BDA Final Project - Evaluation of the relationship between mean income and suicide rate.

Anonymous

Contents

Introduction

Motivation

An oft-wondered and oft-debated question for people, for as long as the concept of money has existed, is whether having money makes a person happier. While happiness is a tricky thing to quantify, extreme sadness can readily be quantified by looking at the suicide rate for a particular group of people. This brings us to the rather macabre question we have chosen to examine for this project: is there a relationship between average income and the suicide rates for groups of people?

The problem and modelling idea

The relationship between the mean income and the suicide rate was modeled as a linear regression problem, and this was done in three different contexts, so that we can try to get some kind of statistically sound answer to the question posed above. The first context is for the 9 different broad groups of professions in the United Kingdom, sourced from the government database for the year 2019. The second is for same, but for 22 groups of professions in the United States. The final context is an international one, the suicide rates were seen for the 93 countries of the world with a population >100k, and the mean incomes for those countries found for the year 2019. In all three cases, the mean income was taken in dollars and standardized by the cost of living for the corresponding country, in order to be able to club such different data together.

Prior art

The US data was examined by the website registerednursing.org, and a single linear trend was fitted with no statistical information or information about the methods. Reviewing past academic publication reveals that similar studies were done with the data in previous years for South Korea [1] and Denmark [2]. The former calculated the suicide hazard ratios for each income bracket to a certain confidence level, while the latter included more factors in the analysis such as age and gender, transforming the problem.

Layout of the report

The report commences with explaining the data used for the analysis. After this, the different models that are applied are explained, and then their respective performances are analyzed in terms of convergence and how well they fit the data. Some different priors are hyper priors are tested in a sensitivity analysis, and finally, the results are interpreted with our overall conclusion.

The data

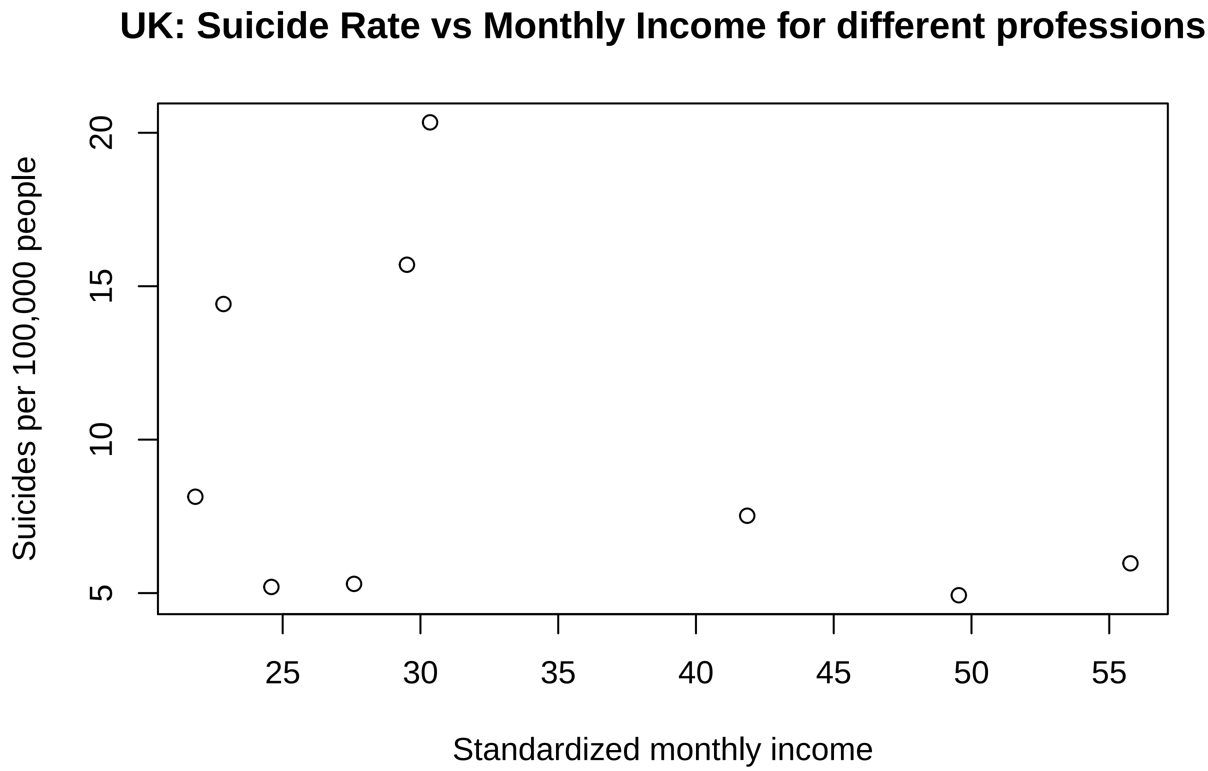
Our data was not retrieved from some common source, but rather from different sources and they were adjusted so that they were coherent together.

