BDA Project

2023-11-18

Introduction

The birth rates in Finland have been decreasing remarkably since 2010. The number of children born in 2010 compared with the number of births in 2022 is 33% smaller. The fertility rate in Finland in 2022 was the lowest ever measured. The decreasing trend has been continuous and thus, the decreased fertility rate cannot be explained by random variation. (Tilastokeskus, 2023) The decreasing trend of the birth rate can be seen in the following figure:

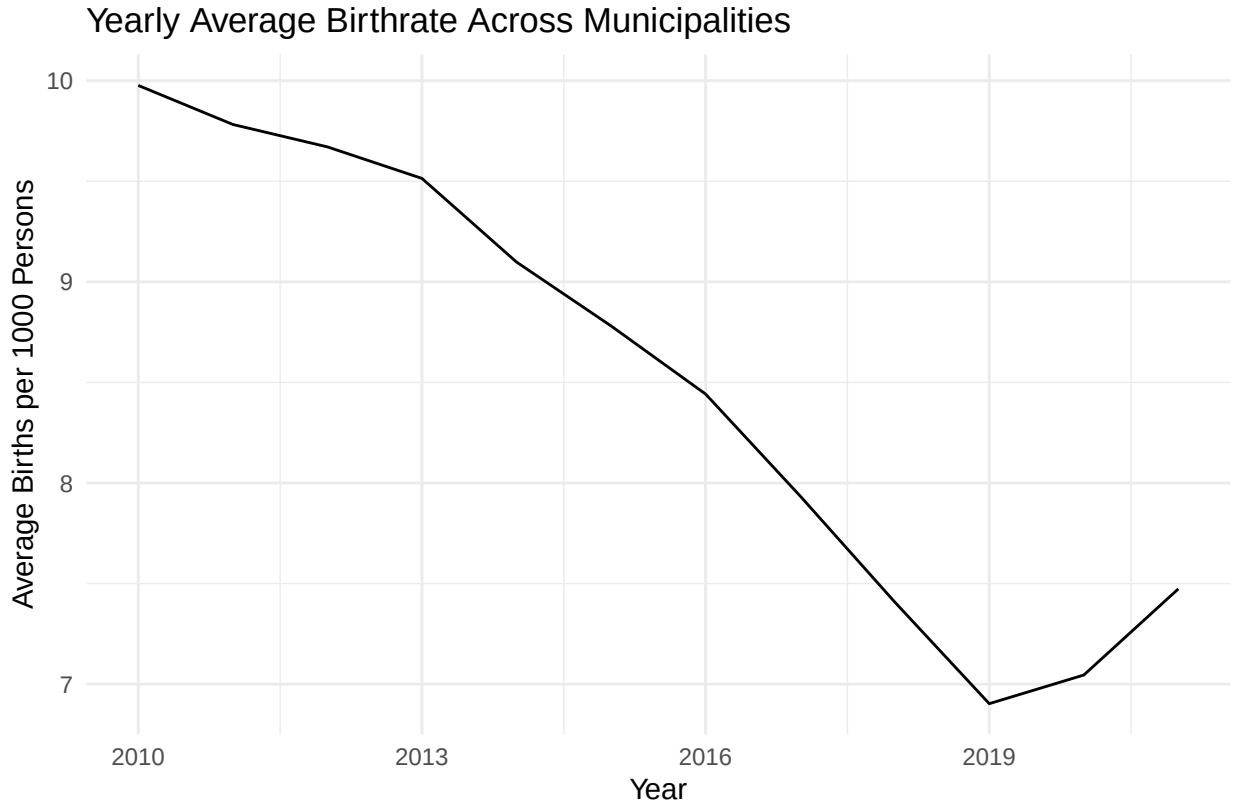

Figure 1. Yearly Average Birthrate Across Municipalities

The number of births has a lagged effect on the proportion of the population of the working age. The proportion of the population of working age has thus an effect on the financing of the well-being society of the state.

The goal of this project is to study and quantify the effects of different variables on the fertility rate in Finland to understand the cause-and-effect relationships behind the phenomenon. The model selected to study this question further is Bayesian linear regression. We will compare two linear models, a pooled model and a hierarchical model with a municipality specific random effect.

Description of the Models

As stated in the introduction, we are using two different regression models: Pooled Bayesian Linear Regression, and Hierarchical Bayesian Linear Regression. Both models use Bayesian inference incorporating the prior knowledge about the influence of the explanatory variables on the dependent variable to the model.

The Pooled Bayesian Linear regression is treating all the data as part of a single, large group. It assumes that a common set of regression coefficients can be estimated to predict the values of the dependent variable. Thus, it assumes that the variation is similar in every group. The pooled model is chosen to examine if the group-level variation is present in the birth rates of Finnish municipalities by comparing its performance with a hierarchical model.

The Hierarchical Bayesian Linear Regression, in turn, acknowledges the existence of groups with different levels of variation in the data set. Its key characteristics is that it includes both regression coefficients that are group-specific, and coefficients that are common across the whole data set. To set the context to our research question of the birth rates of different municipalities in Finland, we are using common regression coefficients for all the independent variables, but a group-level intercepts for every municipality.

