BDA - Project

Anonymous

Contents

1 Introduction

Basic mathematical skills are important in many tasks encountered throughout life such as calculating the monthly payments of a mortgage, balancing a personal budget or understanding inflation and investing. Individuals who do not possess these skills are at a clear disadvantage and can end up making bad financial decisions that are hard to recover from like excessive payday borrowing. These cases are costly both from the perspective of the individual as well as society. We investigate how the proportion of such vulnerable people

Table 1: The analyzed variables, shorthands and their codes in the World Bank data sets

Variable	Code	Shorthand
PISA: Proportion of female/male 15-year-olds with mathematics proficiency below Level 1	LO.PISA.MAT.0.FE/MA	PISA.FE/MA
Current health expenditure per capita, PPP	SH.XPD.CHEX.PP.CD	HEALTH
GNI per capita, PPP	NY.GNP.PCAP.PP.CD	GNI
Fertility rate, total	SP.DYN.TFRT.IN	FERT
Learning-Adjusted Years of School, Female/Male	HD.HCI.LAYS.FE/MA	LEARN
Voice and Accountability: Percentile Rank	VA.PER.RNK	VOICE
Government Effectiveness: Percentile Rank	GE.PER.RNK	GOVERN

vary across OECD and certain non-OECD countries by analyzing the proportion of students falling below the lowest proficiency level in mathematics in the Programme for International Student Assessment (PISA) 2018 study (OECD 2018), and different social and economical factors that contribute to these national differences. We apply two different regression models, normal linear regression and Beta regression to estimate the effects of the different factors. Additionally, we analyze this proportion by sex and thus consider non-hierarchical and hierarchical formulations of some of the models.

2 Data and the analysis problem

Our cross-country data comes from various World Bank databases. To obtain a comprehensive collection of covariates that quantify different social and economical aspects of nations, we used the World Development Indicators (World Bank 2022c), Educational Statistics (World Bank 2022a), Gender Statistics (World Bank 2022b) and Worldwide Governance Indicators (World Bank 2022d) data sets. One issue with the covariate selection was a large proportion of missing data for many covariate candidates, especially when considering years before 2018. Additionally, the countries that take part in the the PISA tests vary year by year which further complicates the analysis. To keep the scope of the project reasonable, we decided to consider the latest available results from 2018 and only consider covariates for which the proportion of missing values was below 10 %. We then removed observations for which there were missing values. This reduced the initial 2018 PISA results data from 77 countries to 67 countries. The covariates that satisfied this selection criterion are shown in Table 1. Additional analysis of the covariates is provided in Section 4.

In order to quantify how many students have fallen substantially behind in mathematics, we use the proportion of PISA test takers whose results are below the lowest proficiency level as described in the PISA technical report (OECD 2018). Students that fall below this proficiency level are likely to fail even elementary real life mathematical problems and thus have the bleakest economical prospects. To analyze whether sex influences the effects of the covariates, we consider the proportion both for girls and boys. These response variables are also shown in Table 1. The main goal of this project is to estimate how well we can model these test proportions using the selected covariates and how large an effect each individual covariate has on the responses. Additionally, we analyze whether the effects of the covariates are different for each sex.

3 Model specifications

The main inference task of this project is regression. A simple initial model for such a task would be a non-hierarchical linear Gaussian model where the expected value of the response is modeled as a linear function of the covariates and the error is assumed to be additive, independent and generated from a zero-mean normal distribution with a homoscedastic error scale. Let $X \in \mathbb{R}^{n \times p}$ be a data matrix where n is the number of observations and p the number of covariates, $\mathbf{y} \in \mathbb{R}^n$ a vector of the responses, $\boldsymbol{\beta} \in \mathbb{R}^p$ the covariate parameters, $\alpha \in \mathbb{R}^n$ the intercept (equal for each observation) and $\sigma_\epsilon \in \mathbb{R}^+$ the residual scale. Then the linear Gaussian model with prior distribution families described in Section 4 is

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_\beta, \Sigma_\beta) \\
\alpha &\sim \mathrm{N}(\mu_\alpha, \sigma_\alpha) \\
\sigma_\epsilon &\sim \text{Half-Normal}(0, \sigma_\sigma) \\
\mathbf{y}|\boldsymbol{\beta}, \alpha, \sigma_\epsilon &\sim \mathrm{N}(X\boldsymbol{\beta} + \alpha, \sigma_\epsilon I)
\end{aligned}
\tag{1}
$$

where we parametrize the univariate normal distribution using the scale instead of variance. We specify the parameter values of the priors in Section 4. We model each proportion separately using shared covariates with the exception of Learning-Adjusted Years of School where we use sex-specific values for both groups. We note here that this linear Gaussian model has certain non-physical aspects to it since the response variable values are constrained to the interval $[0, 100]$ (or $[0, 1]$). Depending on how the observed values behave, the model might have issues near the boundaries of this interval. To compare and see whether these issues are substantial, we also estimate this model using log-transformed covariates and responses (Log-Log model). On this scale, the covariate coefficients can be interpreted as elasticities that estimate what is the proportional change in the response given one percent increase in the covariate. This gives a useful alternative model which is non-linear on the original scale.

While we think that there could exist differences between the effects for the sexes, in general they would probably be roughly quite similar. Thus it makes sense to consider a hierarchical model where the group-specific parameters are sampled from a shared population distribution. It also makes sense to have a shared residual standard deviation. We follow the non-centered parametrization style shown in (Betancourt 2017) where instead of estimating the group-wise parameters directly, a latent variable is used which can then be easily transformed into the group-wise parameters. This approach tends to reduce the number of divergences encountered by the Hamiltonian Monte Carlo (HMC) sampler which indicates possible bias in the estimation. Hierarchical version of the linear Gaussian model is

$$
\begin{aligned}
\mu_\alpha &\sim \mathrm{N}(\mu_{\alpha\text{Hyper}}, \sigma_{\alpha\text{Hyper}}) \\
\tau_\alpha &\sim \text{Half-Normal}(0, \hat{\sigma}_\alpha) \\
\mu_j &\sim \mathrm{N}(\mu_{\text{Hyper}_j}, \sigma_{\text{Hyper}_j}) \\
\tau_j &\sim \text{Half-Normal}(0, \hat{\sigma}_\beta) \\
z^i, \hat{z}^i_j &\sim \mathrm{N}(0, 1) \\
\boldsymbol{\beta}^i_j &= \tau_j \cdot \hat{z}^i_j + \mu_j \\
\alpha^i &= \tau_\alpha \cdot z^i + \mu_\alpha \\
\sigma_\epsilon &\sim \text{Half-Normal}(0, \sigma_\sigma) \\
\mathbf{y}^i|\boldsymbol{\beta}^i, \alpha^i, \sigma_\epsilon &\sim \mathrm{N}(X^i\boldsymbol{\beta}^i + \alpha^i, \sigma_\epsilon I)
\end{aligned}
\tag{2}
$$

where the superscript indicates the group and subscript the covariate in question, and τ_i is the scale of the population distribution of each hierarchical parameter. To keep the model specification simple, we draw each component of the covariate parameter vector independently from the population distribution. We also make similar independence assumptions in the non-hierarchical model as will be discussed in Section 4.

After performing the posterior predictive checks for these linear models in Section 7, we noticed that all of them had issues replicating the existing data set with substantially different summary statistics and systematic issues with the computed distributions. As was noted there, the issues are related to the data being concentrated near the lower bound of the data range. As a possible solution to this, we test a Beta regression model that naturally models the data constraints and is also quite flexible in its shape (Ferrari and Cribari-Neto 2004). To keep the number of models reasonable, we only consider the separate non-hierarchical model structure for the Beta regression. The hierarchical linear models also did not perform well so we suspect that in this case the form of the data generating distribution is more critical than modeling between-group differences of two groups hierarchically.

While the Beta distribution is parametrized by two positive shape parameters α and β, in a regression context it is more useful to use a different parametrization that involves the expected value and some sort of dispersion/precision parameter. (Ferrari and Cribari-Neto 2004) propose using the parametrization $\mu = \alpha/(\alpha + \beta)$ and $\phi = \alpha + \beta$ in which case the original parameters can be computed as $\alpha = \mu\phi$ and $\beta = (1 - \mu)\phi$. In this case, μ is the expected value of the Beta distributed variate and the variance is $\mu(1-\mu)/(1+\phi)$. Thus ϕ can be interpreted as a precision parameter and the model is naturally heteroscedastic with variance being a function of the expected value. In order to enforce the unit interval constraint on μ, a link function $g(\cdot)$ is applied such that $g(\mu_i) = \boldsymbol{\beta}^T \mathbf{x}_i + \alpha$ for each observation i. While there exists manu link functions that satisfy the necessary properties, the logit link $g(\mu) = \log(\mu/(1-\mu))$ has the benefit of having the parameters correspond to changes in log odds ratios of outcomes given a unit increase in the covariates (Ferrari and Cribari-Neto 2004). The formal model description with prior distribution choices discussed in Section 4 is shown in System (3).

$$
\begin{aligned}
&\beta_j \sim \mathrm{T}(\nu_j, \mu_j, \sigma_j) \\
&\alpha \sim \mathrm{T}(\nu_\alpha, \mu_\alpha, \sigma_\alpha) \\
&\phi \sim \text{Half-T}(\nu_\phi, 0, \sigma_\phi) \\
&y_i|\boldsymbol{\beta}, \alpha, \phi \sim \mathrm{Beta}(g^{-1}(\boldsymbol{\beta}^T \mathbf{x}_i + \alpha), \phi)
\end{aligned}
\tag{3}
$$

The T denotes the Student's t distribution with degrees of freedom ν, location μ and scale σ.

4 Constructing weakly informative priors

The chosen covariates have vastly different scales which can make the inference quite slow and inefficient. Thus we standardize the response and the explanatory variables when analyzing the models. However, we estimate the relevant scale of each covariate on the original scale for the linear model since it is much more intuitive and transform that to an equivalent scale parameter on the standardized scale. The sample standard deviations for each variable can be found in Appendix 14.1. Since the Beta regression model models responses on the open unit interval, we naturally don't standardize the response.

4.1 Covariate coefficients

To construct weakly informative priors for the model parameters, we consider a priori how large a change in an explanatory variable would be needed to have an effect on the proportion of boys/girls below proficiency level 1 in the PISA tests. We then set this to be the standard deviation of the prior. For each individual explanatory variable, we chose a prior mean of zero with a symmetric normal distribution to cover both positive and negative effects. The distribution family was chosen for its relationship between probability coverage and standard deviation: three standard deviations from the expected value covers almost all of the probability mass, making prior standard deviation relatively easy to set given a relevant range of values. We first consider the priors for the linear model.

Current health expenditure per capita, Purchasing Power Parity (PPP) (current international $): We would expect large changes to occur when the differences are on the order of thousands since these would indicate either a highly developed healthcare system or a vast system covering large part of the population. 10 % change in the proportions is definitely substantial for a 1000 € change in expenditures. Given the standard deviations, this corresponds to a standard deviation of approximately 1.3 on the standardized scale.

GNI per capita, PPP (current international $): We would expect similar behavior as was discussed for the current health expenditure, though perhaps larger since GNI includes all outputs of a nation rather than just healthcare. Given that 1000 was used previously, 10000 would be quite reasonable here for a 10 % change. Similarly, the scale of the prior is approximately 1.3 on the standardized scale.

Fertility rate, total (births per woman): We would expect substantial changes to occur in the test proportions given even one more birth per woman since additional children will take attention away from other children, and the fertility rate can also indicate more traditional views on the position of women in society which can lead to girls not pursuing schooling. We would estimate that a 10 % change in the proportions for each birth is a reasonable scale estimate. On the standardized scale, this corresponds to a standard deviation of approximately 0.3.

Learning-Adjusted Years of School, Female/Male: We would expect similar kind to effect sizes as for the fertility variable given how much material tends to be covered for each school year. We would estimate that 10 % change in the proportions for each year is a reasonable scale estimate. On the standardized scale, this corresponds to a standard deviation of approximately 1.

Voice and Accountability: Percentile Rank: The effects of the governance indices are harder to estimate. Since the exact ordering is likely to be somewhat fuzzy, we would expect large changes in the percentile rank to reflect substantial differences in the underlying governance. Thus we estimate that a 25 % change in the rank would correspond to a 10 % change in the PISA score proportions. This corresponds to a standard deviation of approximately 0.7.

