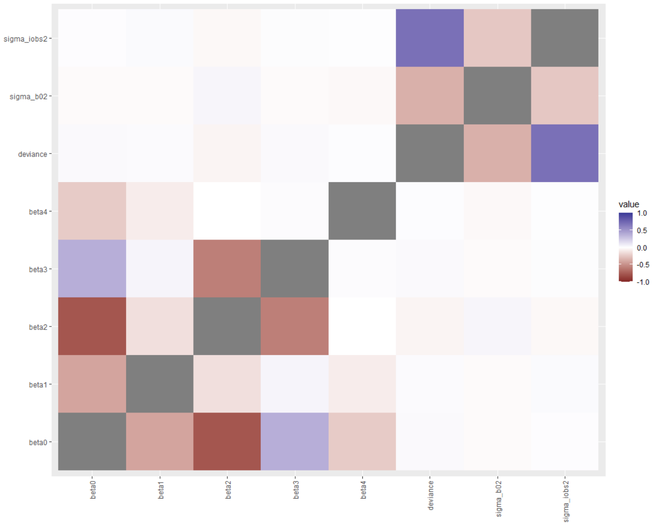

Figure 15: Cross Correlation

3.1.1 Posterior Summaries

Table 10: Postetrior Parameter Summaries For Bayesian Hierarchical Multiple Linear Regression

Parameter	Variable	mu.vect	sd.vect	2.5%	75%	95%	99%	MCse
beta0	Intercept	0.268	0.065	0.141	0.312	0.374	0.420	0.0003807
beta1	Age	-0.002	0.001	-0.003	-0.001	0.000	0.000	0.000004682
beta2	Weight	-0.015	0.001	-0.017	-0.014	-0.014	-0.013	0.000005597
beta3	Sex(male)	0.105	0.016	0.072	0.116	0.132	0.143	0.00009238
beta4	Living Area(urban)	0.098	0.029	0.042	0.117	0.145	0.165	0.0001571
sigma_b02	σ^2_{b0}	0.069	0.006	0.058	0.073	0.080	0.085	0.0001755
sigma_iobs2	σ^2	0.041	0.003	0.036	0.042	0.045	0.047	0.00003812
r	intra-correlation	0.627	0.028	0.571	0.609	0.646	0.677	

Table 10 illustrates the posterior parameter estimates obtained by fitting a Bayesian hierarchical model with random slopes to the log cadmium values with age, weight, sex(male) and living area(urban) as covariates. In order to determine which covariates influence the response, the confidence interval should not contain zero. The table 10 shows the variables Age is not significant. However the variable Weight is seen to have a significant diminishing effect on the levels of cadmium. The columns 2.5%, 75%, 95%, 99% confidence intervals pertains to the plausible values given the data. In order to ensure efficiency estimates are obtained from the MCMC,the Monte-Carlo (MC) error should be less than 5% of the respective standard deviations. The Monte-Carlo error is less than 5% the standard deviation for both parameters. $\sigma^2_{b0} = 0.069$ represents the variability in the random intercept implying little variability is present between households. Also, $\sigma^2 = 0.0.041$ represents the measurement error variability in the data and $r = 0.627$ represents the correlation between within individuals of the same households which is quite high. Since the σ^2 is relatively smaller than the σ^2_{b0} this implies very little shrinkage is observed in the prediction of the mean cadmium of the individuals in the household. The $r = 0.627$ controls the amount of shrinkage of the mean

3.2 Model Diagnosis

3.2.1 Normal Distribution of Random Intercept

Figure 16 and Figure 17 shows the histogram for the random intercept and the theoretical quantile plot for the random intercepts respectively. Figure 17 illustrates a slight deviation at the both tails of the 45 degree line. However, both plots suggest that the distribution of the random intercept can be assumed to be normally distributed.