As our data is time series, the regressors are chosen in a manner that takes the past values into account. In addition to the variables in the data set, we chose to add the lagged version of the dependent variable as an explanatory variable. Moreover, the birth rate in year t is predicted with the explanatory variable values in year t-1 to account to the fact that the decision to have a child is made typically 9 months before the birth, at latest.

In the following paragraphs the mathematical form of the models is presented. In the formulas, Y is representing the birth rate, which is the dependent variable of the model.

Pooled model

$$Y_t = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + \beta_4 X_{4,t-1} + \beta_5 Y_{t-1} + \epsilon_t$$

where, x_1 is the proportion of the population under 15 years of age, x_2 is the proportion of the population with higher education, x_3 is the proportion of the population living in rental apartments, x_4 is the price level adjusted income level, and Y_t and Y_{t-1} are the birth rates this year and last year.

The prior distributions for the parameters $\beta_1, \ldots, \beta_5$ are as follows:

$$\beta_i \sim N(0, 50^2)$$

Hierarchical model

$$Y_t = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + \beta_4 X_{4,t-1} + \beta_5 Y_{t-1} + \epsilon_t + \beta_{municipality} + \epsilon_i$$

The prior distributions for the parameters $\beta_1, \ldots, \beta_5$ are similar as in the pooled model. The municipality-wise intercept term has the following priors:

$$\begin{aligned} \beta_{municipality} &\sim N(0, \tau) \\ \tau &\sim \text{exponential}(0.02) \end{aligned}$$

Description of the data

The data for this study was obtained from Tilastokeskus, which is the national statistics institution of Finland. The dataset used in the analysis was created by merging data from 4 different publicly available datasets about Finnish municipalities. The final dataset consisted of 30 variables and 2 748 rows. The following statistics were collected:

Variables

- **Vuosi:** Represents the year of the measurements
- **Syntyneet:** The amount of births/1000 habitants in the current year
- **Alle_15_vuotiaita:** Percentage of citizens under the age of 15
- **Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus:** Percentage of citizens that hold a degree of higher education
- **Vuokra_Asunnoissa_Asuva_Asuntokuntia:** Proportion of households living on rent -**Tulotaso** Inflation adjusted indexed income level

```
data_print <- data[, c("Vuosi", "Kunta", "Syntyneet", "Alle_15_vuotiaita", "Korkea_Asteen_Tutkinnon_Suo
names(data_print) <- c("Vuosi", "Kunta", "Synt", "Alle15", "KorkeaAste", "Tyottomat", "VuokraAs", "Tulo


print(head(data_print))
```

```
##   Vuosi       Kunta      Synt Alle15 KorkeaAste Tyottomat VuokraAs Tulotaso
## 1  2010       Espoo 14.433198   19.5       43.9       5.8     33.9 25.64021
## 2  2010    Helsinki 11.399221   13.5       37.5       7.8     47.1 19.28784
## 3  2010    Karkkila 10.641763   17.1       18.6       8.9     23.8 19.20103
## 4  2010  Kauniainen  5.984578   18.6       56.0       4.1     29.9 36.17704
## 5  2010      Kerava 12.397176   17.3       30.1       6.6     32.2 21.42090
## 6  2010 Kirkkonummi 14.482161   23.0       38.1       6.2     21.9 27.98686
```

Choosing the Priors

In this analysis we chose to use weakly informative priors. The prior we chose for each regression parameter is a normal distribution with a zero mean and standard deviation of 50. By that choice we assume that before seeing the data, the average effect of each variable on fertility is zero. This assumption implies that, without additional information, we do not expect the variables to have a positive or negative effect on birth rates.

The large variance is stating that we accept a large uncertainty about the effects of the variables. It means that we allow the true effects of the variables to vary widely based on the data. The large variance assumption is reasonable since there is no strong theoretical or empirical basis for expecting a particular effect size.

For the municipality-wise intercept term we chose to use zero mean and exponential(0.02) standard deviation as priors. The use of this prior on the basis that it we believe that the municipal standard terms do not differ significantly from each other, yet there is some variation that is expected.

This choice of prior distribution therefore reflects uncertainty about the effects of the variables and is a typical approach in situations where the data itself provides the main information about the effects of the variables. This is also a common way in Bayesian modeling when you want to avoid making too strong pre-assumptions when analyzing the data.

