Master Thesis in Statistics and Data Mining

Differences in cognitive performance between children of different birth orders and family sizes

Lauren Duff

Abstract

The objective of this thesis was to investigate the relationship between children's cognitive performance, on one hand, and their family sizes and birth order, on the other hand. Frequentist and Bayesian inferences have been used to select variables to arrive to a statistical model explaining variation in children's cognitive performances. Cognitive performance has been measured by a children's math test.

The thesis has used data merged from two different datasets, the National Longitudinal Survey of Youth, NLSY79, and the NLSY97 Child/Young adults which contained 1578 children in 1095 families with 16 explanatory variables.

The models that have been used are a multiple regression model, a multilevel model from a frequentist as well as from Bayesian inference. Several types of variable selection have also been introduced for both inferences such as forward selection, backward elimination, lasso and spike and slab priors.

Results showed that for higher birth orders and larger family sizes, the response variable PIAT score declined. The variable selection method that produced the preferred model was the lasso. The variables that impacted a child's cognitive performance from the estimated lasso were the father's presence, family size, mother's IQ, age of the mother at first birth, income, birth order, age of the child and the mother not being married.

Acknowledgements

I would like to thank my supervisors, Linda Wänström and Bertil Wegmann, for the support and always being available to hear my thoughts and ideas. The support from friends and family has contributed more than they will ever know. My opponent, Keith Lowton, should not be forgotten for the good critic and the pastries during our revision meeting.

Contents

List of tables

1 Introduction

Many studies have found negative correlations between children's cognitive performances (measuring intelligence and related constructs) and their family sizes and birth orders (Blake, 1989). This thesis has looked at these relationships more closely. In the introduction section, the objectives of the thesis will be given, background information on cognitive performance and family size and birth order will be presented, and theories aiming to explain these relationships will be covered as well as a summary of related work within this area.

1.1 Objectives

The main objective of this master thesis has been to investigate the relationship between children's cognitive performance, on one hand, and their family sizes and birth order, on the other hand. In particular, two statistical inferences; the frequentist and the Bayesian, have been used to select variables to arrive at a statistical model explaining variation in children's cognitive performances. It has been of specific interest to see if children's cognitive performances differ between children of different family sizes, and between children of different birth orders. Cognitive performance has been measured by a children's math test. If the frequentist and the Bayesian inference result in different models, reasons to these differences have been explored. Reasons behind investigating different methods are based upon the field of applied research has used different methods and it will be interesting to see if the results differed.

1.2 Background

The field of cognitive performance within and between families is extremely large and there are large amounts of research on the subject. In general, studies have found (Wichman, Rodgers and MacCallum, 2005; Devereux, Black and Salvanes, 2005; Bjerkedal and Kristensen, 2007) that family size is negatively correlated with cognitive performance, such that children with more siblings perform worse, on average, on cognitive performance tests (such as IQ tests and other achievement tests). Studies have also found that children of higher birth orders, i.e. children born later into a family; on average perform worse on cognitive performance tests compared to children of lower birth orders (born earlier into a family). These relationships can be explained by different theories such as the resource dilution theory, confluence model and admixture hypothesis.

The resource dilution theory (Downey, 2001) is defined by how the family structure could benefit or hurt the wellbeing of children. The theory explains that parental resources are finite and with the increase of family members, these resources (which can be parental time, energy, and wealth) decline with each child. Resources are diluted within the family. For example, children in larger families cannot be given as much affection and attention from their parents as children from smaller families (Downey, 2001). Parents provide three types of finite resources: settings, treatments and opportunities (Blake, 1981). Settings include the living situation, life's necessities and developmental objects such as books and music. Treatments are comprised of the parent's teaching as well as the amount of attention provided to the children. The last resource, opportunities, gives the children more experiences to widen their horizons. It is also discussed that students learn more in smaller classes so this theory extends to the school situation. Certain resources are equally shared between the siblings such as toys and books. However, money for college is often split between all of the siblings and studies have shown (Blake, 1981) those parents with many children feel less responsible to pay for their children's college compared to parents with fewer children.

Because parental resources are diluted in families with more children, this theory predicts lower scores on cognitive performance tests for children with more siblings compared to children with fewer siblings, and thus predicts a negative relationship between these variables. According to the theory however, the addition of one sibling has a more drastic effect in smaller families and the negative effect of family size should decrease as family size increased.

The confluence model (Zajonc and Sulloway, 2007) is a mathematical model that explains family size and birth order differences in cognitive performance by specifying that children are influenced by the intellectual levels in their families. The intellectual level of the family is made up of the intellectual levels comprised of each family member. Each parent is assigned an absolute intellectual level value (for example 30). The child's intellectual value is assumed to be zero at birth, and as he/she grows up, his/her intellectual level increases. At age 10, the child's intellectual level can be 10 for example. The average intellectual level of the family increases for the child with age, until a sibling is born into the family. The second born child starts from a lower intellectual value but will shortly pass the first child according to the confluence algebra. The first born child however benefits from tutoring the second born child, which indicates that being born early into a family is more advantageous than being born later (compared to a child's siblings).

According to the model, children in single-parent households experience lower intellectual family levels and children born in families with many adults experience higher levels. The birth gaps are also shown to be important between children. Smaller birth gaps between children point to lower intellectual environments. In conclusion, the confluence model predicts that as family size increases, the overall cognitive performance for the entire family declines (Zajonc and Sulloway, 2007). It is predicted that children from smaller families and children with lower birth orders have a higher cognitive performance compared to children with larger family sizes and/or with high birth orders.

The admixture hypothesis (Rodgers, 2001) explains the relationship between family size and birth order and cognitive performance by stating that it is partly due to other factors such as the parents' cognitive ability and socioeconomic status. In general, parents with lower cognitive ability tend to have more children which can account for lower cognitive ability in children with larger family sizes. Because most studies that have found a relationship between birth order and cognitive performance are conducted on between family data (i.e. comparing a first born in one family to a second born in another family to a third born in yet another family etc.) children of lower birth orders have, on average smaller family sizes than children of higher birth orders (Rodgers, 2001). In summary, this theory assumes that the relationship between family size and birth order to cognitive performance is, at least partly, spurious and caused by other factors.

1.3 Previous studies

Various studies have investigated the relationships between cognitive performance and family size and birth order. Most studies have been conducted on between family data that compared one child in one family to another child in another family. These studies have found negative correlations between family size and cognitive performance and between birth order and cognitive performance (Wichman, Rodgers and MacCallum, 2005; Devereux, Black and Salvanes, 2005; Bjerkedal and Kristensen, 2007; Zajonc and Sulloway, 2007; Downey, 2001).

Wichman, Rodgers and MacCallum (2005) studied within family data from the NLSY79 data using multilevel models, and found that there was a decline in children's cognitive performance for higher birth orders. They however concluded that the birth order effect lies between and not within families. They compared three different multilevel models where an explanatory variable was added in each model. The first model contained only the birth order effect, the second model an age cohort dummy variable specifying two age groups (younger/older children sample) and the third model contained the mother's age at the birth of the first child. The first and second model showed significant negative birth order effects consistent with other research. However, when between family variance was controlled in the third model, the birth order effect was no longer a significant determinant of a child's cognitive performance.

However, Devereux, Black and Salvanes (2005), found a strong birth order effect on a child's cognitive performance both within and between families. The data came from Norwegian families and the variables used were birth order, test year, educational fulfillment for the parents, father's cognitive performance, as well as demographic variables such as gender and age. This study found a strong birth order effect however family size had little or no effect on a child's cognitive performance.

Bjerkedal and Kristensen (2007) conducted a similar study on Norwegian families on birth order and cognitive performance. They included explanatory variables such as income and birth weight. The focus of this study was to compare scores of siblings of close birth orders both within families and between families. Using ordinary least squares (OLS) regression models they found that cognitive performance declined within families for higher birth orders, although within-family effects were small. With higher birth orders, the differences tended to level off and no concrete results could be found. The main differences were found between the first and the second born. This study concluded evidence to support the confluence theory.

In general, the above studies have found that cognitive performance declined with higher birth orders and whether this decline lies within or between families, or both, was uncertain. Previous studies also have found conflicting results on family size. A difference between Wichman et al.´s, Devereux et al.´s and Bjerkedal et al.´s analyses was that the former used multilevel modeling, and the latter used summary statistics and OLS regression. OLS regression does not take into account the different levels of observations (such as child and family) which could contribute to the varied results from these studies. In addition, they used different explanatory variables in the models as well as different measures of cognitive performance. This study has used within family data and has included more explanatory variables that distinguished the siblings and the families to account for the between and within family variance.

2 Data

This section will give the reader an understanding of the data and the variables that have been used for the results shown in section four. First a description of the data will be presented followed by which variables have been chosen and the reasoning behind these specific variables for the thesis.

2.1 Description of data

This thesis has used data merged from two different datasets, the National Longitudinal Survey of Youth, NLSY79, and the NLSY97 Child/Young adults. The NLSY79 data consisted of 12,686 young men and women living in the United States in 1979 and who were thereafter interviewed every year. The participants were between the ages of 14 to 22 when first interviewed. The dataset originated from a multi-stage stratified area probability sample from almost all of the 50 states as well as the District of Columbia in the United States (Zagorsky and White, 1999).

The NLSY97 Child/Young adults were a separate survey performed from 1986 on the children born to the female respondents of the NLSY79 dataset. These children were interviewed on a biannual basis. Because the two datasets had information on the females' (mothers') identification numbers, they could be merged. The data thus contained information both on the children and their mothers, such as cognitive performance tests, labor market behavior, educational experience, family background and health issues.

2.2 Variables

The response variable of interest in this thesis was a child's cognitive performance. Different tests have been used to measure a child's cognitive ability, and in this thesis the Peabody Individual Achievement (PIAT) math subtest (Zagorsky & White, 1999) has been used. The PIAT test was used to obtain an individual's scholastic attainment and included the subtests math, reading comprehension, and reading recognition. However because earlier studies have shown that math scores may be more influenced by family size and birth order, only the math subtest has been analyzed in this thesis (Baker, 1993).

The math subtest contained 85 multiple-choice questions that increased in difficulty. The questions represented a cross-section of various curricula in use in the United States such as numeral recognition, geometry and trigonometry. The PIAT math subtest score was age-standardized and higher scores indicated greater cognitive performance (Zagorsky and White, 1999). Throughout this thesis the PIAT math score has been used as the indicator for a child's cognitive performance.

One of the main explanatory variables of interest was family size, measured by the number of biological children born to the mother and living in the household (from 1 to 11). The other main explanatory variable, birth order, was measured by the order in which the child was born, and it had values from being the first born (1) up to being the tenth born (10) by the mother.

Because there were thousands of variables in the original dataset, variables that have been shown to be important to the relationship between cognitive performance and family size and birth order were selected as control variables. The strong correlation between cognitive performance and family size/birth order got weaker when adding certain control variables (Wänström, 2007), which showed that it was important to add control variables.

Control variables chosen were age of the mother at the birth of the first child (in years) and the reasoning behind this choice was based on previous studies (Wichman, Rodgers and MacCallum, 2005). Total number of weekly hours worked at the mother's current job (included multiple jobs if applicable) was selected because according to the resource dilution theory if a mother worked many hours, she would not have time to give attention to the child compared to a stay at home mother.

Dummy variables were created for nominal variables. For the race of the mother, the first dummy variable (racemom1) was coded one if the race was African American, otherwise zero. The second dummy variable (racemom2) was coded one if the race was Hispanic, otherwise zero. Marital status was measured by four dummy variables measuring whether the mother was never married, separated, divorced or widowed. The variable levels chosen as the reference categories were the ones that had the largest proportion to the variable (non-African American/non-Hispanic and married). Race of the mother was selected based upon previous studies (Wänström, 2007). Father's presence in the household was coded one if present and zero if not. Father's presence and marital status were chosen because according to the confluence model single-parent households had lower intellectual family levels.

Poverty could be considered as a socioeconomic status (SES) indicator which has been found to have an impact on a child's cognitive performance from previous studies (Wänström, 2007). Poverty, was coded one, if the child was considered to live in poverty, zero otherwise. The poverty level was calculated by considering the size of the family unit and the family income (DeNavas-Walt, 2010).

Income was also an indicator of SES, which was the total net income for the past calendar year, which was measured in American dollars. The cognitive performance of the mother (variable named Mother's IQ in this thesis) was measured by the AFQT score. The AFQT score was the Armed forced qualification test which tested the person's word knowledge, paragraph comprehension, arithmetic reasoning and mathematics knowledge (Cascio and Lewis, 2005). The AFQT score was measured in percentiles. Chosen child variables were gender (males=0, females=1) and age (in years). Mother's IQ was chosen because of the admixture hypothesis which explained that cognitive performance was explained by other factors such as a parents' IQ. Descriptive statistics over the variables can be seen in tables 1 and 2.

Variable	Mean	Standard Deviation	Minimum value	Maximum value
PIAT score	105.227	15.001	65	135
Family size	2.944	1.443	1	11
Birth order	2.381	1.291	1	10
Age of mother first birth	25.878	5.479	13	39
Total number of hours worked in a week	31.809	26.288	0	160
Total income	76579.13	75081.41	0	408500
Mother's IQ	44.332	29.009	1	99
Child's age	10.876	2.365	5	14

Table 1: Means, standard deviations, min- and max values for numerical variables, N=1578

Variable	Values	Proportion
Racemom	Non-African American, Non-Hispanic (Reference variable)	0.574
	African American	0.247
	Hispanic	0.179
Marital status	Married (Reference variable)	0.692
	Never married	0.099
	Separated	0.052
	Divorced	0.143
	Widowed	0.014
Poverty	Not in poverty	0.852
	In poverty (Reference category)	0.148
Father's presence	Father not present	0.33
	Father present (Reference category)	0.67
Gender	Male	0.478
	Female (Reference category)	0.522

Table 2: Proportions for categorical variables, N=1578

3 Methods

In this section, the statistical methods used in the thesis have been explained. An introduction to the two different types of inferences, frequentist and Bayesian, have been given. The models that have been used were a multiple regression model and a multilevel model. Several types of variable selection have also been introduced for both inferences.

