Bias in students' evaluations

A. Burzacchi, D. Falco, M. Teodori
Supervised by Prof.ssa A. Guglielmi, Dott. R. Corradin

Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione
Academic Year 2019/2020 - First Semester
Bayesian Statistics
Master of Science in Mathematical Engineering

19 February 2020

Contents

Introduction

The goal of the project is to understand which factors influence more students' evaluation in Italy.

OECD (Organization for Economic Co-operation and Development) dataset is a huge available dataset containing a lot of information about 15 years old students.

Math, reading and science scores are recorded through a suitable test, moreover there are also many characteristics of the students, from socio-economic status to personal habits, interests, family background and so on. There is also some schools' information: the kind of school, the size, some teachers' characteristics. There is information from more than 11000 students of 473 schools.

Since the amount of covariates is too big to be analyzed at the same time, we selected the most interesting. Therefore, from more than 1000 covariates we reduced the analysis to 40.

In the analysis we selected, through the Bayesian method using JAGS, the important characteristics, where "important" means those which are related to students' evaluations.

Firstly we used univariate models, where the response is the math score, secondly we extended the models considering as response math and reading score.

With the selected covariates we tried to predict students' evaluation.

In the original dataset, the regional location of the schools is partially available. The region of the school is indicated only when it is Lombardia, Trentino Alto-Adige or Campania. Therefore, we analyzed separately differences among these three regions.

In the end, we clusterized schools with respect to the data and the students' evaluations.

We removed the observations with at least one NA. Since in each analysis interesting covariates are not always the same, the number of students considered changes.

For clarity, traceplots are put together in Appendix A, while methods of covariate selection are in Appendix B. All codes and results can be found on Github `https://github.com/araiari/Pisa-BurzacchiFalcoTeodori`.

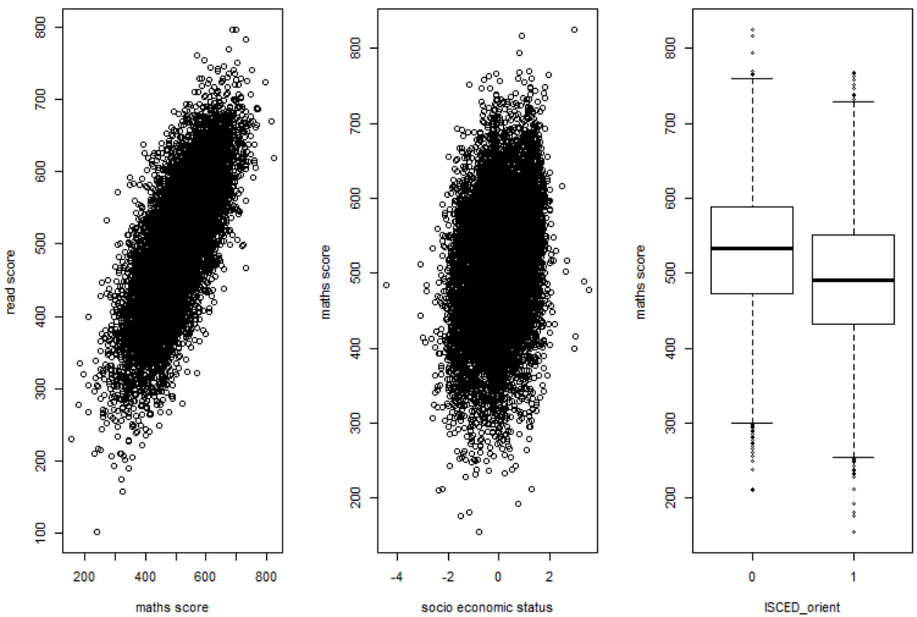


Figure 1: Preliminary plot

1 Univariate model

1.1 Linear model

The goal of this section is to find how to explain students' evaluations (in particular the math score) through students' characteristics. We want to build a linear model of the form:

$$\begin{cases} Y_i \,|\, \mu_i, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_i,\, \sigma^2) & i = 1 : N \\ \quad \mu_i = \beta_0 + \boldsymbol{X}_i^t \boldsymbol{\beta} & \\ \beta_0 \,|\, \tau_0^2 \sim \mathcal{N}(0,\, \tau_0^2) & \tau_0 = 50 \\ \boldsymbol{\beta} \,|, \tau_1^2, \ldots, \tau_p^2 \sim \mathcal{N}_p\big(\mathbf{0},\, \begin{bmatrix} \tau_1^2 & & 0 \\ & \cdots & \\ 0 & & \tau_p^2 \end{bmatrix}\big) & \tau_j = 50 \\ \sigma \sim \mathcal{U}\text{nif}(0, 120) & \end{cases} \tag{1}$$

In order to select properly students' covariates, we performed a Stochastic search variable selection (9) and an Elastic Net (10).

Out of 25 covariates, 17 were kept by both methods, the SSVS (Figure 10) kept also *motivat* and *home possession*, the EN (Figure 11) also *instrum motivat*.

We built eight different linear models with all possible combinations of these three covariates, keeping always the 17 covariates kept by both methods, and we compared the models through WAIC index.

WAIC index suggested to keep *home possession* and *motivat* and to discard *instrum motivat*. Overall these are the 19 covariates kept (17 + *home possession* and *motivat*):

Gender	Index immigration status	Cultural possessions
Escs	Video games	Internet social
Study before school	Study after school	Learning time math
Learning time lang	Learning time science	Disciplinary climate
Enjoy science	Interest broad science	Test anxiety
Edu partent medium	Edu partent high	Home possessions
Motivat		

Unfortunately this kind of approach was too permissive. Indeed building a linear model with all these covariates (sometimes highly correlated) created some problems in the convergence. For example some covariates converged around negative values when they should be positive.

To fix this problem we discarded *a priori* the covariates discarded in the first step, and we built seven liner models always bigger (the first with only six covariates, the last one with all nineteen). The entrance order was decided trying to insert firstly covariates not correlated to the one already in the previous models.

The biggest models (in particular model 6 and 7) had some problem in the convergence, while smaller models had a good convergence and nice traceplots.

Using twice the data, we compared these models through a leave-one-out Cross Validation.

The crossvalidation was built in this way: for each observation i and for each iteration j we computed the error $err_i[j] = |pred_i[j] - y_i|$, where y_i is the true evaluation of student i and $pred_i[j] = \beta_0[j] + \boldsymbol{X}_i^t \boldsymbol{\beta}[j]$. Then we computeted the mean wrt iterations and finally the mean wrt students. $\beta_0[j]$, $\boldsymbol{\beta}[j]$ are realizations of the posterior distribution.

We decided to used the forth model, with 12 covariates. (Figure 12)

Covariates of model 4:

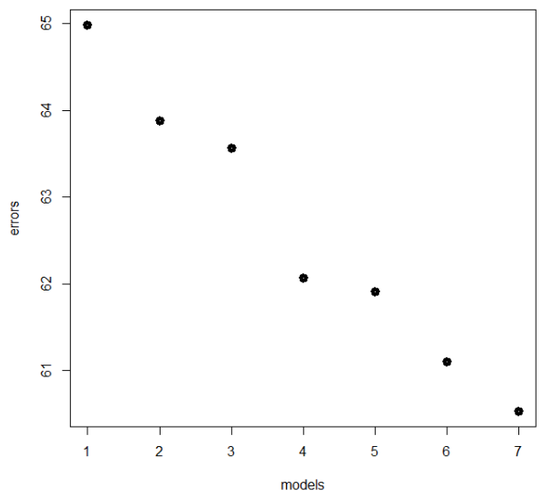


Figure 2: Leave-one-out Crossvalidation

Gender	Cultural possessions	Escs
Video games	Study before school	Study after school
Enjoy science	Learning time science	Disciplinary climate
Motivat	Interest broad science	Test anxiety

All these covariates, but *Video games*, *Study before school* and *Test anxiety*, converge around positive values, which is consistent both with intuition and available data.

1.2 Hierarchical model

After having understood the important covariates related to students, now we want to include also some schools' information and to build a hierarchical model.