Setting the Priors

```
#Setting the priors
priors <- c(
            prior(normal(0, 50), coef = "Alle_15_vuotiaita"),
            prior(normal(0, 50), coef = "Syntyneet"),
            prior(normal(0, 50), coef = "Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus"),
            prior(normal(0, 50), coef = "Vuokra_Asunnoissa_Asuva_Asuntokuntia"),
            prior(normal(0, 50), coef = "Tulotaso"))

priors_h <- c(prior(normal(0, 50), coef = "Alle_15_vuotiaita"),
            prior(normal(0, 50), coef = "Syntyneet"),
```

```
            prior(normal(0, 50), coef = "Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus"),
            prior(normal(0, 50), coef = "Vuokra_Asunnoissa_Asuva_Asuntokuntia"),
            prior(normal(0, 50), coef = "Tulotaso"),
            prior(exponential(0.02), class = "sd", group = "Kunta"))

# Alternative priors for the hierarchical model

priors_h_2 <- c(
            prior(normal(0, 50), coef = "Alle_15_vuotiaita"),
            prior(normal(0, 50), coef = "Syntyneet"),
            prior(normal(-2, 50), coef = "Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus"),
            prior(normal(0, 50), coef = "Vuokra_Asunnoissa_Asuva_Asuntokuntia"),
            prior(normal(100, 50), coef = "Tulotaso"),
            prior(exponential(0.02), class = "sd", group = "Kunta"))

# Alternative prior for the pooled model

priors2 <- c(
            prior(normal(0, 50), coef = "Alle_15_vuotiaita"),
            prior(normal(0, 50), coef = "Syntyneet"),
            prior(normal(-2, 50), coef = "Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus"),
            prior(normal(0, 50), coef = "Vuokra_Asunnoissa_Asuva_Asuntokuntia"),
            prior(normal(100, 50), coef = "Tulotaso"))
```

Fitting the Pooled Model

```
#Fitting the pooled model

suppressMessages(f4 <- brms::brm(
Next_Year_Syntyneet ~ 1 + Alle_15_vuotiaita  + Syntyneet  + Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus
+Vuokra_Asunnoissa_Asuva_Asuntokuntia + Tulotaso,
data = data,
family = gaussian(),
prior = priors,
chains = 4,
iter = 2000,
refresh = 0))
```

Fitting the Hierarchical Model

```
#Fitting the hierarchical model
suppressMessages(f3 <- brms::brm(
  Next_Year_Syntyneet ~ Alle_15_vuotiaita + Syntyneet +
    Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus + Vuokra_Asunnoissa_Asuva_Asuntokuntia +
    Tulotaso +
    (1 | Kunta),
  data = data,
  family = gaussian(),
  prior = priors_h,
  chains = 4,
  iter = 2000,
  refresh = 0
```

```
))
```

MCMC inference

The MCMC simulations with 4 chains and 2000 iterations each. It was considered to be an optimum between computational complexity, with enough iterations to make room for convergence and exploration of the distribution. We used the gaussian family as we assumed that most demographic variables are normally distributed.

Summary of the Pooled Model

```
#Summary of convergence diagnostics for the pooled model
summary(f4, rhat = TRUE)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Next_Year_Syntyneet ~ 1 + Alle_15_vuotiaita + Syntyneet + Korkea_Asteen_Tutkinnon_Suorittan
##    Data: data (Number of observations: 2453)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##                                          Estimate Est.Error l-95% CI
## Intercept                                   -2.87      0.31    -3.49
## Alle_15_vuotiaita                            0.27      0.02     0.24
## Syntyneet                                    0.46      0.02     0.43
## Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus    -0.06      0.01    -0.08
## Vuokra_Asunnoissa_Asuva_Asuntokuntia         0.07      0.01     0.06
## Tulotaso                                     0.15      0.02     0.11
##                                          u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept                                   -2.27 1.00     2748     2687
## Alle_15_vuotiaita                            0.31 1.00     2985     2741
## Syntyneet                                    0.50 1.00     2674     2900
## Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus    -0.05 1.00     2737     2183
## Vuokra_Asunnoissa_Asuva_Asuntokuntia         0.08 1.00     2772     2783
## Tulotaso                                     0.20 1.00     2695     2209
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.61      0.02     1.56     1.65 1.00     3862     2986
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Summary of the Pooled Model

```
summary(f3, rhat = TRUE)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Next_Year_Syntyneet ~ Alle_15_vuotiaita + Syntyneet + Korkea_Asteen_Tutkinnon_Suorittaneita
##    Data: data (Number of observations: 2453)
```

```
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~Kunta (Number of levels: 223)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.59      0.07     0.46     0.74 1.01      515     1248
##
## Population-Level Effects:
##                                           Estimate Est.Error l-95% CI
## Intercept                                    -4.38      0.49    -5.34
## Alle_15_vuotiaita                             0.32      0.02     0.28
## Syntyneet                                     0.33      0.03     0.28
## Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus    -0.11      0.01    -0.14
## Vuokra_Asunnoissa_Asuva_Asuntokuntia           0.10      0.01     0.08
## Tulotaso                                      0.29      0.04     0.21
##                                           u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept                                    -3.43 1.00     1497     2608
## Alle_15_vuotiaita                             0.37 1.00     3294     3087
## Syntyneet                                     0.38 1.00      866     1882
## Korkea_Asteen_Tutkinnon_Suorittaneita_Osuus    -0.09 1.00      696     1807
## Vuokra_Asunnoissa_Asuva_Asuntokuntia           0.12 1.00     1636     2707
## Tulotaso                                      0.36 1.00      815     1826
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.53      0.02     1.48     1.58 1.00     2456     2806
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Convergence Analysis