United Kingdom data

The data for both the suicide rate (per 100,000 people) and the mean weekly income (which was converted into monthly income, into dollars, and standardized by the cost of living) were both retrieved for the same 9 profession groups from UK Office for National Statistics (ons.gov.uk). The 9 occupation groups selected are

1. Managers, directors and senior officials
2. Professional occupations
3. Associate professional and technical occupations
4. Administrative and secretarial occupations
5. Skilled trades occupations
6. Caring, leisure and other service occupations
7. Sales and customer service occupations
8. Process, plant and machine operatives
9. Unskilled occupations / elementary operations

```
ukdata =read.csv('UKdata.csv');
plot(ukdata$wagepercol, ukdata$suicides100k, xlab='Standardized monthly income', ylab = 'Suicides per 1
```

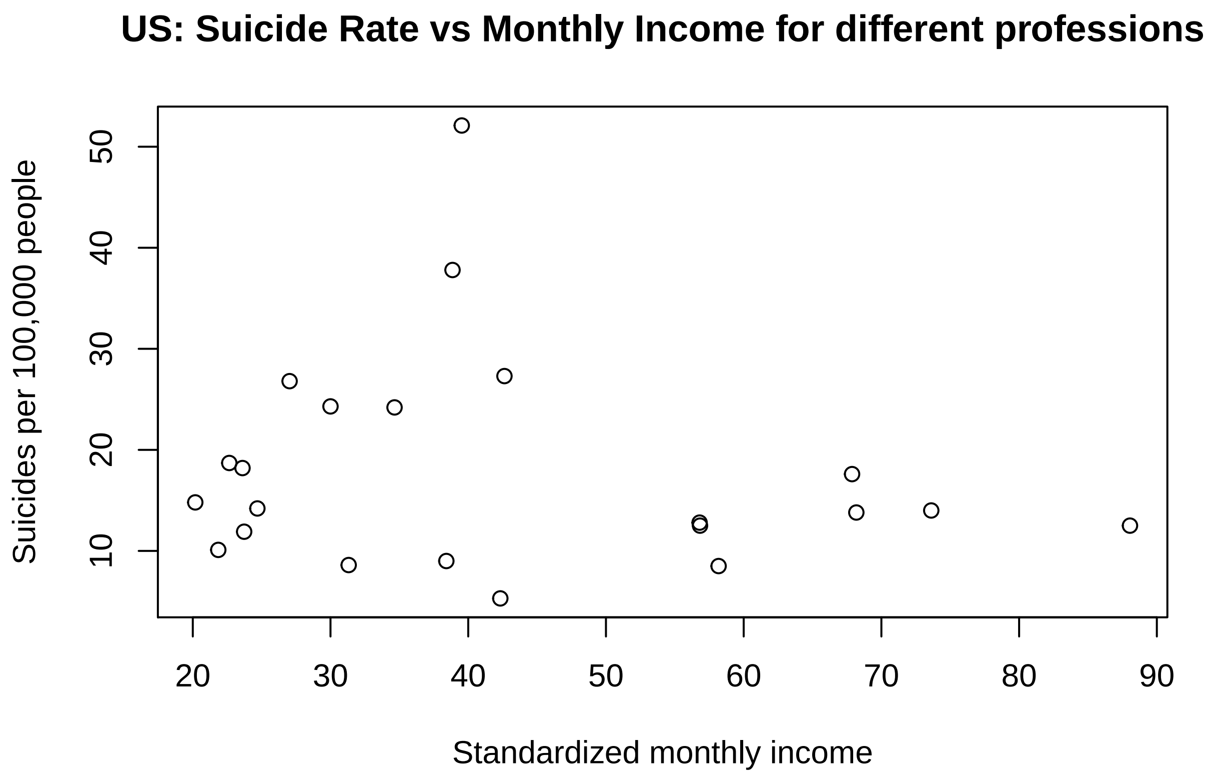


United States data

The data for the suicide rate (per 100,000 people) for 22 broad profession groups was retrieved from the Maerican Centre for Disease Control (CDC), and the mean monthly income (which was standardized by the cost of living) was retrieved for the same 22 profession groups from US Bureau of Labor Statistics (BLS). The 22 occupation groups selected are

