WORLDWIDE ANALYSIS OF SUICIDE RATES

Yichen Zhu
Department of Statistical Science
Duke University
April 27, 2019

ABSTRACT

Suicides have become an unignorable problem in modern society. Researchers collected suicide data in 101 countries and regions between 1985 and 2016, with the goal of studying the relation between suicide rates and covariates including sex, age, generation(birth year), year, country and GDP per capita. After discovering geological pattern of suicide rates among countries in EDA process, we decide to use a conditional autoregression model to incorporate the spatial information. We use a poisson regression framework, while let the country random effects follow a Conditional AutoRegression prior. Additional conditions are enforced to address identifiable problems. Our results show that male and older people are more likely to suicide than female or younger people, and former Suviet Union members and northest European countries have higher suicides rates than most other countries. We also observe higher suicide rates between 1994 and 2000, while the generation(birth year) and GDP has little influence on suicide rates.

1 Introduction

Ever since human being had recorded history, there have been suicides. And it have drawn more and more attention due to the fast development of technologies and new pressure of modern life. A recent study collected suicide data in 101 countries and regions between 1985 and 2016. For each country and each year, the whole population is divided into multiple subsets. People in each subset belong to the same sex, age group and generation(birth year) group. We have 6 age groups and 6 generation groups. It's worthwhile to note that age, generation and year forms a "deterministic loop" in the sense that given any two of them, the remaining variable can be determined. However they have different scocial scientific interpretations, and it make sense to analysis them simultaneously under some conditions. Population size and number of suicided people are recorded on subgroup level, as well two additional covariates: GDP and HDI (human development index). We delete HDI because more than 2/3 of HDI data are missing, and no relation between HDI and suicide can be seen in the observed data. We finally have 26K subgroups, each with 5 categorical covariates and one continuous covariate. The objective of our study is to interpret the relationship between suicide rates and country, year, sex, age, generation (birth year) and GDP. Since the data is very abundant, it's essential to visualize the relationship between suicide rates and each covariates.

2 EDA

We first explore the marginal relatioship between suicide rate and sex, age, generation(birth year) and GDP. We treat each subgroup of a certain country, year as one sample. For sex, age and generation, we plot the boxplots for suicide rates of all samples clustered by these categorical covariates. For GDP, we plot the scatterplot of suicide rates and GDP per capita. The results are shown in Figure 1. We can see there are clear trends in the marginal relationship between suicide rates and covariates. Male, older people and people of earlier generations tend to have higher suicide rates than female, younger people and people of later generations. However we should be aware of the dependency between age and generation: people of later generations, for example generation Z, are impossible to be very old when the survey is conducted. It will be more helpful to explore the conditional relationship, which will be done in the modeling procedure. For GDP, it seems countries of lower GDP per capita tend to have higher suicide rates, as we see all countries with suicide rates higher than 100 per 100k population have GDP per capita lower than 5×10^4. However it's also possible that these higher suicide rates are due to randomness: there are many countries in the world, like island countries in Caribbean Sea, which have both low GDP per capita and small population (as small as 250 in subgroup level). These countries are very likely to produce outliers in suicide rates. In contrast, countries of high GDP per capita usually have considerable amout of population and little randomness in suicide rates.