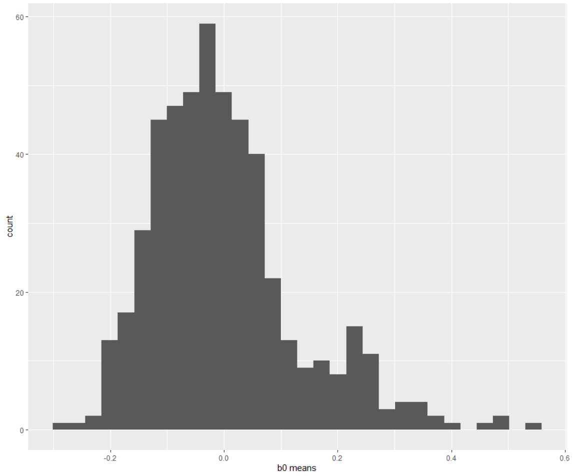

Figure 16: Histogram of Random Intercepts

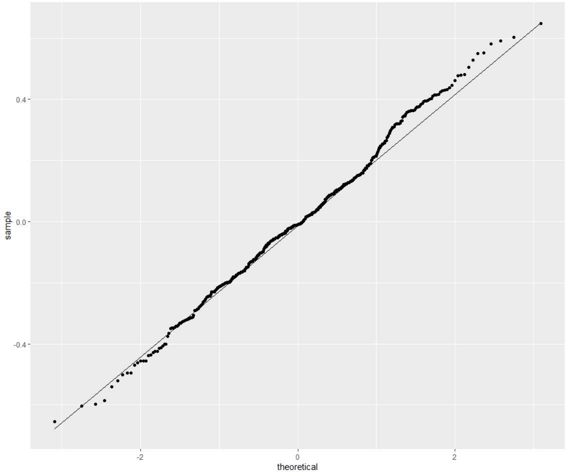

Figure 17: Normal Q-Q Plot

3.2.2 Posterior Predictive Checks

In the Bayesian framework the Posterior Predictive Check (PPC) is a global measure for assessing the goodness of fit of a model. Provided the Markov Chain has converged, the Posterior Predictive Check is based on sampling from the Posterior Predictive Distribution and it compares the extremeness of the sampled values with the observed values. This implies that the statistic of values computed for the observed values and the one for the sampled values from the posterior predictive density are compared. As a result, Table 11 illustrates no evidence against the model thereby indicating the model fits the data properly. Furthermore, a comparison of the hierarchical model ($DIC = 689.6$) and the non-hierarchical model($DIC = 631.84$) in section 2 in terms of DIC points the non-hierarchical model to be the better model in predicting cadmium levels since it has a lower DIC.

Table 11: Posterior Predictive p-value for random intercept with different goodness of fit tests

Statistic	**PPP-value**
Minimum	0.462
Maximum	0.577
Skewness	0.451
Kurtosis	0.489
Sinharay and Stern	0.519

4 Compute the 75%-ile of the median exposure values of cadmium. What is the probability of exceeding this threshold and what are the factors that impact that probability

A binary variable was created where 1 = individuals that exceed the 75 percentile of cadmium levels and 0 for individuals that are below the 75 percentile of cadmium levels. In order to analyse the probability for an individual to exceed the 75% of cadmium levels while taking the clustering of households into account, a Bayesian logistic random intercept model (GLMM) was fitted. The model with covariates is defined as follows

$$logit(\pi_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Weight_{ij} + \beta_3 Sex_{ij} + \beta_4 LivingArea_{ij} + b_{0i}$$

where: $\pi_{ij} = P(Y_{ij} = 1)$ is the probability that the i^{th} individual in the j^{th} household exceeds the 75 percentile of cadmium with $i = 1, ..., 1008$ and $j = 1, ..., 503$. Age represents the age of the individuals (continuous). Weight represents the weight of the individuals (continuous). Sex is the gender of an individual coded (1=Male, 0=Female) and living area represents individuals living area coded Living Area(1=urban, 0=rural). The fixed effects parameters were modelled with $N(0, 1.E-4)$. The random intercept $b_{0i} \sim N(0, \tau_{b0})$ where $\tau_{b0} \sim dunif(0, 100)$ and $\sigma^2_{b0} = \frac{1}{\tau_{b0}}$.

To perform this analyses, three chains, 50000 iterations and a burnin of 15000 is used. This led to 35000 iterations being used to compute the posterior mean and posterior variance of the distribution. To validate the analyses, convergence diagnosis were used such as the trace plot, auto-correlation plot, Brooks-Gelman-Rubin diagnostic plots and a cross correlation plot for the parameters were utilised. The response variable (median exposure value of cadmium) was transformed to log form to get a normal distribution. Furthermore, sensitivity analyses was done by varying the prior of the parameters.