A priori all coefficients are centered on zero, except for the intercepts (group varying), which are centered on the sample mean of math scores. X is the design matrix containing schools' covariates, Z is the design matrix containing students' covariates.

$$
\begin{cases}
Y_{ij} \mid \mu_{ij}, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_{ij},\ \sigma^2) & i = 1:N;\ j = 1:ng \\
\mu_{ij} = \gamma_{0j} + \boldsymbol{X}_j^t \boldsymbol{\theta} + \boldsymbol{Z}_{ij}^t \boldsymbol{\gamma}_j & \\
\gamma_{0j} \mid \hat{Y}, \tau_0^2 \overset{iid}{\sim} \mathcal{N}(\hat{Y},\ \tau_0^2) & j = 1:ng \\
\boldsymbol{\theta} \mid \tau_\theta^2 \sim \mathcal{N}_p\big(\mathbf{0}, \tau_\theta^2\, \mathbb{I}_p\big) & \\
\boldsymbol{\gamma}_j \mid \tau_1^2, \dots, \tau_v^2 \overset{iid}{\sim} \mathcal{N}_v\big(\mathbf{0},\ \begin{bmatrix} \tau_1^2 & \cdots & 0 \\ 0 & & \tau_v^2 \end{bmatrix}\big) & j = 1:ng \\
\sigma \sim \text{Inv-}\mathcal{G}(2, 10) & \\
\tau_0^2, \tau_l^2 \overset{iid}{\sim} \text{Inv-}\mathcal{G}(2, 50) & l = 1:v \\
1/\tau_\theta^2 \sim \mathcal{E}(100) + 0.5 &
\end{cases}
\tag{2}
$$

In the preliminary covariate selection, from the huge amount of covariates we kept seven schools variables. Despite of an important reduction, many of them did not seem interesting. The result of the *hierarchical elastic net* (12) suggested not to keep any schools' covariate (figure 13). However, since we wanted to keep at least one of them, we decided to include *ISCED orientation*, a binary variable equal to zero when the school is a *liceo* and one when it's a *scuola professionale*. Our choice is supported by the preliminary analysis where we can see a slight difference in scores.

We built the hierarchical model using only one schools' covariate ($p = 1$) and the twelve covariates selected before.

We tried to insert also another school variable, *student teacher ratio*, but its coefficient was centered around zero.

Moreover, we tried to use a smaller number of students' covariates (only six), but we saw that there was an important loss of information and so we used all twelve covariated selected before.

Data said that result in a *liceo* are in general better then in a *scuola professionale*. Coherently, the sign of *ISCED orient* is negative. (Figure 14)

Since for each group there are not a lot of observations (20 students per school in mean), in the hierarchical model most of the variability is explained by the varying group intercept γ_{0j}.

2 Bivariate model

2.1 Linear model

This section is about the bivariate model, an extension of the univariate model with a bivariate response, math and reading score (in the first section there was only the math score).

$$
\begin{cases}
\boldsymbol{Y}_i \,|\, \boldsymbol{\mu}_i, \Sigma \overset{ind}{\sim} \mathcal{N}_2(\boldsymbol{\mu}_i,\ \Sigma) & i = 1 : N \\
\boldsymbol{\mu}_i = \begin{bmatrix} \beta_{0,1} + \boldsymbol{X}_i^t \boldsymbol{\beta}_1 \\ \beta_{0,2} + \boldsymbol{X}_i^t \boldsymbol{\beta}_2 \end{bmatrix} & \\
\boldsymbol{\beta}_0 \,|\, \tau_0^2 \ \sim \mathcal{N}_2(\mathbf{0},\ \tau_0^2\, \mathbb{I}_2) & \tau_0 = 50 \\
\boldsymbol{\beta}_1 \,|\, \tau_1^2 \ \sim \mathcal{N}_p(\mathbf{0},\ \tau_1^2\, \mathbb{I}_p) & \tau_1 = 50 \\
\boldsymbol{\beta}_2 \,|\, \tau_2^2 \ \sim \mathcal{N}_p(\mathbf{0},\ \tau_2^2\, \mathbb{I}_p) & \tau_2 = 50 \\
\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} & \\
\sigma_1, \sigma_2 \overset{iid}{\sim} \mathcal{U}\text{nif}(0, 120) & \\
\rho \sim \mathcal{U}\text{nif}(-1, 1) &
\end{cases}
\tag{3}
$$

We proceed as in the first section, selecting before students' covariates, and then building a hierarchical model.

First of all, we extended the Elastic Net to the bivariate case (13). For what we saw before, not surprisingly, only few covariates were discarded. (Figure 15)

Selected covariates by EN:

Gender	Index immigration status	Cultural possessions
Escs	Video games	Internet social
Study before school	Study after school	Learning time math
Subjective well being	Learning time science	Disciplinary climate
Enjoy science	Interest broad science	Test anxiety
Motivat	Sport after school	

We built eight different models always bigger, the first with three covariates, the last one with seventeen. We compared these models through a leave-one-out crossvalidation (in the same way of the first section) and we decided to keep the fifth model, with eleven covariates. (Figure 16)

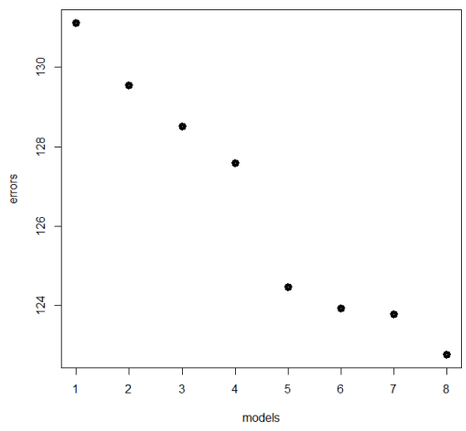


Figure 3: Crossvalidation

Covariates of the fifth model:

Gender	Cultural possessions	Escs
Study before school	Study after school	Video games
Enjoy science	Interest broad science	Test anxiety
Motivat	Learning time science	

As in the first section, the sign of the covariates is not counterintuitive. *Video games*, *Study before school* and *Test anxiety* are negative.

2.2 Hierarchical model

For what we saw in the first section, the only interesting variable related to the school is *ISCED orientation*. Therefore we built a hierarchical model including this information. Since there is only one school's covariate, we have p= 1.

A priori all coefficients are centered on zero, except for the intercepts (group varying), which are centered on the sample mean. X is the design matrix containing schools' covariates, Z is the design matrix containing students' covariates.

$$
\begin{cases}
\boldsymbol{Y}_{ij} \,|\, \boldsymbol{\mu}_{ij}, \Sigma \overset{ind}{\sim} \mathcal{N}_2(\boldsymbol{\mu}_{ij},\, \Sigma) & i = 1:N;\ j = 1:ng \\
\boldsymbol{\mu}_{ij} = \begin{bmatrix} \boldsymbol{\gamma}_{0j,1} + \boldsymbol{Z}_{ij}^t \gamma_{1j} + \boldsymbol{X}_j^t \boldsymbol{\beta}_1 \\ \gamma_{0j,2} + \boldsymbol{Z}_{ij}^t \boldsymbol{\gamma}_{2j} + \boldsymbol{X}_j^t \boldsymbol{\beta}_2 \end{bmatrix} & \\
\boldsymbol{\gamma}_{0j} \,|\, \hat{\boldsymbol{Y}}, \tau_0^2 \overset{ind}{\sim} \mathcal{N}_2(\hat{\boldsymbol{Y}},\, \tau_0^2 \mathbb{I}_2) & j = 1:ng \\
\boldsymbol{\gamma}_{1j,k} \,|\, \tau_{1,k}^2 \overset{ind}{\sim} \mathcal{N}(0,\, \tau_{1,k}^2) & j = 1:ng; k = 1:v \\
\boldsymbol{\gamma}_{2j,k} \,|\, \tau_{2,k}^2 \overset{ind}{\sim} \mathcal{N}(0,\, \tau_{2,k}^2) & j = 1:ng; k = 1:v \\
\boldsymbol{\beta}_1 \,|\, \omega_1^2 \sim \mathcal{N}_p(\boldsymbol{0},\, \omega_1^2 \mathbb{I}_p) & \\
\boldsymbol{\beta}_2 \,|\, \omega_2^2 \sim \mathcal{N}_p(\boldsymbol{0},\, \omega_2^2 \mathbb{I}_p) & \\
\tau_0^2,\, \tau_{1,1}^2, \ldots, \tau_{1,v}^2,\, \tau_{2,1}^2, \ldots, \tau_{2,v}^2 \overset{iid}{\sim} \text{Inv-}\mathcal{G}(2, 50) & \\
1/\omega_1^2,\, 1/\omega_2^2 \overset{iid}{\sim} \mathcal{E}(100) + 0.5 & \\
\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} & \\
\sigma_1, \sigma_2 \overset{iid}{\sim} \mathcal{U}\text{nif}(0, 120) & \\
\rho \sim \mathcal{U}\text{nif}(-1, 1) &
\end{cases}
\tag{4}
$$

In both bivariate models (linear and hierarchical), it is important to underline the role of ρ, which models the correlation between math and reading score. The correlation of data is very high (71%) and the *a posteriori* ρ converges between 65% and 68%. (Figures 17, 18)

The prior distribution of σ is a $\mathcal{U}$niform(0,120). As suggest by Gelman 2006[2] , we tried also using a Troncated Normal and with an Inverse Gamma with small parameter, but with worse results.