In this chapter the convergence analysis is conducted. The trace plots are presented for the pooled model. For the hierarchical model we only conduct the analysis for rhat values to keep the figure sizes reasonable. That is due to the hierarchical model's large number of paramteres. In the following figure are the metropolis chains for the different parameters of the Pooled Model.

```
p = mcmc_trace(f4)
p = p + labs(title = "The Trace Plots of the Metropolis Chains of The Pooled Model Parameters",caption =
print(p)
```

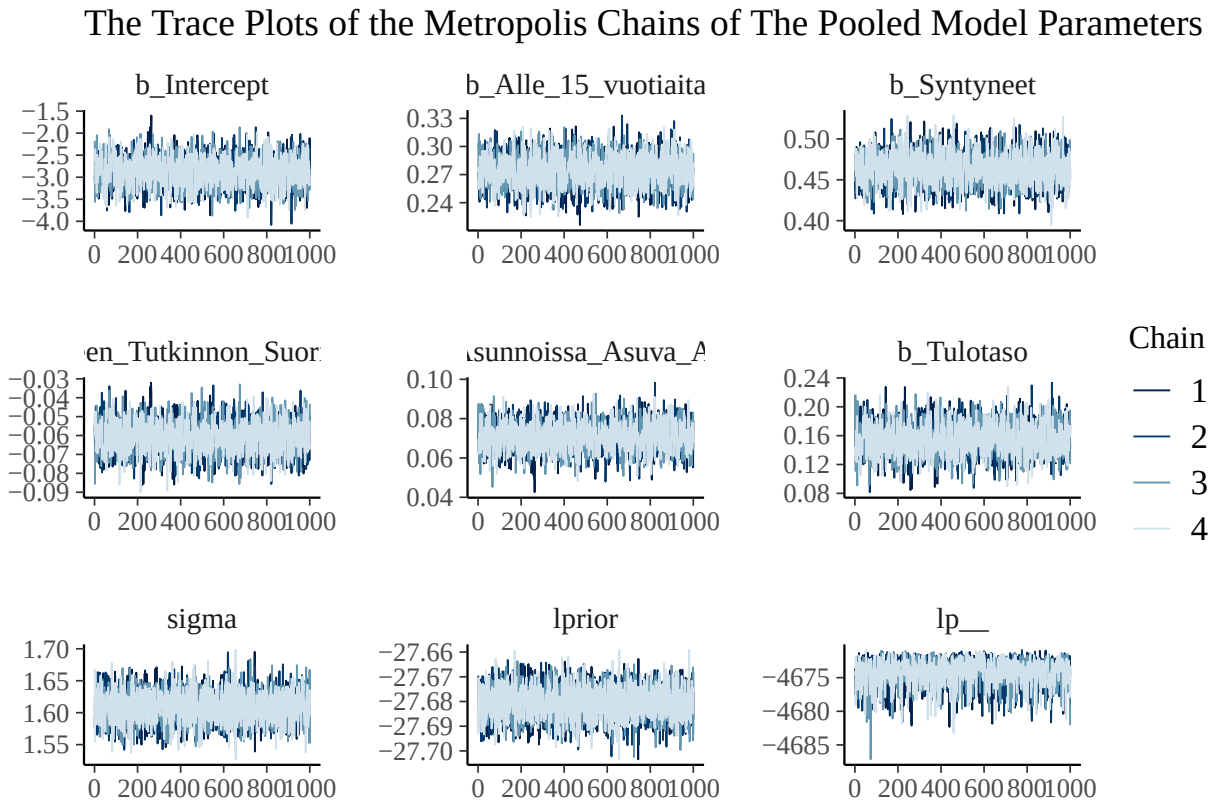

Figure 2. The Trace Plots of the Metropolis Chains of The Pooled Model Parameters

The plot indicates that the chains are converged well since there cannot be seen any chains separate from the others.

The $\hat{R}$ value is a measurement for the convergence of chains. It compares the between and within-chain estimates for the variables. If the value is significantly larger than 1, it can be interpreted as struggle in the chains converging. In both of the models used in this study, the R-Hat values equal to 1.00, which suggests that there are no convergence issues in the chains.

The ESS-value is a measure for the sampling efficiency and it helps to see whether the simulation has explored the distribution properly or not. Higher ESS-values suggest better exploration. The ESS-values are higher in the pooled model, than in the hierarchical model. Still, the ESS-values in the hierarchical model can be considered sufficient, as they exceed the limit of 1000.

Posterior predictive checks and what can be interpreted from them. What was done to improve the model if the checks indicated misspecification.