4.1 Sensitivity analysis

Sensitivity analysis was performed by changing the prior distribution of the beta and variance parameters. The priors of the beta parameters were varied by using N(0.01, 1.0E-4) ,N(0, 1.0E-9) and dt(0,01,100). The prior of the random intercept's variance parameter was varied by using a IG(0.001, 0.001),dunif(0,100) and dunif(0,1000). Changing these priors led to little changes in the beta parameters and no changes for the variance parameters implying that changing the priors had no impact on the analysis and are thus non-informative priors were utilised.

Table 12: Varying Prior of Beta's

Parameter	**dnorm(0.01,1.0E-4)**	**dnorm(0,1.0E-9)**	**dt(0,0.1,100)**
beta0	10.654(1.348)	10.654(1.348)	10.454(1.168)
beta1	-0.014(0.013)	-0.014(0.013)	-0.014(0.013)
beta2	-0.203(0.023)	-0.203(0.023)	-0.251(0.053)
beta3	1.410(0.286)	1.410(0.286)	1.412(0.156)
beta4	1.291(0.387)	1.291(0.387)	1.352(0.255)

Table 13: Varying prior of σ^2_{b0} showing Posterior Mean(posterior standard deviation)

Parameter	**dunif(0, 100)**	**dunif(0, 1000)**	**IG(0.001,0.001)**
σ^2_{b0}	8.260(1.955)	8.260(1.955)	8.260(1.955)

4.2 Convergence diagnosis for Bayesian logistic random intercept model

Figures 18 shows trace plots of the parameters indicating stationarity was achieved with a thick horizontal line an indication of high mixing rate and individual moves are not easily noticeable which falls inline with Gelfand and Smith (1990) proposal. Figure 19 shows an auto-correlation plot indicating that starting values are easily forgotten in estimating the parameters, this implies good and independent sample moves were made during sampling. Figure 20 shows a formal graphical diagnosis using the Brooks-Gelman-Rubin procedure where convergence is attained when difference in chains cannot be noticed. The figure 20 shows all parameters attained convergence after about 3000 iterations since the chains all converged to the horizontal line at 1. Figure 21 shows a cross correlation plot which indicates correlation between the parameters.