In both hierarchical models, it is necessary to set the prior mean of the intercepts equal to the sample mean (and not to zero). Indeed there are some schools were there are only males (*gender* = 1), therefore in the design matrix there are two columns always equal to one (the intercept and the gender). In this case, not considering as prior mean the sample mean, we would get strange results, such as an intercept close to zero and the gender coefficient very high (keep in mind that there is an intercept and a gender coefficient for each group). Since the results are good just using as prior mean the sample mean, we did not consider necessary to make jittering on gender or to remove this kind of observations. In linear models there is not this problem because there are no groups and there is not a column always equal to one.

2.3 Prediction

One of the aims of our project was to predict the evaluation of a new student either from an existing school or from a new school. We decided to use the hierarchical model presented in subsection 2.2, which allows us to get information about math and reading scores.

We started considering a new student from an existing school j. The predictive distribution of the evaluation is described by the formula below. The parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the fixed coefficients of the school covariates, while Θ_j is the collection of all the random parameters depending on the groups, i.e. $\Theta_j = \{\boldsymbol{\gamma}_{0j}, \boldsymbol{\gamma}_{1j}, \boldsymbol{\gamma}_{2j}\}$.

$$\mathcal{L}(\boldsymbol{Y}_j^{new} \,|\, \text{data}) = \int \mathcal{L}(\boldsymbol{Y}_j^{new} \,|\, \Theta_j, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{X}^{new}, \boldsymbol{Z}^{new}, \Sigma)\, \mathcal{L}(d\Theta_j, d\boldsymbol{\beta}_1, d\boldsymbol{\beta}_2, d\Sigma \,|\, \text{data})$$

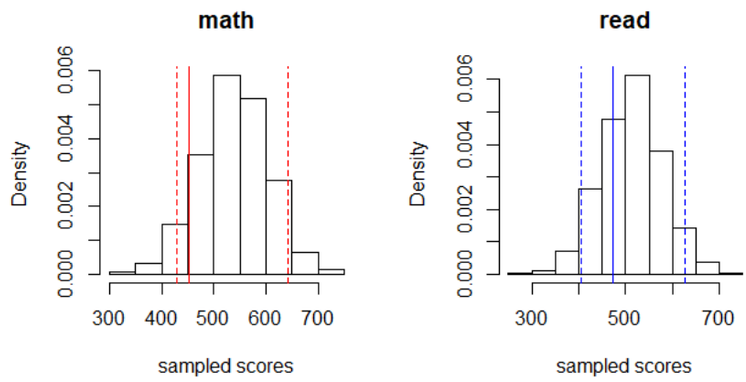
To sample from this distribution, we considered the sampling of the parameters from their joint posterior distribution. The sample $\{\boldsymbol{\gamma}_{0j}[i], \boldsymbol{\gamma}_{1j}[i], \boldsymbol{\gamma}_{2j}[i], \boldsymbol{\beta}_1[i], \boldsymbol{\beta}_2[i], \Sigma[i]\}_{i=1:Niter}$ was the result of a MCMC sampling method using JAGS as in subsection 2.2.

With these values, we computed a sample of $\boldsymbol{Y}_j^{new}$ in the following way:

$$\forall i = 1, ..., Niter \qquad \boldsymbol{Y}_j^{new}[i] \overset{ind}{\sim} \mathcal{N}_2\left(\boldsymbol{\gamma}_{0j}[i] + \begin{bmatrix} \boldsymbol{Z}_{new}^t \gamma_{1j}[i] + \boldsymbol{X}_{new}^t \boldsymbol{\beta}_1[i] \\ \boldsymbol{Z}_{new}^t \boldsymbol{\gamma}_{2j}[i] + \boldsymbol{X}_{new}^t \boldsymbol{\beta}_2[i] \end{bmatrix},\ \Sigma[i]\right)$$

We built a pointwise estimation and a credible interval of level 0.90 of $\boldsymbol{Y}_j^{new}$ for some students of our dataset. Here we report the results of one of them, whose evaluations in maths and reading were [452.047; 471.684]. The real evaluations fits into the interval, which is very wide, but it is usually far from the sample mean.

	Sample quantile 0.05	Sample mean	Sample quantile 0.95
Maths	429.479	538.496	642.729
Reading	405.105	513.252	625.691



Histogram of the components $Y_{j,maths}$ and $Y_{j,read}$ of the new student's evaluation

We tried also to predict the evaluation of a new student coming from a new school s. The random parameters in Θ_s are now unknown and must be predicted along with the evaluation. With the symbol T we indicate the collection of the variances of the random parameters in $\Theta_j \forall j$, i.e. $\mathrm{T} = \{\tau_0^2, \tau_{1,1}^2, ..., \tau_{1,p}^2, \tau_{2,1}^2, ..., \tau_{2,p}^2\}$.

The form of the predictive distribution is the following:

$$\mathcal{L}(\boldsymbol{Y}_s^{new}, \Theta_s \,|\, \text{data}) = \int \mathcal{L}(\boldsymbol{Y}_s^{new} \,|\, \Theta_s, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{X}^{new}, \boldsymbol{Z}^{new}, \Sigma)\, \mathcal{L}(\Theta_s \,|\, \mathrm{T})$$
$$\times\, \mathcal{L}(d\Theta_1, ..., d\Theta_{ng}, d\boldsymbol{\beta}_1, d\boldsymbol{\beta}_2, d\mathrm{T}, d\Sigma \,|\, \text{data})$$

As before, we considered the samplings of the parameters T, $\Theta_1, ..., \Theta_{ng}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and Σ from their joint posterior distribution. The sample was the result of a MCMC sampling method using JAGS.

With these values, we computed a sample of Θ_s and then $\boldsymbol{Y}_s^{new}$ in the following way: $\forall i = 1, ..., Niter$

$$\boldsymbol{\gamma}_{0s}[i] \overset{ind}{\sim} \mathcal{N}_2(\hat{\boldsymbol{Y}},\ \tau_0^2[i]\mathbb{I}_2)$$

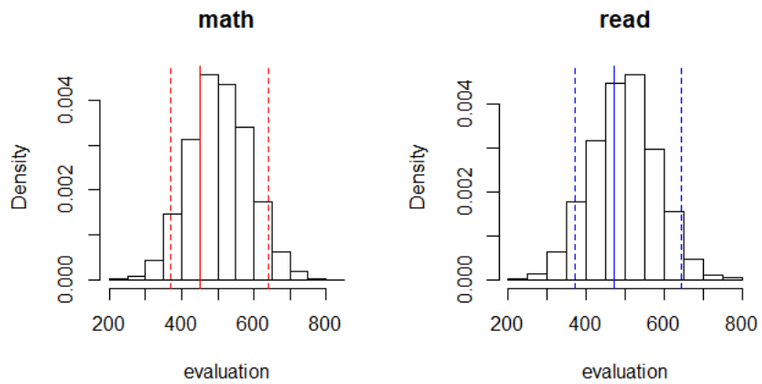
$$\gamma_{1s,k}[i] \overset{ind}{\sim} \mathcal{N}(0,\ \tau_{1,k}^2[i]) \quad \forall k = 1, ..., p$$

$$\gamma_{2s,k}[i] \overset{ind}{\sim} \mathcal{N}(0,\ \tau_{2,k}^2[i]) \quad \forall k = 1, ..., p$$

$$\boldsymbol{Y}_s^{new}[i] \overset{ind}{\sim} \mathcal{N}_2\left(\boldsymbol{\gamma}_{0s}[i] + \begin{bmatrix} \boldsymbol{X}_{new}^t \gamma_{1s}[i] + \boldsymbol{Z}_{new}^t \boldsymbol{\beta}_1[i] \\ \boldsymbol{X}_{new}^t \boldsymbol{\gamma}_{2s}[i] + \boldsymbol{Z}_{new}^t \boldsymbol{\beta}_2[i] \end{bmatrix},\ \Sigma[i]\right)$$

We built a pointwise estimation and a credible interval of level 0.90 of $\boldsymbol{Y}_s^{new}$ for some students of out dataset. We reported the result of the same student as before.