Posterior predictive checks are used to compare the draws obtained from the model to the actual observed data. From the plots we can then obtain information about possible issues in the model, like its ability to catch certain features in the data or a difference in summary statistics like mean and standard deviation.

```
par(mfrow=c(1, 2))
p1 <- pp_check(f4) + ggtitle("Pooled model")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
p2 <- pp_check(f3) + ggtitle("Hierarchical model")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
grid.arrange(p1, p2, ncol = 2)

grid.text("Figure 3. Posterior Predictive Checks for Both Models",
          x = 0.5, y =0, just = "bottom", gp = gpar(fontface = "italic", fontsize = 10))
```

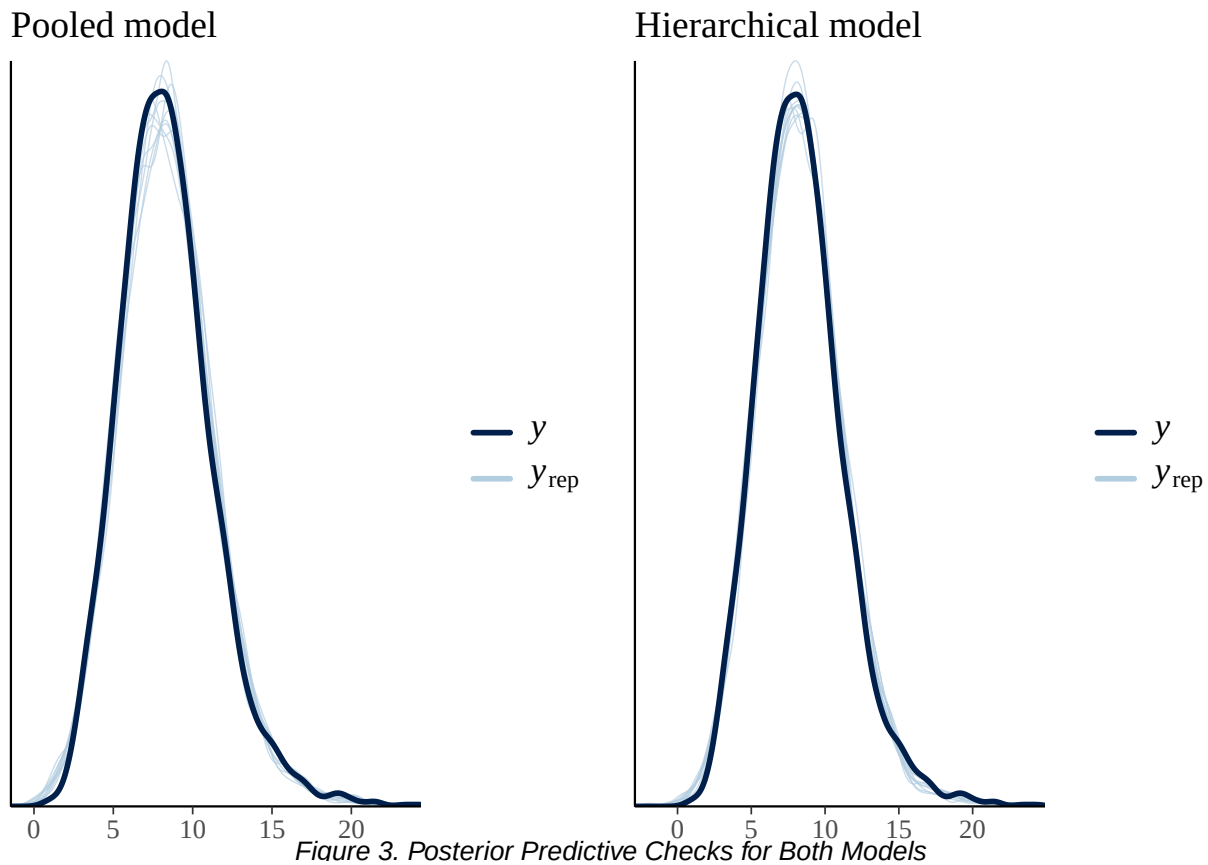

Figure 3. Posterior Predictive Checks for Both Models

From the posterior predictive check for the pooled model, it can be seen that the draws fit to the observed data quite well. The draws follow the distribution of the observed data fairly well in all places, but around the mean the deviation is noticeably larger.

The posterior predictive plot for the hierarchical model seems to catch the properties of the data almost equally as the pooled model. The mean of the draws is a bit smaller in the hierarchical model.

HMC diagnostics - Divergence and tree depth analysis

```
num.max_depth_f4 = rstan::check_treedepth(f4$fit)
```

```
## 0 of 4000 iterations saturated the maximum tree depth of 10.
```

```
num.divergent_f4 = rstan::check_divergences(f4$fit)
```

```
## 0 of 4000 iterations ended with a divergence.
```

```
num.max_depth_f3 = rstan::check_treedepth(f3$fit)
```

```
## 0 of 4000 iterations saturated the maximum tree depth of 10.
```

```
num.divergent_f3 = rstan::check_divergences(f3$fit)
```

```
## 0 of 4000 iterations ended with a divergence.
```

```
cat(num.divergent_f3, num.max_depth_f3, num.divergent_f4, num.max_depth_f4)
```

According to the divergence statistics there are no problems in terms of the chains diverging. These results suggest that there were no issues in exploration of the parameter space divergence. The results received from the R-hat values and ESS suggest the same thing, so we can be quite confident that there are no issues in the model related to this topic.

Sensitivity analysis with respect to prior choices

In this chapter the sensitivity analysis with respect to prior choices is presented. As previously stated, we are using weakly informative priors due to the lack of expert or scientific prior knowledge about the explanatory variables' influence on the response variable. The sensitivity analysis is thus conducted by choosing new priors representing our thoughts and best guesses about the influence of the explanatory variables on the birth rate. However, we only set non-zero-mean priors for those variables which influence we knew or were able to guess anything about. Rest of the prior means for the variables means we kept as zero.