3.1 A frequentist vs a Bayesian inference

Statistical inference is made using two different types of methods: frequentist and Bayesian inference. Most statisticians are familiar with the frequentist way of thinking. Frequentist inference was developed by Fisher, Neyman and Pearson and was based on the ideology that probability is a limiting frequency (Wagenmakers, Lee, Lodewyckx, Iverson, 2008). Inferences are based exclusively on the sampling distribution and a frequentist does not condition on the observed data and views the parameters as fixed. The data is a repeatable random sample and the parameters are constant during this repeatable process. A frequentist's 95 percent confidence interval for a parameter does not imply that the parameter has a 95 percent probability of being within that interval. Their interpretation is instead that in 95 percent of the cases, the interval has covered the true parameter.

Bayesian inference was introduced by Thomas Bayes whom introduced Bayes' theorem in 1763 and was primarily used until 1920 and was used once again with the advancement of computers. Pierre-Simon Laplace developed Bayes' theorem eleven years after Bayes unaware of his publication. Laplace introduced inductive reasoning based on probability (Laplace, 1774). The Bayesian approach views the data as fixed. The data is observed from the realized sample, and the parameters are unknown and portrayed as probabilistic. The Bayesian approach uses prior probability distributions and likelihood functions. The prior probability function expresses the uncertainty about a parameter before having a look at the data. The likelihood function is a function of the parameters of the statistical model given the data. Combining the prior and likelihood the posterior is obtained which is the probability of the parameters given the data (Casella, 1990). Instead of using confidence intervals, a Bayesian uses a credibility interval. For example, a 95 percent credibility interval is interpreted as a 95 percent probability that this interval contains the parameter (Jaynes and Kempthone 1976).

3.2 Multilevel model

When analyzing the effect from several explanatory variables on a continuous response variable, a multiple regression model is often used and can be modeled as:

$$y_i = \alpha + \beta x_{ki} + \varepsilon_i, \qquad (3.1)$$

where $x_{1i}, \dots, x_{ki}$ represents the predictor variables, α is an intercept and the vector $\beta = (\beta_1, \dots, \beta_k)$ are the regression parameters, i = 1,…,n, and ε_i is the error term. A multiple regression model usually assumes $y_i \sim N\left(\alpha + \beta x_{ki}, \sigma_y^2\right)$ and ε_i are independent of each other.

When the data is hierarchical, i.e. there are different levels in the data, a multilevel model is appropriately used to analyze the effect on one response variable and several explanatory variables (Goldstein, 1999). In this case, children are nested within families so that children represent the first level of the data and families represent the second level. If a basic regression model is used, the variation between and within families cannot be detected and important group effects would be lost. In addition, if children within families are more correlated with each other (which is often the case because they share similar genes and environment) a basic regression model using OLS estimates does not provide correct standard errors (Goldstein, 1999). The response for the ith child in the jth family can be modeled by:

$$y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{ij}, \qquad (3.2)$$

where $\alpha_j = \alpha_0 + \mu_{0j}$

The multilevel model usually assumes the following $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$ and $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$. The $x_{1ij}, \dots, x_{kij}$ are the n explanatory variables which can be measured at the first level. α_j is a random intercept that is allowed to vary between families. $\alpha_0$is thus the general family intercept and μ_{0j} is the residual for family j. $\beta = (\beta_1, \dots, \beta_k)$ is a vector of fixed regression parameters assumed to be the same across families. In a multilevel model, any of these parameters are assumed to be random, however; only in the frequentist inference iss the intercept assumed to be random, in order to account for the correlation between children in the same family.

Intraclass correlation

In a multilevel model, the intraclass correlation is important to look at because it explains how strongly individuals in the same group resemble each other, and it is calculated as followed:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}, \qquad (3.3)$$

where σ_y^2 is the variance for within the families and σ_α^2 is the variance for between families. The correlation can also be interpreted as the amount of unexplained variation in y that is between families. The value of the correlation lies between zero and one and if the value is larger than zero it was an indication that there exists some kind of social grouping. Basic regression models should in that case not be used because different levels of variance are not taken into account and standard errors become inflated. If the value is close to one there is however no variance to explain at the individual level (Snijders and Bosker, 2000).

3.3 Frequentist inference

A general description of a multilevel model has been given in 3.2. Inferences from a frequentist perspective will be given here.

3.3.1 Parameter estimation

When estimating the model in equation 3.2 from a frequentist perspective, the procedure PROC Mixed in SAS is used. This procedure uses restricted/residual maximum likelihood estimation (REML) to estimate the parameters (Littel, 2006). REML partitions the likelihood into two parts; the first component is the regression residuals for the observations for the fixed parameters and the second component is the residual likelihood that contains the variance parameters of the random effects. The first component then has no fixed effects and all the residuals have a mean value of zero. Maximum likelihood estimation on the residuals to get estimates of the variance components is done where the likelihood is maximized for each component is done separately (O'Neill, 2010). REML produces less biased estimates of random effects variances and is less sensitive to outliers compared to maximum likelihood estimation (MLE).

3.3.2 Model comparison

A likelihood ratio test is used to test the null hypothesis that σ_{α}^{2}=0. If this hypothesis is rejected, a model with a random intercept is preferred over a model with a fixed intercept, i.e. a multilevel model is preferred over a regular regression model. The likelihood ratio test is computed as followed:

$$D_{01} = -2log_e(\frac{\lambda_0}{\lambda_1}), \qquad (3.4)$$

where λ_0 is the likelihood of the model under the null hypothesis, and λ_1 is the likelihood of the model under the alternative hypothesis. In the case of comparing a model with a random and a fixed intercept, the λ_0 is the likelihood for the model with the fixed intercept and the λ_1 is the likelihood for the model with the random intercept. D_{01} is χ^2 distributed with q degrees of freedom, where q is the difference in the number of estimated parameters between the two models.

Frequentists tend to use two different comparative fit indices to determine which of some competing model was preferred: the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Kadane and Lazar, 2004). Both of these fit indices have been used for model comparison in this thesis.

AIC is a measure of the quality of the model being tested. This test takes into account the complexity of the model as well as the goodness of fit of the model (Akaike, 1974). AIC is calculated by:

$$AIC = -2LL + 2p, \qquad (3.5)$$

where p is the number of maximum likelihood estimated parameters and LL is the log likelihood. The preferred model is the one with the lowest AIC value. AIC is mainly used when the primary goal of the modeling is prediction and penalizes models with many parameters. The AIC value is looked at when using the lasso.

Another way to compare models is by looking at the Bayesian Information Criterion (BIC) and which is sometimes also called for the Schwarz information criterion (Schwartz, 1978). The BIC favors models where the posteriori is most probable. The BIC is calculated by:

$$BIC = -2LL + \ln{(n)} * p, \qquad (3.6)$$

where p is the number of maximum likelihood estimated parameters and LL is the log likelihood. The BIC is mainly used when the primary goal of the modeling is for descriptive purposes, i.e., to feature the most important explanatory variables that influences the response variable which makes this criterion more critical to this thesis. The model with the minimum value of BIC is selected and this model also penalizes a model with many parameters. The BIC value is looked at in all of the variable selection methods.

The difference between the AIC and the BIC is that BIC tends to favor models with fewer explanatory variables compared to AIC. AIC tends to overestimate the number of parameters in the model.

3.3.3 Variable selection

Variable selection is used to select the best subset of explanatory variables that explains the response variable. The advantages of variable selection (over estimating a model that includes all possible explanatory variables) is that unnecessary explanatory variables add noise to the estimation of other variables that are of interest. Multicolinearity may also be a problem if there are many explanatory variables, and they are highly correlated with each other (Kutner, Nachtsheim, Neter and Li, 2005).

In regular regression analysis, variable selection is often done using some type of stepwise procedure, adding or removing explanatory variables depending on whether or not they significantly contribute to explaining the variance in the response variable (Miller, 2002). Variable selection has not been as frequently used for multilevel models. When it has been used, it is usually done by comparing the different models using tests such as the likelihood ratio test or by comparing BIC values. The problem in finding a good method for variable selection for this type of model is that the explanatory variables can be selected for each level or across levels. Variable selection has only been done on the fixed explanatory variables in this thesis (Dedrick, Ferron, Hess, Hogarty, Kromrey, Lang, Niles and Lee, 2009).

Two well-known variable selection methods are first tested in this thesis. Starting with a multilevel model with all of the variables and then removing one variable at a time that is not significant, otherwise known as backward elimination, has been used. To determine if a variable should be removed a t-test is performed with the test statistic:

$$t_g^* = \frac{b_g}{s\{b_g\}}, \qquad (3.7)$$

where b_g is the chosen g estimated parameters value and $s\{b_g\}$ is the standard deviation for the chosen estimated parameter b_g. The variable with the smallest test statistic is removed first if the p-value iss above the significance level. The significance level is chosen to ten percent to avoid missing important variables in the multilevel model. After the removal of a variable the BIC value is checked to see if the value is lower than the model before. If the BIC value decreased, another variable is removed based upon the same criteria until the BIC value is higher than the model estimated before. This process is repeated until all of the variables with a p-value above the significance level are removed and the model has the lowest BIC value of the models compared (Kutner, Nachtsheim, Neter and Li, 2005).

Alternatively, starting with a null model and adding explanatory variables one at a time to see whether the new variable is significant, also known as forward selection was also used (Raudenbush and Bryk, 2002). The variable that has been chosen first to be added to the model is the variable that has the highest t-value in models with only one explanatory variable as well as based upon whether the p-value is below the level of significance. The level of significance is chosen to ten percent here as well for comparison. Once the new variable is added, the BIC value is recorded. This process continues until the model with the lowest BIC value is found.

3.4 Lasso

Another method for variable selection is to do a lasso regression. The lasso is viewed as either a frequentist or Bayesian method of inference. Lasso stands for least absolute shrinkage and selection operator. A lasso regression is a shrinkage and selection method, which minimizes the usual sum of squared errors with a bound on the sum of the absolute values of the coefficients (Tibshirani, 1996). This method can be seen as a combination of subset selection and ridge regression, where it shrinks some coefficients and sets others to zero. The lasso estimator in equation 3.8 is for an ordinary regression model and is presented for illustrative as well as intuitive reasons. The estimator is calculated as:

$$\tilde{\beta}_L = \arg\min(\tilde{y} - X\beta)'(\tilde{y} - X\beta) + \ t\sum_{k=1}^{g} |\beta_k| \ , \qquad (3.8)$$

where $\tilde{y} = y - \ \bar{y}1_n$, X is the number of observations times number of parameters (n x p) matrix of the standardized explanatory variables, β_k are the regression parameters ranging from k=1…g, and $t \geq 0$ is the shrinking parameter (Park and Casella, 2008). If t is equal to zero then this became a simple OLS estimator. Hence, the first part of the equation in 3.8 is the OLS estimator.

When estimating the lasso, a value for the shrinking parameter, t, is essential since it controls the degree of shrinkage of the estimator. t can be chosen in two different ways and to select t in this thesis cross-validation has been done. The method used for cross-validation was the k-fold (Chand, 2012). The value for t was chosen to eight and the BIC value (see equation 3.6) has also confirmed that when t was eight we got the best model. Lowering the t more than this value, the BIC started to increase.

The computation of the lasso was done in R using the lmmlasso package. This package was created based from the work of Schelldorfer, Buhlmann and Van De Greer (2011) which fits linear mixed-effects models with lasso penalty for the fixed effects. The residual sum of squares,$(\tilde{y} - X\beta)'(\tilde{y} - X\beta)$, is minimized with respect to the non-differentiable constraint expressed in terms of the L_1 norm regularization of the coefficients. Regularization is used by penalizing models with a process of introducing additional information in order to solve ill-posed problems or to prevent overfitting (Park and Casella, 2008).

When checking for which parameters are significant, it is common to use the estimated standard errors of the estimated parameters and/or the AIC/BIC measures that are defined in equations 3.5 and 3.6. However, standard errors are difficult to estimate since the lasso estimates are non-linear and non-differentiable functions of the response values (Tibshirani, 1996). Because the lasso shrinks some parameters close to zero, we consider these parameters to be non-significant and, as such, remove the corresponding variables when comparing the variable selection between the different types of methods.

3.5 Bayesian inference

Bayesian inference is used to estimate the data using the multilevel model in equation 3.2. The difference with the Bayesian approach is that unknown quantities are treated as random variables and a prior probability distribution is needed to be assigned to each of them. Bayes theorem combines the likelihood function with the prior distribution to a posterior distribution of the unknown parameters and inference on the parameters is created by estimating the posterior (Lunn, Thomas, Best and Spiegelhalter, 2000).

The posterior distribution for parameters β is written as:

$$p(\beta|y) \; \propto p(y|\,\beta) * p(\beta), \qquad (3.9)$$

where the posterior is proportional to the likelihood function multiplied by the prior parameters (Lunn, Thomas, Best and Spiegelhalter, 2000).

Calculating the intraclass correlation (equation 3.3) with Bayesian inference is a little bit more complicated. For every parameter draw and in our case 30000 draws, the intraclass correlation was calculated. The result was then 30000 intraclass correlations and the posterior mean of these draws could be viewed as an estimate of the intraclass correlation.

Estimating the posterior has been done in the statistical program BUGS, which stands for Bayesian inference Using Gibbs Sampling (Lunn, Jackson, Best, Thomas and Spiegelhalter, 2012).

Gibbs sampling is a certain Markov chain Monte Carlo (MCMC) method which simulates from the given distributions and once the process has converged, it provides a sample from the desired posterior distribution (Hastie and Tibshirani, 2009).

This results in a Markov chain with draws of β from the posterior. In each step of the Gibbs sampler, a new parameter is drawn given the previously accepted draws of the other parameters. The Gibbs sampler only consideres univariate conditional distributions (all of the parameters but one is assigned a fixed value) (Guo and Gu, 2011).

