Imputation of missing demographic count data for refugee populations

Model

Goal: impute the number of refugees in each sex/age bracket for population groups (population of origin country i in country of asylum j). For this model, we will look at the number of children only in a binomial model and subsequently extend to a multinomial model for all age/sex brackets.

Binomial model for the number of children from each country of origin i:

We assume the counts of children to be binomially distributed with the to be estimated population parameter p_j, the probability of a refugee from country i in country j to be a child under the age of 18.

$$y_j \sim Bin(n_j, p_j)$$

n_j, the total number of refugees in country of asylum j, is completely observed in UNHCR's end-year population statistics and is assumed to be measured without error. The subscript j on p_j indicates that we assume the proportion of refugee children from a country of origin to come from a common population distribution with overall intercept p, but to differ from one country of asylum to the other. We will furthermore not model the probability p_j directly, but rather in a binomial hierarchical regression model the logit of p_j:

$$\theta_j = log(\frac{p_j}{1-p_j})$$

We then assume that the θ_js in the different countries of asylum come from a joint global population normal prior distribution with mean μ and standard deviation σ:

$$\theta_j \sim Normal(\mu, \sigma)$$

For μ we use the logit of the proportion of children in the national country of origin population as per the 2020 medium variant projection of the DESA World Population Prospects 2019 (https://population.un.org/wpp/). In 2020, 49.2% of the Afghan national population were children. While we have to assume this might differ among displaced populations, it is a sensible starting point as the mean of a weakly informative prior distribution on the population intercept θ and, combined with a standard deviation σ of 1.5, results in a relatively flat prior on the probability space between 0 and 1 for p, the probability of a refugee being below the age of 18 in the binomial model.

To create multiple imputations, we are ultimately interested in obtaining draws from the posterior predictive distribution on the level of the response variable y_j (i.e. counts of children in country of asylum j given the known population size n_j) for countries of asylum with missing demographic data, that is, draws from $p(\tilde{y}|y)$. As an additional step and since we are interested in draws from the posterior predictive distribution for countries without observed demographic data that would consequently not have been part of the original model fit, for each country of asylum with missing data we first take draws from the posterior distribution of the population parameter θ_j for a new country j and label these newly simulated parameters $\tilde{\theta}$. We then draw $\tilde{y}$ given $\tilde{\theta}$.

Model test with end-2020 data on refugees from Afghanistan

Prior predictive check

The prior distribution on the global intercept p_j, or more precise on its logit θ_j, encodes our subject matter knowledge of the proportion of children in the global Afghan refugee population. We do not want to make overly strong assumptions and choose the above-mentioned Normal prior with mean at the WPP proportion of children and a large enough standard deviation that results in a relatively flat distribution on the [0,1] parameter space of p_j. There is slightly less probability on the tails close to 0 and 1, mirroring our belief that especially values close to 1, that is, 100% of the Afghan refugees in a country are children, are unlikely. We might want to slightly improve this in future iterations to reflect the fact that proportions close to 0% are slightly more likely, however with this flat distribution the prior should play a relatively minor roles and has a mostly regularising function for countries with missing demographic values.

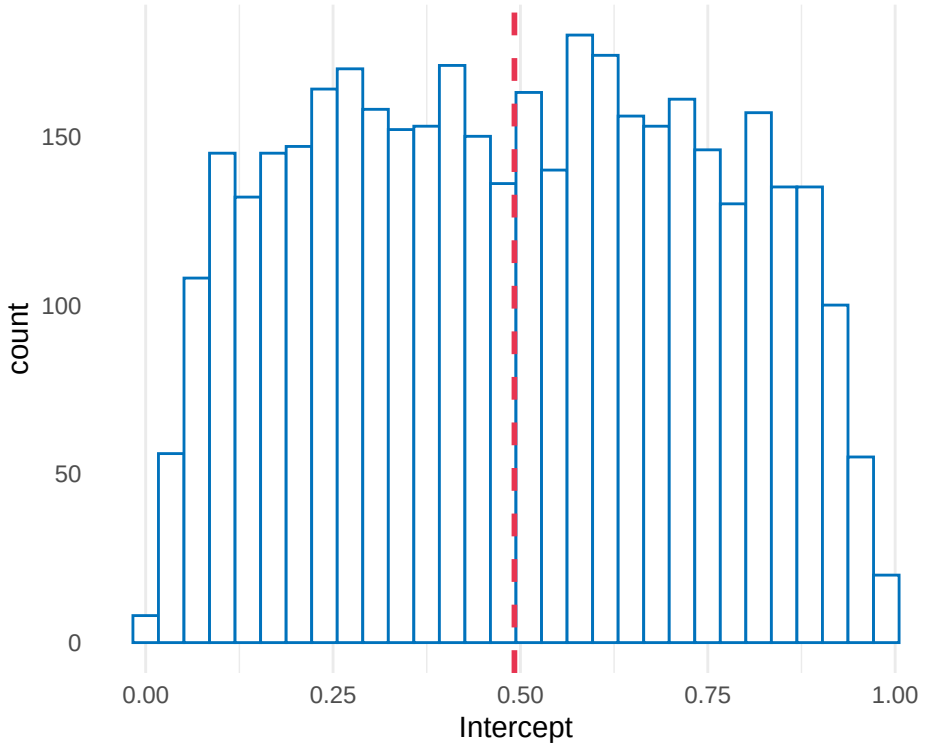

Figure 1: Simulation of global intercepts from the prior distribution

The prior simulation shows us the knowledge about the proportion of children in the global Afghan refugee population before we see any data. The distribution is plausible, and we can choose similar prior model distributions for the multinomial version of the models for all sex/age brackets.

