BDA - Project Work

Jacopo Losi, Nicola Saljoughi

Contents

Introduction

This project is based on a study carried out in 2015 by a group of researchers to estimate the incidence of serious suicide attempts in Shandong, China, and to examine the factors associated with fatality among the attempters.
We have chosen to examine a dataset on suicides because it is a really important but often underconsidered problem in today's society. Not only this problem reflects a larger problem in a country societal system but it can also be a burden for hospital resources. We think that by being able to talk about it more openly and by truly trying to estimate its size and impact we can start to understand where the causes are rooted and what can be done to fight it.
We invite the reader to check the source section to further read about the setting and results of the named paper.

Analysis Problem

The objective of the project is to use the bayesian approach to develop models to evaluate the most influential factors related to serious suicide attempts (SSAs, defined as suicide attempts resulting in either death or hospitalisation) and being able to make predictions for the years following the period where the study was set.

Data

Data from two independent health surveillance systems were linked, constituted by records of suicide deaths and hospitalisations that occured among residents in selected countries during 2009-2011.
The data set is constituted by 2571 observations of 11 variables:

- `Person_ID`: ID number, $1, ..., 2571$
- `Hospitalised`: *yes* or *no*
- `Died`: *yes* or *no*
- `Urban`: *yes*, *no* or *unknown*
- `Year`: 2009, 2010 or 2011
- `Month`: $1, ..., 12$
- `Sex`: *female* or *male*
- `Age`: years
- `Education`: *iliterate*, *primary*, *Secondary*, *Tertiary* or *unknown*
- `Occupation`: one of ten categories
- `method`: one of nine methods

It is important to notice that the population in the study is predominantly rura and that the limitation of the study is that the incidence estimates are likely to be underestimated due to underreporting in both surveillance systems.

Source

Sun J, Guo X, Zhang J, Wang M, Jia C, Xu A (2015) "Incidence and fatality of serious suicide attempts in a predominantly rural population in Shandong, China: a public health surveillance study," BMJ Open 5(2): e006762. https://doi.org/10.1136/bmjopen-2014-006762

Data downloaded via Dryad Digital Repository. https://doi.org/10.5061/dryad.r0v35

Analysis

In this section we will carry out our analysis following the bayesian approach, first developing two different models to analyse the data, assessing their convergence, doing posterior predictive checking, comparing them to choose the one that performs best and eventually use the obtained model to select the most influencial factors and answering our analysis problem.

Model description

In order to evaluate the factors which influence the probability of SSA the most it is an obvious chioice to develop a multiple logistic regression model. Two different models have been implemented which will then be compared in the following analysis:

- **simple logistic regression model** with uniform priors and with no distinction between years and
- **hierarchical logistic regression model** where we divide our data into three groups (one for each year) and then develop our model defining priors in a hierarchical manner.

Simple Logistic Regression Model

Hierarchical Logistic Regression Model

Prior choices

Stan Code

Here we implement our models using `Stan`.

```
## Create Stan data
dat <- list(N        = nrow(mydata),
            p        = ncol(mydata) - 2,
            died     = as.numeric(mydata$Died),
            urban    = as.numeric(mydata$Urban),
            year     = as.numeric(mydata$Year),
            season   = as.numeric(mydata$Season),
            sex      = as.numeric(mydata$Sex),
            age      = as.numeric(mydata$Age),
            edu      = as.numeric(mydata$Education),
            job      = as.numeric(mydata$Occupation),
            method   = as.numeric(mydata$method))
```

Simple model

```
## SIMPLE LOGISTIC REGRESSION MODEL

## Load Stan Model
fileNameOne <- "./logistic_regression_model.stan"
stan_code_simple <- readChar(fileNameOne, file.info(fileNameOne)$size)
cat(stan_code_simple)
```

```
## data {
##   // Define variables in data
##   // Number of observations (an integer)
##   int<lower=0> N;
##
##   // Number of parameters
##   int<lower=0> p;
```

```
##
##   // Variables
##   int  died[N];
##   int<lower=0>  urban[N];
##   int<lower=0>  year[N];
##   int<lower=0>  season[N];
##   int<lower=0>  sex[N];
##   int<lower=0>  age[N];
##   int<lower=0>  edu[N];
##   int<lower=0>  job[N];
##   int<lower=0>  method[N];
## }
##
## parameters {
##   // Define parameters to estimate
##   real beta[p];
## }
##
## transformed parameters  {
##   // Probability trasformation from linear predictor
##   real<lower=0> odds[N];
##   real<lower=0, upper=1> prob[N];
##   for (i in 1:N) {
##     odds[i] = exp(beta[1] + beta[2]*urban[i]  + beta[3]*year[i] +
##                           beta[4]*season[i] + beta[5]*sex[i]  +
##                           beta[6]*age[i]    + beta[7]*edu[i]  +
##                           beta[8]*job[i]    + beta[9]*method[i] );
##     prob[i] = odds[i] / (odds[i] + 1);
##   }
## }
##
## model {
##   // Prior part of Bayesian inference (flat if unspecified)
##
##   // Likelihood part of Bayesian inference
##     died ~ bernoulli(prob);
## }
```

Hierarchical model

```
## HIERARCHICAL LOGISTIC REGRESSION MODEL

## Load Stan Model
fileNameTwo <- "./logistic_regression_model.stan"
stan_code_hier <- readChar(fileNameTwo, file.info(fileNameTwo)$size)
cat(stan_code_hier)
```

```
## data {
##   // Define variables in data
##   // Number of observations (an integer)
##   int<lower=0> N;
##
##   // Number of parameters
##   int<lower=0> p;
```

```
##
##   // Variables
##   int  died[N];
##   int<lower=0>  urban[N];
##   int<lower=0>  year[N];
##   int<lower=0>  season[N];
##   int<lower=0>  sex[N];
##   int<lower=0>  age[N];
##   int<lower=0>  edu[N];
##   int<lower=0>  job[N];
##   int<lower=0>  method[N];
## }
##
## parameters {
##   // Define parameters to estimate
##   real beta[p];
## }
##
## transformed parameters  {
##   // Probability trasformation from linear predictor
##   real<lower=0> odds[N];
##   real<lower=0, upper=1> prob[N];
##   for (i in 1:N) {
##     odds[i] = exp(beta[1] + beta[2]*urban[i]  + beta[3]*year[i] +
##                           beta[4]*season[i] + beta[5]*sex[i]  +
##                           beta[6]*age[i]    + beta[7]*edu[i]  +
##                           beta[8]*job[i]    + beta[9]*method[i] );
##     prob[i] = odds[i] / (odds[i] + 1);
##   }
## }
##
## model {
##   // Prior part of Bayesian inference (flat if unspecified)
##
##   // Likelihood part of Bayesian inference
##     died ~ bernoulli(prob);
## }
```

Convergence Analysis

Posterior Predictive Checking

Model Comparison

Sensitivity Analysis

Conclusions

Problems encountered

Potential improvements