3.5.1 Prior distributions

Prior distributions for the unknown parameters in equation 3.9 need to be specified. The unknown parameters are the vector of $\beta = (\beta_1, \beta_2, \dots, \beta_k)$ for all parameters, μ_α, and the variances σ_α^2 and σ_y^2.

We assume a multivariate normal distribution for $\beta \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where Σ is $\sigma_\beta^2 * I$ and I is the identity matrix. Hence, the elements of β are assumed uncorrelated apriori. Noninformative priors were chosen for the parameters, since no previous information was known about the parameters (Gelman and Hill, 2007). For the prior to be called noninformative, its range of uncertainty should be wider than the range of reasonable values of the parameters (Gelman and Hill, 2007). We typically choose $\sigma_\beta^2 = 100$ to be a good choice for a sufficiently large variance of the noninformative prior distribution. For the variances, the following priors have been chosen: $\sigma_y^2 \sim U(0, 100)$ for the within family variance and $\sigma_\alpha^2 \sim U(0, 1000)$ for the between family variance. Gelman and Hill (2007) argue that an inverse-gamma prior distribution for the variances will create problems close to zero. Therefore, we found the uniform prior distributions to be a better choice in our case.

The prior distribution for μ_α was chosen to $\mu_\alpha \sim N(0, 1000)$. We performed a prior sensitivity analysis to check that these choices of priors were non-informative.

3.5.2 Parameter estimation

For parameter estimation, random initial values are used in the MCMC and the MCMC runs until the draws converge to the posterior distribution. Three MCMC chains were used for the estimation. To see whether the MCMC chains mixed well and converged to the posterior distribution, we interpreted MCMC trajectories of the parameter draws graphically. Testing for convergence has been done by investigating $\hat{R}$, which is an indicator for whether convergence is achieved when the value was close to one on all parameters. For every parameter, $\hat{R}$ is the square root of the variance of the mixture of the three chains, divided by the mean within-chain variance (Gelman and Hill, 2007). If the draws are too correlated to each other then it was hard to get convergence. Another measure that was of interest was the number of efficient draws, defined n_{eff}. n_{eff} is a crude measure of the number of effective sample sizes for the given parameter and should be at least 100 for estimations (Gelman and Hill, 2007).

One of the issues for an MCMC in estimating the posterior distribution is the burn-in period. The burn-in period consists of the first set of MCMC draws that are discarded in favor to the latter draws that optimally converged to a stationary distribution of the posterior. The burn-in period was chosen to one thousand, which discarded a fifth of the iterations as the number of iterations has been chosen to 5000 (Gelman and Hill, 2007).

3.5.3 Model comparison

When choosing the best subset of covariates for the Bayesian multilevel model in BUGS, we choose the model with the lowest DIC value. The DIC is the deviance information criterion and is used instead of AIC and BIC for model selection. Deviance is defined by -2 times the log-likelihood and the lowest expected deviance has the highest posterior probability. However, the deviance does not take model complexity into account and this was why the DIC is a better model fit criterion. This criterion is only valid when the posterior distribution is approximately multivariate normal and is calculated by:

$$DIC = \bar{D} + p\_D. \qquad (3.10)$$

It uses two measurements of how well the model fitted the data (represented as $\bar{D}$) as well as a measurement of the models complexity (represented as p_D). $\bar{D}$ is the posterior expectation of the deviance whereas p_D is the effective number of parameters with the following equation:

$$p\_D = \bar{D} - D(\bar{\beta}), \qquad (3.11)$$

which is the expected deviance minus the deviance evaluated at the posterior expectations (Spiegelhalter, Best and Carlin, 1998).

3.5.4 Variable selection

We used the spike and slab algorithm for variable selection in the Bayesian multilevel model using the R package, spikeSlabGAM. This package implements a Bayesian variable selection to select covariates using blockwise Gibbs sampling for MCMC inference (Scheipl, 2011).

The spike component concentrates its mass at values close to zero, which allows shrinkage of small effects to zero, whereas the slab component has its mass over a wide range of plausible values for the regression coefficient. Variable selection is based on the posterior probability of assigning the corresponding regression coefficient to the slab component, called the posterior inclusion probability of the covariate. The posterior inclusion probability of a covariate equals the proportion of draws that the covariate has been included in the MCMC iterations (Malsiner-Walli and Wagner, 2011).

All of the covariates are represented as a linear combination of basis functions (B-splines). This package splits the continuous covariates into two parts: linear and smooth forms. This is done to resolve issues of overfitting by imposing smoothness by adding a penalty function (Marx and Eilers, 1996). The penalty is a P-spline which stands for a penalized B-spline. The B-spline represents the coefficients that are determined partially by the data to be fitted and partially by the penalty function. The formula defines the candidate set of model parameters that comprised the model of maximal complexity under consideration.

This process introduced a binary dormant variable,γ_j, that is associated with the coefficients of each model parameter that the contribution of a parameter to the predictor is forced to be zero or extremely small (this is the spike). The binary dormant variable, γ_j, is unchanged in another state (the slab) where the parameter has little shrinkage. The posterior distributions of these dormant variables are represented as the marginal posterior probabilities for inclusion of the model. The priors used for this model was a normal-mixture of inverse gammas (NMIG), which used a bimodal prior on the variance of the coefficients. This proceeded in a spike-and-slab type on the coefficients.

Variables are selected based upon the posterior probability inclusion as well as term importance which is defined by $\pi_l = \bar{\eta}_l^T * \bar{\eta}_{-1} / \bar{\eta}_{-1}^T \bar{\eta}_{-1}$, where $\bar{\eta}_l$ is the posterior expectation of the predictor associated with the lth term and $\bar{\eta}_{-1}$ is the predictor minus the intercept. The sum of π is equal to one which provides a rough percentage decomposition of the sum of squares of the linear predictor (Gu, 1992).

For this package, the default values for the MCMC have been manipulated. The values that have been changed are the number of chains, the length of the chains and thinning. Because we only used the parameter draws for the estimation of the posterior, we choose no thinning and as such saved all parameter draws. We have increased the length of the chain as well as the number of chains to be sure convergence was achieved.

4 Results

This chapter will present the results obtained in this thesis. First, the parameters of the multilevel model have been estimated (equation 3.2), and variables are selected using different selection methods from frequentist inference. Then parameters are estimated and variables are selected using different selection methods from Bayesian inference. Then a comparison is made between the two different inferences and the different variable selection methods.

4.1 Descriptive statistics

The dataset contained 1576 children in 1095 families with non-missing values on all variables. The average PIAT math score declined for higher birth orders (see figure 1). Family size was also of interest to explain cognitive performance and this link can be seen in figure 2.

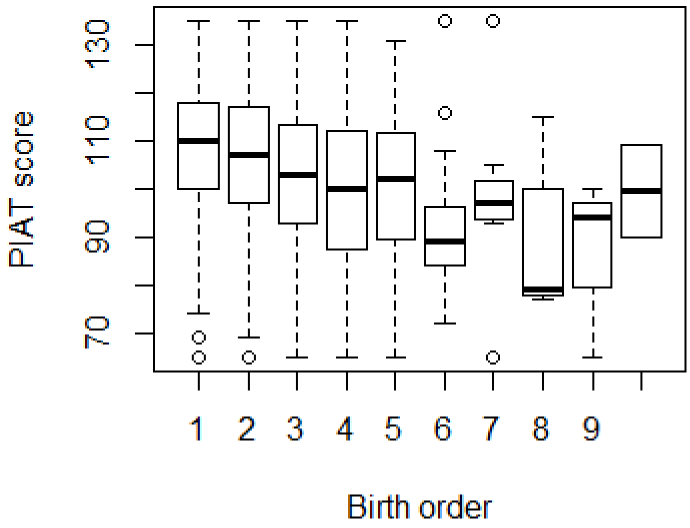


Figure 1: Boxplot for math score for each birth order

From figure 1 it can be seen that the average PIAT score decreased with higher birth orders. The trend seemed to be fairly linear; however some fluctuations exist after the fifth child. The boxplot contained a few outliers, where a child has scored a very low or high PIAT score. This can be explained by the lower number of children in these birth orders which can be seen in table 3.

Birth order	Frequency	Percent
1	411	26
2	568	36
3	359	23
4	139	9
5	63	4
6	21	1
7	7	0
8	5	0
9	3	0
10	2	0

Table 3: Frequency, percent for each birth order sample

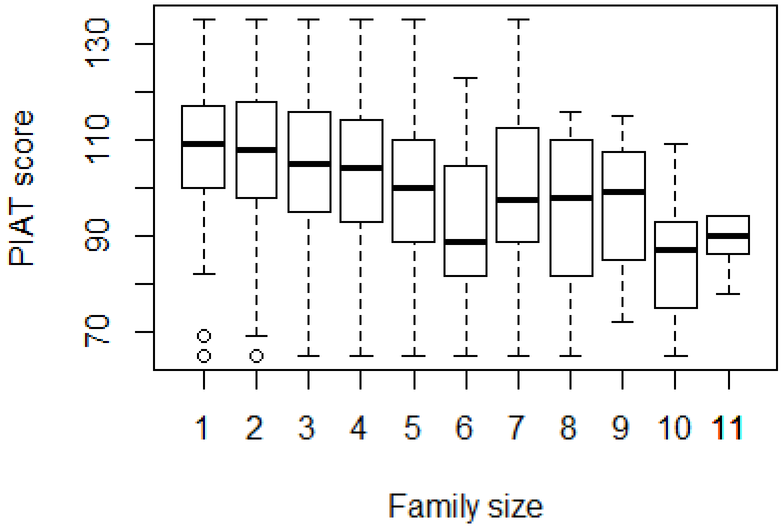


Figure 2: Boxplot for math score for different family sizes

Figure 2 showed a fairly decaying linear trend over the PIAT scores with an increase in the number of siblings. Even here, there seemed to be a few outliers depending on a high or low PIAT score. Fluctuations in the linear trend existed here as well; however the results were less reliable for large family size. The different family sizes can be seen in table 4.

Number of siblings	Frequency	Percent
1	129	8
2	584	37
3	456	29
4	237	15
5	99	6
6	28	2
7	20	1
8	11	1
9	4	0
10	5	0
11	2	0

Table 4: Frequency, percent for each family size

4.2 Frequentist inference

Results from the multilevel model and variable selection from the frequentists inference will be presented here.

4.2.1 Regression model

A regression model (equation 3.1) has been estimated with PIAT math score as a response variable with family size, father's presence, racemom1, racemom2, mother's IQ, age of mother at first birth, total income, poverty, birth order, child's age, total hours worked, gender, martial1, marital2, marital3, and marital4 as explanatory variables. Multicollinearity has been checked for both backward elimination and forward selection and there was no indication that the predictor variables were highly correlated with each other.

Backward elimination resulted in the following explanatory variables: father's presence, racemom1, mother's IQ, total income, gender, birth order, child's age and total hours worked in a week (see table 5).

Variable	Parameter estimate	Standard error	Pr > F
Intercept	107.888	2.071	<0.001
Father's presence	1.481	0.769	0.055
Racemom1	-3.456	0.851	<0.001
Mother's IQ	0.178	0.014	<0.001
Total income	0.00001	0.0001	0.009
Gender	-1.098	0.649	0.091
Birth order	-1.278	0.265	<0.001
Child's age	-0.680	0.139	<0.001
Total hours worked	-0.021	0.012	0.085
Variance ($\boldsymbol{\sigma_y^2}$)	11852	165.742	<0.001

Table 5: Results from backward elimination (multiple regression)

The results showed that having a father present, a mother with a higher IQ, and a high income family was beneficial for a child's PIAT score. The higher the birth order, the more the mother works, and the older the child was, the lower the child's score. We can also see that children with African/American mothers had lower scores, on average, compared to children with non-African American/non-Hispanic mothers. Diagnostic plots for the regression model showed that the residuals followed a normal distribution and that R^2 has a value of 0.267. This means that 26.7 percent of the variation in the response variable could be explained by the explanatory variables, which was relatively low. All of the diagnostic plots for the regression model have been shown in appendix A. However, since the data has two levels it was appropriate to check whether a multilevel model would be more efficient.

4.2.2 Multilevel model

A multilevel model (equation 3.2) without explanatory variables was first estimated as seen in table 6.

	Estimate	**Standard error**	**Pr > |t|**
Intercept	105.030	0.419	<0.001
Variance between families ($\widehat{\boldsymbol{\sigma}_{\alpha}^{2}}$)	96.458	9.582	<0.001
Variance within families ($\widehat{\boldsymbol{\sigma}_{y}^{2}}$)	127.110	7.728	<0.001

Table 6: Estimates of variance for the multilevel model with random intercept and without explanatory variables

Approximately 43% of the variance in cognitive performance was between families. The mean PIAT score for all families was approximately 105.03. The variance between families was 96.458 and the variance within a family was 127.11. To examine whether a multilevel model was more appropriate, the intraclass correlation (equation 3.3) was estimated $\frac{96.458}{96.458+127.11}$ at 0.43 suggesting that a multilevel model was more appropriate than a regular regression model.

In addition, a likelihood ratio test (equation 3.4) indicated that the variance in cognitive performance between families was significantly different from zero, ($D_{01} =$ 119.2, p < 0.05), see also table 7.

Model	-2 Res Log Likelihood
with random intercept	12904.9
without random intercept	13024.1

Table 7: -2 Reg log likelihood for model with and without intercept

Based upon the intraclass correlation as well as the log likelihood test, the more appropriate model for this data was a random intercept multilevel model. The random intercept allowed the intercept to vary, in this case, between the different families.

The full random intercept multilevel model was:

$Score_{ij} = \alpha_j + \beta_1$*Family size + β_2*Father's presence + β_3* Racemom (1) + β_4*Racemom (2) + β_5*Mother's IQ + β_6*Age of mother at first birth + β_7*Total income + β_8*Poverty + β_9*Never married + β_{10}*Separated + β_{11}*Divorced + β_{12}*Widowed + β_{13}*Gender + β_{14}*Birth order + β_{15}*Child's age + β_{16}* Total hours worked + ε_{ij}

Table 8 has presented the estimated parameters. As seen, several variables were not significant.