The priors we chose to use in the sensitivity analysis are as follows:

```
##                            Category Original_Prior Alternative_Prior
## 1                       Under 15 Yo     N(0, 50^2)        N(0, 50^2)
## 2             Birth Rate This Year     N(0, 50^2)        N(0, 50^2)
## 3 Population with Higher Education     N(0, 50^2)       N(-2, 50^2)
## 4        Population Living on Rent     N(0, 50^2)        N(0, 50^2)
## 5                      Income level     N(0, 50^2)      N(100, 50^2)
```

The results of our sensitivity analysis show that the predictions of the models have not been remarkably changed due to the changes in prior distributions. The analysis was conducted separately for both of the models.

Pooled model

There are not any significant differences between the models trained with different priors but however some differences. The pooled model with the alternative priors seemed a bit less accurate than the original model. The peak of the distribution were closer to the obsered distribution of the data.

In the following plot is the posterior predictive checks, where can be seen that there is not any significant differencies between the models.

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

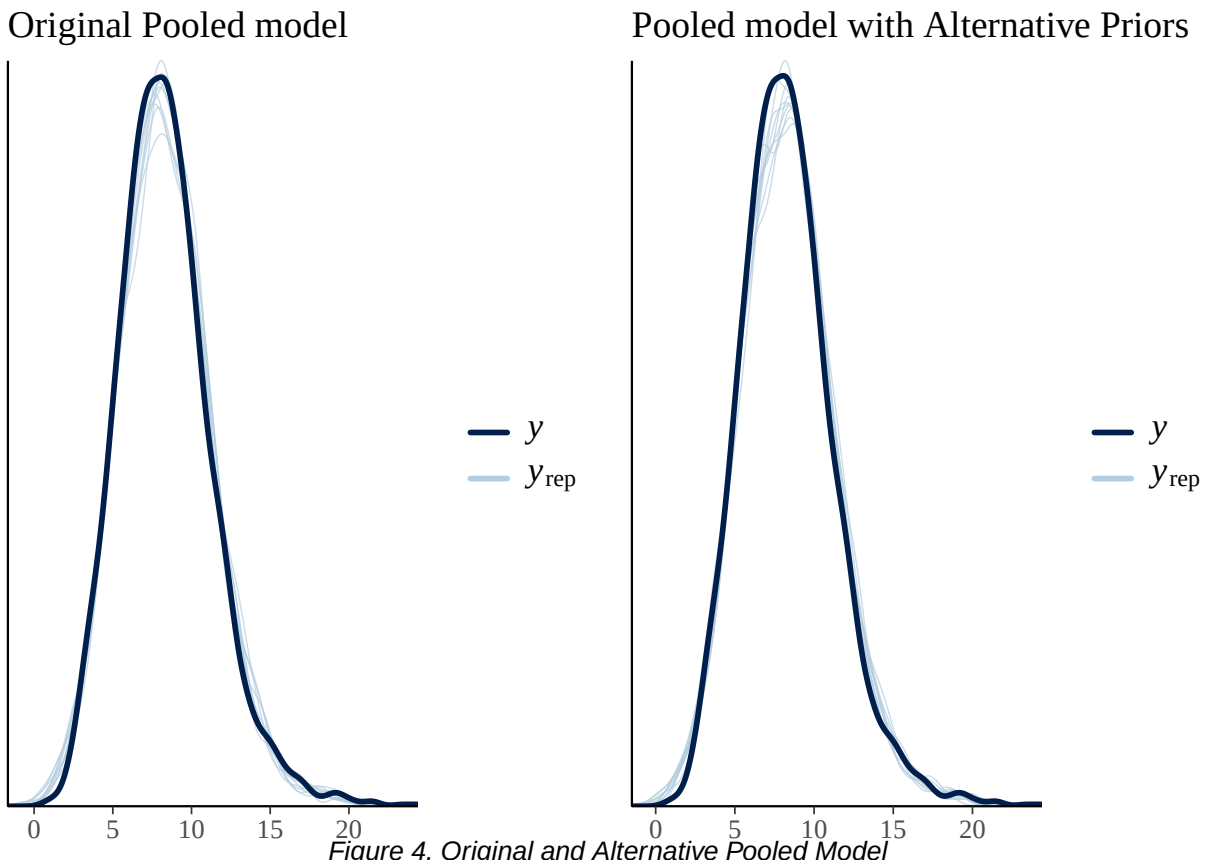

Figure 4. Original and Alternative Pooled Model

Hierarchical model

There are as well not any significant differences between the models trained with different priors. The main differences between the alternative and original models are that the alternative model has slightly better rhat values (difference of 0.01 on some variables), and its predictions seem somewhat more accurate. In the original model, the posterior predictive check indicates that the predictions are slightly biased to the right as the mean of the predictions is a little larger than the mean of the observations. In addition, the alternative model seems a bit more uncertain based on the posterior predictive plot, but on the other hand, based on the standard errors from the regression diagnostics, the uncertainty has remained the same.