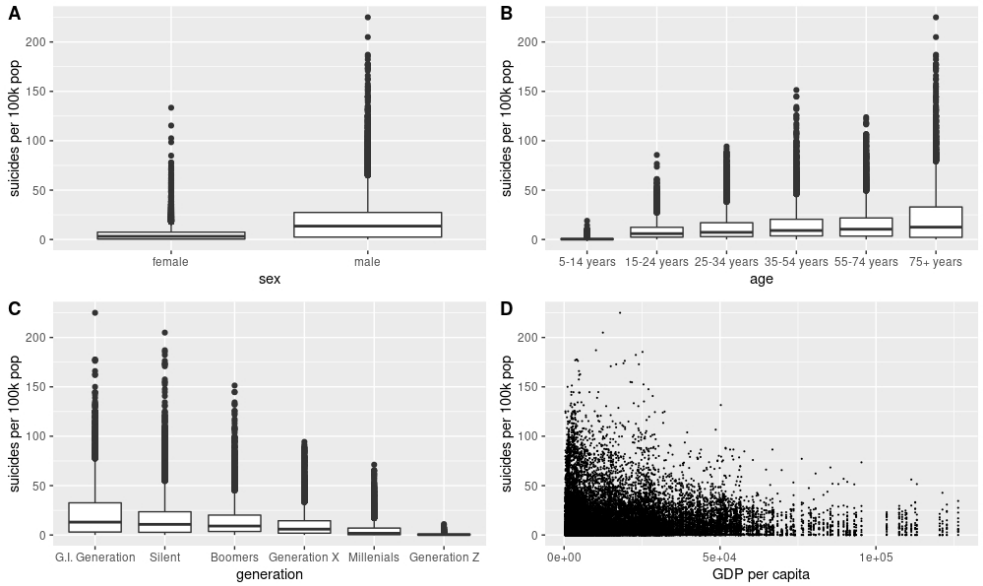

Figure 1: Marginal relationship between suicides rates and sex (figure A), age (figure B), generation (figure C) and GDP (figure D). Each dot/sample in the figure represent one subgroup.

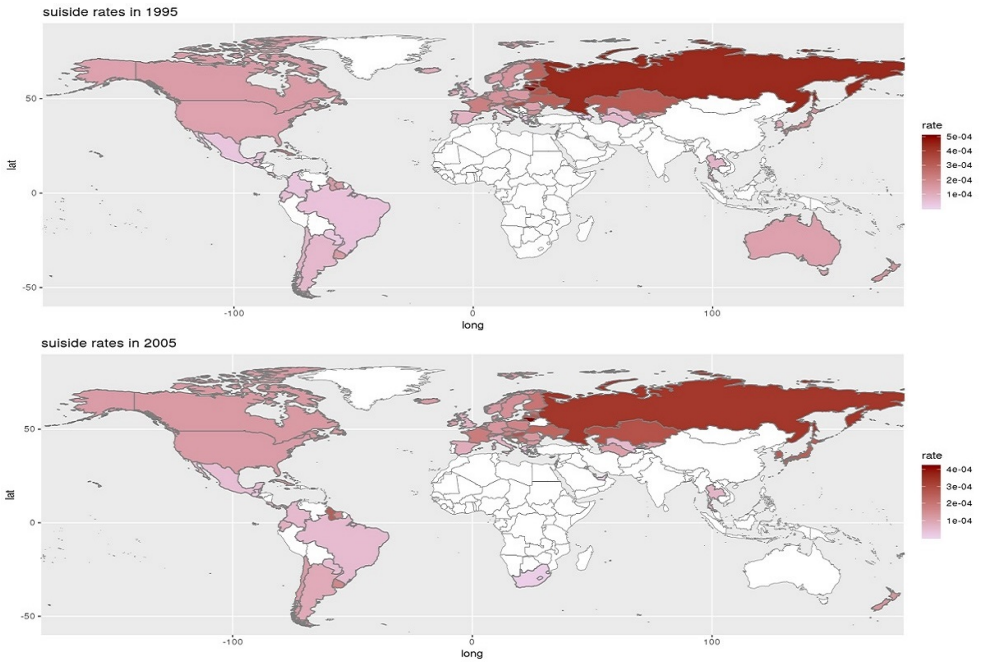

Figure 2: Country and year specific suicide rates for all countries in 1995 and 2005.

We then examine the relationship between suicide rate and country, year. We aggregate subgroups of the same country, year and compute their combined suicide rates, then visualize them by their geological location in the world. The results for year 1995 and 2005 are shown in Figure 2. We can see Russian and eastern European countries tend to have a higher than average suicide rates. There are some minor differences among years, but the major geological pattern is consistent in these two years. In fact, most years between 1985-2016 yield similar patterns. This indicates the necessity of a spatial model which we will describe in the next section.

3 Conditional Autoregression Model

As we have done in the EDA part, we make each subgroup data as one sample. Denote the suicide rate for ith sample as y_i, its country as $s(i)$, its year as $t(i)$, all other covariates as x_i. Specifically, x_i includes constant 1, sex, age group, generation and GDP per capita. Since binomial converges to poisson when $np \to \lambda, n \to \infty$, we can model the suicide rate as

$$y_i \sim \text{Poi}(E_i \exp(x_i^T \beta + u_{s(i)} + v_{t(i)})),$$

Where E_i is the base rate of suicide, proportional to the population in that sample. Since we have constant 1 in x_i, multiplying all E_i by any constant won't make any difference. $v_{t(i)}$ is year-specific random effect. For any v_j ($j = t(i)$ for some i. Note multiple samples can have the same $t(i)$), We simply assume independent gaussian prior