Variable	Parameter estimate	Standard Error	Pr > |t|
Intercept	110.960	4.461	<0.001
Family size	-0.539	0.446	0.227
Father's presence	0.694	1.185	0.558
Racemom1	-3.428	1.056	0.001
Racemom2	-0.816	1.062	0.443
Mother's IQ	0.169	0.017	<0.001
Age of mother at first birth	-0.026	0.100	0.794
Total income	0.000011	5.693E-6	0.060
Poverty	-1.031	1.238	0.406
Never married	-1.543	1.602	0.336
Separated	-2.288	1.938	0.238
Divorced	-0.585	1.459	0.689
Widowed	-0.355	3.128	0.909
Gender	-1.249	0.635	0.049
Birth order	-0.748	0.527	0.156

Child’s age	-0.694	0.157	<0.001
Total hours worked	-0.025	0.0138	0.068
Variance between families $(\widehat{\sigma_{\alpha}^{2}})$	41.182	7.385	<0.001
Variance within families $(\widehat{\sigma_{y}^{2}})$	125.05	7.462	<0.001

Table 8: Parameter estimate, standard error and p-values for multilevel model with all explanatory variables

4.2.3 Variable selection

A backward elimination suggested a model with the explanatory variables father’s presence, the mother not being married, the mother’s IQ, total income, birth order, child’s age and gender. Parameter estimates have been given in table 9.

Variable	**Parameter estimate**	**Standard Error**	**Pr >|t|**
Intercept	106.280	2.079	<0.001
Father’s presence	1.659	0.827	0.045
Never married	-2.214	1.259	0.079
Mother’s IQ	0.192	0.015	<0.001
Total income	0.000013	5.524E-6	0.018
Birth order	-1.277	0.282	<0.001
Child’s age	-0.714	0.138	<0.001
Gender	-1.233	0.635	0.053
Variance between families $(\widehat{\sigma_{\alpha}^{2}})$	42.380	7.382	<0.001
Variance within families $(\widehat{\sigma_{y}^{2}})$	125.130	7.444	<0.001

Table 9: Parameter estimates, standard errors and p-values for backward elimination

As seen, a child’s cognitive performance was positively affected by father’s presence, the mother’s IQ, and the family’s income. Children with higher birth orders, older children, girls and children where their mother never married had on average lower scores.

Forward selection resulted in a slightly different model, choosing father’s presence, mother not being married, mother’s IQ, total hours worked, birth order, child’s age, and gender as explanatory variables.

Variable	Parameter estimate	Standard Error	Pr >|t|
Intercept	107.540	2.109	<0.001
Father’s presence	1.990	0.808	0.014
Never married	-2.472	1.257	0.049
Mother’s IQ	0.202	0.013	<0.001
Total hours worked	-0.027	0.013	0.047
Birth order	-1.345	0.281	<0.001
Child’s age	-0.705	0.138	<0.001
Gender (child)	-1.226	0.636	0.054
Variance between families ($\widehat{\boldsymbol{\sigma}_{\alpha}^{2}}$)	42.765	7.371	<0.001
Variance within families ($\widehat{\boldsymbol{\sigma}_{y}^{2}}$)	124.970	7.425	<0.001

Table 10: Parameter estimates, standard errors and p-values for forward selection

The parameter estimates for the forward selection have been shown in table 10. The only variable that was excluded compared to the backward elimination was the family’s income. The variable that was now included in the model was the total amount of hours the mother works, which decreased the PIAT score of the child. The other parameter estimates as well as standard errors were similar to the backward elimination.

4.2.4 Lasso

Variables have been selected using the shrinking method lasso and these variables were family size, father’s presence, mother not being married, mother’s IQ, age of mother at first birth, total income, birth order and child’s age. Table 11 has displayed the parameter estimates.

Variable	Estimate
Intercept	101.095
Family size	-0.078
Father’s presence	0.689
Never married	-2.049
Mother’s IQ	0.172
Age of mom at first birth	0.053
Total income	0.001
Birth order	-0.543
Child’s age	-0.344
Variance between families ($\widehat{\boldsymbol{\sigma}_{\alpha}^{2}}$)	12.309
Variance within families ($\widehat{\boldsymbol{\sigma}_{y}^{2}}$)	146.309

Table 11: Parameter estimates for the lasso

Compared to previous results the lasso also included the following variables to some extent: the age of the mother at first birth and the number of biological children. The older the mother was at first birth, the higher PIAT score was expected for a child, unlike the expected decline in the PIAT score for a child in a family with a large family size. In general, the parameter estimates have all been lowered compared to both the forward selection and backward elimination, but the estimated signs of the parameters were the same. Diagnostic plots for the lasso have been presented in appendix D where normality could be seen.

4.2.5 Model criteria

Model criteria as well as the estimated variances within and between families for the different models in the frequentist inference part have been presented in table 12.

Model	**AIC value**	**BIC value**	**-2 Res LL**	**Variance between families**	**Variance within families**
Intercept only model	12908.9	12918.9	12904.9	96.459	127.11
Full model	12521.5	12531.5	12517.5	41.182	125.05
Forward selection model	12533.0	12543.0	12529.0	42.765	124.97
Backward elimination	12547.0	12557.0	12543.0	42.380	125.13
Lasso	12562.2	12626.6	12538.2	12.309	146.309

Table 12: Difference between the models based on different criteria

The AIC- and BIC- values did not differ much between the full model and the models chosen by the forward and backward selection procedures. According to the AIC criterion, the best model was estimated by the forward selection, and according to the BIC criterion the multilevel model with all of the explanatory variables. According to the -2 Res LL criterion, the best model was estimated from the multilevel model with all of the explanatory variables. The unexplained variance between and within families did not differ very much between the full model and the models chosen by forward and backward selection. However, the model chosen from the lasso method had more unexplained within family variance and less unexplained between family variance compared to the models chosen from forward and backward selection. This could be explained by the variables chosen in the model were variables that were considered to be within families and not between families (such as family size).

4.3 Bayesian inference

Results from the Bayesian inference will be presented for the full multilevel model with all variables included and the spike and slab variable selection method in this subsection.

4.3.1 Full multilevel model

A multilevel model with Bayesian inference (equation 3.2) without explanatory variables was first estimated as seen in table 13.

	Estimate	Standard error	2.5 percentile	97.5 percentile
Intercept	105.105	0.329	0.010	105.7
Variance between families ($\boldsymbol{\sigma_{\alpha}^2}$)	63.857	45.831	0.001	113.1
Variance within families ($\boldsymbol{\sigma_{y}^2}$)	160.067	46.640	0.002	237.1

Table 13: Estimates of variance for the Bayesian inference multilevel model with random intercept and without explanatory variables

Table 13 displayed the results from the Bayesian estimation of the full multilevel model. The mean of the posterior draws of the intraclass correlation ρ was 0.428, which indicates that the multilevel model was a better choice than the multiple regression model for estimation of the data as only 43 percent of the variance in cognitive performance was between families. The intraclass correlation was tested on the Bayesian multilevel model with no explanatory variables. The marginal posterior distributions for all of the parameters were symmetric and bell-shaped. All of the $\hat{R}$'s were equal to one, which indicated that the estimated distributions of the parameters converged to the posterior distribution.

Parameter	Mean	Standard Deviation	2.5 percentile	97.5 percentile	n_{eff}
Intercept	103.362	0.769	101.9	104.9	120
Family size	-0.214	0.093	-0.4	0.0	3400
Father's presence	0.024	0.099	-0.2	0.1	12000
Mother's IQ	0.197	0.014	0.2	0.2	470
Age of mother first birth	-0.041	0.042	-0.1	0.0	4701
Total income	0.00001	0.00005	0.0	0.0	4000
Poverty	-0.036	0.100	-0.2	0.2	4400
Gender	-0.041	0.100	-0.2	0.2	2400
Birth order	-0.192	0.098	-0.4	0.0	210

Child’s age	-0.402	0.092	-0.5	-0.3	8100
Total hours worked	-0.042	0.071	-0.1	0.0	12000
Never married	-0.015	0.013	-0.2	0.2	12000
Separated	-0.009	0.099	-0.2	0.2	12000
Divorced	-0.214	0.099	-0.2	0.2	12000
Widowed	-0.001	0.099	-0.2	0.2	12000
Racemom (1)	-0.057	0.098	-0.3	0.1	12000
Racemom (2)	-0.001	0.098	-0.2	0.2	12000
Variance $\boldsymbol{\sigma_y^2}$	127.0	7.602	112.7	142.6	310
Variance $\boldsymbol{\sigma_\alpha^2}$	43.3	7.419	38.3	58.0	160
Deviance	12122.22	75.472	11980.0	12270.0	220
Intraclass correlation ρ	0.428	0.034	0.346	0.392	

Table 14: Bayesian multilevel model, all parameters

The variables that were of importance in the model were the variables that did not include zero in the credibility interval seen in table 14. The significant parameters that influenced a child’s cognitive performance positively were mother’s IQ and income. The variables that negatively influenced a child’s cognitive performance were family size and the child’s age. The last six variables in table 14 were dummy variables, so the decline was compared to being married as well as being non-African American/non-Hispanic. The Bayesian multilevel model parameters have converged (see MCMC trajectories in appendix B). n_{eff} represented the number of efficient draws from the total of 12000 draws and all parameters fulfill the minimum limit of 100 number of efficient draws.

For the multilevel with Bayesian inference, many priors have been tested before the final priors were chosen. Since there were so many observations, the priors did not seem to impact the model very much. Prior sensitivity was tested with 10, 100 and 1000 and the posterior means changed slightly. The prior sensitivity analysis showed that there were very small differences between the posterior means when choosing other priors distributions which means that the posterior distributions were not sensitive to the prior distributions. This could be caused by the large number of observations (the likelihood) influencing the posterior distribution more than the prior distributions.

4.3.2 Variable selection

From the slab and spike method, the following information was given for each parameter for the Bayesian multilevel model. The parameters that were of significance were chosen based upon term importance as well as the marginal posterior inclusion probabilities. Three chains of 500 samples were run with a burn-in period of 100 samples.

Parameter	Posterior inclusion probability	Parameter importance
Intercept	1	0.290
Racemom (1)	1	0.102
Racemom (2)	0.783	0.013
Family size (lin)	0.894	0.034
Family size (sm)	0.649	0
Father's presence	0.530	0.008
Poverty	0.526	0.004
Mother's IQ (lin)	1	0.354
Mother's IQ (sm)	0.790	0.003
Age of mother first birth (lin)	0.523	0.003
Age of mother first birth (sm)	0.557	0.001
Total income (lin)	0.972	0.054
Total income (sm)	0.966	0.047
Never married	0.494	0.003
Separated	0.567	0.008
Divorced	0.489	0
Widowed	0.675	0
Gender	0.699	0.001
Birth order (lin)	0.614	0.014
Birth order (sm)	0.787	0.001
Child's age (lin)	1	0.035
Child's age (sm)	0.872	0.005
Total hours worked (lin)	0.805	0.007
Total hours worked (sm)	0.867	0.014

Table 15: Slab and spike variable selection

From table 15, most of the variables have gotten a high posterior inclusion probability except for the variable Divorced which posterior inclusion probability was below 0.5. Variables with posterior inclusion probabilities above 0.9 have been included into the model. The variables that have got a high value for parameter importance was the mother not being married, mother's IQ, total income, and the child's age. A high posterior inclusion probability as well as term importance concluded which variables that should be included in the model. It was apparent that the variables racemom (1), mother’s IQ, and the age of the child should be included in the model since the inclusion probability was equal to one. The 95 percent credibility intervals for the parameters of the model have been shown graphically in appendix C. The variables that have been included in the model have slightly smaller intervals compared to non-significant variables.

4.4 Comparison of frequentist vs. Bayesian variable selection

After estimating five different variable selection methods from two different inferences, differences have occurred. In table 16 variables that have been chosen in each method are shown. These variables were represented with a star. For the Bayesian multilevel model, the chosen significant variables in the table were for the parameters whose credibility interval do not include zero.

Comparison of the different variable selection methods

Variable included in the final model	Frequentist: Forward selection	Frequentist: Backward elimination	Frequentist: Lasso	Bayesian: Spike and slab	Bayesian: Multilevel Model
Family size			*	*	*
Father’s presence	*	*	*		
Mother’s IQ	*	*	*	*	*
Age of mother first birth			*		*
Total income		*	*		*
Poverty					
Gender	*	*			
Birth order	*	*	*		*
Child’s age	*	*	*	*	
Total hours worked	*			*	*
Never married	*	*	*		
Separated					
Divorced					
Widowed					
Racemom (1)				*	
Racemom (2)					

Table 16: Comparison of the five different models of variable selection

The variable that was important for a child’s cognitive performance was the mother’s IQ which was present in all of the five models. The variables that were important for the frequentist method were father’s presence, age of mother at first birth, total income, gender, birth order, total hours worked, child’s age and the mother not being married. The lasso (which was both frequentist and Bayesian) suggested family size, the father’s presence, age of the mother at first birth, total income, birth order, child’s age and the mother not being married as the variables that influenced a child’s cognitive performance. From the Bayesian inferences, the variables that influenced a child’s cognitive performance were family size, the father’s presence, age of the mother at first birth, total income, child’s age, total hours worked, birth order and the mother being African American.

5 Discussion

The purpose of this thesis was to investigate the relationship between children's cognitive performance based on their family size and their birth order. It was also of interest to use two different types of statistical inferences; the frequentist and the Bayesian to arrive to a statistical model explaining variation in children's cognitive performances.

To achieve the objective, a total of four different models for variable selection were estimated from both the frequentist and the Bayesian inferences. Explanatory variables selected from forward selection were the father's presence, the mother not being married, mother's IQ, total hours worked, birth order, age of the child, and gender of the child. From the backward elimination the explanatory variables that were selected were the father's presence, the mother not being married, mother's IQ, income, birth order, age of the child and the gender of the child. Lasso selected the father's presence, family size, mother not being married, mother's IQ, age of the mother at first birth, income, birth order and age of the child as the explanatory variables in the model. Spike and slab chose family size, mother's IQ, age of the child, total number of hours worked and if the mother was African American as the explanatory variables for the model.