In the following plot is the posterior predictive checks, where can be seen that there is not any significant differencies between the models.

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

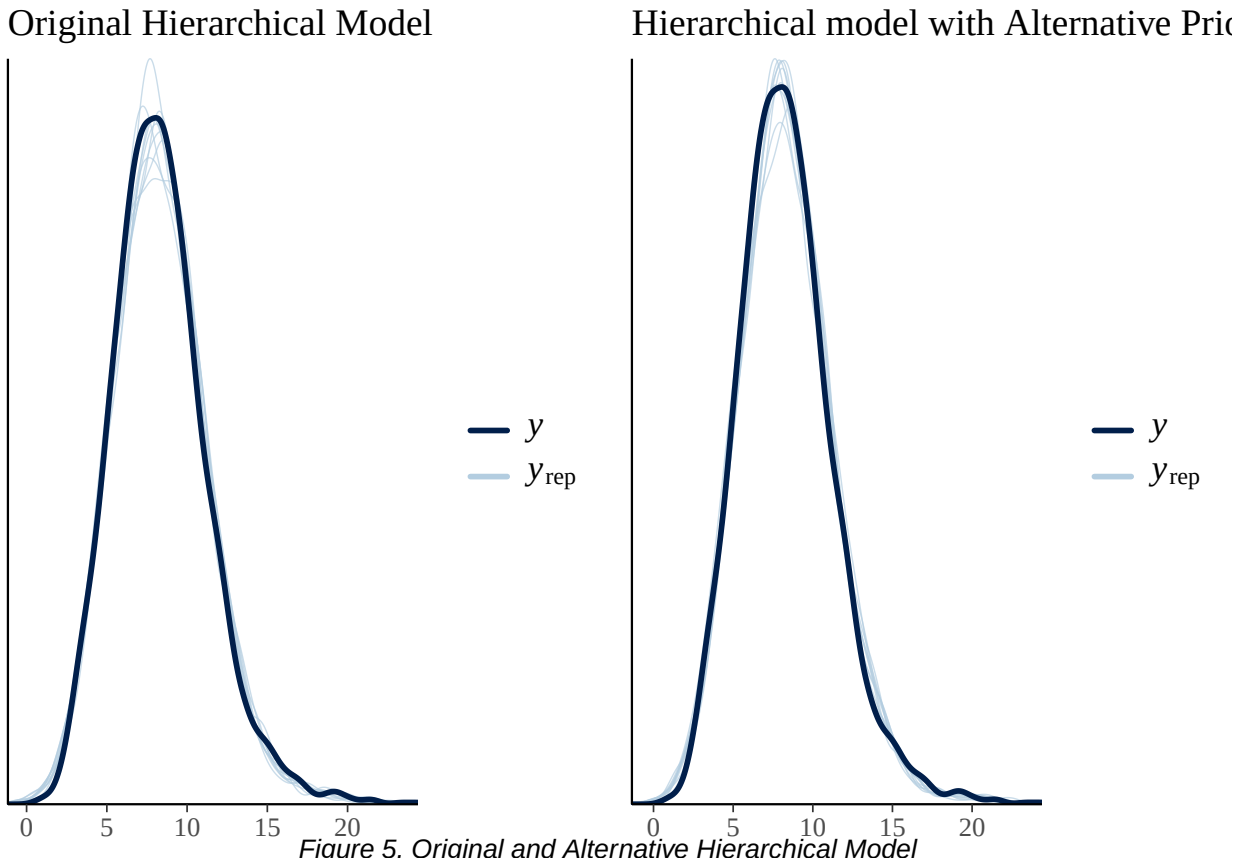

Figure 5. Original and Alternative Hierarchical Model

All in all, the conclusion of the sensitivity analysis is that the changes in the priors did not make any significant changes to the predictions. It is also good to note that we did not make any large adjustments to set the alternative priors, meaning that larger changes could influence the model more. The alternative priors were our best guesses about the explanatory variables' influence on the response variable and thus is in our view the most realistic alternative for the priors we originally set.

Model comparison with Leave-one-out Cross-Validation (LOO-CV)

```
loof4 <- loo(f4)
loof3 <- loo(f3)
```

LOO-CV for the Pooled Model

```
loof4
```

```
##
## Computed from 4000 by 2453 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -4651.6 44.6
## p_loo         9.1  0.6
## looic      9303.2 89.2
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
```

```
## See help('pareto-k-diagnostic') for details.
```

LOO-CV for the Hierarchical Model

```
loof3
```

```
##
## Computed from 4000 by 2453 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -4594.5 45.5
## p_loo       142.5  5.3
## looic      9189.0 91.0
## ------
## Monte Carlo SE of elpd_loo is 0.2.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

Model Comparison