Government Effectiveness: Percentile Rank: We use the same prior scale as for the Voice and Accountability: Percentile Rank variable. On the standardized scale, this corresponds to a standard deviation of approximately 0.5.

We use the same values for the covariate parameter population mean hyperpriors. Since the covariances between the covariate parameters are not the primary quantities of interest, we set the prior covariance matrix as diagonal. For setting the prior parameter value for the hierarchical scale parameters, we follow the recommendation of (Stan Dev Team 2020) by using a half-normal distribution with a mean of zero. We adjust the scale of this prior to be 0.2. As they explain, a low number of groups tends to require stronger priors to regularize the inference. Additionally, we expect the group differences to be at most moderate so on the standardized scale, we would expect the 0.2 scale to cover all relevant differences between the two groups.

Constructing the priors for Beta regression model is more difficult given that they represent changes in log odds ratios. To ease this task, we follow the recommendation given in (Stan Dev Team 2020) for logistic regression parameters. They recommend using $\mathrm{T}(\nu, 0, \sigma)$ as a weakly informative prior where $3 < \nu < 7$ and σ is chosen to provide weak information. While we don't exactly perform logistic regression, the link function is the same and the covariate parameters have the same interpretation. Given that we are estimating effects on the log odds ratio scale, a scale of 1 for the t distributions should cover most of the meaningful effects, especially since we standardize the covariates. In addition, we set $\nu = 4$ to obtain some robustness against the misscalibration of the prior scale. We set the location to zero to cover both positive and negative effect sizes.

4.2 Intercept and residual/precision scale

Since we standardize the data, we expect the intercept term to be quite close to zero. The standardization can also be used to infer a reasonable prior scale for the residual scale parameter since if the covariate coefficients are zero, the standard error of the constant prediction is the standard deviation of the data which is 1, the maximal value (Stan Dev Team 2022b). Thus we select 0.5 to be the prior scale of the residual scale parameter. As in the covariate parameter case, we use the same values for the hyperprior of μ_α and for the hierarchical scale parameter τ_α.

For the Beta regression model, we use the same t distribution prior for the intercept as in the covariate parameter case. The precision parameter ϕ needs to have a wide enough prior to cover possible shapes of the Beta distribution, possibly reaching on the order of hundreds (Ferrari and Cribari-Neto 2004). To include some possibility of large values into the prior, we use a t distribution with 2 degrees of freedom and a scale parameter of 10.

The linear models using the aforementioned prior parameter values are shown in Systems (4) and (5). We also use the same prior parameter values for the Log-Log models since we standardize the log-transformed variables. They are also reasonable with respect to recommendations of (Stan Dev Team 2020) for elasticities.

$$
\begin{aligned}
&\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(1.3^2, 1.3^2, 0.3^2, 1^2, 0.7^2, 0.5^2)) \\
&\alpha \sim \mathrm{N}(0, 0.5) \\
&\sigma_\epsilon \sim \text{Half-Normal}(0, 0.5) \\
&\mathbf{y}|\boldsymbol{\beta}, \alpha, \sigma_\epsilon \sim \mathrm{N}(X\boldsymbol{\beta} + \alpha, \sigma_\epsilon I)
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
&\mu_\alpha \sim \mathrm{N}(0, 0.5) \\
&\tau_\alpha \sim \text{Half-Normal}(0, 0.2) \\
&\mu_j \sim \mathrm{N}(0, \sqrt{[\Sigma_\beta]_{jj}}) \\
&\tau_j \sim \text{Half-Normal}(0, 0.2) \\
&z^i, \hat{z}^i_j \sim \mathrm{N}(0, 1) \\
&\boldsymbol{\beta}^i_j = \tau_j \cdot \hat{z}^i_j + \mu_j \\
&\alpha^i = \tau_\alpha \cdot z^i + \mu_\alpha \\
&\sigma_\epsilon \sim \text{Half-Normal}(0, 0.5) \\
&\mathbf{y}^i|\boldsymbol{\beta}^i, \alpha^i, \sigma_\epsilon \sim \mathrm{N}(X^i\boldsymbol{\beta}^i + \alpha^i, \sigma_\epsilon I)
\end{aligned}
\tag{5}
$$

The Beta regression model with the aforementioned prior parameter values is shown in System (6).

$$
\begin{aligned}
&\beta_j \sim \mathrm{T}(4, 0, 1) \\
&\alpha \sim \mathrm{T}(4, 0, 1) \\
&\phi \sim \text{Half-T}(2, 0, 10) \\
&y_i|\boldsymbol{\beta}, \alpha, \phi \sim \mathrm{Beta}(g^{-1}(\boldsymbol{\beta}^T\mathbf{x}_i + \alpha), \phi)
\end{aligned}
\tag{6}
$$

5 Stan implementations

In this section we only include the parameters and the model part of the Stan source codes. The full source codes for each model can be found in Appendix 14.5.

5.1 Linear model

```
parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  vector<lower=0>[2] sigma; // Measurement SD per group
}

transformed parameters {
  matrix[N, 2] mu_std; // Linear model for the mean
  for (i in 1:2) {
    mu_std[,i] = alpha[i] + to_matrix(X_std[,,i]) * beta[,i];
```

```
  }
}

model {
  // Priors
  for (i in 1:2) {
    beta[,i] ~ multi_normal(beta_mu_prior, beta_sigma_prior);
    alpha[i] ~ normal(alpha_mu_prior, alpha_sd_prior);
    sigma[i] ~ normal(0, sigma_sd_prior);
  }
  // Likelihood
  for (i in 1:2) {
    y_std[,i] ~ normal(mu_std[, i], sigma[i]);
  }
}
```

5.2 Hierarchical linear model

```
parameters {
  real mu_j[P]; // Covariate param. population means
  real<lower=0> tau_j[P]; // Covariate param. population scales
  real mu_alpha; // Intercept population mean
  real<lower=0> tau_alpha; // Intercept population scale
  real<lower=0> sigma; // Shared measurement scale

  real z_hat[P,2]; // Latent variables for the covariate parameters
  real z[2]; // Latent variables for the intercepts
}

transformed parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  matrix[N, 2] mu_std; // Linear model for the mean

  // Generation of the group-wise parameters from the latent variables
  for (i in 1:2) {
    alpha[i] = tau_alpha * z[i] + mu_alpha;
    for (j in 1:P) {
      beta[j,i] = tau_j[j] * z_hat[j,i] + mu_j[j];
    }
  }

  for (i in 1:2) {
    mu_std[,i] = alpha[i] + to_matrix(X_std[,,i]) * beta[,i];
  }
}

model {
  // Hyperpriors
  mu_alpha ~ normal(alpha_mu_prior, alpha_sd_prior);
  tau_alpha ~ normal(0, alpha_tau_scale);
```

```
  for (j in 1:P) {
    mu_j[j] ~ normal(beta_mu_prior[j], beta_sigma_prior[j]);
    tau_j[j] ~ normal(0, beta_tau_scale);
  }

  // Prior
  sigma ~ normal(0, sigma_sd_prior);

  // Latent variables
  for (i in 1:2) {
    z[i] ~ normal(0, 1);
    for (j in 1:P) {
      z_hat[j,i] ~ normal(0, 1);
    }
  }

  // Likelihood
  for (i in 1:2) {
    y_std[,i] ~ normal(mu_std[, i], sigma);
  }
}
```

5.3 Beta regression model

```
parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  vector<lower=0>[2] phi; // Precision per group
}

transformed parameters {
  matrix[N, 2] mu; // Expected values of the Beta variables
  for (i in 1:2) {
    mu[,i] = inv_logit(alpha[i] + to_matrix(X_std[,,i]) * beta[,i]);
  }
}

model {
  // Priors
  for (i in 1:2) {
    alpha[i] ~ student_t(nu_params_prior, loc_params_prior, scale_params_prior);
    phi[i] ~ student_t(nu_precision_prior, 0, scale_precision_prior);
    for (j in 1:P) {
      beta[j,i] ~ student_t(nu_params_prior, loc_params_prior, scale_params_prior);
    }
  }
  // Likelihood
  for (i in 1:2) {
    y[,i] ~ beta(mu[,i] * phi[i], (1 - mu[,i]) * phi[i]);
  }
}
```

6 Inference and convergence diagnostics

We report the choices and diagnostics separately for the different model types.

6.1 Linear models

We used the default sampler settings for the separate model since we did not encounter any issues when using them. We also initially used the default settings for the hierarchical models. However, we observed a few divergent transitions. We can identify whether these is any systematic behavior in the divergences by using the HMC diagnostics plots from the bayesplot package and running the sampler for more iterations, in this case 10000 post warm-up iterations per chain. Figure 1 represents each MCMC draw as a line segment where the divergences are highlighted in red. We only plot the hierarchical parameters since divergence issues in hierarchical models are often related to the hierarchical scale parameters. As we can see, there is no apparent pattern to the divergences and they do not differ systematically from the non-divergent draws. In this case, we managed to eliminate them completely by decreasing the step size of the sampler. Additional HMC and No-U-Turn Sampler (NUTS) (Hoffman, Gelman, et al. 2014) diagnostics for all fitted models can be found in Appendix 14.2.

```
separ_model <- stan(file = 'PISA_2018_Sep.stan', data = stan_data_sep,
                    seed = seed)

hier_model <- stan(file = 'PISA_2018_Hier.stan', data = stan_data_hier,
                   seed = seed,
                   control=list(adapt_delta=0.98))
```

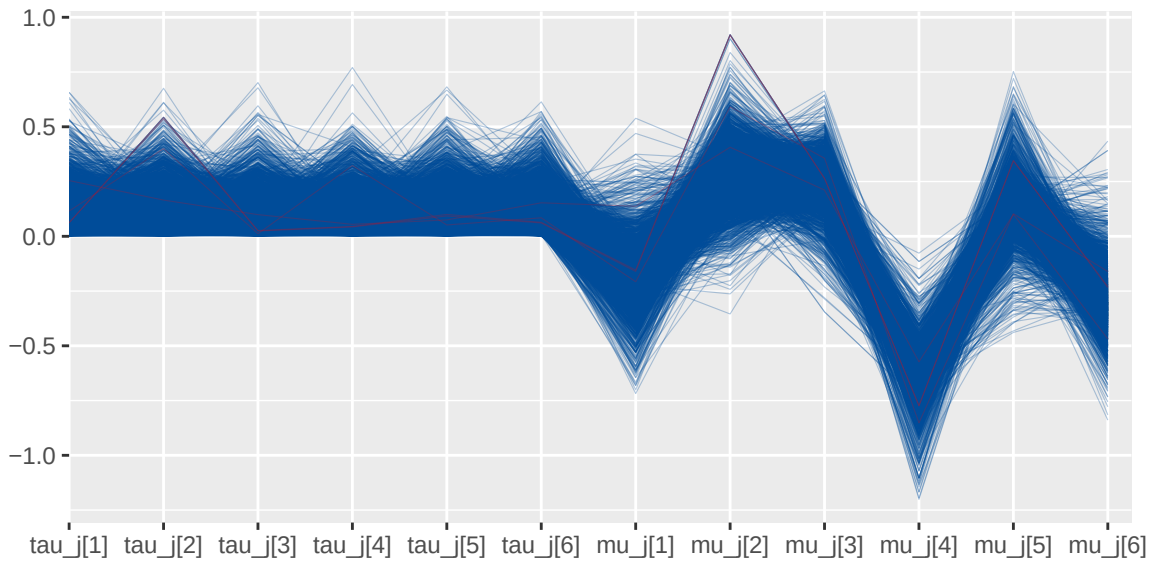

Figure 1: Divergences of the hierarchical linear model

The $\hat{R}$ and effective sample size (ESS) ratio estimates for selected parameters are shown in Figure 2 for the separate model and in Figure 3 for the hierarchical model. For the separate model, we consider all parameters, and the first parameter value is always the girl group. For the hierarchical model, we consider the diagnostics for the shared parameters. Given the recommended thresholds of $\hat{R}$ being below 1.01 and the bulk-ESS being above 400 for 4 chains (Stan Dev Team 2022a), we can be quite certain that the Markov chains have mixed well and are stationary, meaning that they are likely estimating the target posterior

distribution. The ESS ratios exceeding 1 can indicate that the estimates are converging faster to the true mean compared to i.i.d. sampling from the posterior (Gabry and Modrak 2022). Thus location statistics are likely to be very reliable for these normal models.