1	Management occupations	12	Protective service occupations
2	Business and financial operations occupations	13	Food preparation and serving related occupations
3	Computer and mathematical occupations	14	Building and grounds cleaning and maintenance occupations
4	Architecture and engineering occupations	15	Personal care and service occupations
5	Life, physical, and social science occupations	16	Sales and related occupations
6	Community and social service occupations	17	Office and administrative support occupations
7	Legal occupations	18	Farming, fishing, and forestry occupations
8	Educational instruction and library occupations	19	Construction and extraction occupations
9	Arts, design, entertainment, sports, and media occupations	20	Installation, maintenance, and repair occupations
10	Healthcare practitioners and technical occupations	21	Production occupations
11	Healthcare support occupations	22	Transportation and material moving occupations

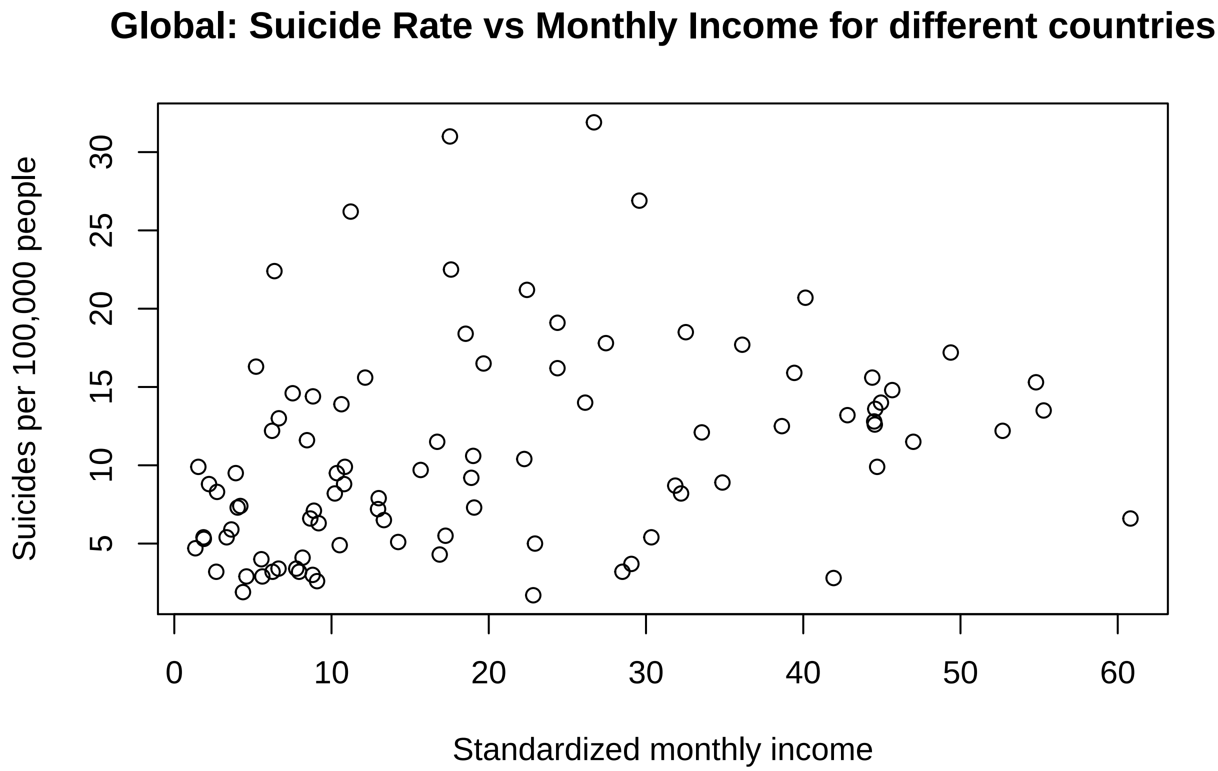
```
usdata =read.csv('USdata.csv');
plot(usdata$wagepercol, usdata$suicides100k, xlab='Standardized monthly income', ylab = 'Suicides per 1(
```