Getting different results in my forward and backward selection could be dependent on the order of which variables were added or removed. This was one of the disadvantages when using the backward and forward methods for variable selection, that the optimal model could be missed depending on the order of the variables being added and/or removed.

After estimating many different models from both frequentist and Bayesian inferences, there were few differences between the models. Reasons for why the frequentist and the Bayesian inferences did not differ that much could be explained that the priors did not seem to impact the posteriors very much. This could be because of the large number of observations that were being tested and noninformative prior distributions were used. In future research using more informative priors based upon this study or other studies could be used.

It was important to investigate variable selection in multilevel models since there were few studies on the subject. Most studies on cognitive performance (Wichman, Rodgers and MacCallum, 2005; Devereux, Black and Salvanes, 2005; Bjerkedal and Kristensen, 2007) used few variables in a multilevel model or many variables but only used multiple regression models. Most studies have gotten different conclusions as well what influenced cognitive performance that support different theories. From this thesis, using different methods of variable selection on multilevel models has been used and the model that I would prefer to use was the lasso for variable selection. The reason for this was that it can be used both from a frequentist as well as from Bayesian inferences as well as the optimal model would not be missed (which can happen in forward selection/backward elimination). Spike and slab could also be a good variable selection model, but lasso estimated a model that seems to be more similar with earlier studies conclusions.

In summary, variables in all of the models had the same influence on the PIAT score either negatively or positively. The parameters that increased the PIAT score were the father’s presence, mother’s IQ and the family’s income. All other variables contributed to the decline in the PIAT score.

The resource dilution theory (Downey, 2001) states that parental resources decline with larger family sizes. Four of the variable selection methods chose birth order to be significant in estimating a child’s cognitive performance and the higher the birth order the lower the PIAT score the child would receive. Three of the methods showed that the larger family sizes decreased the PIAT score as well.

The confluence model (Zajonc and Sulloway, 2007) explains that a child’s cognitive performance is based upon a mathematical model influenced by the intellectual levels in the family. The variable, mother’s IQ, has been shown to play a role in the child’s cognitive performance in all of the models. If the mother never married it seemed to decrease a child’s cognitive performance and if the father was present the PIAT score was higher.

The admixture hypothesis (Rodgers, 2001) assumes that other factors influence the cognitive performance. This could be seen in this thesis, where other variables were significant in the models such as the mother's IQ. However, birth order and/or family size have been significant in these models as well, so this theory could not be completely supported in this thesis. This could be explained that other control variables that were important for cognitive performance have been missed. There was no variable available for the father's IQ for instance. The mother's highest education level could have been of interest as well as location since education is different depending on the state.

This thesis has gotten similar results as Wichman, Rodgers, MacCallum (2005) that there was a decline in children's cognitive performance for higher birth orders and that with the inclusion of between family variance variables, the birth order effect became no longer significant in the spike and slab model.

Including more variables in the study was done by Devereus, Black and Salvanes (2005) where a strong birth order effect was found to influence a child's cognitive performance. However, in this study family size did not contribute to the decline in a child's cognitive performance. This thesis has found both variables (birth order and family size) to influence a child's cognitive performance negatively. Including more variables decreased this influence and in some models, the variables became in-significant.

Bjerkedal and Kristensen (2007) found a decline in cognitive performance with higher birth orders, but could not state if it was between or within families. This was because of the method used was an OLS regression which did not capture the variance in the two levels. It is important in the future for applied researchers that multilevel models are used when two levels are known so variances between and within groups can be studied.

In conclusion, this thesis supports that there was an existent decline in cognitive performance when it comes to larger family sizes and with higher birth orders.

However, other variables seem to influence a child’s cognitive performance as well, specifically the mother’s IQ.

Throughout the process of this thesis some limitations with the data as well as methods occurred. Since the dataset was very large, there was a large amount of missing values in the data. However, the focus of this study was not to analyze the missing values since this would have taken too much time. So for future studies, it might be of interest to manage the missing values and incorporate them into the analysis instead of just removing them as it was done for this study. Many families only included one sibling in the data so it was difficult to compare within a family whether the birth order accounts for a higher cognitive performance. Siblings were excluded since they did not have a PIAT math score for that year. Higher birth orders and large family sizes (over five) did not have many children represented and this makes the estimations slightly inaccurate. So because of this, the results might be misleading.

From the beginning, the question whether alcohol and drug abuse could influence a child’s cognitive performance was something that was of interest to investigate. However, this could not be investigated due to the fact there was only fourteen mothers that had answered the question. This would be too small of a sample to get concrete conclusions and this question was then discarded. Future research might look into this.

The output for the lasso which was completed using the R-package did not include any more information except for the estimated parameters. This can be misleading with the variables selected since there were no standard error estimates to control the precision of the estimations are. However, significance could be seen by looking at the convergence as well as the values of AIC, BIC and -2res LL. One can also control by looking at the diagnostics plots in appendix D where the residuals look good, but it would have been nice to be able to check this more.

The response variable in all models, PIAT score, was just one indicator for cognitive performance and this variable cannot capture the whole dimension of cognitive performance so this needs to be kept in mind.

In the future, it could be interesting to investigate variable selection with this dataset but with a three-level multilevel model. This would help when comparing between the family size relationships within a family, as this was hard to do with few within family observations. Interaction variables were thought to be tested and eventually used in this thesis but because of the time issue, this could not be investigated which could be interesting for future research. Birth gaps could also be investigated in future studies.

6 Conclusion

Based on the two different methods of variable selection from a frequentist inference, these variables influenced the child's cognitive performance negatively: mother not being married, total number of hours the mother worked, birth order, age of the child, being a girl and family size. The variables that influenced a child's cognitive performance positively was when the father was present, the mother's IQ was high, if the mother had the child later in life, and a higher income. The difference between the forward selection and the backward elimination was forward selection included total number of hours the mother worked in the model and excluded the total income.

The lasso, which can be seen from a frequentist as well as Bayesian inference, stated that the father's presence, mother's IQ, age of the mother at first birth and total income influenced a child's cognitive performance positively. Family size, mother never being married, birth order and the child's age influenced a child's cognitive performance negatively.

From a Bayesian inference, there was also a decline in the PIAT score depending on family size as well as birth order. The variables that influenced a child's cognitive performance positively in the multilevel model were when the mother's IQ was high and the family had a higher income. The variables that influenced a child's cognitive performance negatively were when the mother never married, if age of the mother at first birth, the age of the child, family size and if the mother worked many hours. The spike and slab model chose family size, total hours worked and if the mother was of African-American descent as negatively influential parameters. Mother's IQ, and child's age were positively influential parameters in the model.

A multilevel model should be used in analyzing this type of data (children nested within families). Deciding on which type of inference (frequentist vs. Bayesian) does not seem to impact the final results. When determining which type of variable selection method to use, applied researchers need to be careful since the results differ depending on which method is chosen. The lasso can be a good method to choose since it can be used as both from a frequentist as well as a Bayesian inference.

References

Baker, P. C. (1993). NLSY child handbook: a guide to the 1986-1990 National Longitudinal Survey of Youth child data.

Bjerkedal, T., Kristensen, P., Skjeret, G. A., & Brevik, J. I. (2007). Intelligence test scores and birth order among young Norwegian men (conscripts) analyzed within and between families. *Intelligence*, *35*(5), 503-514.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2005). The more the merrier? The effect of family size and birth order on children's education. *The Quarterly Journal of Economics, 120*(2), 669-700.

Blake, J. (1989). *Family size and achievement* (Vol. 3). Univ of California Press.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Casella, G., & Berger, R. L. (1990). *Statistical inference* (Vol. 70). Belmont, CA: Duxbury Press

Cascio, E. U., & Lewis, E. G. (2005). *Schooling and the AFQT: Evidence from school entry laws* (No. w11113). National Bureau of Economic Research.

Chand, S. (2012, January). On tuning parameter selection of lasso-type methods-A Monte Carlo study. In *Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on* (pp. 120-129). IEEE.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., ... & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69-102.

DeNavas-Walt, C. (2010). *Income, Poverty, and Health Insurance Coverage in the United States (2004)*. DIANE Publishing.

Downey, D. B. (2001). Number of siblings and intellectual development: The resource dilution explanation. *American Psychologist, 56*(6-7), 497

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science,* 89-102.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldstein, H. (2011). Multilevel statistical models (Vol. 922). John Wiley & Sons.

Guo, Y., & Gu, S. (2011, July). Multi-label classification using conditional dependency networks. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two* (pp. 1300-1305). AAAI Press.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2, No. 1). New York: Springer.

Jaynes, E. T., & Kempthorne, O. (1976). Confidence intervals vs Bayesian intervals. In Foundations of probability theory, statistical inference, and statistical theories of science (pp. 175-257). Springer Netherlands.

Jeffreys H (1961). *Theory of Probability*. Third edition. Oxford University Press, Oxford, England.

Karlsson, S. (2012). *Forecasting with Bayesian Vector Autoregressions* (No. 2012: 12).

Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, *99*(465), 279-290.

Kutner, M., Nachtsheim, C., & Neter, J. W. li. 2005. Applied linear Statistical Models.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369-411.

Laplace P (1812). *Theorie Analytique des Probabilites*. Courcier, Paris. Reprinted as "Oeuvres Completes de Laplace", 7, 1878-1912. Paris: Gauthier-Villars

Littell, R. C. (2006). *SAS for mixed models*. SAS institute.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325-337.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: a practical introduction to Bayesian analysis*. CRC Press.

Malsiner-Walli, G., & Wagner, H. (2011). Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, *40*(4), 241-264.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS* (Vol. 698). John Wiley & Sons.

O'Neill, M. (2010) *Anova and Reml, A guide to linear mixedmodels in an experimental design context*, statistical advisory & training service pty ltd.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681-686.

Rodgers, J. L. (2001). What causes birth order–intelligence patterns? The admixture hypothesis, revived. *American Psychologist*, *56*(6-7), 505.

Scheipl, F. (2011). spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *arXiv preprint arXiv:1105.5253*.

Schelldorfer, J., Bühlmann, P., DE, G., & VAN, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization. *Scandinavian Journal of Statistics*, *38*(2), 197-214.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, *23*(4), 323-355.

Snijders Tom, A. B., & Bosker Roel, J. (2000). Multilevel analysis. *An introduction to Basic and Advanced Multilevel Modeling*.

Spiegelhalter, D. J., Best, N., Carlin, B. P., & Van der Linde, A. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Research Report, 98-009.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181-207). Springer New York.

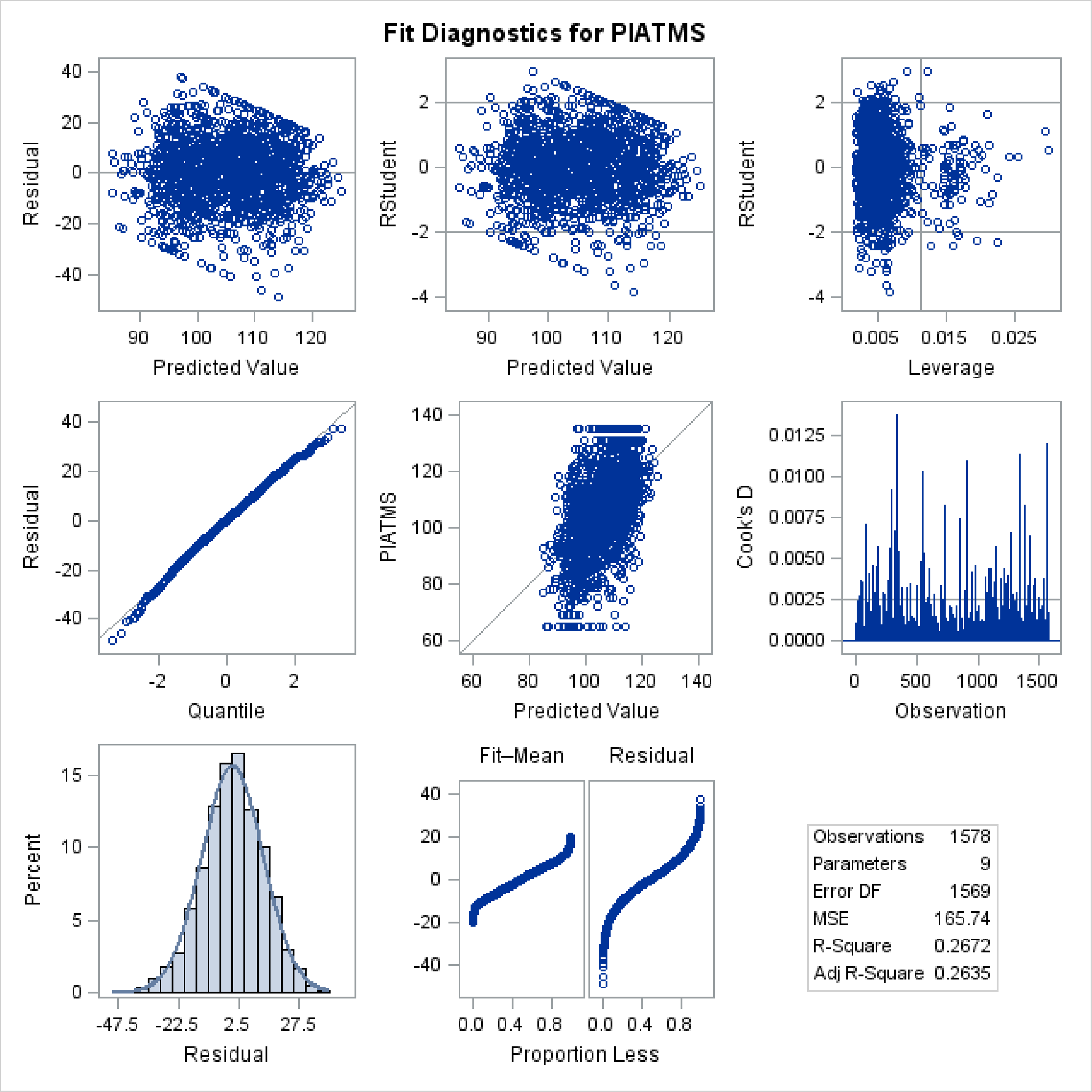
Wichman, A. L., Rodgers, J. L., & MacCallum, R. C. (2006). A multilevel approach to the relationship between birth order and intelligence. *Personality and social psychology bulletin*, *32*(1), 117-127.