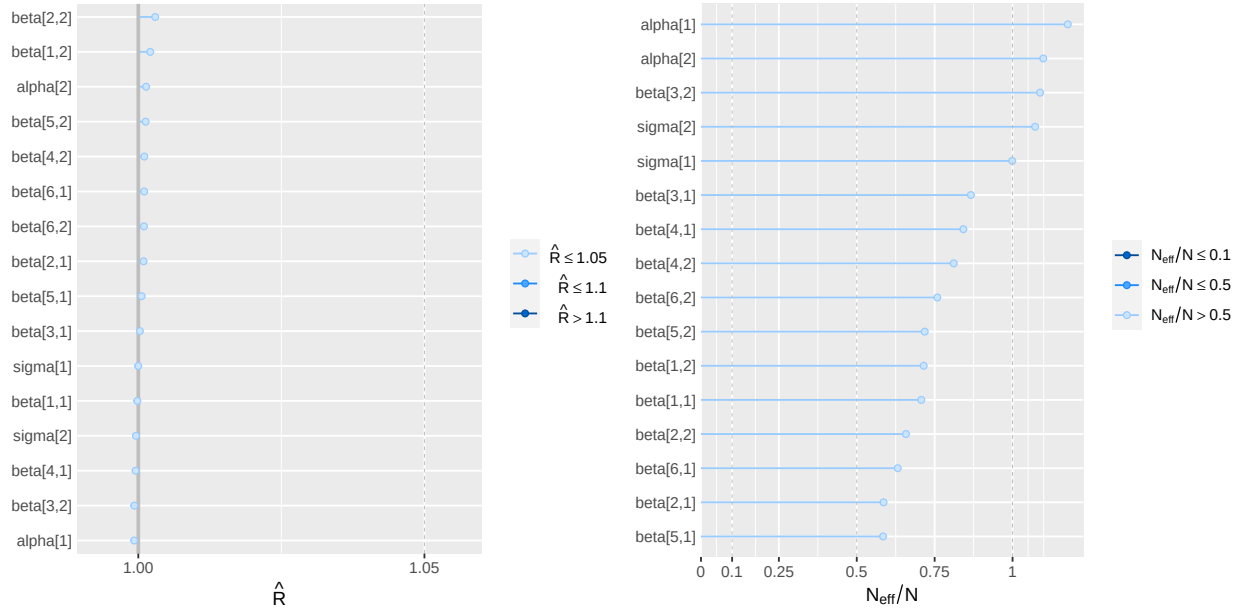

Figure 2: Convergence diagnostics for the separate linear model

6.2 Log-Log models

The behavior of the sampler was very similar for the Log-Log models where we could eliminate the divergences from the hierarchical model by decreasing the HMC step size. Since the results are quite similar, we include the diagnostic results for the hierarchical model in Appendix 14.2. The convergence diagnostics can be seen in Figures 4 and 16 with the hierarchical model seeming to be slightly more efficient in terms of ESS ratios when using the logarithmic scale.

```
separ_model_log <- stan(file = 'PISA_2018_Sep.stan', data = stan_data_sep_log,
                        seed = seed)
hier_model_log <- stan(file = 'PISA_2018_Hier.stan', data = stan_data_hier_log,
                       seed = seed,
                       control=list(adapt_delta=0.98))
```

6.3 Beta regression

While the data generating distribution is different in the Beta regression model, the sampler diagnostics have not changed much as can be seen in Figure 5. Thus we can again be quite certain that the draws made by the sampler estimate well the true posterior distribution of the parameters.

```
beta_model <- stan(file = 'PISA_2018_Beta.stan', data = stan_data_beta,
                   seed = seed)
```

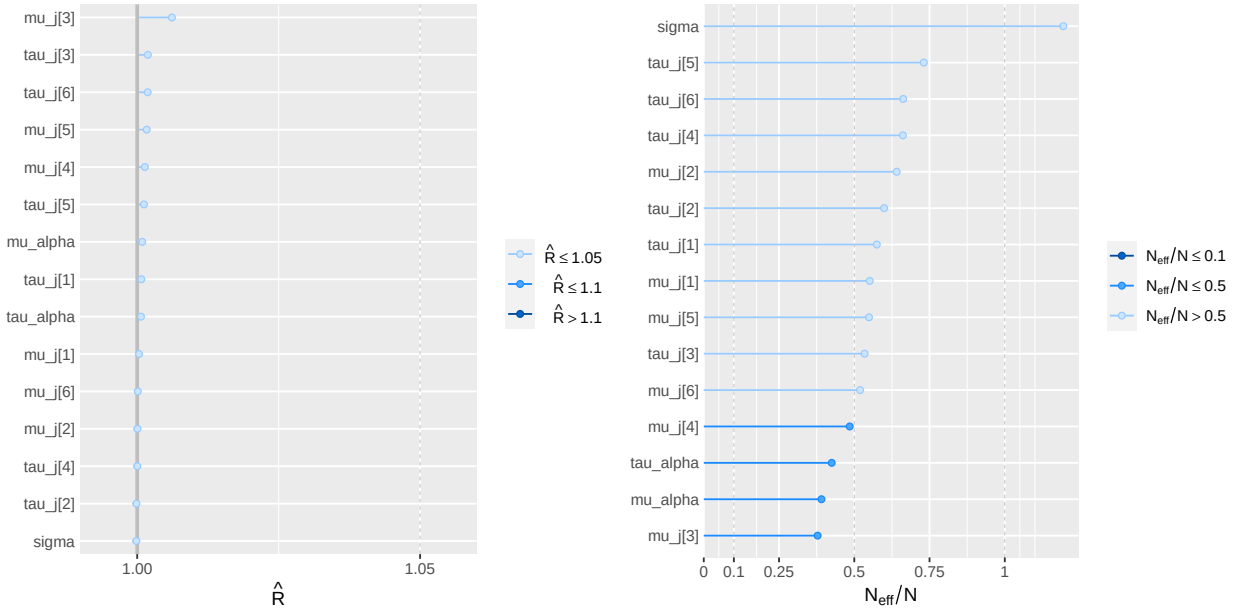

Figure 3: Convergence diagnostics for the hierarchical linear model

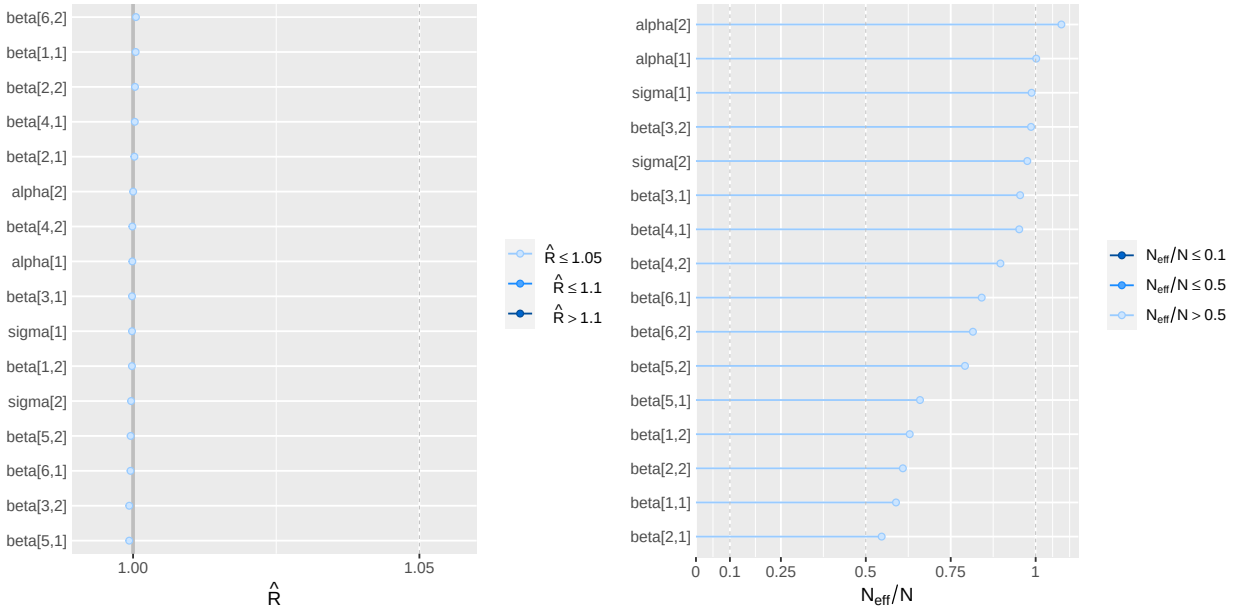

Figure 4: Convergence diagnostics for the separate Log-Log model

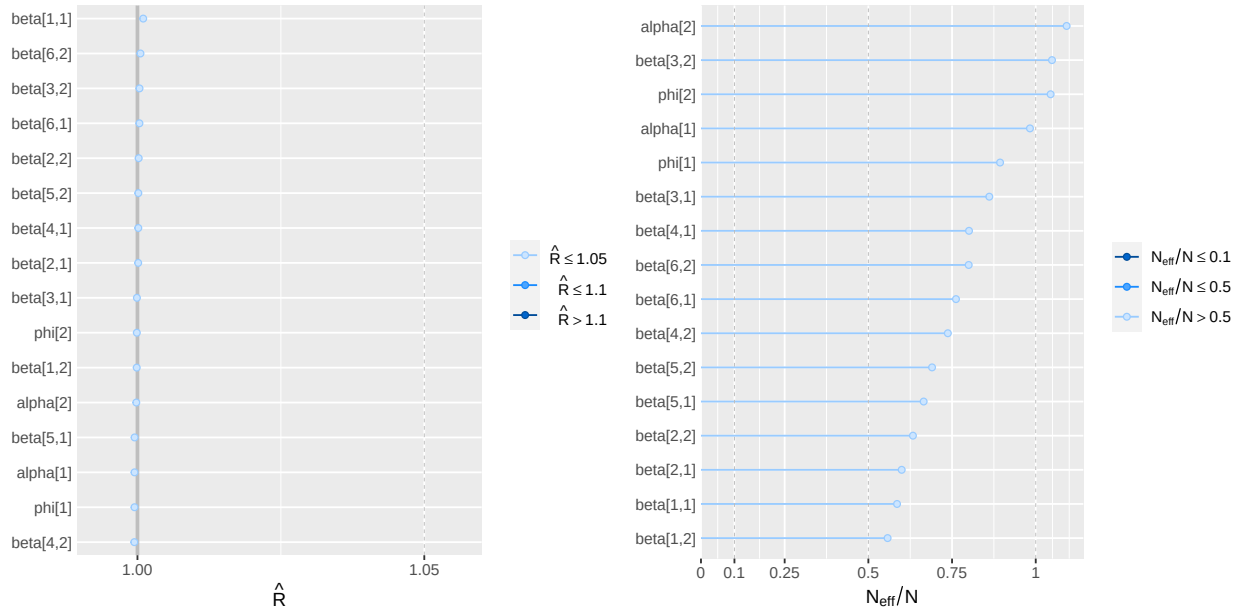

Figure 5: Convergence diagnostics for the Beta regression model

7 Posterior predictive checking

We perform posterior predictive checks (PPC) to evaluate whether our models have misspecification issues by analyzing how well the posterior predictive distributions can generate data that is similar to the observed data set (the proportion of boys/girls below proficiency level 1 in mathematics). Systematic deviations from the observed data can indicate issues with model specification. To get a comprehensive overview of the differences between the data set and the replications produced from the posterior predictive distributions, we visualize the kernel density estimates, the empirical cumulative distribution functions and summary statistics that are ancillary with respect to the model (Gabry et al. 2017). For the normal model, skewness is such a summary statistic as well as minimum and maximum which help to analyze how well the model handles possible boundaries. We note that skewness is not an ancillary statistic for the Beta distribution so it is not as reliable for detecting possible misspecifications (Gabry et al. 2017). We don't believe that this is an issue since we consider multiple different checks that evaluate different aspects of the replications.