Chains

No visible issues with chain convergence.

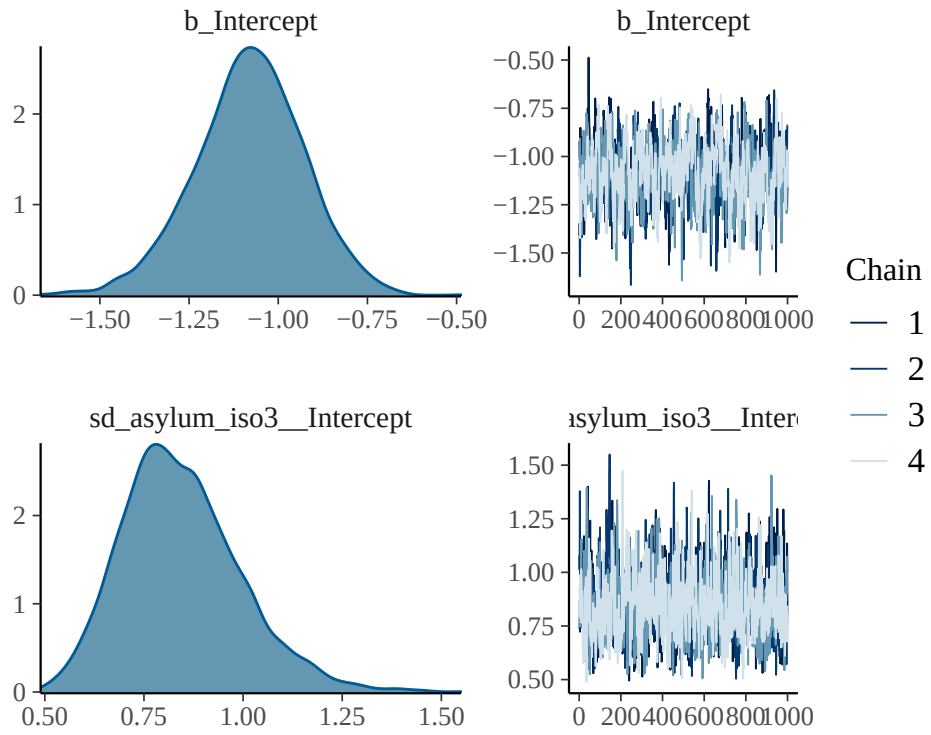

Figure 2: Trace and density

Posterior predictive check

A posterior predictive check on the level of the response variable y_j for each country of asylum j can show us the range of values the fitted model proposes for each country with observed data. By comparing it to the actually observed data, we can assess whether the model fits the data.

We see (Figure 3) that the model fit is generally plausible and includes the observed values. We will have to look at the uncertainty however - the uncertainty bands seem narrow, indicating that potentially we are overfitting or mis-specifying the variance.

Leave-one-out cross-validation

Leave-one-out cross-validation can help us decide whether the model has sufficient out-of-sample predictive power. For this, we re-fit the model leaving out one country with observed data at a time and compare the prediction for that country from the model fit to the actually observed data. This check is particularly important for our models since we are aiming to make predictions of the number of children for countries with unobserved data.

We evaluate the cross-validation with a measure called Pareto k-estimates that tells us how influential each data point (country of asylum) is. Values over 0.7 indicate model misspecification. We can see (Figure 4) a lot of values over that threshold, potentially confirming the impression from the posterior predictive check that our model might be overfitting.

Figure 3: Posterior predictive distribution (y-rep) of the number of children per country of asylum with observed data (y)

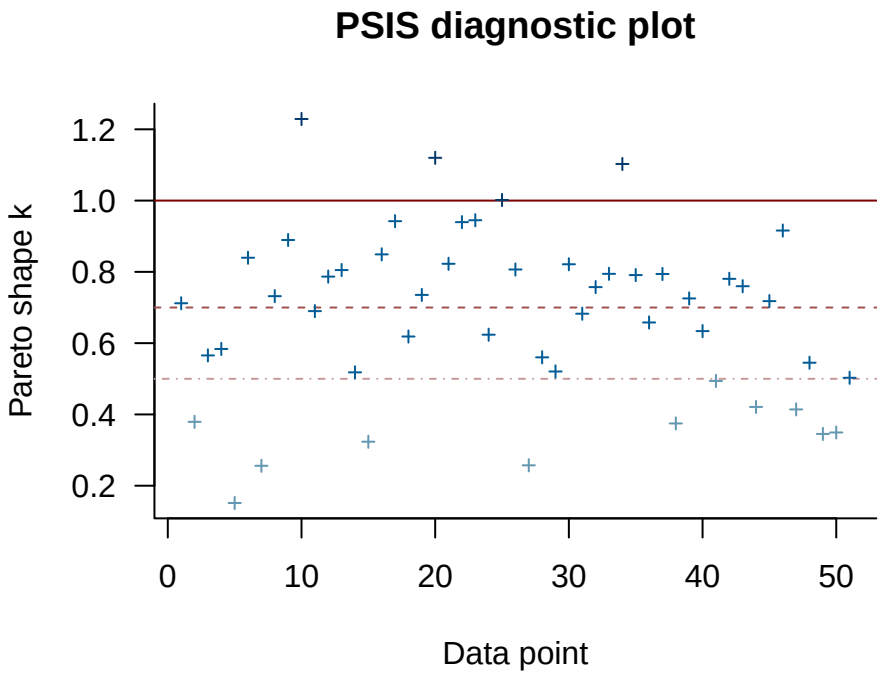

Figure 4: Pareto k estimates