Global data

The data for both the suicide rate (per 100,000 people) and the mean monthly income (which was converted into dollars, and standardized by the cost of living of each country) were both retrieved for the 93 countries which have a population greater than 100,000 from the World Population Review website (worldpopulationreview.com). This is quite a different context from the profession-wise data for the countries above, as it is instead countrywise. The standardizing by the cost of living of each country, which was also obtained from the World Population Review website, was our attempt to render the comparison as a fair one.

```
countrydata =read.csv('countrydata.csv');
plot(countrydata$wagepercol, countrydata$suicides100k, xlab='Standardized monthly income', ylab = 'Suic
```



This is the most complicated context in which this relationship is examined. It must be noted that there are numerous factors contributing to the suicide rates of different countries besides the income standardized by cost of living (Which is similar to the Purchasing Power of the country). In addition to the level of poverty, there are cultural factors (more emphasis on personal happiness, varying sense of familial responsibilities), geographical factors (failing agricultural produce, larger periods of darkness) and several possible other factors (political unrest, war). Therefore the problem in this context is quite simplified, and on cursory viewing of the figure above, it actually appears that there is positive correlation between wealth and income, which on first thought might seem unexpected.

Methodology

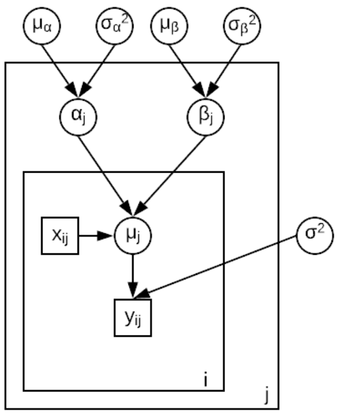
Models used

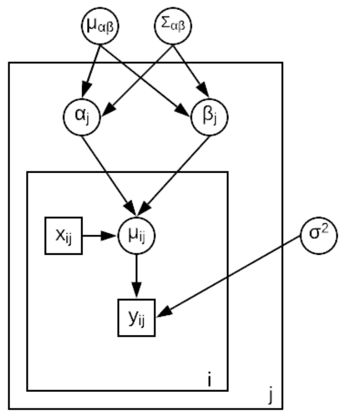
The standardized monthly income and the suicide rates are modeled together as a linear regression problem, such that the mean of the suicide rate R for a particular income level I would be given by $R = \alpha + \beta I$. The values of α and β for this purpose are output through the written Stan codes. For this modeling problem, a few kinds of linear regression models were tested.

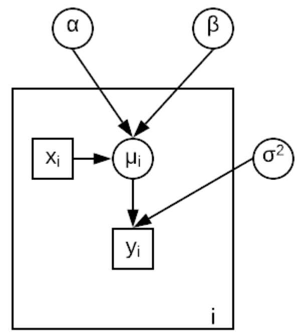
First, hierarchical models are tested. A common σ^2 is used for all three contexts, with weakly informative priors. The values of α and β are drawn from hyperpriors common to all contexts. This is depicted by the first image below.