Zagorsky, J. L., & White, L. (1999). NLSY79 user's guide: A guide to the 1979–1998 National Longitudinal Survey of Youth data. *Washington, DC: US Department of Labor*.

Zajonc, R. B., & Sulloway, F. J. (2007). The confluence model: Birth order as a within-family or between-family dynamic?. *Personality and Social Psychology Bulletin, 33*(9), 1187-1194.

Appendix A

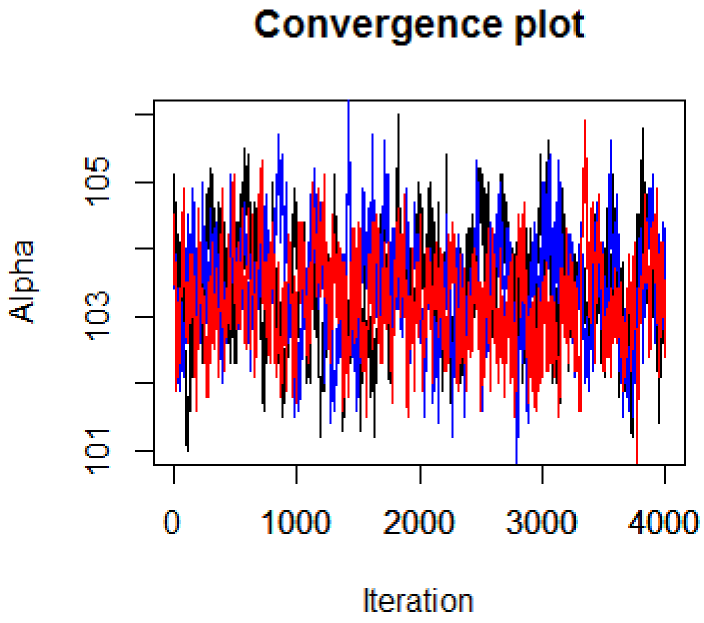
Fit Diagnostics for multiple regression model

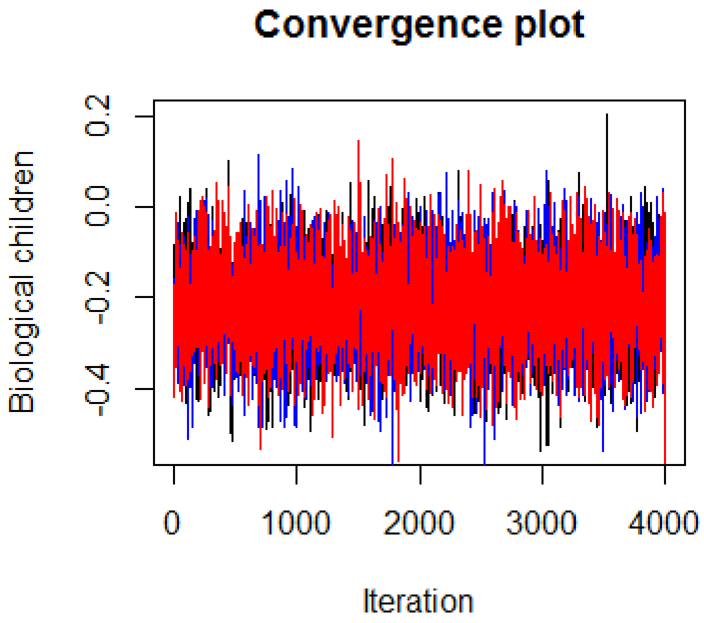


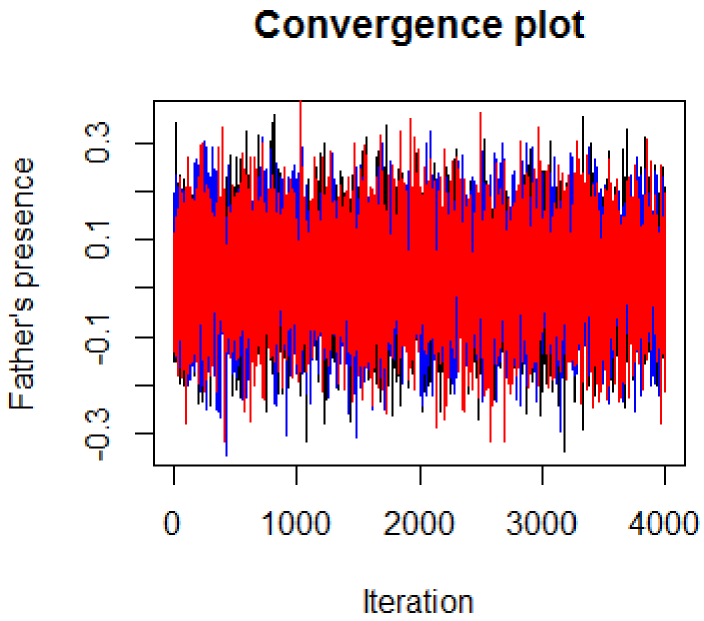
Observations	1578
Parameters	9
Error DF	1569
MSE	165.74
R-Square	0.2672
Adj R-Square	0.2635

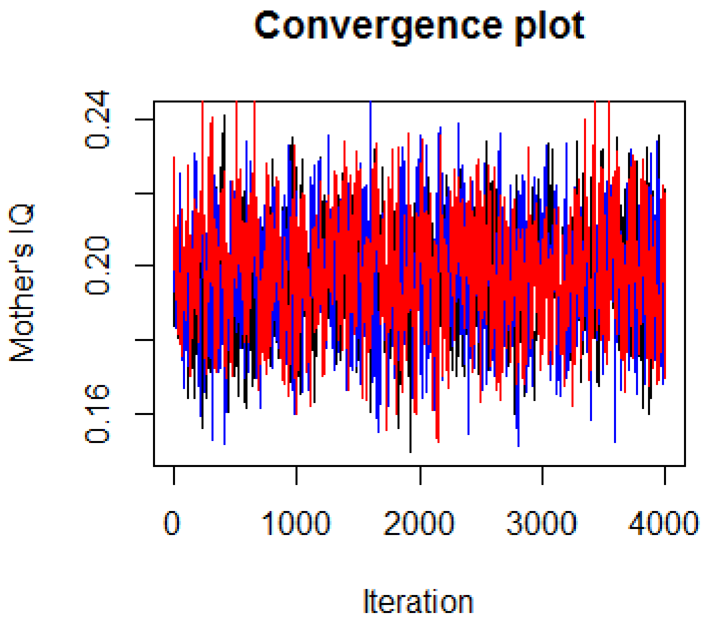
Appendix B

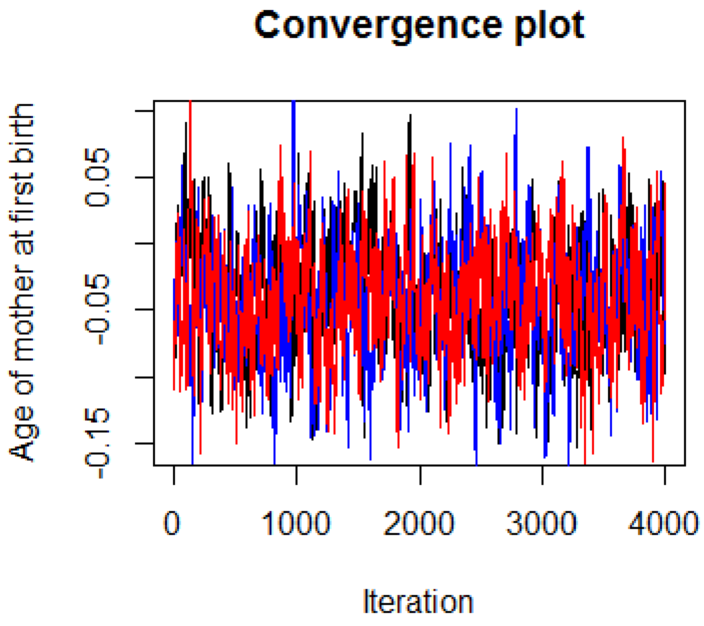
Convergence plots (trace plots) for all of the parameters in the Bayesian multilevel model

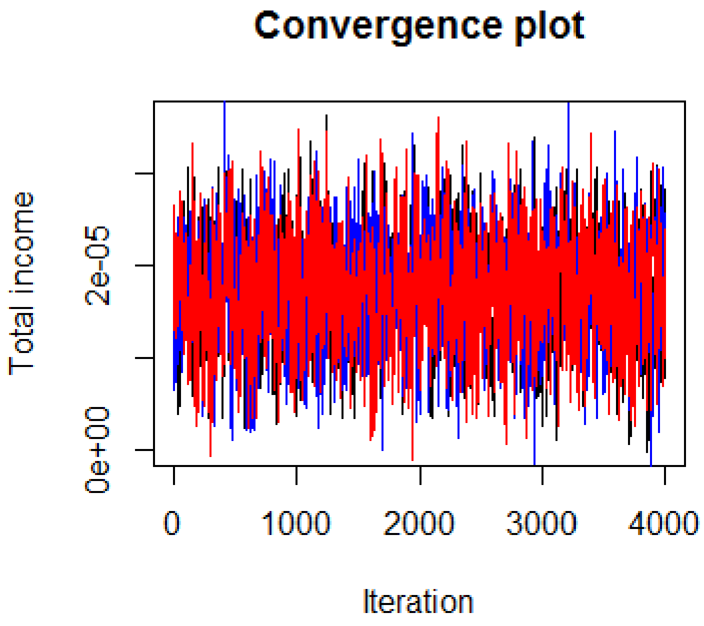


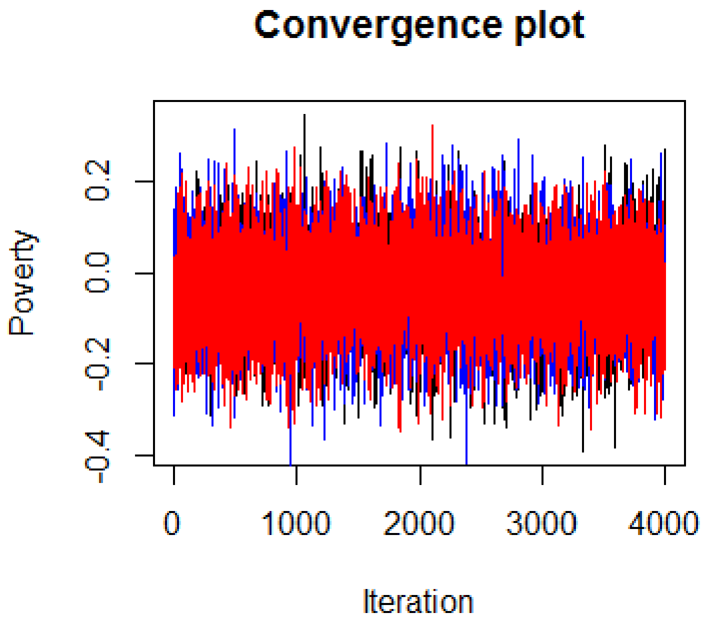


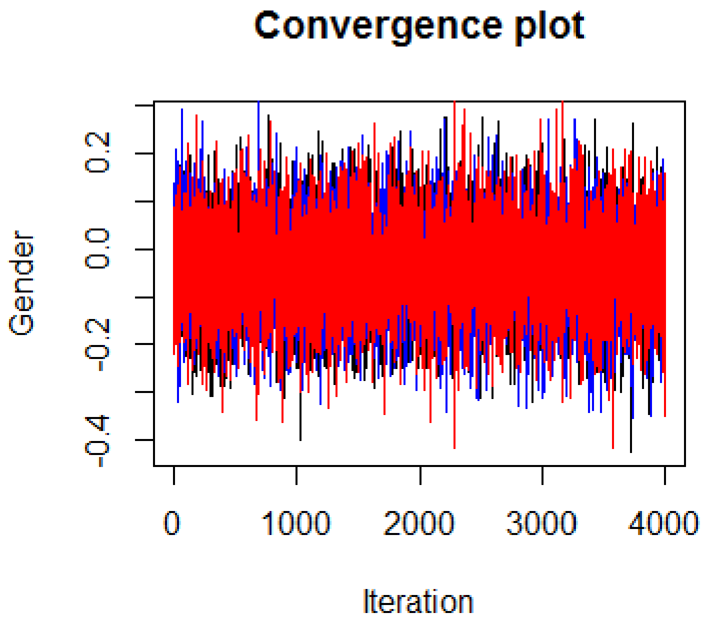

Convergence plot
Father's presence
-0.3
-0.1
0.1
0.3
0
1000
2000
3000
4000
Iteration