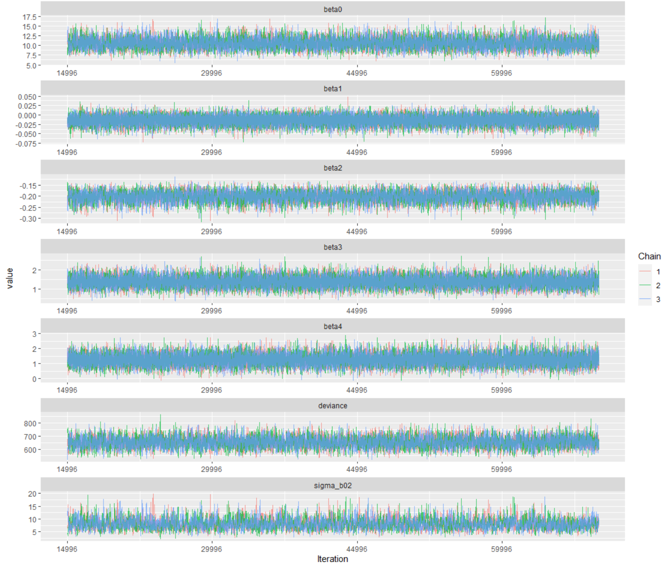

Figure 18: Trace Plots

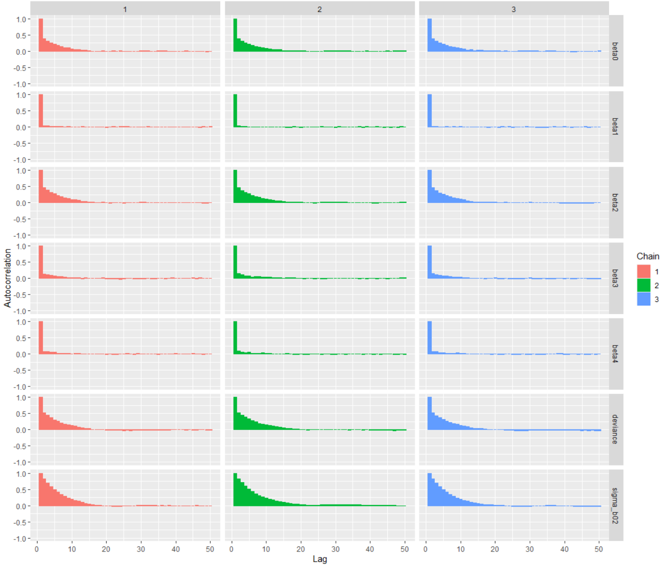

Figure 19: Auto correlation

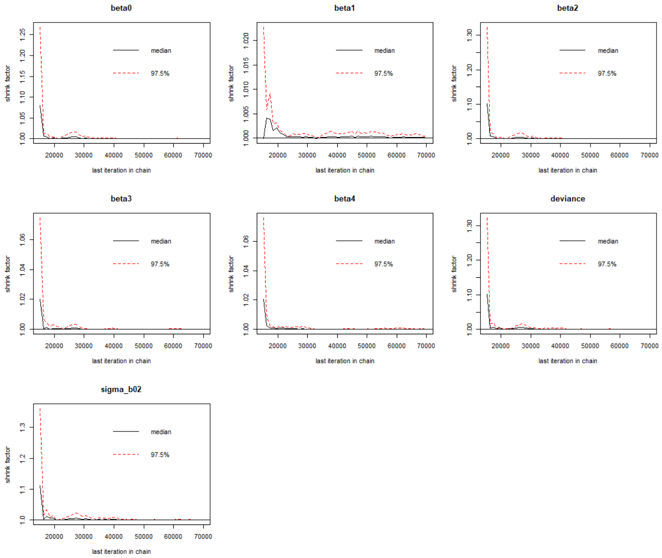

Figure 20: Brooks-Gelman-Rubin

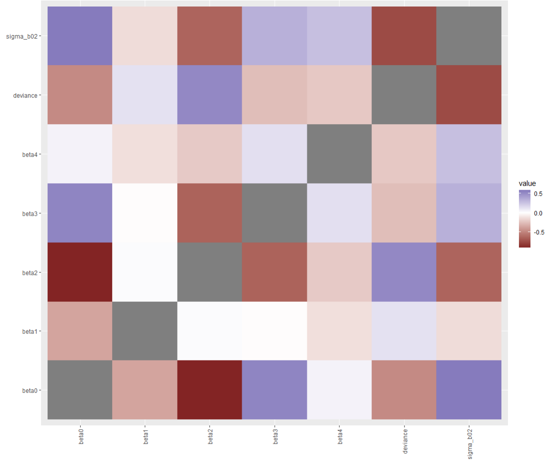

Figure 21: Cross Correlation

4.3 Posterior Parameter Estimates

Table 14: Posterior Parameter Estimates

Parameter	**Variable**	**mean**	**Std dev**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	**MCse**
beta0	Intercept	10.654	1.348	8.100	9.728	10.613	11.553	13.387	0.01776
beta1	Age	-0.014	0.013	-0.040	-0.023	-0.014	-0.006	0.011	0.0000845
beta2	Weight	-0.203	0.023	-0.248	-0.219	-0.203	-0.187	-0.160	0.0003244
beta3	Sex (Male)	1.410	0.286	0.872	1.208	1.400	1.602	1.977	0.002439
beta4	Living Area (Urban)	1.291	0.387	0.562	1.026	1.284	1.549	2.071	0.002851
σ^2_{b0}		8.260	1.955	4.683	6.861	8.226	9.633	12.148	0.005829

Table 14 illustrates the posterior parameter estimates obtained by fitting a Bayesian logistic model with random slopes to model the probability that the i^{th} individual in the j^{th} household exceeds the 75 percentile of cadmium with age, weight, sex(male) and living area(urban) as covariates. In order to determine which covariates influence the response, the confidence interval should not contain zero. The table 14 shows the variables Age is not significant. However the variable Weight is seen to reduce the probability to exceed the 75 percentile. The columns 2.5%, 75%, 95%, 99% confidence intervals pertains to the plausible values given the data. In order to ensure efficient estimates are obtained from the MCMC,the Monte-Carlo (MC) error should be less than 5% of the respective standard deviations. The Monte-Carlo error is less than 5% the standard deviation for both parameters. $\sigma^2_{b0} = 8.26$ represents the variability in the random intercept implying quite some variability is present between households. The intra-class correlation was calculated to be 0.715 implying quite some similarity within individuals of the same household.