The second kind of hierarchical model assumes that there is a correlation between α and β and therefore the hyperpriors for the two have a vector μ and a covariance matrix Σ. Generation of a covariance priors from the hyperpriors requires the separate generation of the correlation matrix Σ_0 (using the LKJ correlation distribution) and the scaling factor σ_α and σ_β. A diagonal matrix D is generated from the two σs, and then the covariance matrix is obtained as a transformed parameter as $\Sigma = D \times \Sigma_0 \times D$. This model is depicted in the second figure below.

The two hierarchical models are compared with both pooled and separate models, which have priors defined for α and β as well as for σ^2. For the pooled model, a single estimate is made for α, β and for σ^2 on combining the data from all three contexts. For the separate model, three completely separate estimates are made for each of the contexts for α, β and for σ^2. The model layout for both pooled and separate modes look similar, and is depicted as the third figure below.







To summarize, the four models tested are:

- Model 1: Hierarchical model with uncorrelated α and β (Figure 1)
- Model 2: Hierarchical model with correlated α and β. (Figure 2)
- Model 3: Separate model (Figure 3)
- Model 4: Pooled model (Figure 3)

Stan code

The Stan code for all of the three models are presented below

Initializing

```
library(loo)
library(rstan)
library(shinystan)

SEED <- 2 # set random seed for reproducability
```

Model 1

```
file_name1 = "Model_1.stan"
writeLines(readLines(file_name1))
```

```
## data {
##   int <lower=0> N1;    //  number  of data  points, uk
##   int <lower=0> N2;    //  number  of data  points, us
##   int <lower=0> N3;    //  number  of data  points, country
##   vector[N1] y1;
##   vector[N2] y2;
```

```
##   vector[N3] y3;
##   vector[N1] x1;
##   vector[N2] x2;
##   vector[N3] x3;
## }
##
## parameters {
##   vector[3] alpha;
##   vector[3] beta;
##   real mu_alpha;
##   real mu_beta;
##   real <lower=0>sigma_alpha;
##   real <lower=0>sigma_beta;
##   real <lower=0> sigma;
## }
##
## transformed parameters {
##   vector[N1] mu1 = alpha[1] +beta[1]*x1;
##   vector[N2] mu2 = alpha[2] +beta[2]*x2;
##   vector[N3] mu3 = alpha[3] +beta[3]*x3;
##
## }
##
##
## model {
##   //  priors
##   mu_alpha ~ normal(15,15);
##   mu_beta ~ normal(0,2);
##   sigma_alpha ~ gamma(1,1);
##   sigma_beta ~ gamma(1,1);
##   sigma ~ gamma(1,1);
##
##   alpha[1] ~ normal(mu_alpha, sigma_alpha);
##   beta[1] ~ normal(mu_beta, sigma_beta);
##   y1 ~ normal(mu1, sigma);
##
##
##   alpha[2] ~ normal(mu_alpha, sigma_alpha);
##   beta[2] ~ normal(mu_beta, sigma_beta);
##   y2 ~ normal(mu2, sigma);
##
##
##   alpha[3] ~ normal(mu_alpha, sigma_alpha);
##   beta[3] ~ normal(mu_beta, sigma_beta);
##   y3 ~ normal(mu3, sigma);
##
##  // theta[J+1] ~ normal(mu, sigma_p); // Getting the mean for the seventh machine
##
## }
## generated quantities {
##   vector[N1] log_lik1;
##   vector[N2] log_lik2;
##   vector[N3] log_lik3;
##
```

```
##
##   for (n in 1:N1){
##     log_lik1[n] = normal_lpdf(y1[n] | mu1[n], sigma);
##   }
##   for (n in 1:N2){
##     log_lik2[n] = normal_lpdf(y2[n] | mu2[n], sigma);
##   }
##   for (n in 1:N3){
##     log_lik3[n] = normal_lpdf(y3[n] | mu3[n], sigma);
##   }
##  }
```