	Sample quantile 0.05	Sample mean	Sample quantile 0.95
Maths	371.021	504.705	640.818
Reading	359.726	496.536	632.868



Histogram of the components $Y_{j,maths}$ and $Y_{j,read}$ of the new student's evaluation

The predictions of a new observation from an existing and a new school are not accurate. We expected this kind of result. In fact, there is a large number of groups and only few observations in each group.

A way to improve the prediction would be adding more covariates about the students. However, this would require a modification in the MCMC algorithm, to make it more accurate and avoid numerical problems.

3 ANOVA

We want to study how the region can impact on the evaluations of the students. The region of residence is known only for the students coming from Campania, Lombardia and Trentino Alto Adige. To accomplish our goal, we used an ANOVA.

Firstly, a classical ANOVA was performed: the response variable is the vector of marks in mathematics and reading and its mean is the parameter $\boldsymbol{\gamma}_j$, different in each region j. The model is the following:

$$\begin{cases} \boldsymbol{Y}_{ij} \,|\, \boldsymbol{\gamma}_j, \Sigma \overset{ind}{\sim} \mathcal{N}_2(\boldsymbol{\gamma}_j,\, \Sigma) \quad i = 1:N; j = 1:3 \\ \boldsymbol{\gamma}_j \,|\, \tau^2 \overset{iid}{\sim} \mathcal{N}_2(\mathbf{0},\, \tau^2\, \mathbb{I}_2) \quad j = 1:3 \\ \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ \sigma_1, \sigma_2 \overset{iid}{\sim} \mathcal{U}\text{nif}(0,1) \\ \rho \sim \mathcal{U}\text{nif}(-1,1) \end{cases} \tag{5}$$

We wanted to compare the parameters $\boldsymbol{\gamma}_j$, j=1, 2, 3 to find differences in the evaluation of the students between the three regions.

For this analysis the data were standardized. The parameter τ^2 was fixed equal to 5. We sampled from the model using JAGS with 12000 iterations, after 1000 iterations of burnin, and thin equal to 5. For all the parameters, the traceplots are wide and the posterior density plot are regular. Both the Gewake diagnostic and the plots show that convergence was reached. However, the autocorrelation goes to zero slowly.

The posterior densities of the mean parameters $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ are shown in figure 4. It is evident that there is an high difference in the scores depending on the region. In particular, students from Campania seem to have the worst evaluations both in mathematics and reading between the analyzed regions.

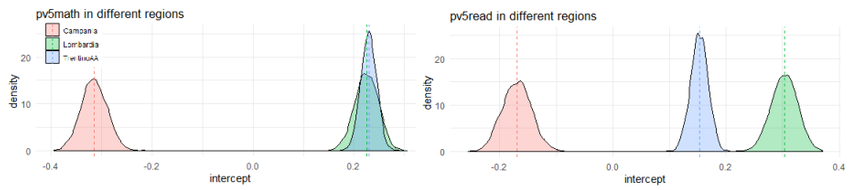


Figure 4: Posterior density plot of $\boldsymbol{\gamma}_{j,\text{math}}$ and $\boldsymbol{\gamma}_{j,\text{read}}$ in model (5)

The model previously described does not consider the available information about the students, such as the gender and the economic-social-cultural status. Thus, the result cannot be considered as precise as we would like.

We decided to build another model which uses, besides the region, the important data of the students. This second model is a linear mixed effects model and, as before, the response is the evaluation of the students. Its mean depends both by the covariates of the student, multiplied by appropriate coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and on the random parameter

$\boldsymbol{\gamma}_j$, different for every region.

$$
\begin{cases}
\boldsymbol{Y}_{ij} \,|\, \boldsymbol{\mu}_{ij}, \Sigma \overset{ind}{\sim} \mathcal{N}_2(\boldsymbol{\mu}_{ij},\, \Sigma) \quad i = 1:N; j = 1:3 \\
\boldsymbol{\mu}_{ij} = \boldsymbol{\gamma}_j + \begin{bmatrix} X_i^t \boldsymbol{\beta}_1 \\ X_i^t \boldsymbol{\beta}_2 \end{bmatrix} \\
\boldsymbol{\gamma}_j \,|\, \tau^2 \overset{iid}{\sim} \mathcal{N}_2(\mathbf{0},\, \tau^2\, \mathbb{I}_2) \quad j = 1:3 \\
\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \,|\, \omega^2 \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0},\, \omega^2\, \mathbb{I}_p) \\
\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\
\sigma_1, \sigma_2 \overset{iid}{\sim} \mathcal{U}\text{nif}(0,1) \\
\rho \sim \mathcal{U}\text{nif}(-1,1)
\end{cases}
\tag{6}
$$

The only parameter depending on the region is the intercept. Therefore, to study the marks between the regions, we compared the three parameters $\boldsymbol{\gamma}_j, j$=1, 2, 3.

As before, the data were standardized and the variances parameters τ^2 and ω^2 were both fixed equal to 5. The data of the students in the matrix $\boldsymbol{X}$ are the most important ones, chosen with the covariate selection described in subsection 2.2. We sampled from the model using JAGS with 12000 iterations, after 1000 iterations of burnin, and thin equal to 5. From the diagnostic plots, we deduced that the convergence is reached. Moreover, with this model, the autocorrelation of the parameters goes to zero more quickly.

Figure 5 shows the posterior densities of the random intercepts $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$. The relation between the regions is the same as before and Campania has worse scores than Lombardia and Trentino. However, the values for the three intercepts are now higher than before. For instance, the reading marks in Campania have positive mean. This happens because there are other factors penalizing the mean of $\boldsymbol{Y}$, such as the covariates which appear with negative coefficients: anxiety for tests, orientation of the institute, use of video games and gender (only for reading).

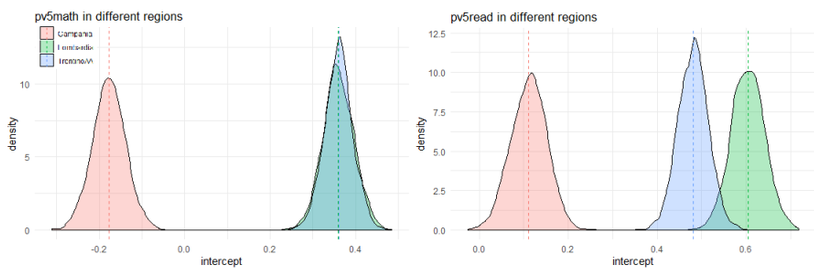


Figure 5: Posterior density plot of $\gamma_{j,\text{math}}$ and $\gamma_{j,\text{read}}$ in model (6)

We wanted to order the three region with respect to the evaluation of the students. From the plots, the order from the best to the worst is the following: Lombardia, Trentino, Campania. Two hypothesis tests were performed to prove our guess. In both cases, we used the model (6) and we compared the parameters $\boldsymbol{\gamma}_j$.

We started studying whether Campania has actually worse marks than Lombardia and Trentino:

$$\text{H}_0\text{: } \{\boldsymbol{\gamma}_1 < \boldsymbol{\gamma}_2\} \wedge \{\boldsymbol{\gamma}_1 < \boldsymbol{\gamma}_3\} \quad \text{H}_1\text{: otherwise}$$

We computed the Bayes factor of the test by evaluating the posterior odd and the prior odd. The two ratios were estimated by counting how many times H_0 and H_1 occur, respectively, in a sample from the prior distribution and from the posterior distribution of the parameters. In particular, we sampled each parameter 12000 times from a bivariate gaussian with zero mean and covariance matrix $\tau^2\, \mathbb{I}_2$. Then, we used the sample from the posterior distribution given as an output of JAGS.