```
loo_compare(loof4, loof3)
```

```
##    elpd_diff se_diff
## f3   0.0       0.0
## f4 -57.1      12.9
```

The LOO-CV process measures the performance of the models on new and unseen data, using observations one at a time as test data and the rest as training data. Smaller LOO-CV values can be interpreted as better model performance. The results from our models can be suggest that the hierarchical model is performing slightly better, in terms of the LOO-values. From the loo_compare function we can see that the differences in the ELPD-LOO values are smaller than the standard errors of the estimates, which might mean that one model has xactually a better predicting power.

Discussion of issues and potential improvements

The issues in the modeling process are related to the following aspects: the lack of prior knowledge about the variables influencing birth rates and their impact on the variable selection process.

Predicting birth rates involves considering a multitude of factors, including socio-economic, cultural, and environmental elements. Without robust prior knowledge about which factors are most influential, selecting the appropriate variables for your model becomes challenging. This lack of information can lead to the inclusion of irrelevant variables that introduce noise and uncertainty into the predictions, or to the omission of crucial factors that significantly impact birth rates.

At the beginning of this project, we considered several variables as well as those we eventually selected for our models. Prior to the Bayesian modeling process, we used the traditional backward selection to exclude variables from the large dataset we collected and combined from the Tilastokeskus database.

As we did not have any scientific or expert information about the influence of the variables on the response variable, we chose not to conduct too much exploratory analysis for the dataset before setting the priors. However, the backward selection led us to continue with some significant variables, from which we still had to exclude some due to their insignificance in the Bayesian models. By excluding the statistically insignificant variables from the models, we improved their performance to reliably estimate the influence of each variable.

Despite the measures we took to improve the models, there is still room for improvement. One suggestion for further improvement is to study existing information regarding the variables that influence birth rates.

Another suggestion is to investigate further the type of association between the variables: many of the chosen explanatory variables may not have a linear, but perhaps a polynomial correlation with the birth rate. Studying these relationships further could help improve the models.

Conclusion

Like mentioned in the introduction, the overall trend of birth rates has been decreasing in Finland. In the last 10 years the birth rate/1000 persons has decreased from 10 to around 7.5 children. This of course has severe effects on the size of the working force and thus on the public economy. If the birth rate decreases enough, the public economy won't have enough income to match the costs.

From the regression results it has been learned that most of the variables in our model have an increasing effect on the birth rate, with the only exception being the proportion of citizens with a higher education. This might be due to the fact that highly educated people tend to be more career focused on average, and thus less eager to have children.

The variable with the most predicting power was the current year's birth rate, with a much higher predicting power in the pooled model. Other important factors in predicting the birth rate were the proportion of under 15-year-olds and the income level. These findings suggest that municipalities with wealthy households and a lot of children have will have the highest birth rates. Regarding the variable proportion of households living on rent, we considered whether we should include it in the model or not. In the end we decided to leave it into the model, as it turned out to be statistically significant and have decent prediction power. We expected it to have negative effect on the birth rate, but our expectations were extremely vague regarding the variable, so we set the prior as 0.

If the topic would be researched more, some additional variables could be included in the model. The potential variables could be mapped performing qualitative studies by performing interviews of people from different backgrounds and life situations, to examine the decision making process of having children. Also, future studies could factor in economic situations and the effect of public policies regarding the support of having children, since birth rate is a vital part of ensuring a healthy public economy in the future.

Reflection of Learnings

The project was a great opportunity to deepen the skills learned in the exercises. In particular, the interpretation of the diagnostics becomes more important in self-made models, because there is no model answer behind it, as in practice tasks.

While doing the project, we noticed that setting priori distributions in particular is challenging and fitting models can take a lot of time when looking for the right variables and parameters. In addition, it is sometimes more difficult to find coding help for Bayesian methods on the Internet than what we have previously experienced when doing frequentist analysis.

Overall, the project was a good first step to Bayesian data analysis. We learned to fit the models and deepened our diagnostic skills, and we learned where we still need to improve.

References

Götmark, F., Andersson, M. Human fertility in relation to education, economy, religion, contraception, and family planning programs. BMC Public Health 20, 265 (2020). https://doi.org/10.1186/s12889-020-8331-7

Tilastokeskus (2023) Väestön ennakkotilasto.

Tilastokeskus (2023, a). Kuntien avainluvut muuttujina Alue 2023 ja Tiedot. PxWeb. Available on: https://pxdata.stat.fi/PxWeb/pxweb/fi/Kuntien_avainluvut/Kuntien_avainluvut___uusin/kuntien_avainluvut_viimeisin.px/ [Accessed on 16.10.2023].

Tilastokeskus (2023, b). Yhdenmukaistettu kuluttajahintaindeksi (YKHI) ja Yhdenmukaistettu kuluttajahintaindeksi kiintein veroin (YKHI-KIVE) (2005=100) muuttujina Vuosi, Indeksisarja ja Tiedot. PxWeb.

Saatavilla osoitteessa: https://pxdata.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin___khi/statfin_khi_pxt_11xk.px/ [Accessed on 16.10.2023].

Tilastokeskus (2023, c). Väestö 31.12. muuttujina Alue, Ikä, Sukupuoli, Vuosi ja Tiedot. PxWeb. Saatavilla osoitteessa: https://pxdata.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin___vaerak/statfin_vaerak_pxt_11re.px/ [Accessed on 16.10.2023].