Model 2

```
file_name2 = "Model_2.stan"
writeLines(readLines(file_name2))
```

```
## data {
##   int <lower=0> N1;    //  number  of data  points, uk
##   int <lower=0> N2;    //  number  of data  points, us
##   int <lower=0> N3;    //  number  of data  points, country
##   vector[N1] y1;
##   vector[N2] y2;
##   vector[N3] y3;
##   vector[N1] x1;
##   vector[N2] x2;
##   vector[N3] x3;
## }
##
## parameters {
##   vector[2] theta1;
##   vector[2] theta2;
##   vector[2] theta3;
##   vector[2] mu_theta;
##   corr_matrix[2] sig_corr; //Correlation matrix
##   vector<lower=0>[2] sig_scale; // Scale
##   real <lower=0> sigma;
## }
##
## transformed parameters {
##   matrix[2,2] sig_hyp = diag_matrix(sig_scale)*sig_corr*diag_matrix(sig_scale); //Creating the covari
##   vector[N1] mu1 = theta1[1] +theta1[2]*x1;
##   vector[N2] mu2 = theta2[1] +theta2[2]*x2;
##   vector[N3] mu3 = theta3[1] +theta3[2]*x3;
## }
##
##
## model {
##   sigma ~ gamma(1,1); // Uninformative prior
##   sig_corr ~ lkj_corr(2); //Prior for the correlation matrix
##   sig_scale ~  multi_normal([10,10], [[100,10],[10,100]]) ; // Creating prior from uninformative hype
##   mu_theta ~ multi_normal([10,10], [[100,10],[10,100]]) ; // Creating prior from uninformative hyper
```

```
##
##   theta1 ~ multi_normal(mu_theta, sig_hyp);
##   y1 ~ normal(mu1, sigma);
##
##
##   theta2 ~ multi_normal(mu_theta, sig_hyp);
##   y2 ~ normal(mu2, sigma);
##
##
##   theta3 ~ multi_normal(mu_theta, sig_hyp);
##   y3 ~ normal(mu3, sigma);
##
##  // theta[J+1] ~ normal(mu, sigma_p); // Getting the mean for the seventh machine
##
## }
## generated quantities {
##   vector[N1] log_lik1;
##   vector[N2] log_lik2;
##   vector[N3] log_lik3;
##
##
##   for (n in 1:N1){
##     log_lik1[n] = normal_lpdf(y1[n] | mu1[n], sigma);
##   }
##   for (n in 1:N2){
##     log_lik2[n] = normal_lpdf(y2[n] | mu2[n], sigma);
##   }
##   for (n in 1:N3){
##     log_lik3[n] = normal_lpdf(y3[n] | mu3[n], sigma);
##   }
##  }
```