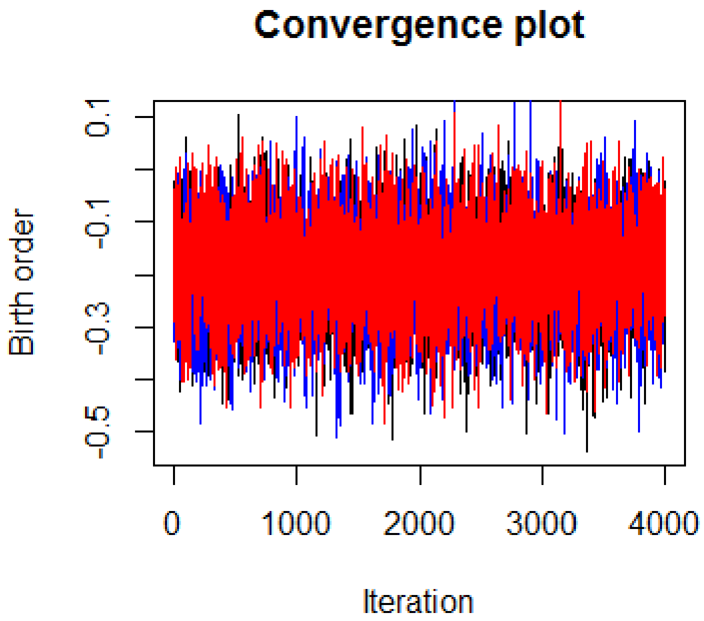

Convergence plot
Mother's IQ
0.16
0.20
0.24
0
1000
2000
3000
4000
Iteration

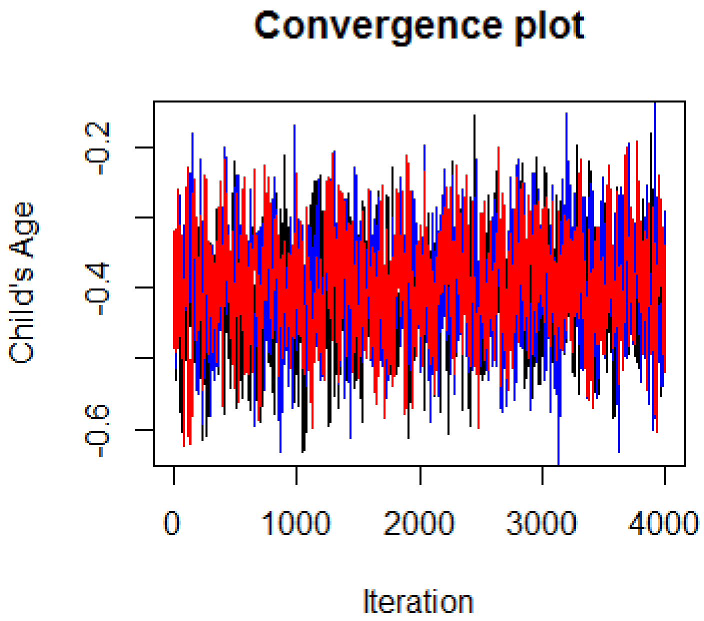

Convergence plot
Age of mother at first birth
0.05
-0.05
-0.15
0
1000
2000
3000
4000
Iteration

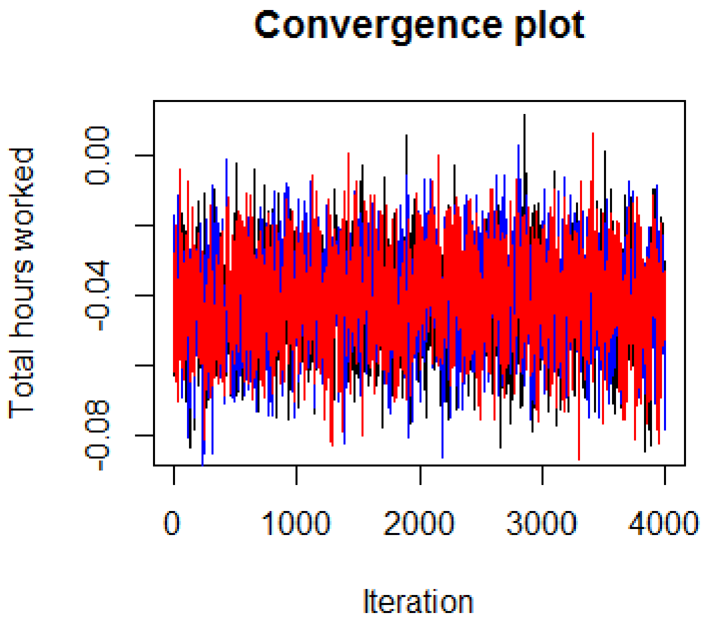

Convergence plot
Total income
2e-05
0e+00
0
1000
2000
3000
4000
Iteration

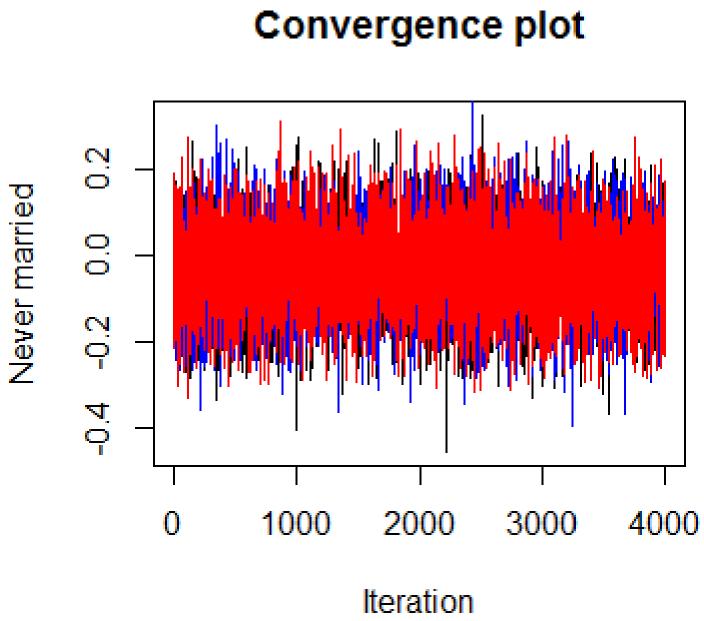

Convergence plot
Poverty
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration

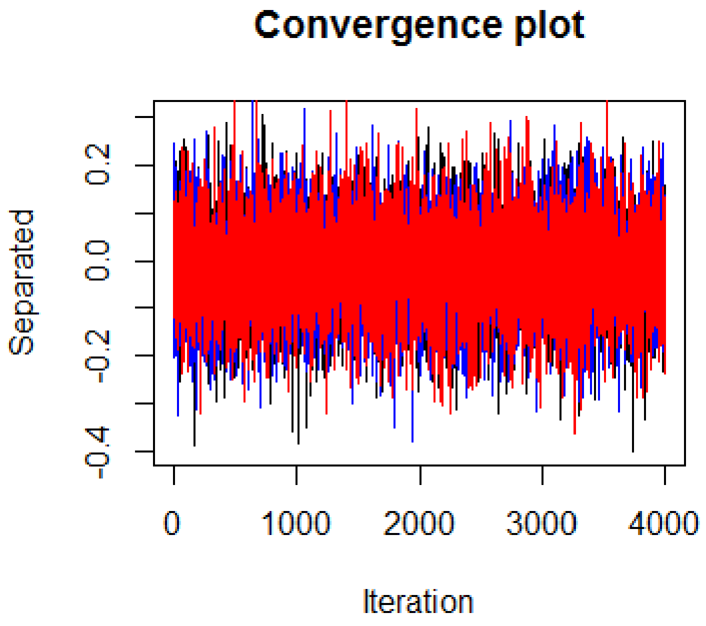

Convergence plot
Gender
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration

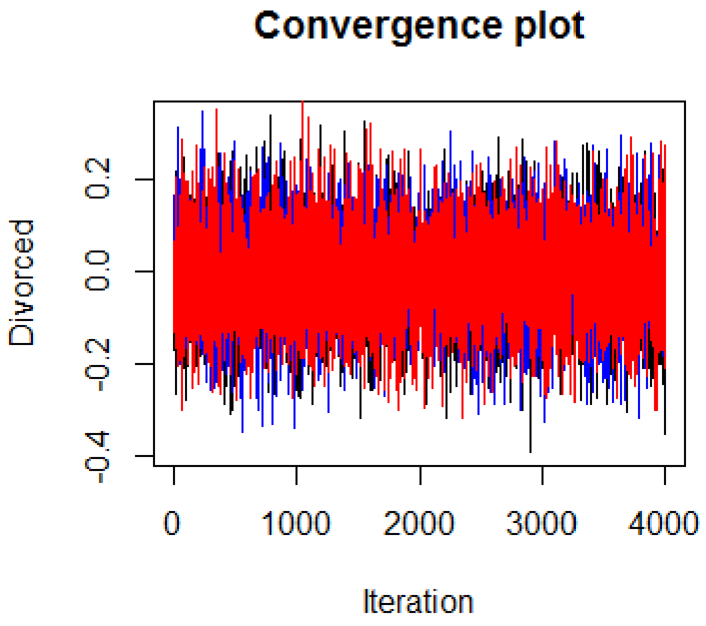

Convergence plot
Birth order
0.1
-0.1
-0.3
-0.5
0
1000
2000
3000
4000
Iteration

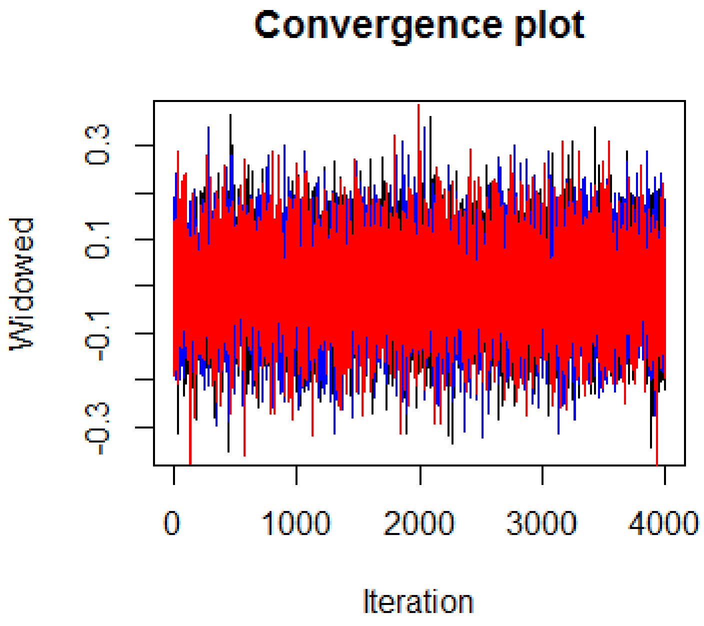

Convergence plot
Child's Age
-0.2
-0.4
-0.6
0
1000
2000
3000
4000
Iteration

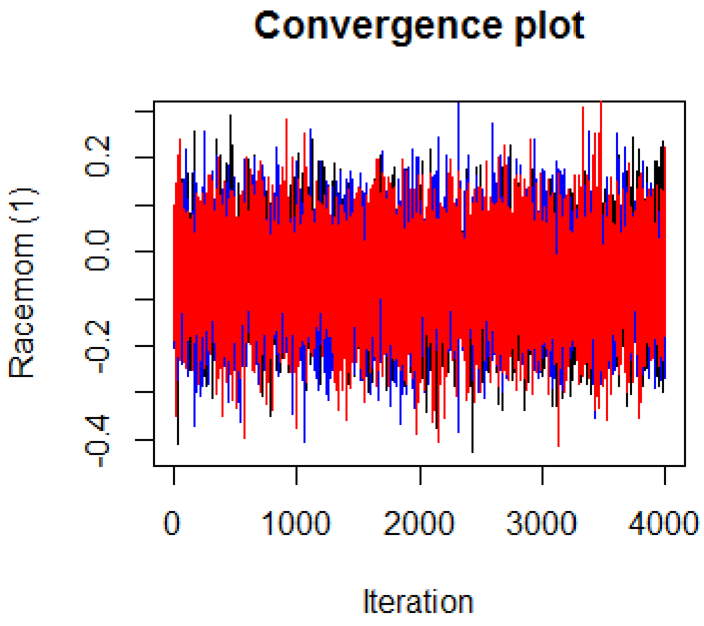

Convergence plot
Total hours worked
0.00
-0.04
-0.08
0
1000
2000
3000
4000
Iteration

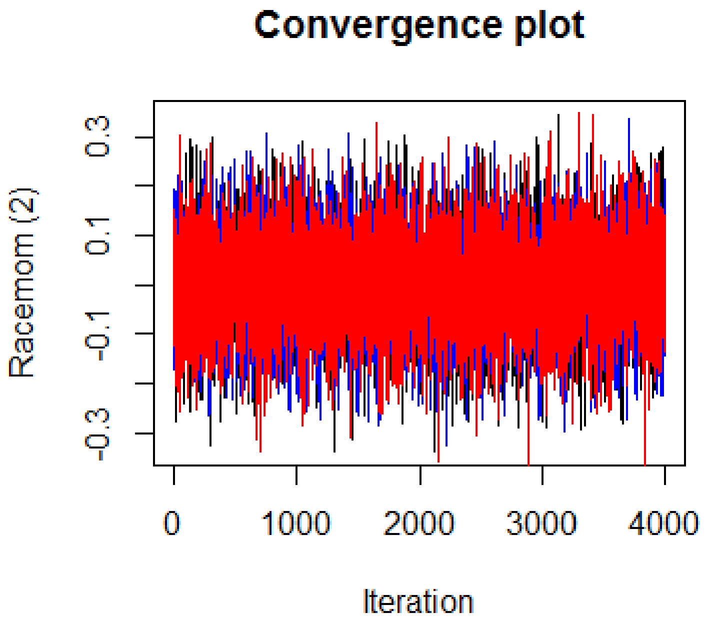

Convergence plot
Never married
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration

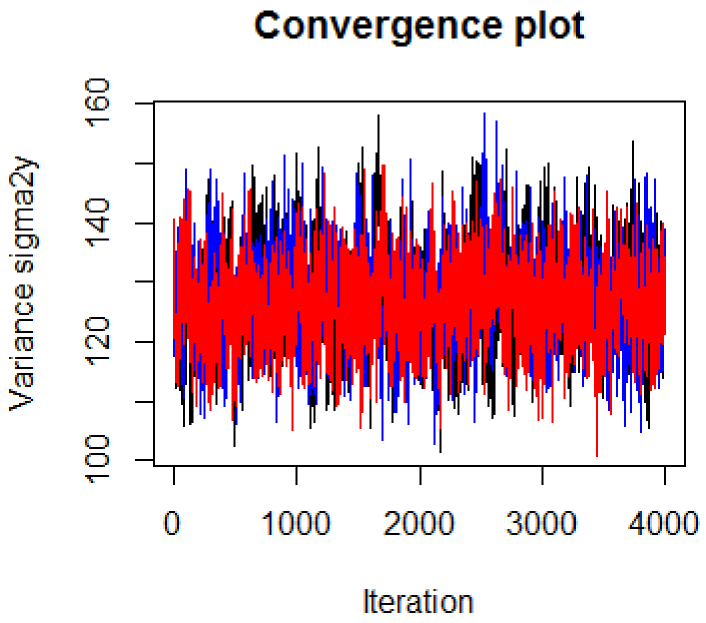

Convergence plot
Separated
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration

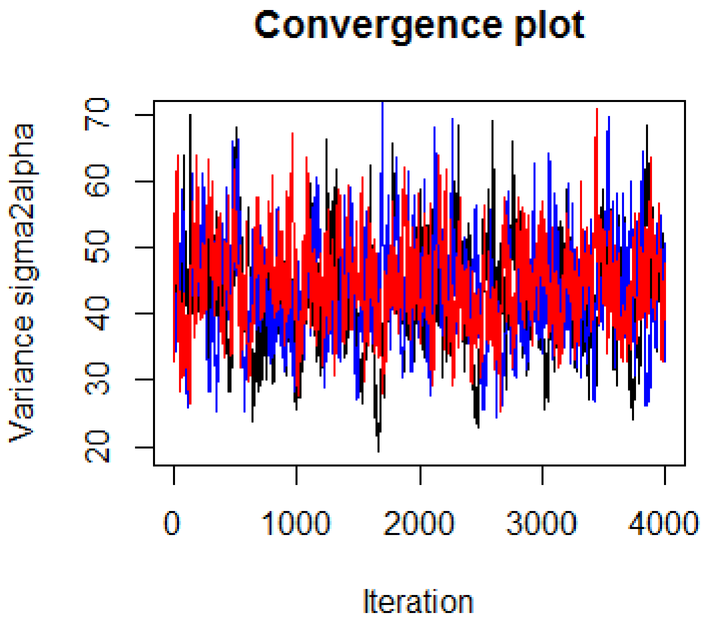

Convergence plot
Divorced
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration


Convergence plot
Widowed
0.3
0.1
-0.1
-0.3
0
1000
2000
3000
4000
Iteration


Convergence plot
Racemom (1)
0.2
0.0
-0.2
-0.4
0
1000
2000
3000
4000
Iteration

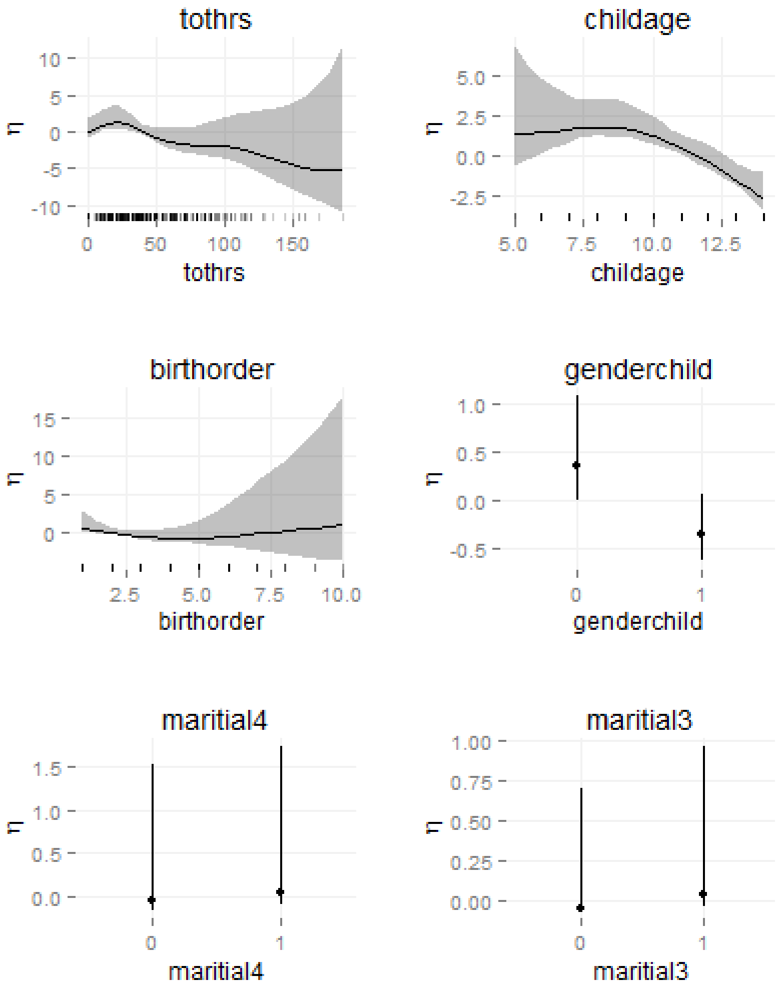

Convergence plot
Racemom (2)
0.3
0.1
-0.1
-0.3
0
1000
2000
3000
4000
Iteration

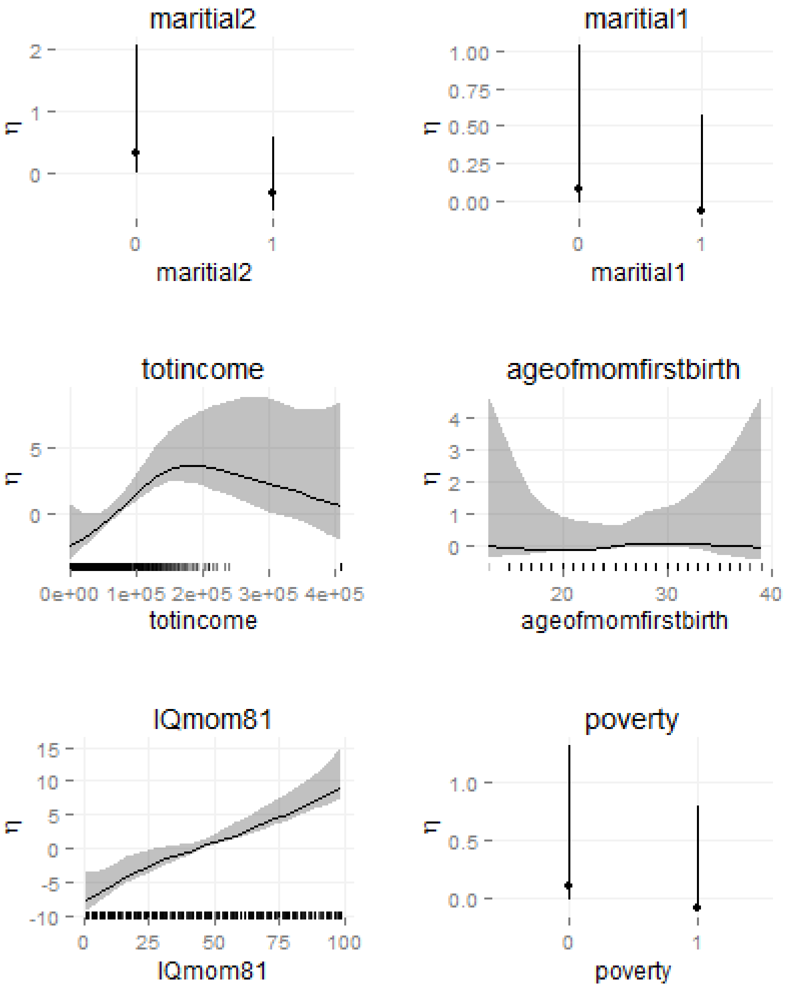

Convergence plot
Variance sigma2y
160
140
120
100
0
1000
2000
3000
4000
Iteration

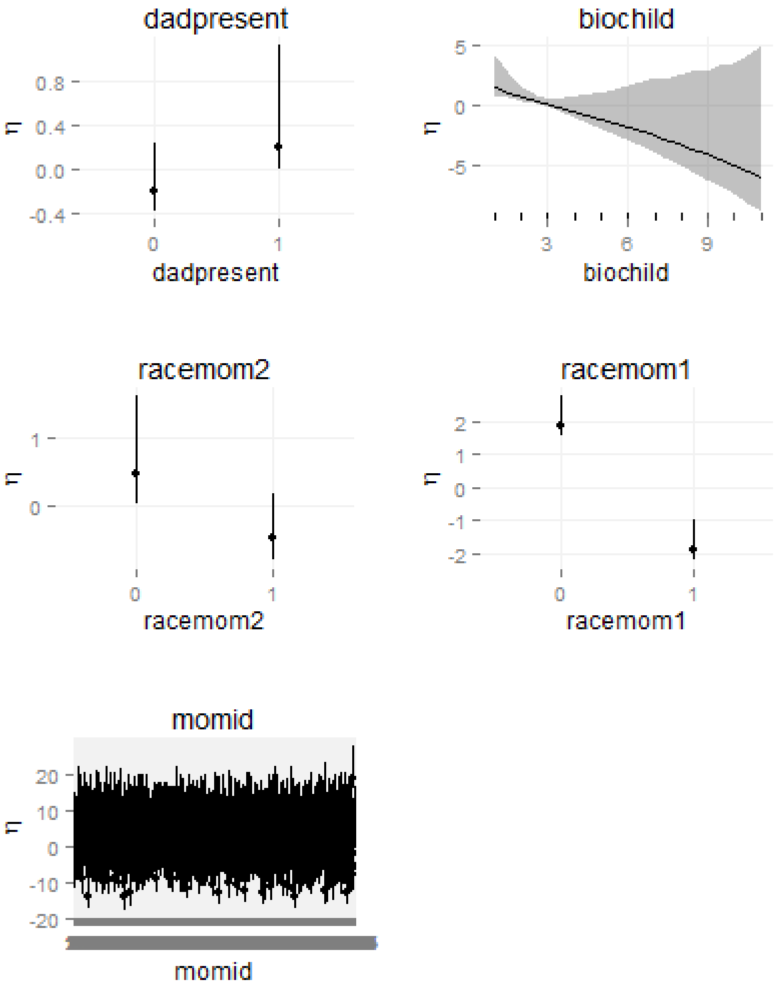

Convergence plot
Variance sigma2alpha
20
30
40
50
60
70
0
1000
2000
3000
4000
Iteration

Appendix C

95 percent credibility intervals over each parameter in the spike and slab variable selection

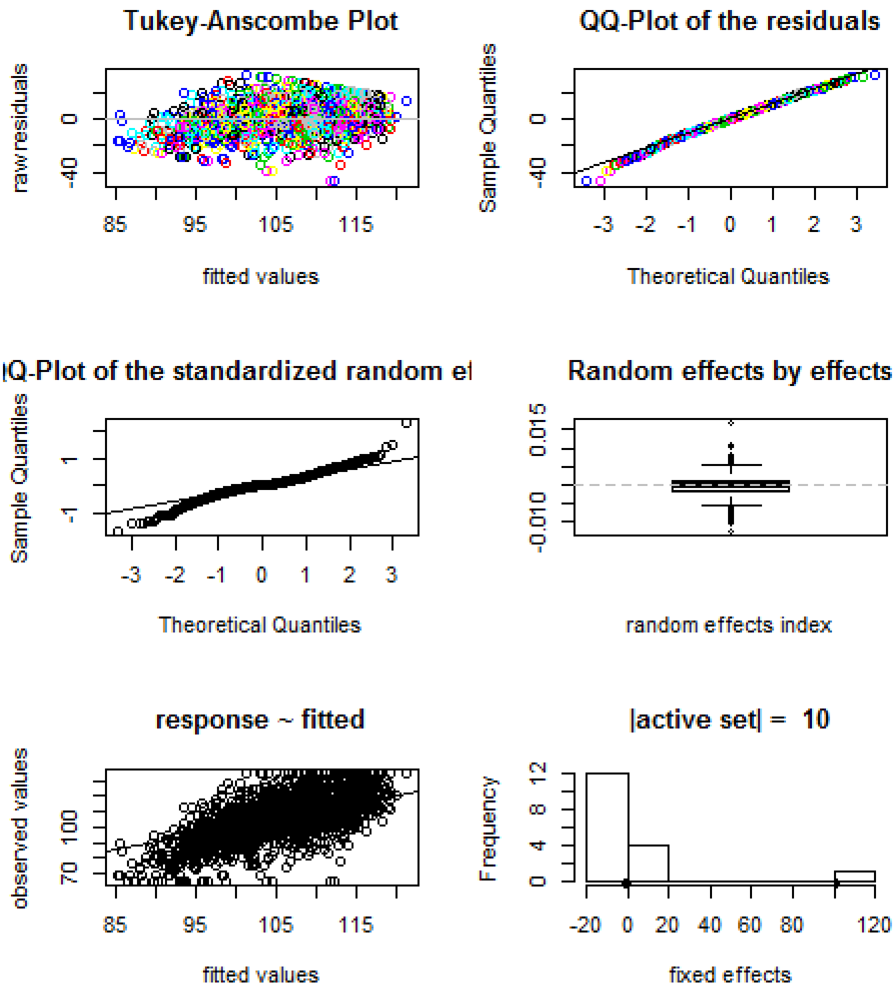



maritial2
2
1
0
η
0
1
maritial2
maritial1
1.00
0.75
0.50
0.25
0.00
η
0
1
maritial1
totincome
5
0
η
0e+00
1e+05
2e+05
3e+05
4e+05
totincome
ageofmomfirstbirth
4
3
2
1
0
η
20
30
40
ageofmomfirstbirth
IQmom81
15
10
5
0
-5
-10
η
0
25
50
75
100
IQmom81
poverty
1.0
0.5
0.0
η
0
1
poverty


dadpresent
0.8
0.4
0.0
-0.4
η
0
1
dadpresent
biochild
5
0
-5
η
3
6
9
biochild
racemom2
1
0
η
0
1
racemom2
racemom1
2
1
0
-1
-2
η
0
1
racemom1
momid
20
10
0
-10
-20
η
momid

Appendix D

Diagnostic plots from the lmmlasso package



LiU-IDA--SE