In our sample, evaluations from Campania are always lower than the ones from Lombardia and Trentino and the posterior odd tents to $+\infty$. Thus, there is a strong evidence

in favour of the null hypothesis.

We made a comparison between Lombardia and Trentino to understand if the former is better than the latter:

$$\mathrm{H}_0\text{: } \{\boldsymbol{\gamma}_3 < \boldsymbol{\gamma}_2\} \quad \mathrm{H}_1\text{: otherwise}$$

The Bayes factor was computed in the same way as before. We obtained a prior odd of 0.33 and a posterior odd of 0.96, thus $2\log(BF) = 2.125$. From the result we had a weak evidence in favour of H_0. This uncertainty was already clear from the posterior density plot, where we can notice an equality between the scores in Lombardia and in Trentino.

4 Clustering

The goal of this section is to find a partition of the schools with respect to the data and maths evaluations of the students. Therefore, we would like to divide the schools into clusters.

For this purpose, we build first a BNP mixture model, using the (standardized) student's evaluations and choosing the d most representative covariates coming from the previous models.

This method (requiring JAGS) samples the school-specific random effects parameters $(b_1, \dots, b_{ng})$ from a Dirichlet Process. The random effects enter in the linear model as intercept, while the linear coefficients $(\beta_1, \dots, \beta_d) = \boldsymbol{\beta}$ are treated as fixed effects.

$$\begin{cases} Y_i \,|\, p_i \overset{ind}{\sim} \mathcal{N}(p_i,\ 1) & i = 1 : N \\ p_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + b_{j[i]} & i = 1 : N; j = 1 : ng \\ \boldsymbol{\beta} \perp \{b_j,\ j = 1 : ng\} & \\ \boldsymbol{\beta} \sim \mathcal{N}_6(\mathbf{0},\ 1000\mathbb{I}_6) & \\ b_1, \dots, b_{ng} \,|\, P \overset{iid}{\sim} P & \\ P \sim DP(\alpha, P_0) & \end{cases} \tag{7}$$

Here, $j[i]$ denotes the school j of student i.

Our first covariates choice took factors regarding both students and schools, written in the table below. With them, we obtained a fair result, even though *Student teacher ratio* resulted non significant.

Gender	Escs
ISCED orientation	Student teacher ratio

We built a more representative and complete model using more covariates:

Gender	Escs	Video games
Interest broad science	Test anxiety	ISCED orientation

For this structure, in particular, the complete data contained $N = 10353$ students in $ng = 462$ schools. For the algorithm, we set up an initial *update* of 20000 iterations and a total of 50000 iterations, with *thin* $= 20$ (and so, $M = 2500$).

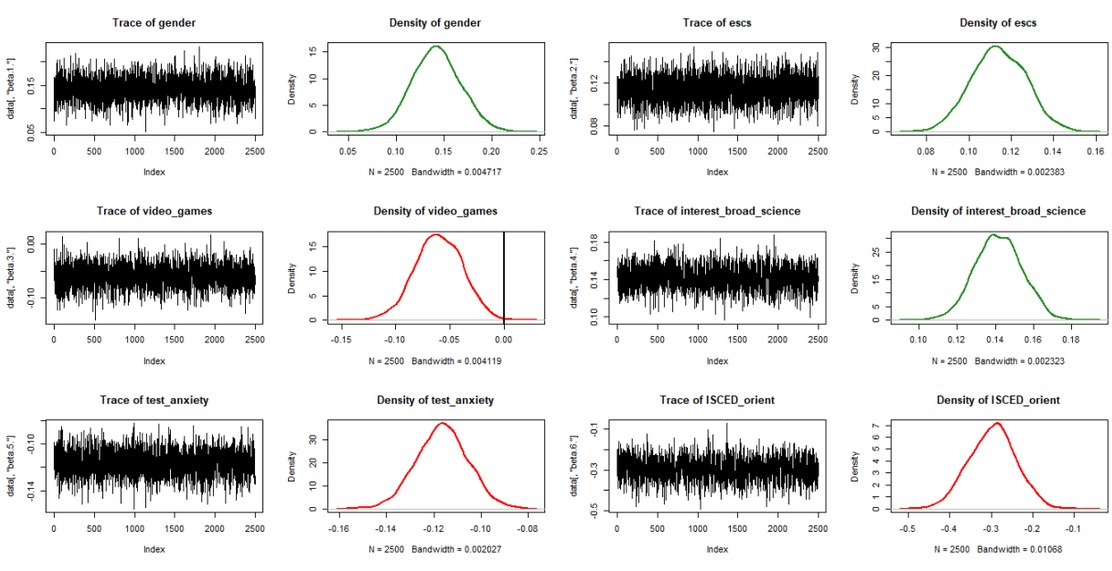


Figure 6: Traceplots of the coefficients of the model (7)

As result, all of the covariates are significant for the model: in particular, *Gender*, *Escs* and *Interest broad science* with positive coefficients, and *Video games*, *Test anxiety* and *ISCED orientation* with negative ones.

The only coefficent whose distribution overlaps the value 0 is *Video games*, but we have that the probability to be at the right of that line is just equal to 0.0028. We expected such result because in our previous models those coefficients were equally significant with the same sign.

A further check on some of the traceplots of the random effects (one for each school) shows a pattern independent from the order of interactions, thus we can conclude that the model is validated.

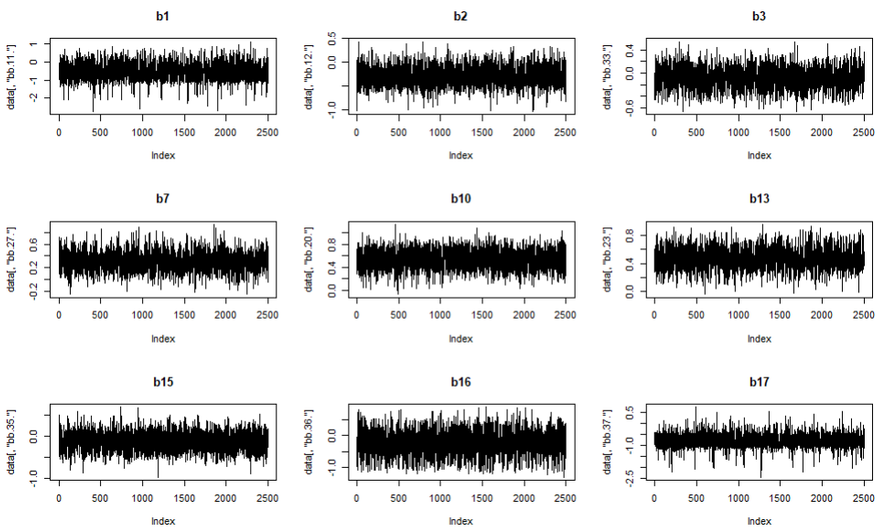


Figure 7: Random effects of some of the schools.

For every iteration $m = 1 : M$ of the algorithm, we can extract a partition ρ of the schools, induced by the posterior distribution of $(b_1^{(m)}, \dots, b_{ng}^{(m)})$. But how can we obtain the best one?

Given a vector of allocation variables $(c_1^{(m)}, \dots, c_{ng}^{(m)})$ such that $b_i^{(m)} = b_j^{(m)} \Leftrightarrow c_i^{(m)} = c_j^{(m)}$, the best partition $\hat{\rho}$ is the one minimizing the Binder's loss function.

$$\begin{cases} \hat{m} = \underset{m}{\operatorname{argmin}}(LF_m) & \\ \pi_{ij} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{\{c_i^{(m)} = c_j^{(m)}\}} & i, j = 1 : ng, i < j \\ \mathrm{LF}_m = \sum_{i<j} (K - \pi_{ij}) \mathbf{1}_{\{c_i^{(m)} = c_j^{(m)}\}} & i, j = 1 : ng, i < j \end{cases} \tag{8}$$

Here, $K \in [0; 1]$ represents a misclassification parameter:

$$K = \frac{b}{a + b}$$

where b can be seen as cost of a wrong assignment to the same cluster, and a as the cost of a wrong assignment to different clusters.