Model 3

```
file_name3 = "Model_3.stan"
writeLines(readLines(file_name3))
```

```
## //separated model
## data {
##   int <lower=0> N1;    //  number  of data  points, uk
##   int <lower=0> N2;    //  number  of data  points, us
##   int <lower=0> N3;    //  number  of data  points, country
##   vector[N1] y1;
##   vector[N2] y2;
##   vector[N3] y3;
##   vector[N1] x1;
##   vector[N2] x2;
##   vector[N3] x3;
##
##
## }
##
```

```
## parameters {
##   vector[2] t1;
##   vector[2] t2;
##   vector[2] t3;
##   real <lower=0> sigma1;
##   real <lower=0> sigma2;
##   real <lower=0> sigma3;
##
## }
##
## transformed parameters {
##   vector[N1] mu1 = t1[1] + t1[2] * x1;
##   vector[N2] mu2 = t2[1] + t2[2] * x2;
##   vector[N3] mu3 = t3[1] + t3[2] * x3;
## }
##
## model {
##   //  priors
##   sigma1 ~ gamma(1,1);
##   sigma2 ~ gamma(1,1);
##   sigma3 ~ gamma(1,1);
##   t1 ~ multi_normal([0, 0], [[100,10],[10,100]]);
##   t2 ~ multi_normal([0, 0], [[100,10],[10,100]]);
##   t3 ~ multi_normal([0, 0], [[100,10],[10,100]]);
##
##
##   for (n in 1:N1){
##     y1[n] ~ normal(mu1[n], sigma1);
##   }
##
##   for (n in 1:N2){
##     y2[n] ~ normal(mu2[n], sigma2);
##   }
##
##   for (n in 1:N3){
##     y3[n] ~ normal(mu3[n], sigma3);
##   }
## }
##
## generated quantities {
##   vector[N1] log_lik1;
##   vector[N2] log_lik2;
##   vector[N3] log_lik3;
##
##
##   for (n in 1:N1){
##     log_lik1[n] = normal_lpdf(y1[n] | mu1[n], sigma1);
##   }
##   for (n in 1:N2){
##     log_lik2[n] = normal_lpdf(y2[n] | mu2[n], sigma2);
##   }
##   for (n in 1:N3){
##     log_lik3[n] = normal_lpdf(y3[n] | mu3[n], sigma3);
##   }
```

```
##  }
```

Model 2

```r
file_name4 = "Model_4.stan"
writeLines(readLines(file_name4))
```

```
## //pooled model when
## data {
##   int <lower=0> N;
##   vector[N] y;
##   vector[N] x;
## }
##
## parameters {
##   vector[2] t;
##   real <lower=0> sigma;
##
## }
##
## transformed parameters {
##   vector[N] mu = t[1] + t[2] * x;
## }
##
## model {
##   //  priors
##   t ~ multi_normal([0, 0], [[100,10],[10,100]]);
##   sigma ~ gamma(1,1);
##
##
##   for (n in 1:N){
##     y[n] ~ normal(mu[n], sigma);
##   }
## }
##
## generated quantities {
##   vector[N] log_lik;
##
##   for (n in 1:N){
##     log_lik[n] = normal_lpdf(y[n] | mu[n], sigma);
##   }
##  }
```

Sensitivity analysis

Separate model

To test the prior sensitivity, we change α and β prior as following

Experiment1

$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}$ and $\sum = \begin{bmatrix}15^2 & 67.5\\ 67.5 & 15^2\end{bmatrix}$

Result:

All Rhat values = 1, elpd_loo = -396.4, 122 k-values are between (-Inf, 0.5] (good) and 2 k-values are in (0.7, 1] (bad)

Experiment2

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\sum = \begin{bmatrix} 3^2 & 9 \\ 9 & 3^2 \end{bmatrix}$

Result: All Rhat values = 1, elpd_loo = -401.4, 121 k-values are between (-Inf, 0.5] (good),1 k-value is in (0.5, 0.7] and 1 k-value is in (0.7, 1] (bad)

Pool model

To test the prior sensitivity, we change α and β prior as following

Experiment1

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\sum = \begin{bmatrix} 15^2 & 67.5 \\ 67.5 & 15^2 \end{bmatrix}$

Result:

All Rhat values = 1, elpd_loo = -431.3, 122 k-values are between (-Inf, 0.5] (good) and 1 k-value is in (0.7, 1] (bad)

Experiment2

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\sum = \begin{bmatrix} 3^2 & 9 \\ 9 & 3^2 \end{bmatrix}$

Result: All Rhat values = 1, elpd_loo = -431.7, 121 k-values are between (-Inf, 0.5] (good), 1 k-value is in (0.7, 1] (bad)

Finalized priors + Results

References:

[1] Qin, Ping, Esben Agerbo, and Preben Bo Mortensen. "Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: a national register–based study of all suicides in Denmark, 1981–1997." American journal of psychiatry 160.4 (2003): 765-772.

[2] Lee, Sang-Uk, et al. "Suicide rates across income levels: retrospective cohort data on 1 million participants collected between 2003 and 2013 in South Korea." Journal of epidemiology 27.6 (2017): 258-264.