7.1 Linear models

Figures 6 and 7 contain the diagnostics plots for the separate and the hierarchical models using the proportion of female students as the response for 200 replications. It is apparent that the normal model can't really handle the non-negativity constraint well, resulting in a left tail that reaches far into the negative values. The right tail fares somewhat better, though looking at the maximum values we can see that the data has a slightly longer right tail. This is also reflected in the skewness values. The behavior of the replications is also similar for male students so we omit the figures. Since the data distribution is quite skewed and concentrated near the lower bound of the data range, it seems necessary to use a model that naturally constrains the predictions to the unit interval range. One possible choice is Beta regression which can be used to model response variables that are constrained to the open unit interval (Ferrari and Cribari-Neto 2004). We consider this regression model after evaluating the Log-Log models. Since the results are so similar between the separate and the hierarchical model, we include subsequent figures for hierarchical models in Appendix 14.3.

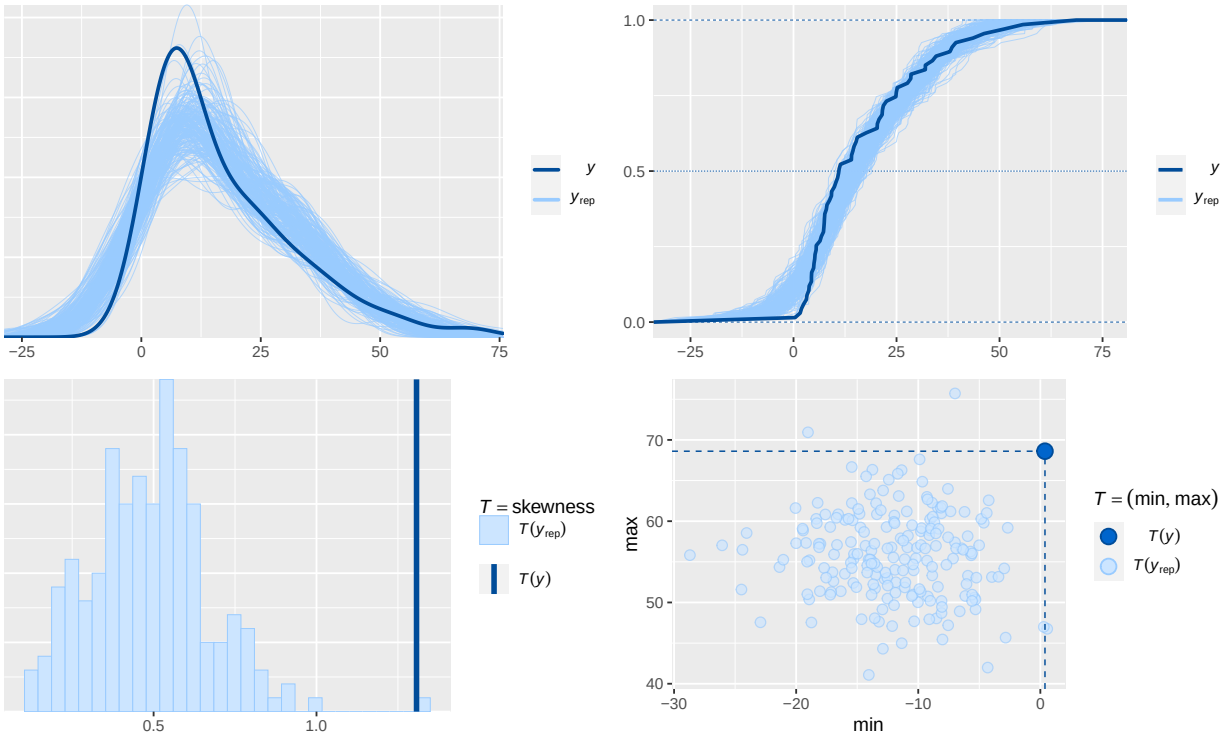

Figure 6: PPC diagnostic figures for the separate linear model

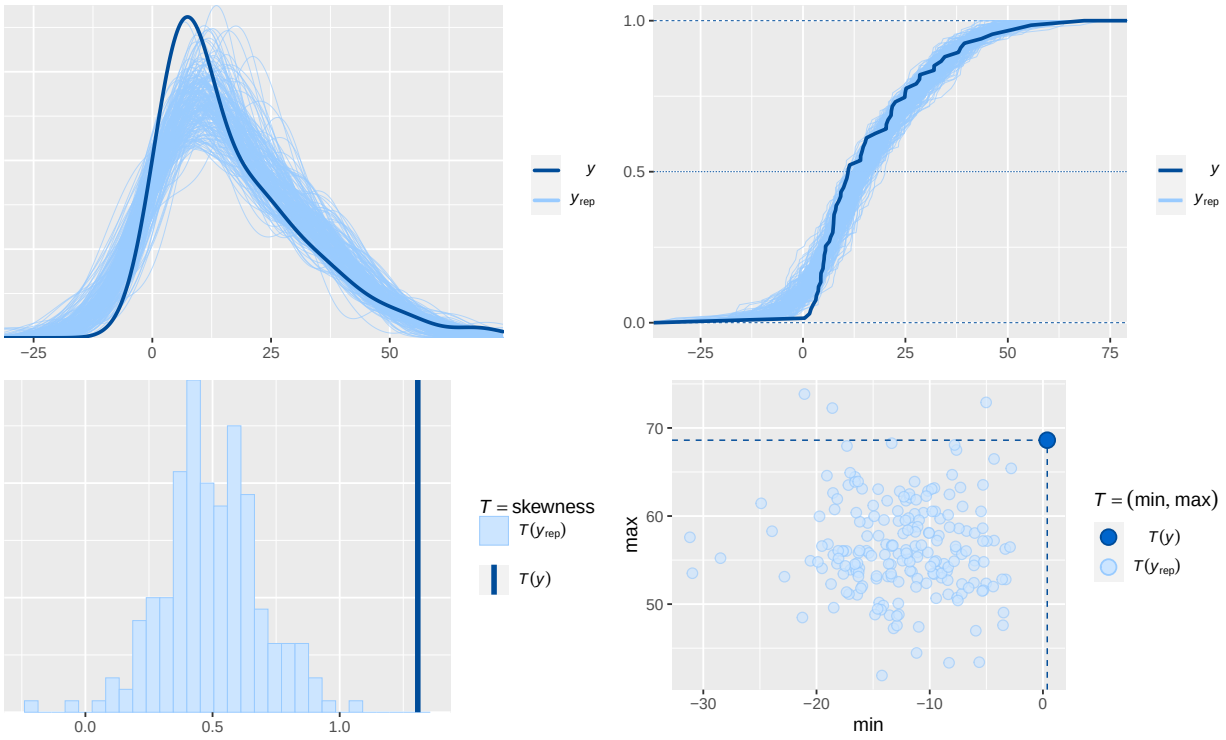

Figure 7: PPC diagnostic figures for the hierarchical linear model

7.2 Log-Log models

The Log-Log models do not seem to fare much better as can be seen in Figures 8 and 17. It seems that the issues have just become inverted with the right tail being the more challenging one to model. Additionally, the logarithmic data distribution is much wider in the center so it is unlikely that the normal model can work even on the logarithmic scale. This further motivates testing the aforementioned Beta regression model.

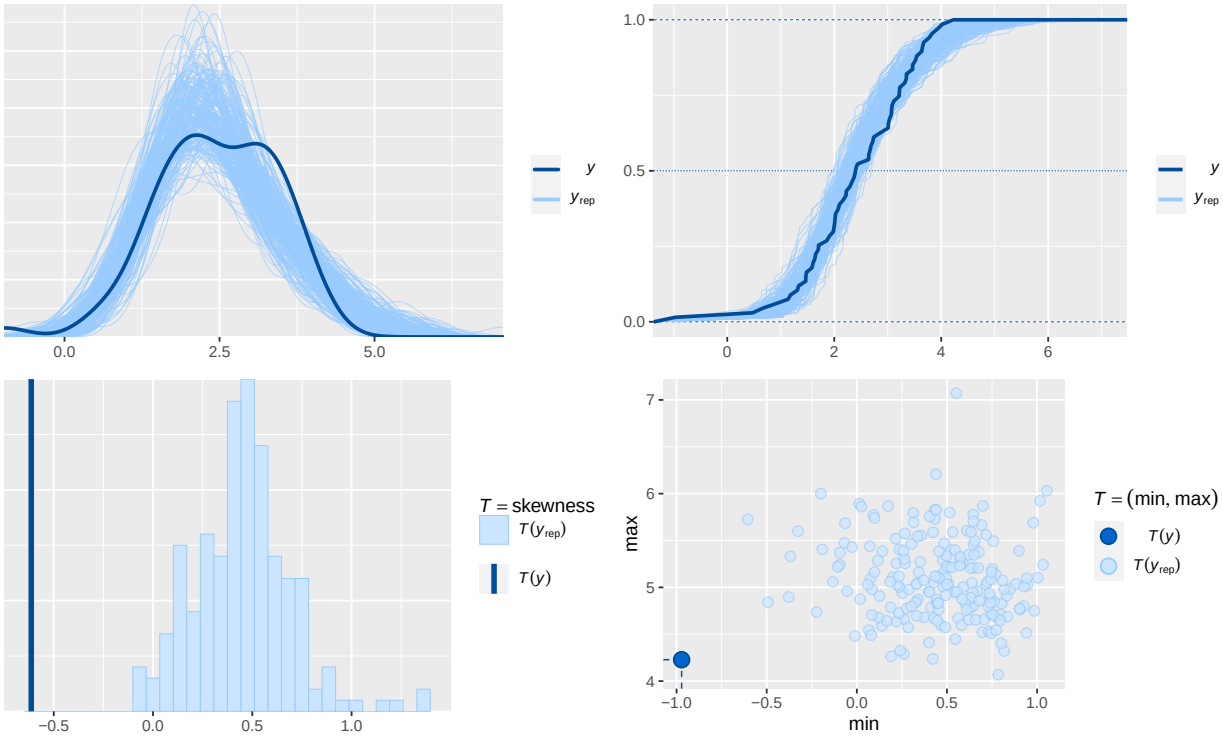

Figure 8: PPC diagnostic figures for the separate Log-Log model

7.3 Beta regression

As we can see in Figure 9, the Beta regression model performs substantially better than the previous models. It handles the open unit interval naturally and has no substantial systematic differences with the data set in terms of posterior predictions. We do see that there is some spread in the maximum values which not surprising since most of the observations are concentrated on the lower part of the unit range. The usage of t distributed priors with relatively low degrees of freedom might also contribute to the uncertainty. However, the performance seems excellent near the lower bound and there are no apparent signs of model misspecification.

8 Predictive diagnostics

In addition to the posterior predictive checking of Section 7, we can evaluate and compare the models using Expected Log Predictive Density (ELPD) (Vehtari, Gelman, and Gabry 2017). We note that this can only be compared easily for models that have the same response, so we can only compare separate models to their corresponding hierarchical extensions. However, since we are computing the Leave-One-Out (LOO) Cross-Validation (CV) estimates of ELPD using Pareto-Smoothed Importance Sampling (PSIS) (Vehtari, Gelman, and Gabry 2017), we can derive useful diagnostic quantities that complement the posterior predictive checks