Setting a value of $K = 0.18$, we obtain a partition of the schools in 7 clusters, of which 4 numerous enough to be analysed.

Partition	1	2	3	4	5	6	7
Schools	174	81	118	67	14	6	2

A higher value of K would have led to a minimizing partition with a higher number of clusters, questioning their interpretability.

Here we can see that schools belonging to different clusters have random intercepts with consistent differences in mean.

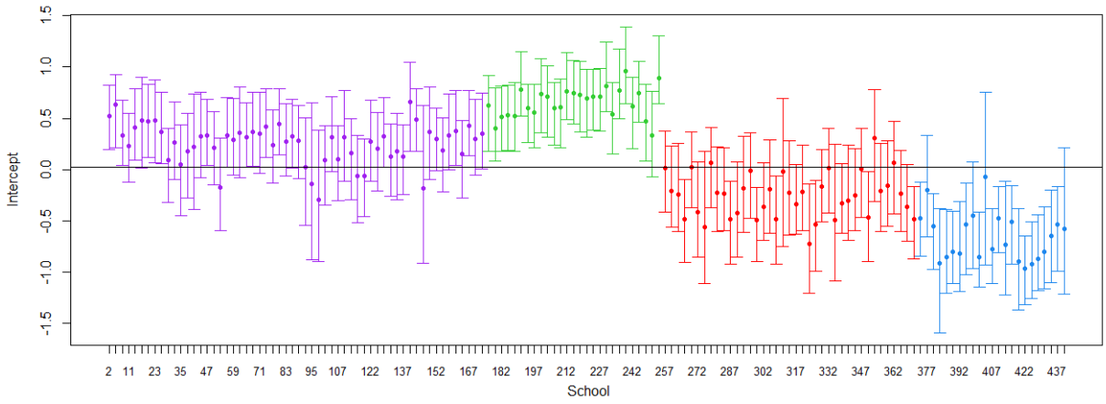


Figure 8: 95 % credible intervals for some of the random intercepts of the schools in different clusters.

Finally, we made an exploratory analysis of the obtained clusters to see their behaviour with respect to the covariates of the model. As result, shown in Figure 9, we can see that the students' evaluation in math have the most significant difference among the clusters, as expected. Indeed, *cluster 2* is the one with the highest mean evaluation, followed with order by *clusters 1, 3* and *4*.

We noticed that clusters with higher evaluations of the students have also higher values of the features which appear with positive coefficients (i.e. escs). On the other hand, the ones with a negative coefficient are smaller in the clusters with higher maths scores.

An exception to this trend is the *gender* covariate, for which we observe a greater proportion of males in *cluster 2* (as expected), but *clusters 1* and *3* present a larger percentage of females than cluster *4* (the barplot in Figure 9 shows the percentage of male students).

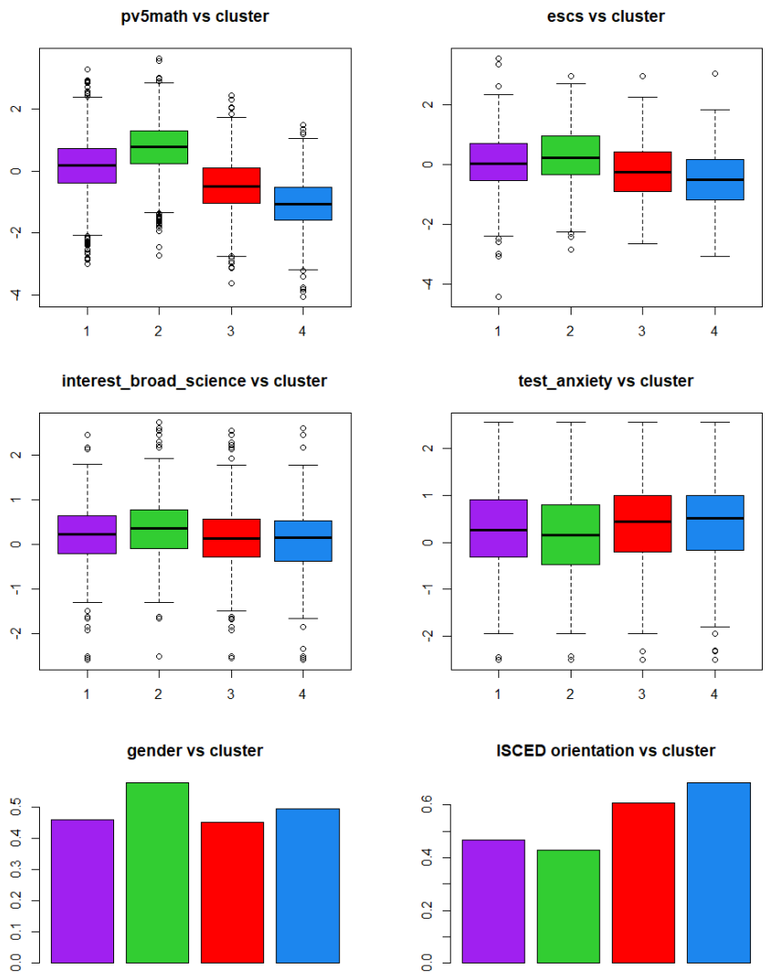


Figure 9: Comparison graphs of the clusters of schools.

Appendix A

In this section there are some example of traceplots and posterior distributions of parameters of models developed in previous sections.

Since the number of parameters is huge, it would be impossible to report all of them, therefore here there are just some examples.
The others can be found on https://github.com/araiari/Pisa-BurzacchiFalcoTeodori

