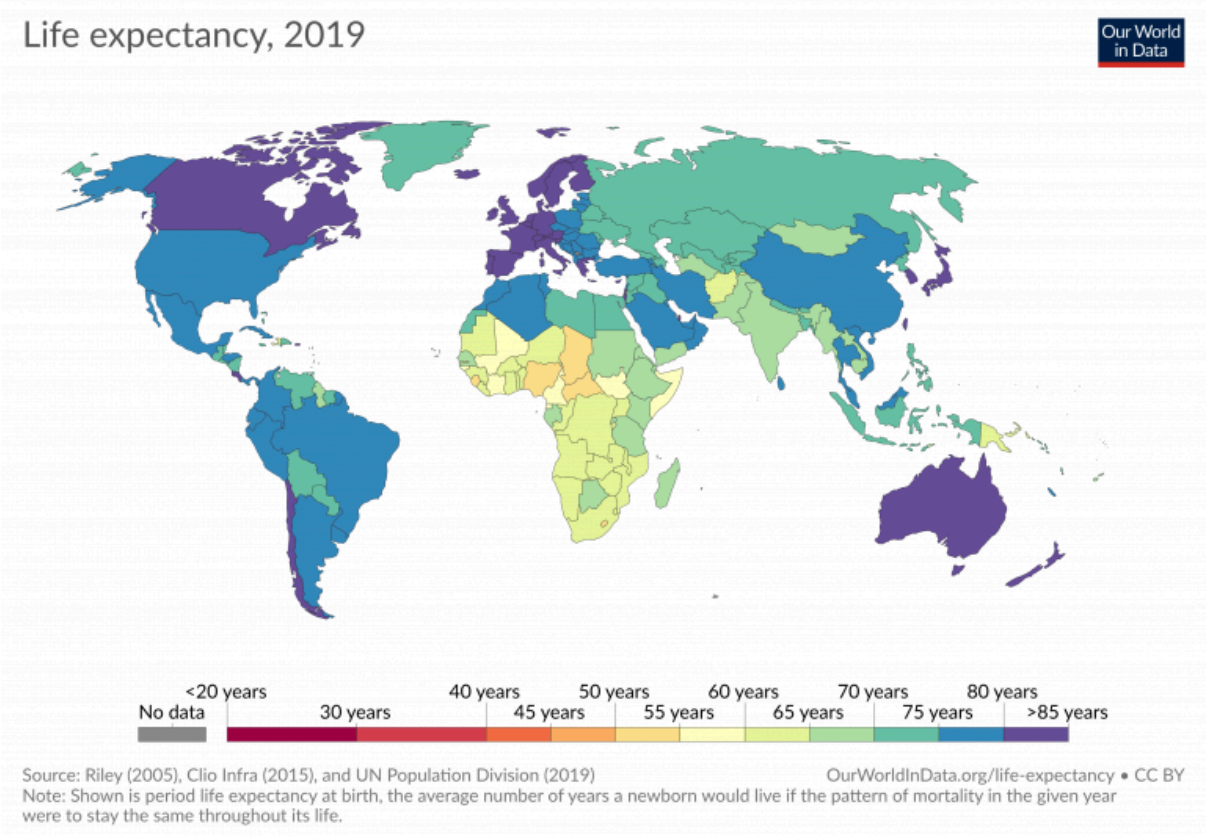
Analysis of life expectancy

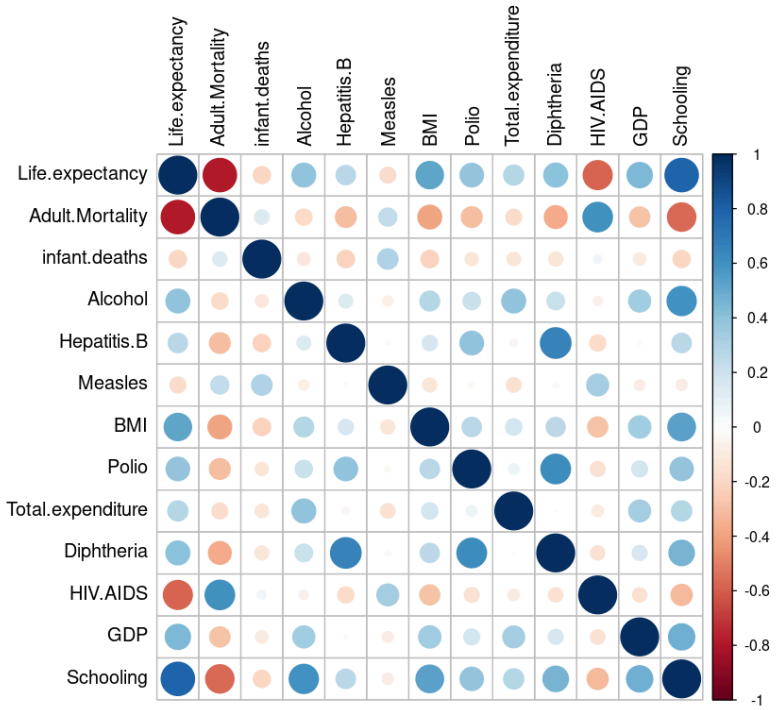
Introduction

The motivation is to study the problem of life expectancy and find which factors most correlate with it. Life expectancy is a complicated problem which is affected by many factors and it is not certainly known what these factors are. The main modelling idea is to study which variables affect life expectancy the most.



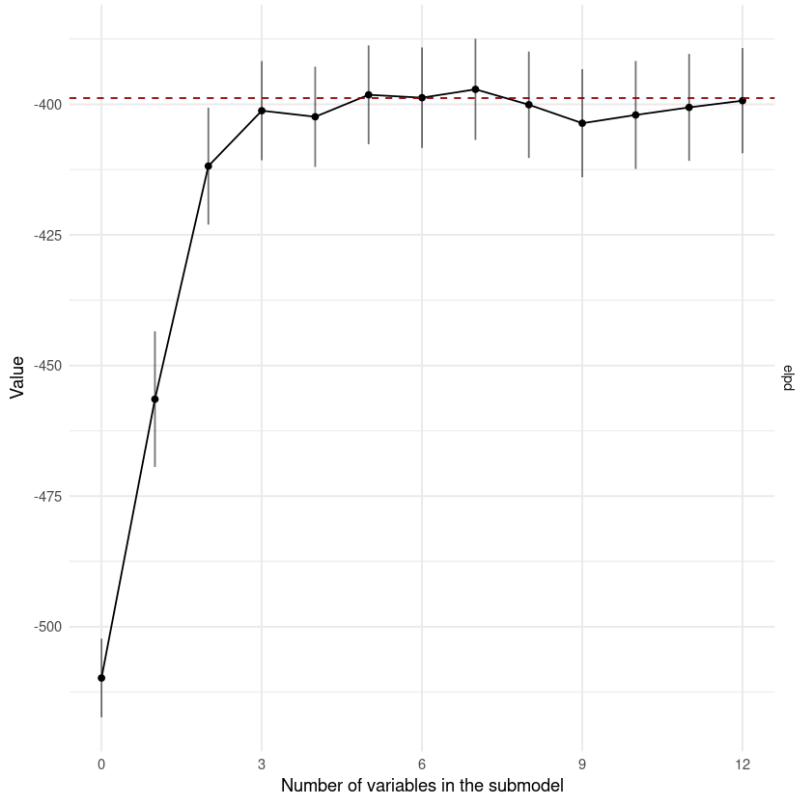
Description of the data

The data used in the analysis is a dataset from WHO published in kaggle datasets. It contains data from all the world's countries from the year 2001 to 201.. For this analysis only the year 2010 is chosen, since time series modeling is not wanted. The data contains in total 18 explanatory variables which is pruned to 12 since it contains some overlapping variables. Also the name of the country, year and status whether the country is developing or developed is omitted. The selected variables for this analysis are adult mortality, infant deaths, alcohol, hepatitis B, measles, BMI, polio, total expenditure, Diphtheria, HIV/AIDS, GDP, Schooling.



Description of models

The analysis is conducted using a linear model with variable selection. The first model is a linear model with all the variables in the dataset. The second model is also a linear model but using only the variables selected by the variable selection algorithm. After conducting variable selection only three variables are selected: adult mortality, schooling and hiv/aids.



Priors

The prior choices for this model are standard normal distributions which are weakly informative. All the population level effects have been given this same prior.