4.4 Model Diagnosis

4.4.1 Normal Distribution of Random Intercept

Figure 22 and Figure **??** shows the histogram for the random intercept and the theoretical quantile plot for the random intercepts respectively. Figure **??** illustrates deviations at the both tails of the 45 degree line and the points not lying on the line. Figure **??** implies the true distribution of the random intercept is not normally distributed.

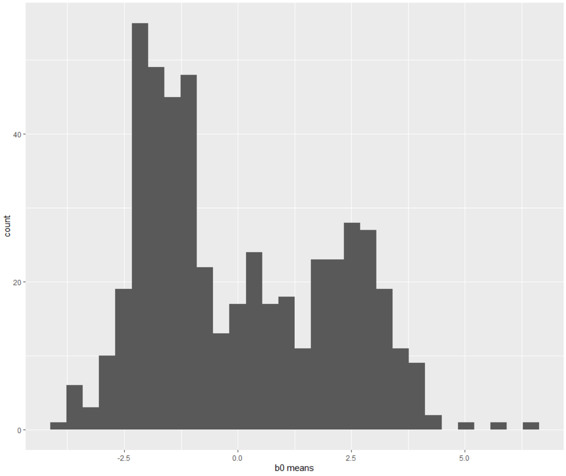

Figure 22: Histogram of Random Intercepts

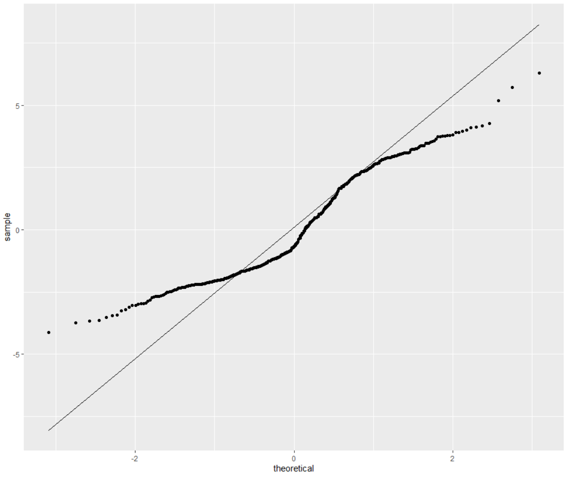

Figure 23: Normal Q-Q Plot

4.4.2 Posterior Predictive Checks

Table 15: Posterior Predictive p-value for random intercept with different goodness of fit tests

Statistic	**Posterior mean(p_D)**
Minimum	0.497
Maximum	0.461
Skewness	0.446
Kurtosis	0.475
Sinharay and Stern	0.463

In the Bayesian framework the Posterior Predictive Check (PPC) a is global measure for assessing the goodness of fit of a model. Provided the Markov Chain has converged, the Posterior Predictive Check is based on sampling from the Posterior Predictive Distribution and it compares the extremeness of the sampled values with the observed values. This implies that the statistic of values computed for the observed values and the one for the sampled values from the posterior predictive density are compared. As a result, Table 15 illustrates no evidence against the model thereby indicating the model fits the data properly.

References

[1] Gelfand, A. E. and Smith, A. F. (1990). Sampling-based ap-proaches to calculating marginal densities.Journal of the American statistical associa-tion, 85(410):398–409.

[2] Lesaffre, E. and Lawson, A. B. (2012).Bayesian biostatistics.John Wiley Sons.

[3] Ntzoufras, Ioannis. 2009. Bayesian Modeling Using WinBUGS. Wiley. $http://books.google.com?id = QfMPDvLpV5cC$.

[4] Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2003). “WinBUGS Version 1.4 Users Manual.” MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/.