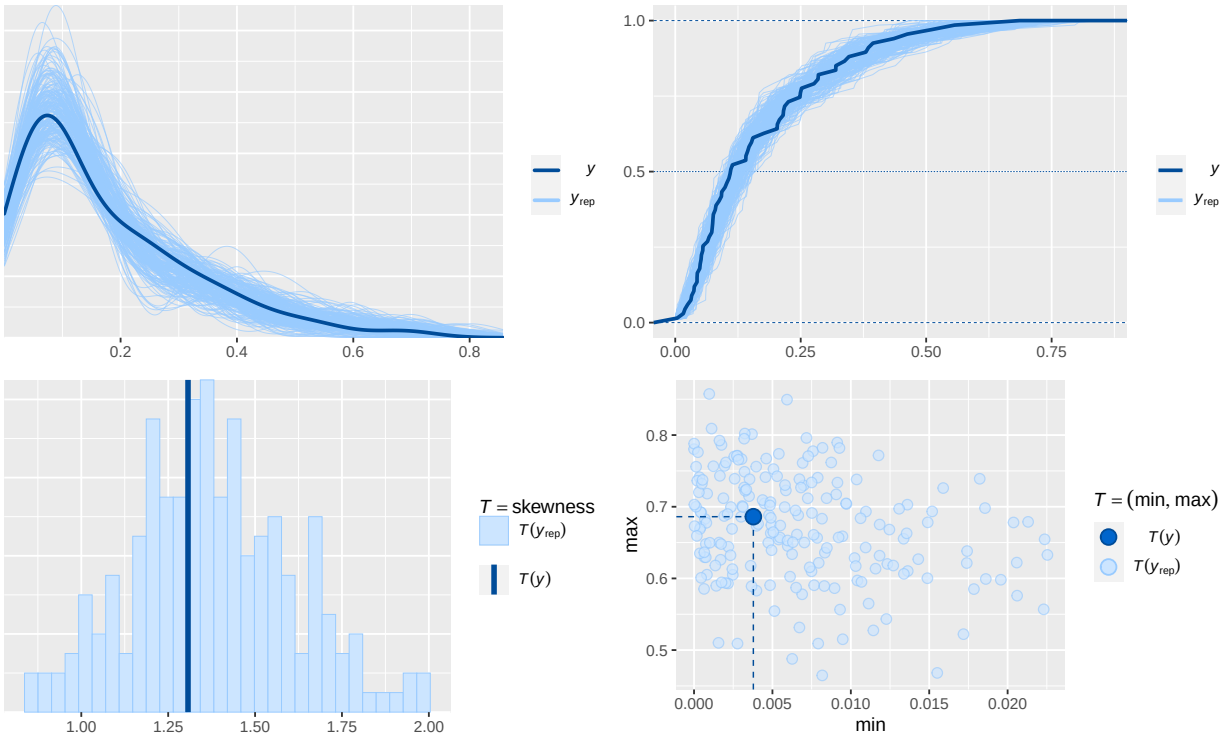

Figure 9: PPC diagnostic figures for the Beta regression model

of Section 7. Mainly, the effective number of parameters (p_{eff}) derived from the difference between the LOO ELPD estimate and its non-cross-validated counterpart and the Pareto-$\hat{k}$ estimates from PSIS. The former can be compared to the actual number of parameters to diagnose possible misspecifications (Vehtari 2022). The latter can be compared to thresholds established in (Vehtari, Gelman, and Gabry 2017) to diagnose misspecification and also to detect influential observations (Gabry et al. 2017).

The LOO ELPD estimates and the effective number of parameters are shown in Table 2 for each model, rounded according to the Monte Carlo Standard Error (MCSE) estimates provided by the method. The table also includes the standard error of the ELPD difference between the model pairs. When we initially computed the estimates, we encountered 2 or 3 observations for which the $\hat{k}$ estimates exceeded the threshold of 0.7 (and 1.0 for some) for all but the linear models, see Figure 10. The observations are given first for girls and then for boys. It seems that there might be a few highly influential observations since the maximum $\hat{k}$ values correspond to the same observation (China) across the three models. Looking at the observed proportions, we can see that China has the lowest proportions of male and female students below level 1 proficiency, below 1 percent each with no other country reaching that threshold. Thus it is not surprising that leaving this observation out can change the posterior substantially for the Log-Log model due to the logarithmic transformation, and for the Beta regression due to the natural boundary. However, the $\hat{k}$ values of these observations became substantially smaller after applying the moment matching algorithm (Paananen et al. 2021) and no issues with them were detected.

```
model_elpds_T <- lapply(models, rstan::loo, moment_match = T, save_psis = T)
loo_comp_lin <- loo_compare(model_elpds_T$Linear, model_elpds_T$`Linear (Hier)`)
loo_comp_log <- loo_compare(model_elpds_T$`Log-Log`, model_elpds_T$`Log-Log (Hier)`)
```

The first important observation from Table 2 is that the effective number of parameters for all non-hierarchical models are larger than the actual number of parameters in the models (16 to be specific). While for the linear and Log-Log models this is further evidence of misspecfication when combined with the posterior

Table 2: LOO-ELPD estimates and p_{eff} for each model using moment matching

Model	ELPD	SE of $\text{ELPD}_{\text{diff}}$	p_{eff}
Linear	-96	1	18
Linear (Hier)	-91	-	14
Log-Log	-147	5	30
Log-Log (Hier)	-136	-	18
Beta	176	-	23

predictive checks of Section 7, it is more difficult to interpret the significance for the Beta regression model since it is possible that even for a well-specified model p_{eff} exceeds the actual number of parameters (Vehtari, Gelman, and Gabry 2017). Since the posterior predictive checks were reasonable, it might be that the model is too flexible and the priors weak enough to allow the inference to change substantially when computing the LOO-ELPD estimates, especially for challenging observations. This is illustrated in Figure 11 where the LOO predictive distributions are plotted for each observation. The point is the median, the inner range covers 50 % and the outer range 90 % of the probability mass. The discrepancy between the LOO predictive distributions and the observed values for China are notable, though it is not the only country that exhibits this kind of behavior. Secondly, we observe that the hierarchical models perform better than their non-hierarchical counterparts in terms of ELPD, and the differences are above the rule of thumb value of 4 when taking the standard errors into account (Sivula, Magnusson, and Vehtari 2020). However, we note that the misspecification detected during the posterior predictive checks can cause the standard error to be unreliable for estimating the uncertainty in the difference, especially since the PPC results were almost identical for both types of models. Thus we suspect that there are no substantial differences between the predictive performances of the separate and hierarchical models.

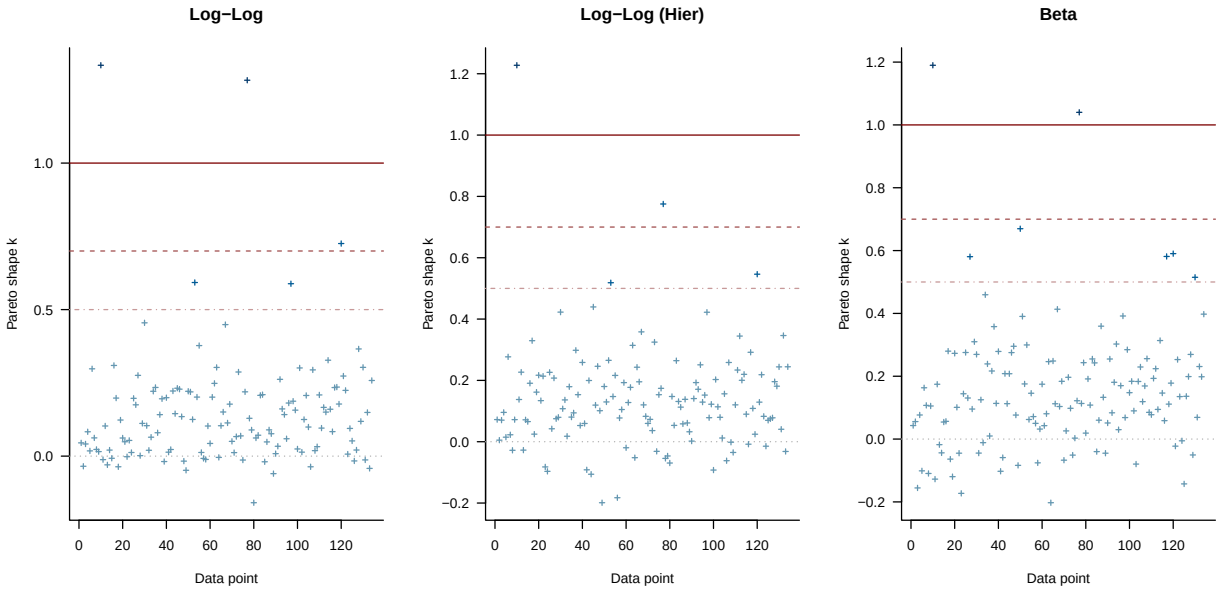

Figure 10: Pareto-$\hat{k}$ diagnostics plots without moment matching

9 Sensitivity analysis

While the prior descriptions of Section 4 are relatively weakly informative, there is a risk that some of the prior distributions used are not robust enough to handle possible outliers and influential observations that

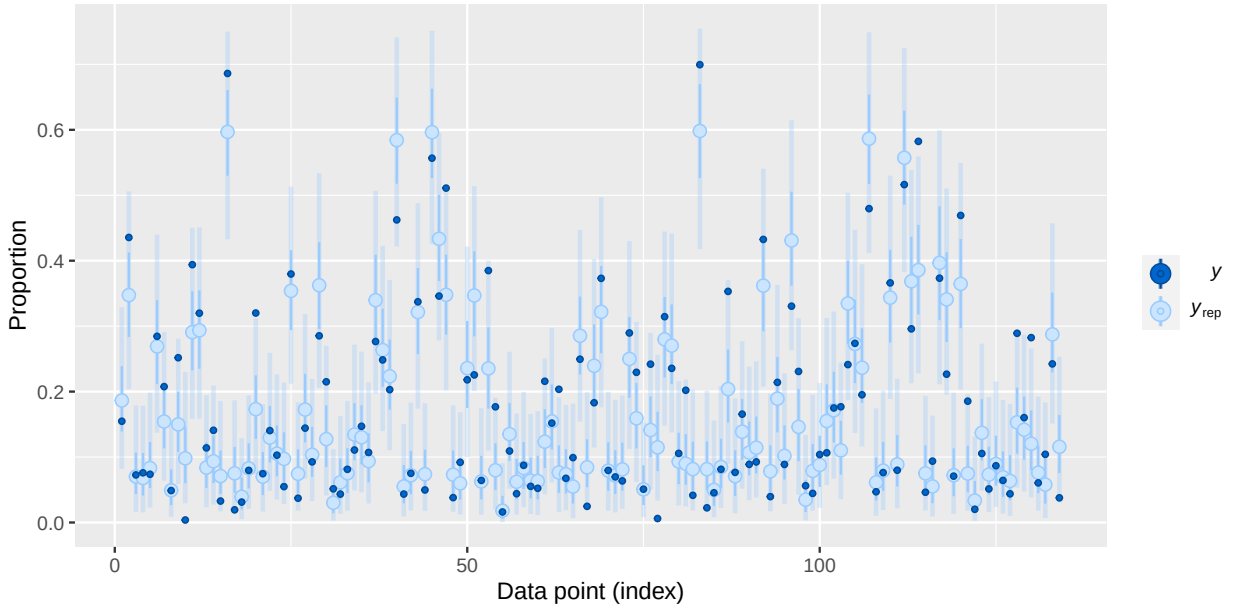

Figure 11: LOO predictive check for the Beta regression model

Table 3: Changes made to the model structures

Model	New parameters	Priors	Likelihood
Linear	$\nu_i \sim \text{Gamma}(2, 0.1)$	$\text{T}(3, \ldots)$	$\text{T}(\nu_i, \ldots)$
Linear (Hier)	$\nu \sim \text{Gamma}(2, 0.1)$	$\text{T}(3, \ldots)$	$\text{T}(\nu, \ldots)$
Beta	-	$\nu_j = \nu_\alpha = 2, \nu_\phi = 1$	-

were identified in Section 8. The same also applies to the data generating distribution. There is also a risk of misscalibration when setting the prior parameter values. To check how large an effect these choices have on the inference, we change all the normal distributions of the linear models to t distributions with low degrees of freedom to better cover possible misspecification issues while using the same scale and location parameters as before. For the Beta regression model, we reduce the degrees of freedom of each t distribution prior. Since our primary quantities of interest are the regression parameters, we investigate visually their marginal posterior distributions before and after changing the prior distributions. An overview of the changes are shown in Table 3. The choice of prior for the degrees of freedom parameters as well as the lower bound is based on (Stan Dev Team 2020).

Figures 12 and 13 contain the comparison for the separate and hierarchical linear models, respectively. They show the median, 80 % and 100 % central probability intervals. There aredifferences in the locations of the posterior distributions, mainly that for the robust model, the marginal posteriors tend to be closer to zero than those for the original model. One large difference between the original and the robust models can be seen in Figure 13 where the hierarchical scale parameter has a much longer right tail when using the robust model. This indicates that the data does not provide much useful information on the group differences and thus does not exclude these large values. Since these results are quite similar across the different models, the rest of the sensitivity analysis plots are given in Appendix 14.4. For the Beta regression, the differences are negligible since the priors were already quite robust as can be seen in Figure 19. Given this and the results in Section 7, we would select the Beta regression model as the best model to use for inference, with some caveats discussed in Section 8 with regards to outlying observations.