Brms code

The model was implemented with brms and the code for the first full linear model is below.

```
fit_multiple_variables <- brm(Life.expectancy ~ Adult.Mortality + Alcohol + infant.deaths +  Hepatitis.B + Measles + BMI  + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + Schooling , data = data, family = gaussian(), prior = set_prior("normal(0,1)", class = "b"), iter=4000, control = list(max_treedepth = 20))
```

And the second linear model with variable selection.

```
fit_varsel_variables <- brm(Life.expectancy ~ Adult.Mortality + HIV.AIDS + Schooling , data = data,family = gaussian(), prior = set_prior("normal(0,1)", class = "b"), iter=4000, control = list(max_treedepth = 20))
```

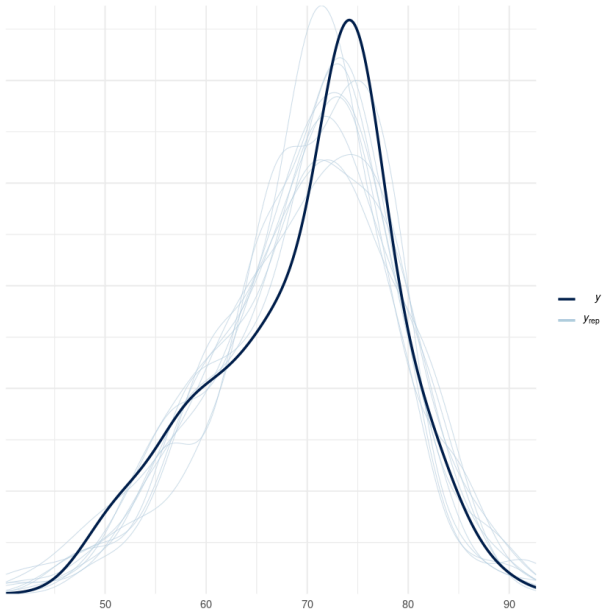
Convergence diagnostics

For converge diagnostics the r_hat value was used. Both of the models converged nicely to r_hat value 1 but some changes to the default parameters have to be made. Different values for the iter parameter were tried and the convergence was obtained with iter=4000. Also the tree depth value had to be tested with multiple values and convergence was obtained with max_treedepth=20.

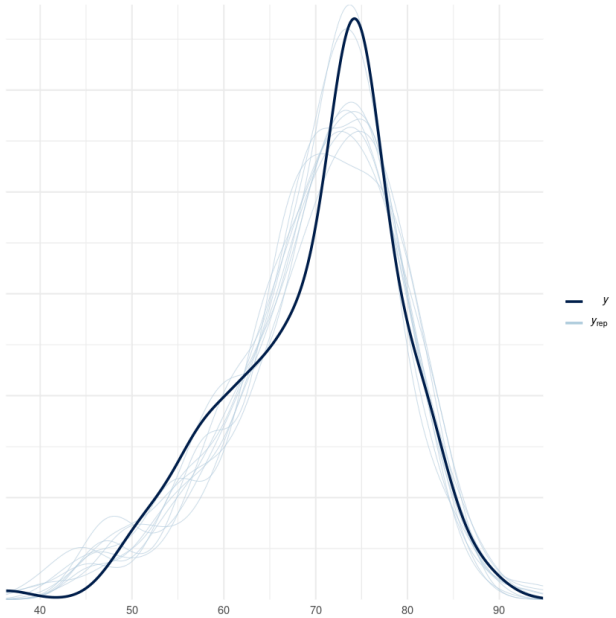
Posterior predictive checks

Fairly nice and accurate posterior predictive checks were obtained.

First model



Second model



Model comparison

Model comparison was performed with loo package in brms and the elps_loo values were used. The first model performs somewhat better independent of the prior choices.

Sensitivity analysis

Both models were tested with different priors. The normal distribution with parameters 0 mean and 10 sd was used for all population level effects. Both of the models perform almost the same independent of the prior choice and the first model still performs better than the second one.

Issues and improvements

Biggest issue faced while creating this analysis was to fit the model since it didn't converge at first. Different parameter values had to be tried and at the end the models did converge. Perhaps the biggest improvement for the model could be to use a non-linear model which could fit to the data better. Another improvement could be to collect more data on more specific groups within a country since this data only has 193 countries which is not very much and the life expectancy can also vary inside a country.

What was learned

During the data analysis I learned a lot about writing a data analysis report and communicating the results. It was educational to create a complete data analysis according to the Bayesian workflow since it unified all the learned things during the course. I also got quite familiar with the brms library and its offerings to bayesian modelling.