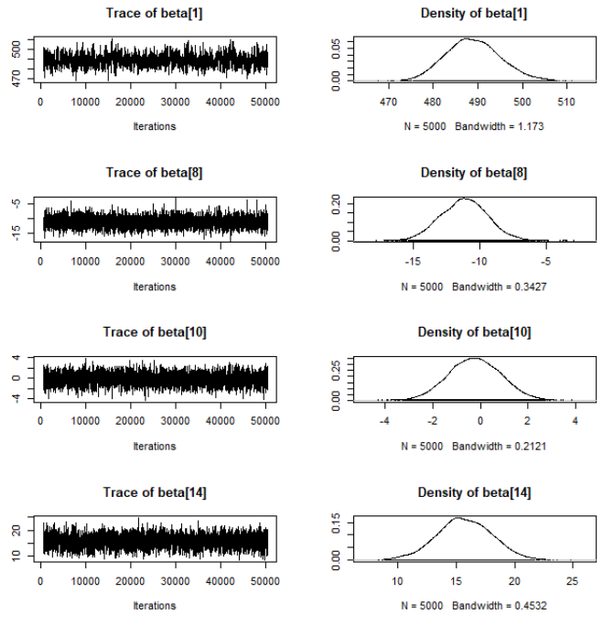


Figure 10: SSVS for the univariate linear model

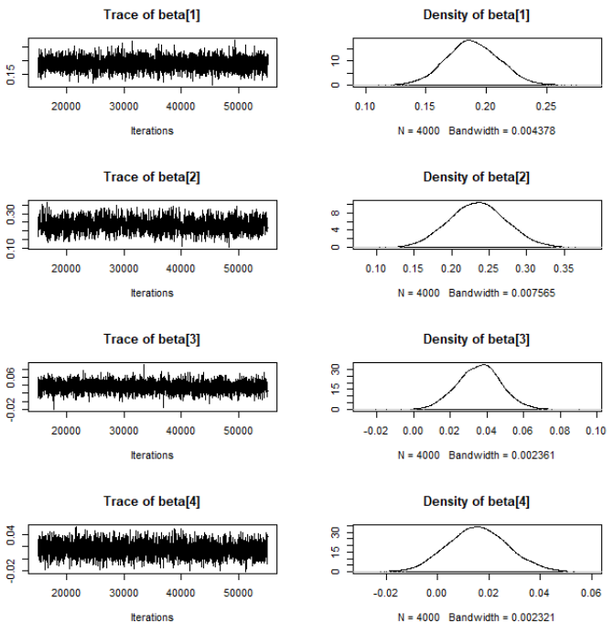


Figure 11: Univariate elastic net

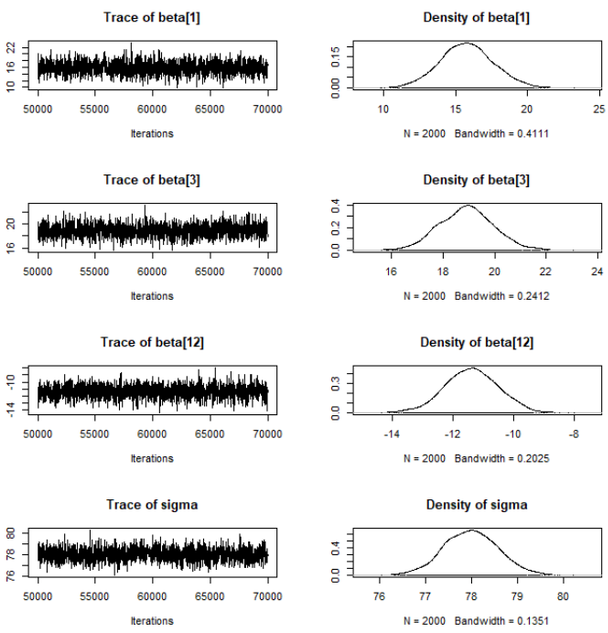


Figure 12: Univariate linear model

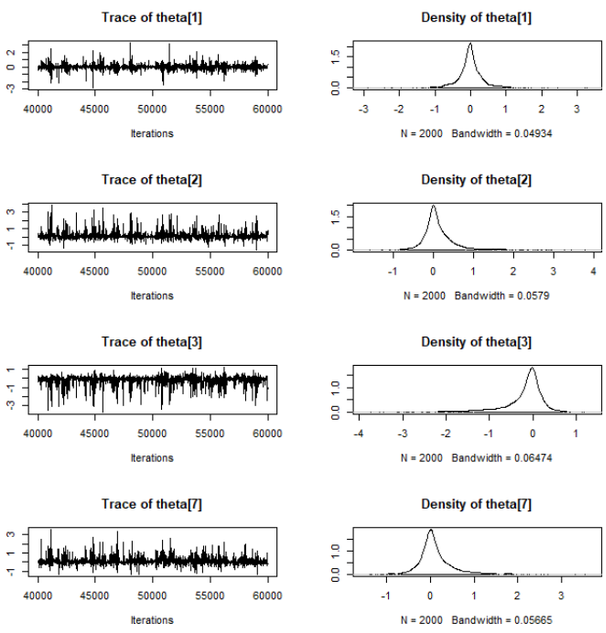


Figure 13: Univariate Hierarchical Elastic Net

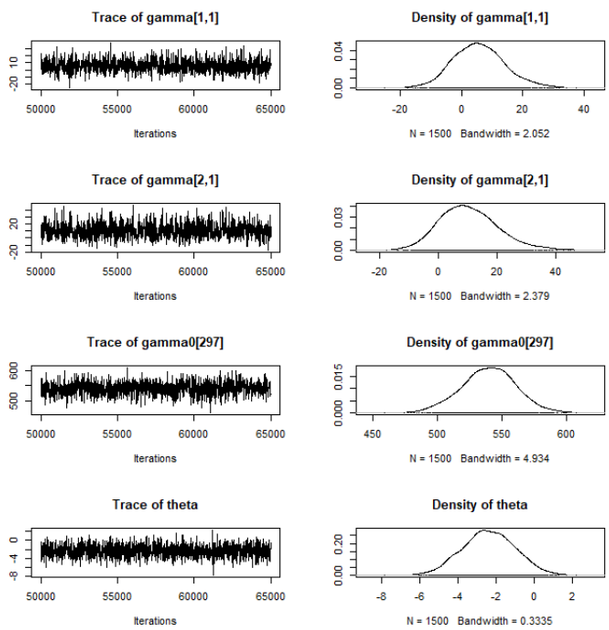


Figure 14: Univariate Hierarchical model

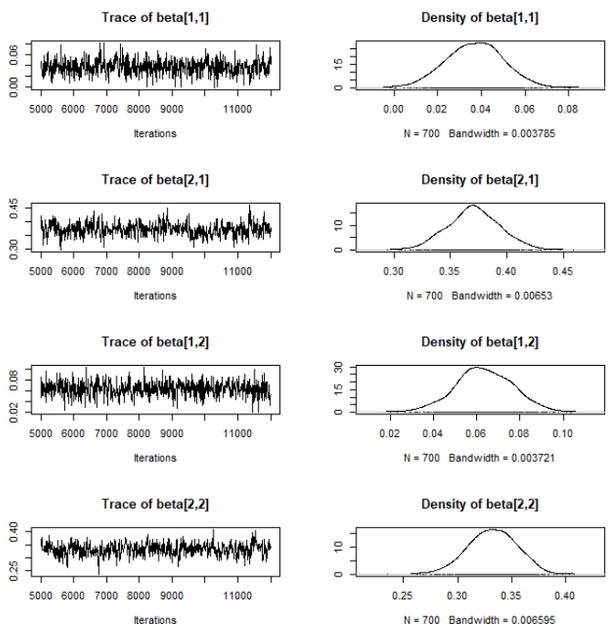


Figure 15: Bivariate elastic net

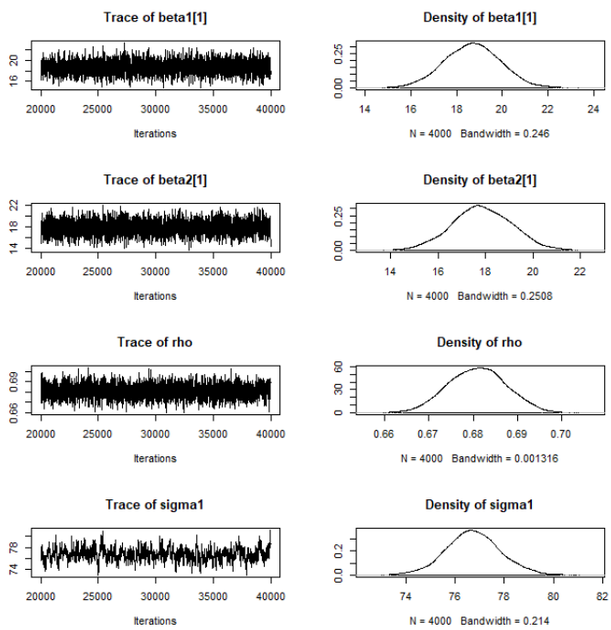


Figure 16: Bivariate linear model

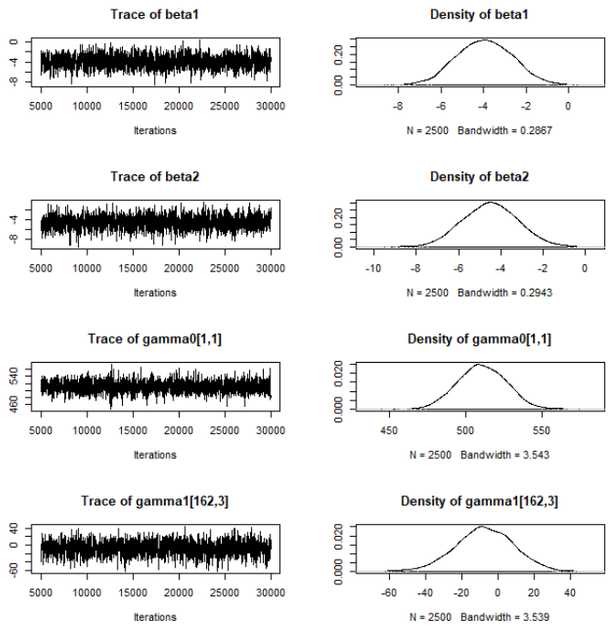


Figure 17: Bivariate hierarchical model

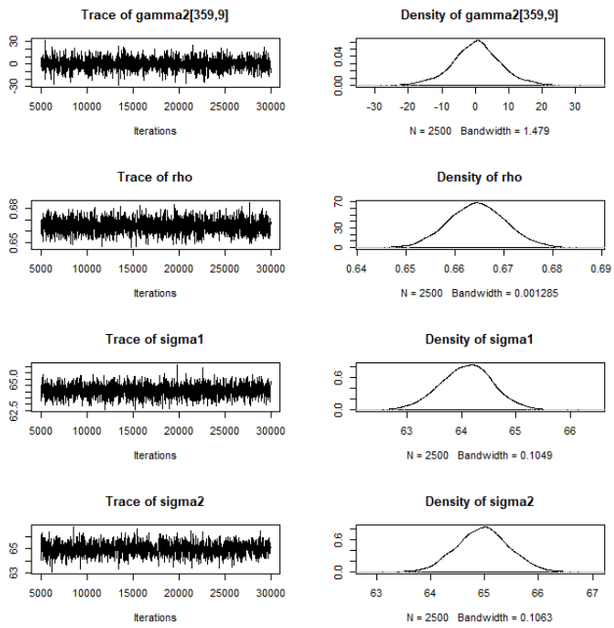


Figure 18: Bivariate hierarchical model

Appendix B

Methods of covariate selection, univariate case

SSVS (Stochastic search variable selection $c_1 = 2.0 \cdot 10^{-4}$, $c_2 = 50.3$, $\lambda = 0.5$)