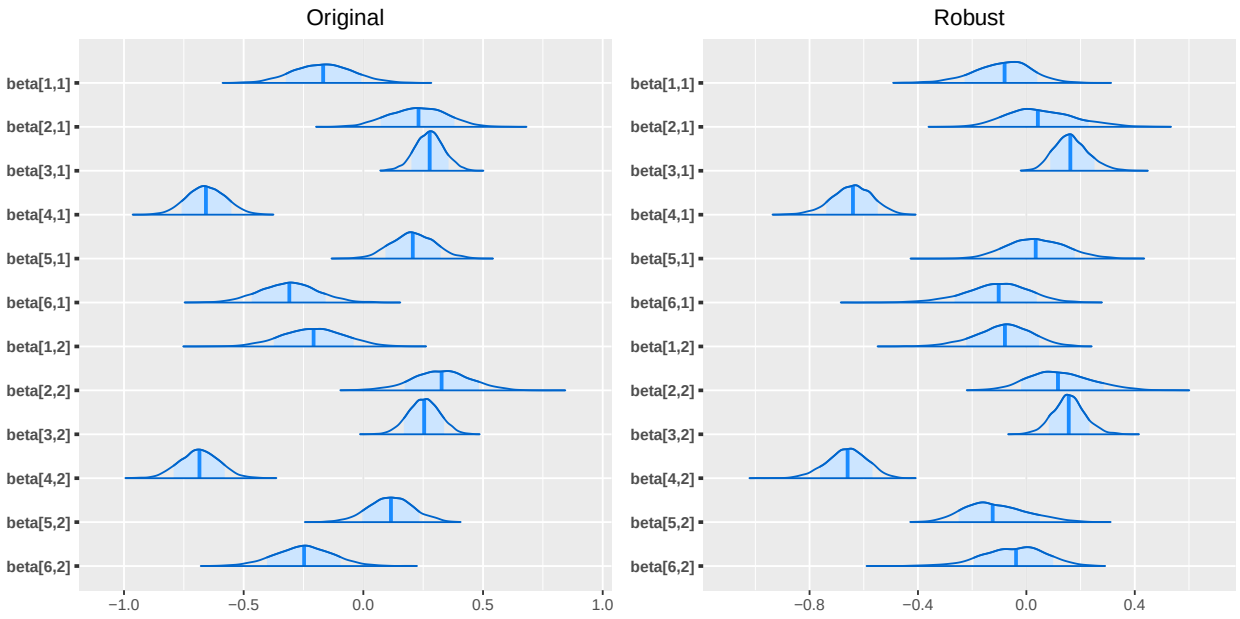

Figure 12: Sensitivity analysis for the linear model

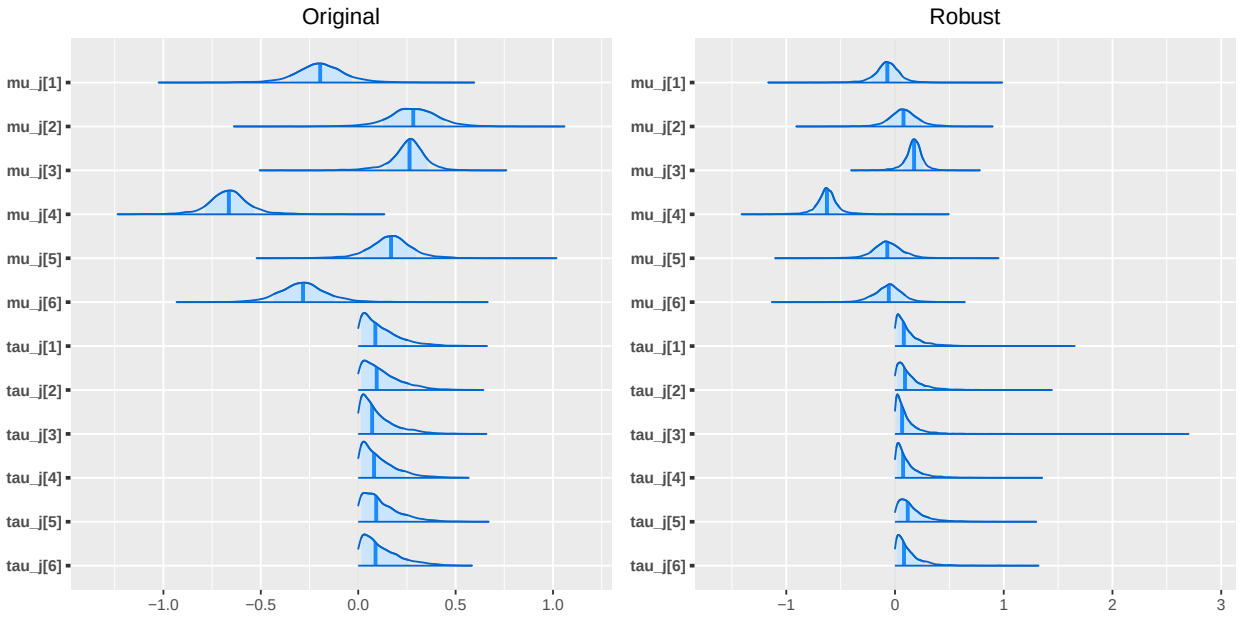

Figure 13: Sensitivity analysis for the hierarchical linear model

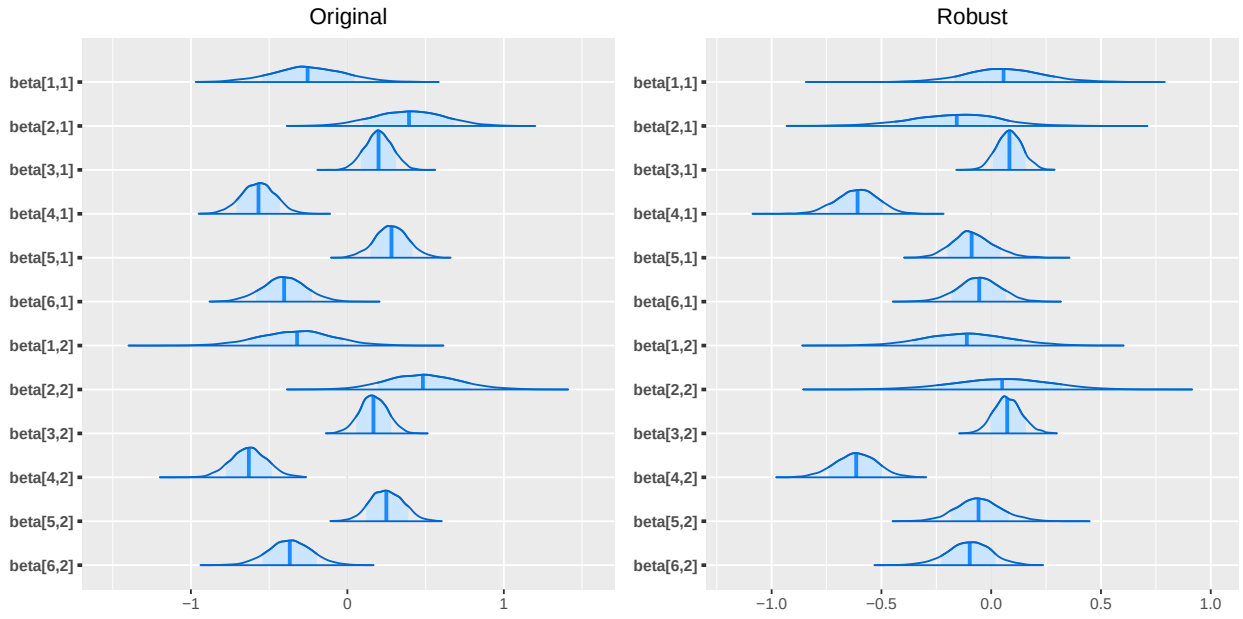

Figure 14: Sensitivity analysis for the Log-Log model

10 Potential improvements

As was noted in the previous section, having just two groups in a hierarchical model makes inferring the group differences difficult. One possible way to solve this is to use results from previous PISA test years and possibly use a two-way grouping with respect to sex and year. However, this analysis is complicated by two factors. Firstly, the countries that partake in the PISA studies varies by year. Secondly, the more years we include in the analysis, the larger the proportion of countries that have missing data for some of the covariates. Missing data itself is not a problem for Bayesian inference since it can be modeled as an unknown parameter just like any other parameter of the model (Stan Dev Team 2022b). Depending on the proportion of missing observations, this could make it possible to include useful covariates, thereby improving precision, or it could substantially increase the uncertainty and make the model sampling difficult to perform. Careful analysis would be required in order to determine what covariates with missing values should be modeled.

While adding the additional years might make the hierarchical model more informative, it might also introduce autocorrelative structure between the subsequent observations. Thus a possible alternative modeling approach would be to add a time series structure to the model, possibly something within the ARMA family of stochastic processes for an initial model. Then we could again incorporate data across multiple years to improve the precision of our estimates, though we should note that the PISA results time series are quite short so inferring the ARMA components with reasonable precision could be difficult.

11 Conclusion

Our main goal was to infer the effect of social and economical covariates on the proportion of girl/boys who are below the PISA proficiency level 1. From the five models considered, the Beta regression model seems to be the best calibrated one according to the posterior predictive checks. The other models have clear indicators of misspecfication and are unlikely to be reliable. While there are observations, especially near the boundaries, that have very high influence on the inference, the model performs reasonably well for many other observations. Figure 15 illustrates the marginal posterior distributions of the covariate parameters. We can see that there are no substantial differences between the groups in terms of the distributions. All

covariates seem to have some sort of effect on the response, though the covariate LEARN has a somewhat larger negative effect size which is not that surprising. Other things that behave as one might expect are the negative relationship between the proportion and HEALTH and GOVERN. A positive relationship with respect to FERT is also expected. The remaining relationships are not as straightforward, especially the positive relationship associated with VOICE. We suspect that there is some kind of interaction between GOVERN and VOICE that leads to the negative relationship with the response. As a whole, the model estimates seem quite reasonable and behave as one would believe taking into account prior knowledge not incorporated into the model.

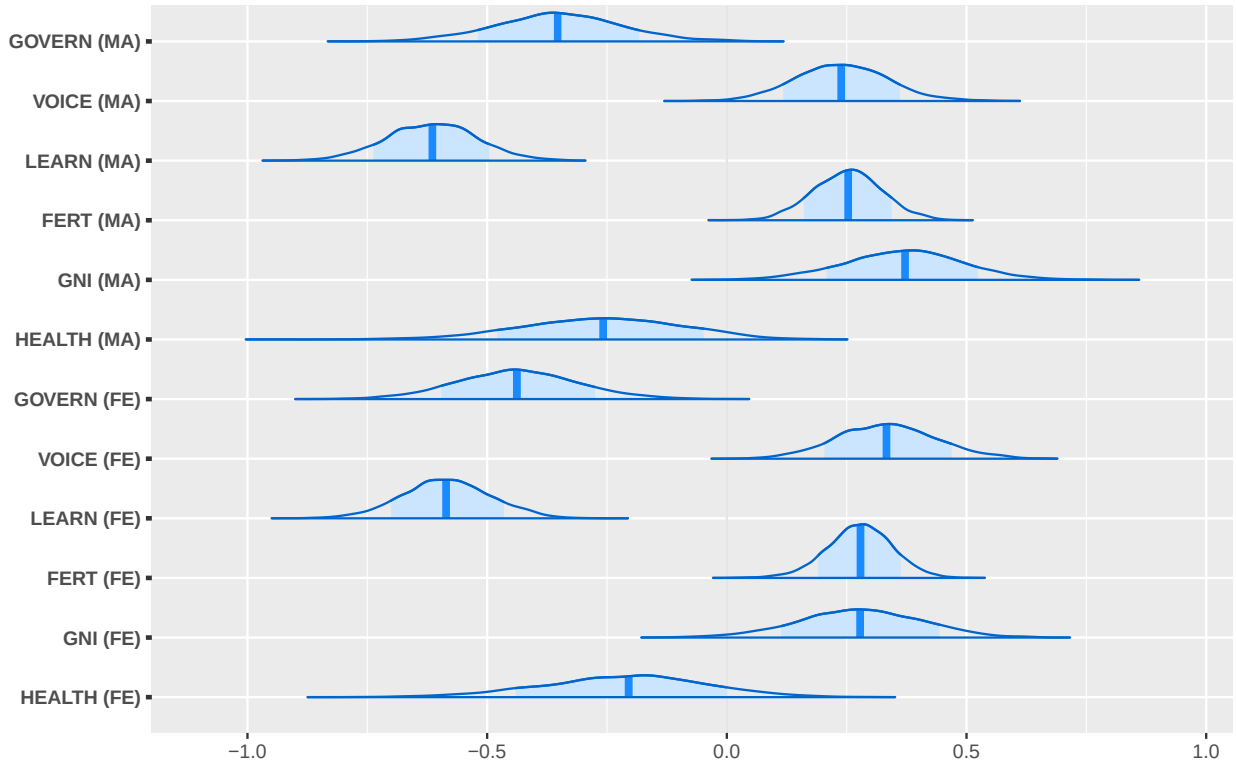

Figure 15: Marginal posterior distributions for the covariate parameters in Beta regression

12 Self-reflection

When considering the priors, it became quite apparent just how much researcher degrees of freedom can effect the results of the analysis, even in an observational study just as this one. However, unlike traditional frequentist inference methods that often involve assumptions that tend to not be mentioned due to them being "standard practices", I feel that the way Bayesian inference was presented in the course and applied in the project emphasized being more explicit about the beliefs and how to set them up reasonably, especially when using Stan where we don't need to restrict the analysis to conjugate models. However, there definitely exist things such as model parametrization and independence that can go unnoticed and cause issues down the line. The diagnostic tools presented in the course and relevant papers do seem quite comprehensive, and so maybe for this project the two model minimum requirement was actually quite close to being optimal given the amount of checks and diagnostics one can perform for a single model.