$$\begin{cases} Y_i \mid \mu_i, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_i,\ \sigma^2) & i = 1 : N \\ \mu_i = \beta_0 + \boldsymbol{X}_i^t \boldsymbol{\beta} & \\ \beta_j \mid \tau_j \overset{ind}{\sim} \mathcal{N}(0,\ \tau_j) & j = 0 : p \\ \tau_j = c_1(1 - \gamma_j) + c_2 \gamma_j & \\ \gamma_j \mid \lambda \overset{iid}{\sim} \mathcal{B}(\lambda) & j = 0 : p \end{cases} \tag{9}$$

Elastic net ($\tau_0 = 50$, $\lambda = 0.1$)

$$\begin{cases} Y_i \mid \mu_i, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_i,\ \sigma^2) & i = 1 : N \\ \mu_i = \beta_0 + \boldsymbol{X}_i^t \boldsymbol{\beta} & \\ \beta_0 \mid \tau_0^2 \sim \mathcal{N}(0,\ \tau_0^2) & \\ \beta_j \mid a_2, \tau_j \overset{ind}{\sim} \mathcal{N}\big(0,\ \frac{\tau_j - 1}{a_2 \tau_j}\big) & j = 1 : p \\ \tau_j \mid a_1, a_2 \overset{iid}{\sim} \text{tr-}\mathcal{G}(0.5, \frac{a_1^2}{8\, a_2},\ 1, \infty) & j = 1 : p \\ a_1, a_2 \overset{iid}{\sim} \mathcal{E}(\lambda) + 0.5 & \end{cases} \tag{10}$$

Lasso (used only at the beginning, not in the final version. $\tau_0 = 50$, $\lambda = 0.1$, $\alpha_1 = 2$, $\alpha_2 = 10$)

$$\begin{cases} Y_{ij} \mid \mu_{ij}, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_{ij},\ \sigma^2) & i = 1 : N;\ j = 1 : ng \\ \mu_{ij} = \gamma_{0j} + \boldsymbol{X}_j^t \boldsymbol{\theta} + \boldsymbol{Z}_{ij}^t \boldsymbol{\gamma}_j & i = 1 : N;\ j = 1 : ng \\ \gamma_{0j} \mid \tau_0^2,\ \hat{Y} \overset{iid}{\sim} \mathcal{N}(\hat{Y},\ \tau_0^2) & j = 1 : ng \\ \boldsymbol{\gamma}_{kj} \mid s_k^2 \overset{ind}{\sim} \mathcal{N}(0,\ s_k^2) & k = 1 : v;\ j = 1 : ng \\ s_k \overset{iid}{\sim} \mathcal{U}\text{nif}(0,\ 100) & k = 1 : v \\ \theta_k \mid l_2 \overset{iid}{\sim} \text{double-}\mathcal{E}(0,\ l_2^{-1/2}) & k = 1 : p \\ l_2 \sim \mathcal{E}(\lambda) + 0.5 & \\ \sigma \sim \text{Inv-}\mathcal{G}(\alpha_1, \alpha_2) & \end{cases} \tag{11}$$

Hierarchical Elastic net ($s_0 = 50$, $\alpha_1 = 2$, $\alpha_2 = 10$)

$$\begin{cases} Y_{ij} \mid \mu_{ij}, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\mu_{ij},\ \sigma^2) & i = 1 : N;\ j = 1 : ng \\ \mu_{ij} = \gamma_{0j} + \boldsymbol{X}_j^t \boldsymbol{\theta} + \boldsymbol{Z}_{ij}^t \boldsymbol{\gamma}_j & i = 1 : N;\ j = 1 : ng \\ \gamma_{0j} \mid s_0^2,\ \hat{Y} \overset{iid}{\sim} \mathcal{N}(\hat{Y},\ s_0^2) & j = 1 : ng \\ \boldsymbol{\gamma}_{kj} \mid s_k^2 \overset{ind}{\sim} \mathcal{N}(0,\ s_k^2) & k = 1 : v;\ j = 1 : ng \\ s_k \overset{iid}{\sim} \mathcal{U}\text{nif}(0,\ 100) & k = 1 : v \\ \theta_k \mid \tau_k, a_2 \overset{ind}{\sim} \mathcal{N}\big(0,\ \frac{\tau_k - 1}{\tau_k a_2}\big) & k = 1 : p \\ \tau_k \mid a_1, a_2 \overset{iid}{\sim} \text{tr-}\mathcal{G}\big(0.5, \frac{a_1^2}{8\, a_2^2},\ 1, \infty\big) & k = 1 : p \\ a_1, a_2 \overset{iid}{\sim} \mathcal{E}(0.1) + 0.5 & \\ \sigma \sim \text{inv-}\mathcal{G}(\alpha_1, \alpha_2) & \end{cases} \tag{12}$$

Methods of covariate selection, bivariate case

Elastic net ($\tau_0 = 50$)

$$
\begin{cases}
\boldsymbol{Y}_i \,|\, \boldsymbol{\mu}_i, \Sigma \overset{ind}{\sim} \mathcal{N}_2(\boldsymbol{\mu}_i,\, \Sigma) & i = 1:N \\
\boldsymbol{\mu_i} = \boldsymbol{\beta_0} + X_{i,1}\boldsymbol{\beta_1} + ... + X_{i,p}\boldsymbol{\beta_p} & \\
\quad \boldsymbol{\beta}_0 \,|\, \tau_0^2, \sim \mathcal{N}_2(\mathbf{0},\, \tau_0^2 \mathbb{I}_2) & \\
\boldsymbol{\beta}_j \,|\, \tau_j, a_2 \overset{ind}{\sim} \mathcal{N}_2(\mathbf{0},\, \frac{\tau_j - 1}{\tau_j a_2} \mathbb{I}_2) & j = 1:p \\
\tau_j \,|\, a_1, a_2 \overset{iid}{\sim} \text{trunc-}\mathcal{G}(\frac{1}{2},\, \frac{a_1^2}{8a_2},\, 1, \infty) & j = 1:p \\
\quad a_1, a_2 \overset{iid}{\sim} \mathcal{E}(0.1) + 0.5 & \\
\quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} & \\
\quad \sigma_1, \sigma_2 \overset{iid}{\sim} \mathcal{U}\text{nif}(0, 100) & \\
\quad \rho \sim \mathcal{U}\text{nif}(-1, 1) &
\end{cases}
\tag{13}
$$

References

[1] OCED, "PISA Data Analysis Manual", *SPSS, Second Edition*, 2009

[2] Gelman, "Prior distributions for variance parameters in hierarchical models", *Bayesian Analysis*, 2006

[3] Gelman, Hill, "Data Analysis Using Regression and Multilevel/Hierarchical Models", *Cambridge University Press*, 2006

[4] Gelman et Al., "Bayesian Data Analysis", *CRC Press*, 2013

[5] Park, Casella, "The Bayesian Lasso", *Journal of the American Statistical Association*, 2008

[6] Mueller, Quintana, Jara, Hanson, "Bayesian Nonparametric Data Analysis", *Springer*, 2015

[7] Jackman, "Bayesian analysis for the Social Sciences", *Wiley*, 2009