13 Bibliography

Betancourt, Michael. 2017. "Diagnosing Biased Inference with Divergences." https://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html.

Ferrari, Silvia, and Francisco Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7): 799–815.

Gabry, Jonah, and Martin Modrak. 2022. "Visual MCMC Diagnostics Using the Bayesplot Package." https://mc-stan.org/bayesplot/articles/visual-mcmc-diagnostics.html.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2017. "Visualization in Bayesian Workflow." *arXiv Preprint arXiv:1709.01449*.

Hoffman, Matthew D, Andrew Gelman, et al. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–623.

OECD. 2018. "Programme for International Student Assessment (PISA) PISA 2018 Technical Report." OECD Publishing.

Paananen, Topi, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. 2021. "Implicitly Adaptive Importance Sampling." *Statistics and Computing* 31 (2): 1–19.

Sivula, Tuomas, Måns Magnusson, and Aki Vehtari. 2020. "Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison." *arXiv Preprint arXiv:2008.10296*.

Stan Dev Team. 2020. "Prior Choice Recommendations." https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

———. 2022a. "Runtime Warnings and Convergence Problems." https://mc-stan.org/misc/warnings.html.

———. 2022b. *Stan User's Guide*. https://mc-stan.org/docs/stan-users-guide/index.html.

Vehtari, Aki. 2022. "Cross-Validation FAQ." https://mc-stan.org/loo/articles/online-only/faq.html.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27 (5): 1413–32.

World Bank. 2022a. "Education Statistics - All Indicators." https://databank.worldbank.org/source/education-statistics-%5E-all-indicators.

———. 2022b. "Gender Statistics." https://databank.worldbank.org/source/gender-statistics.

———. 2022c. "World Development Indicators." https://databank.worldbank.org/source/world-development-indicators.

———. 2022d. "Worldwide Governance Indicators." https://databank.worldbank.org/source/worldwide-governance-indicators.

14 Appendix

14.1 Variable scales

Table 4 provides the sample standard deviations for each variable which were used to transform the original scale priors to the standardized scale.

Table 4: Sample standard deviations of the variables

PISA.FE	PISA.MA	HEALTH	GNI	FERT	LEARN.FE	LEARN.MA	VOICE	GOVERN
15	15	2200	21000	0.4	1.5	1.6	27	19

14.2 HMC/NUTS sampler diagnostics

Table 5 contains various diagnostics given by the Stan NUTS sampler. The numbers indicate how many iterations diverged, saturated the default maximum tree depth and how many chains had a low E-BFMI value (Stan Dev Team 2022a). As we can see, there are no apparent issues with any of the models with regards to sampling.

Table 5: Various HMC/NUTS diagnostic measures for each fitted model

Model	Divergence	Maximum tree depth	E-BFMI
Linear	0	0	0
Linear (Hier)	0	0	0
Log-Log	0	0	0
Log-Log (Hier)	0	0	0
Beta	0	0	0

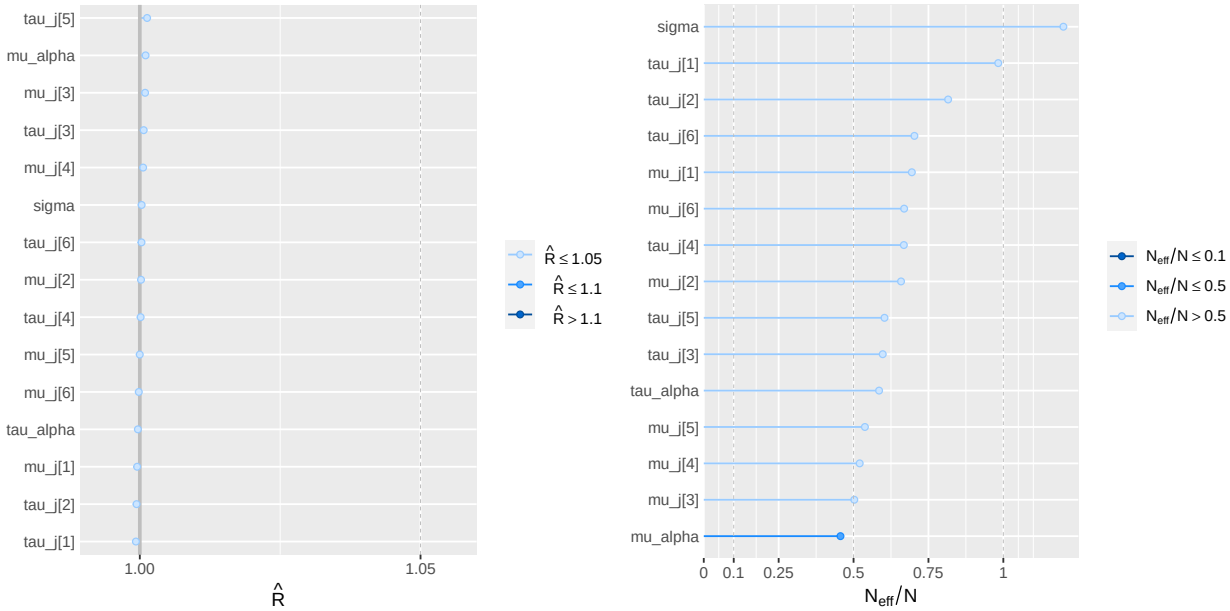

Figure 16: Convergence diagnostics for the hierarchical Log-Log model

14.3 PPC diagnostic plots

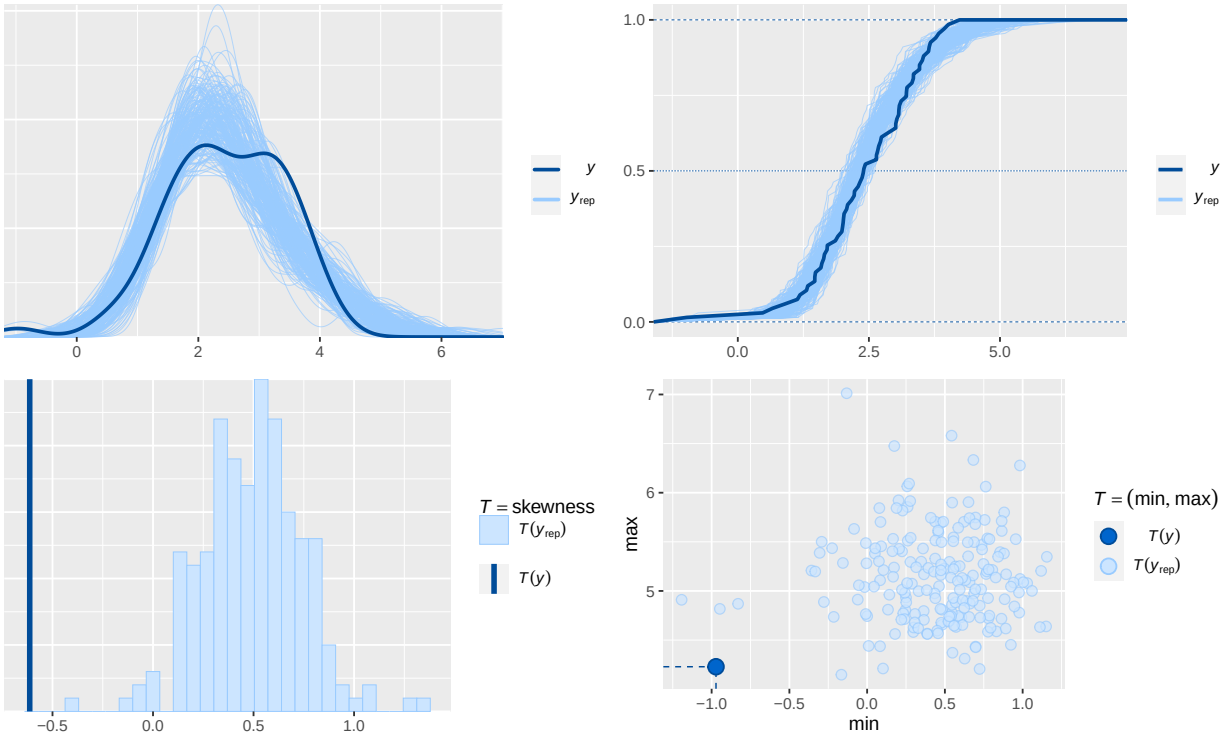

Figure 17: PPC diagnostic figures for the hierarchical Log-Log model

14.4 Sensitivity analysis figures

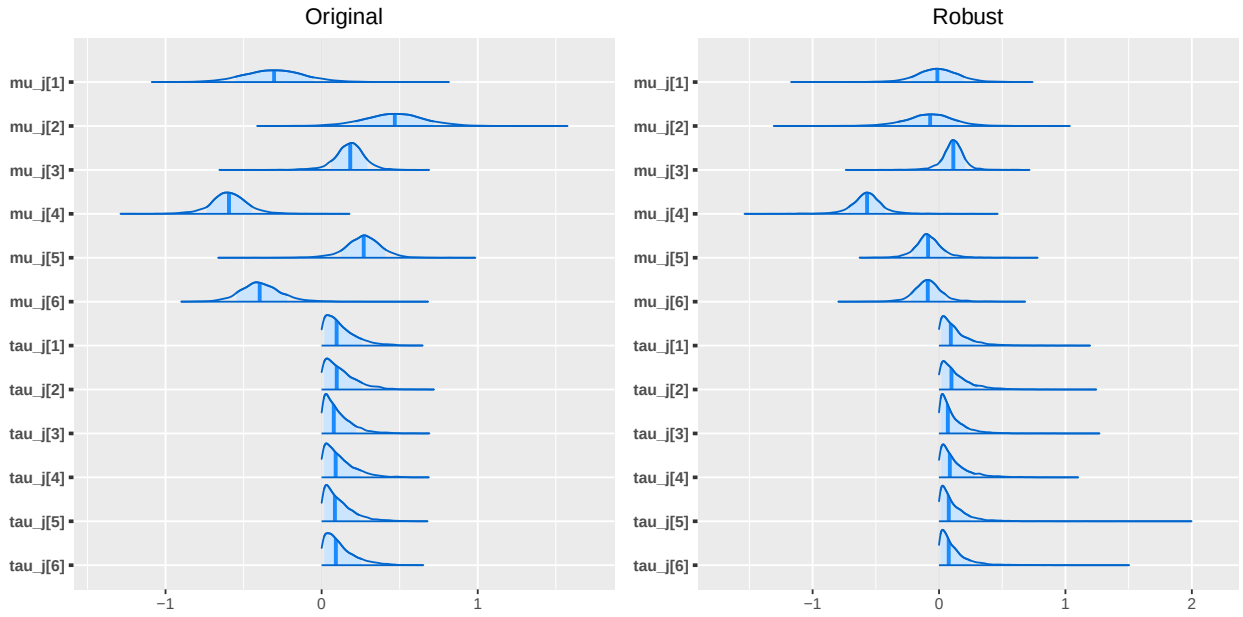

Figure 18: Sensitivity analysis for the hierarchical Log-Log model

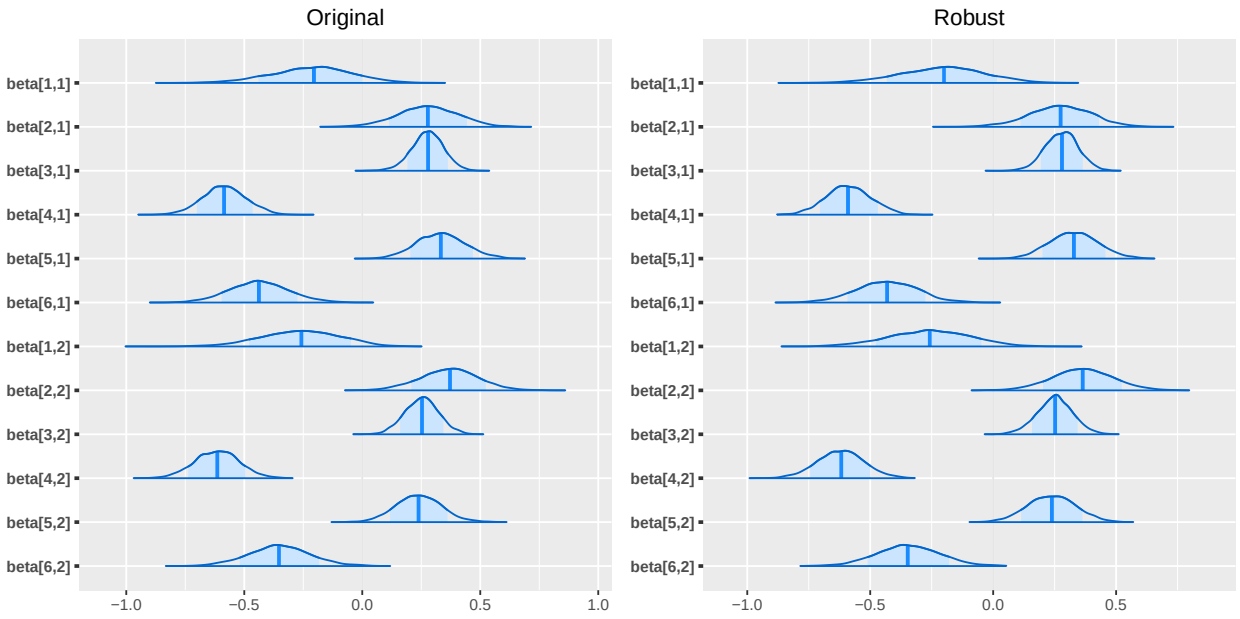

Figure 19: Sensitivity analysis for the Beta regression model, covariate parameters

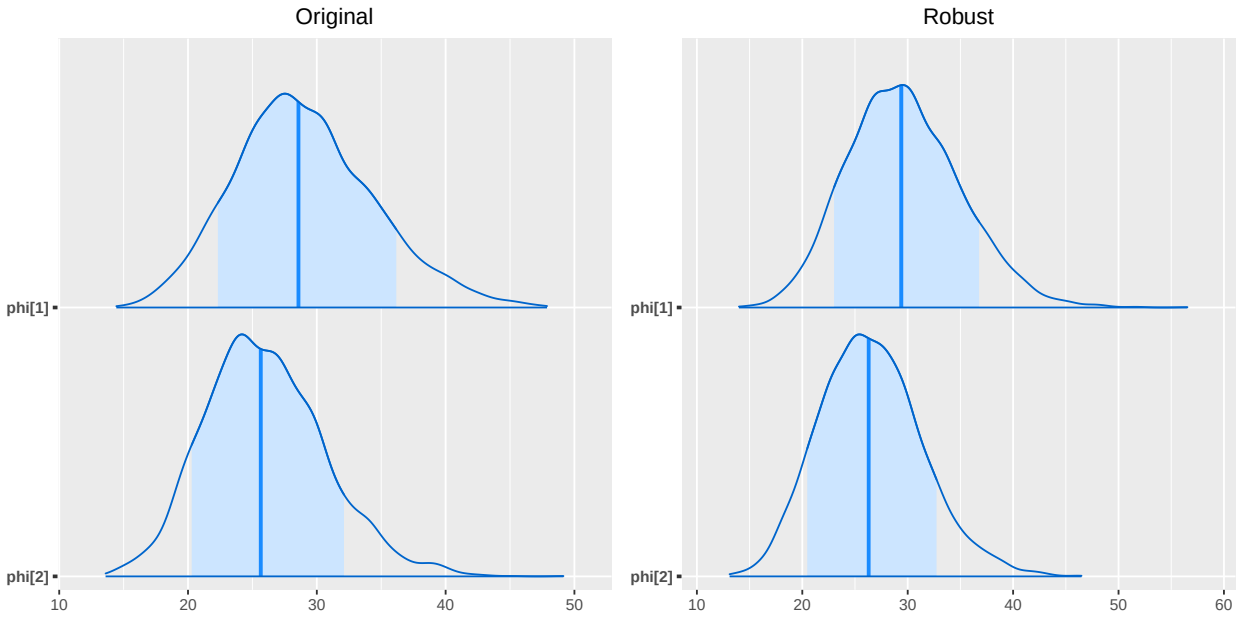

Figure 20: Sensitivity analysis for the Beta regression model, precision parameter

14.5 Stan model source codes

14.5.1 Linear model

```
data {
  int<lower=0> N; // Number of observations
  int<lower=0> P; // Number of expl. variables
  matrix[N,2] y_std; // Response variable (PISA for boys and girls) (standardized)
  real X_std[N,P,2]; // Data matrix for both groups (standardized)

  real y_mean[2]; // Mean of the response variables
  real y_sd[2]; // Standard deviation of response variables

  real alpha_mu_prior; // Mean of the prior dist. of alpha
  real<lower=0> alpha_sd_prior; // Scale of the prior dist. of alpha

  vector[P] beta_mu_prior; // Mean vector of the prior dist. of beta
  matrix[P,P] beta_sigma_prior; // Cov. matrix of the prior dist. of beta

  real<lower=0> sigma_sd_prior; // SCale of the prior dist. of sigma
}

parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  vector<lower=0>[2] sigma; // Measurement SD per group
}

transformed parameters {
  matrix[N, 2] mu_std; // Linear model for the mean
  for (i in 1:2) {
    mu_std[,i] = alpha[i] + to_matrix(X_std[,,i]) * beta[,i];
  }
}

model {
  // Priors
  for (i in 1:2) {
    beta[,i] ~ multi_normal(beta_mu_prior, beta_sigma_prior);
    alpha[i] ~ normal(alpha_mu_prior, alpha_sd_prior);
    sigma[i] ~ normal(0, sigma_sd_prior);
  }
  // Likelihood
  for (i in 1:2) {
    y_std[,i] ~ normal(mu_std[, i], sigma[i]);
  }
}

generated quantities {
  matrix[N, 2] log_lik;
  matrix[N, 2] y_rep;

  for (i in 1:2) {
    for (j in 1:N) {
```

```
      log_lik[j,i] = normal_lpdf(y_std[j,i] | mu_std[j,i], sigma[i]);
      y_rep[j,i] = normal_rng(mu_std[j,i] * y_sd[i] + y_mean[i], sigma[i] * y_sd[i]);
    }
  }
}
```

14.5.2 Hierarchical model

```
data {
  int<lower=0> N; // Number of observations
  int<lower=0> P; // Number of expl. variables
  matrix[N,2] y_std; // Response variable (PISA for boys and girls) (standardized)
  real X_std[N,P,2]; // Data matrix for both groups (standardized)

  real y_mean[2]; // Mean of the response variables
  real y_sd[2]; // Standard deviation of response variables

  real alpha_mu_prior; // Prior mean for the intercept pop. mean
  real<lower=0> alpha_sd_prior; // Prior scale for the intercept pop. mean
  real<lower=0> alpha_tau_scale; // Prior scale for the intercept pop. scale

  vector[P] beta_mu_prior; // Prior means for the covariate param. pop. means
  vector<lower=0>[P] beta_sigma_prior; // Prior scales for the covariate param. pop. means
  real<lower=0> beta_tau_scale; // Prior scale for the  covariate param. pop. scales

  real<lower=0> sigma_sd_prior;
}

parameters {
  real mu_j[P]; // Covariate param. population means
  real<lower=0> tau_j[P]; // Covariate param. population scales
  real mu_alpha; // Intercept population mean
  real<lower=0> tau_alpha; // Intercept population scale
  real<lower=0> sigma; // Shared measurement scale

  real z_hat[P,2]; // Latent variables for the covariate parameters
  real z[2]; // Latent variables for the intercepts
}

transformed parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  matrix[N, 2] mu_std; // Linear model for the mean

  // Generation of the group-wise parameters from the latent variables
  for (i in 1:2) {
    alpha[i] = tau_alpha * z[i] + mu_alpha;
    for (j in 1:P) {
      beta[j,i] = tau_j[j] * z_hat[j,i] + mu_j[j];
    }
  }
```

```
  for (i in 1:2) {
    mu_std[,i] = alpha[i] + to_matrix(X_std[,,i]) * beta[,i];
  }
}

model {
  // Hyperpriors
  mu_alpha ~ normal(alpha_mu_prior, alpha_sd_prior);
  tau_alpha ~ normal(0, alpha_tau_scale);

  for (j in 1:P) {
    mu_j[j] ~ normal(beta_mu_prior[j], beta_sigma_prior[j]);
    tau_j[j] ~ normal(0, beta_tau_scale);
  }

  // Prior
  sigma ~ normal(0, sigma_sd_prior);

  // Latent variables
  for (i in 1:2) {
    z[i] ~ normal(0, 1);
    for (j in 1:P) {
      z_hat[j,i] ~ normal(0, 1);
    }
  }

  // Likelihood
  for (i in 1:2) {
    y_std[,i] ~ normal(mu_std[, i], sigma);
  }
}

generated quantities {
  matrix[N, 2] log_lik;
  matrix[N, 2] y_rep;

  for (i in 1:2) {
    for (j in 1:N) {
      log_lik[j,i] = normal_lpdf(y_std[j,i] | mu_std[j,i], sigma);
      y_rep[j,i] = normal_rng(mu_std[j,i] * y_sd[i] + y_mean[i], sigma * y_sd[i]);
    }
  }
}
```

14.5.3 Beta regression model

```
data {
  int<lower=0> N; // Number of observations
  int<lower=0> P; // Number of expl. variables
  matrix<lower = 0, upper = 1>[N, 2] y; // Response variable (PISA for boys and girls)
  real X_std[N,P,2]; // Data matrix for both groups (standardized)
```

```
  real loc_params_prior; // Location of the prior dist. of covariate params. and intercept
  real<lower=0> nu_params_prior; // dof of the prior dist. of covariate params. and intercept
  real<lower=0> scale_params_prior; // Scale of the prior dist. of covariate params. and intercept

  real<lower=0> nu_precision_prior; // dof of the prior dist. of precision
  real<lower=0> scale_precision_prior; // Scale of the prior dist. of precision
}

parameters {
  matrix[P,2] beta; // Parameters for the expl. variables per group
  vector[2] alpha; // Intercept per group
  vector<lower=0>[2] phi; // Precision per group
}

transformed parameters {
  matrix[N, 2] mu; // Expected values of the Beta variables
  for (i in 1:2) {
    mu[,i] = inv_logit(alpha[i] + to_matrix(X_std[,,i]) * beta[,i]);
  }
}

model {
  // Priors
  for (i in 1:2) {
    alpha[i] ~ student_t(nu_params_prior, loc_params_prior, scale_params_prior);
    phi[i] ~ student_t(nu_precision_prior, 0, scale_precision_prior);
    for (j in 1:P) {
      beta[j,i] ~ student_t(nu_params_prior, loc_params_prior, scale_params_prior);
    }
  }
  // Likelihood
  for (i in 1:2) {
    y[,i] ~ beta(mu[,i] * phi[i], (1 - mu[,i]) * phi[i]);
  }
}

generated quantities {
  matrix[N, 2] log_lik;
  matrix[N, 2] y_rep;

  for (i in 1:2) {
    for (j in 1:N) {
      log_lik[j,i] = beta_lpdf(y[j,i] | mu[j,i] * phi[i], (1 - mu[j,i]) * phi[i]);
      y_rep[j,i] = beta_rng(mu[j,i] * phi[i], (1 - mu[j,i]) * phi[i]);
    }
  }
}
```