$$v_j \sim N(0, \sigma_t^2).$$

$u_{s(i)}$ is country specific random effect. u_j ($j = v(i)$ for some i) follows a Conditional AutoRegression Model (CAR):

$$u_j | u_{k \neq j} \sim \text{N}\left(\frac{\sum_{k \neq j} c_{kj} u_k}{\sum_{k \neq j} c_{kj}}, \frac{1}{\tau \sum_{k \neq j} c_{kj}} \right), \tag{1}$$

$$\tau \sim \text{Gamma}(a, b). \tag{2}$$

Where c_{kj} is a measure of geological closeness of two countries j, k. Conditional distributions alone can't specify a valid joint distribution as the covariance matrix is degenerated. We need an additional constraint $\sum_j u_j = 0$. We assume no prior (or flat prior) of β, that is $p(\beta) \propto 1$.

Denote total number of countries as n_c, then it remains to specify all the constants c_{kj}, which forms an Adjacency Matrix $C = (c_{kj})_{n_c \times n_c}$. We enforce its diagonal elements to be zero, and set the off diagonal elements as (it's symmetrical by definition):

- $c_{jk} = 1$ if both countries j, k are small and geologically adjacent, e.g., Estonia and Latvia.
- $c_{jk} = 0.5$ if both countries j, k are small and close (not adjacent), or at least one country is large and adjacent. E.g., Estonia and Lithuania, Estonia and Russia.
- $c_{jk} = 0$ if countries j, k are too far away.
- We will enforce $\sum_{k \neq j} c_{kj} \geq 0.5, \ \forall j$. This bring $c_{jk} = 0.5$ to some remote countries, like Maldives and Mauritius.
- Some region needs special attention, like island countries in Caribbean Sea.

Model Fitting Procedure As a standard Bayesian sampling procedure, we employ Gibbs Sampling framework while use Random Walk Metroplis Hasting in each Gibbs Update. In general, when updating parameter α (α can any one dimension of β, v, u), we adopt the following proposal distribution

$$\alpha^* | \alpha \sim N\left(\alpha, \frac{2.4}{I_{\alpha_0}(X, Y)} \right),$$

where $I_{\alpha_0}(X, Y)$ is the fisher information with respect to α, but computed using a prespecified value α_0. For any one dimensional of β, u, v, we set $\alpha_0 = 0$. We then apply the constraints $\sum u_j = 0$ "on the fly" by subtracting mean of u_j after metroplis hasting of all u_j. The subtracted term is added to intercept coefficient to ensure the invariance of likelihood. Because the CAR prior is constructed on the relative differences of u_j, the prior is also unchanged, which finally reaches an unchanged posterior.

Deterministic "Loop" As mentioned in Introduction, age, generation, year forms a deterministic loop, but we still have individual effect for each of them. Mathematically, our model makes sense as long as their effects are not additive. This is true in practice. For example, two people of different ages suicide at the same year. It's not only their age at that year matters, but also their generations: the person from Silent generation killed himself in his sixties maybe because his dark childhood memories of World War Two finally grow unbearable, while another person from Generation X ended his own life may due to the pressure of modern world.

4 Results

This section will show the relationship between suicide rates and all the covariates from our model, with tuning parameters $a = 5, b = 10$. All the suicide rates are displayed in relative scale, with the base suicide rate being 5.051×10^{-5}. That is, relative rate r means the true suicide rate is $r \times 5.051 \times 10^{-5}$. We first focus on fixed effects for sex, age, generation and GDP. The posterior mean and 95% confidence interval for all the coefficients are shown in Table 1. For sex and age, what we get in our model is nearly the same we get from EDA: male and older people are more likely to suicide than female and younger people. There is a dramastic increase between 5-14 YO and 15-24 YO, which matches common sense as infants and childs are rarely heard to kill themselves. The relation between suicide rates and age looks strictly increasing, which worths further investigations. The relation between GDP and generations are not quite the same as we have seen in EDA. The exponential form coefficient of GDP is so close to one compared to others, indicating the suicide rate is nearly independent of GDP. The relation between suicide rates and generation are much more complicated as we have seen in the marginal EDA plot. This means after accounting for the correlation between generation and age, later generations, especially Generation Z, have a higher suicide rate than earlier generation.

	mean	96% confidence interval
GDP	0.9971	(0.9970, 0.9972)
male	1.9454	(1.9438, 1.9471)
female	0.5140	(0.5136, 0.5145)
5-14 YO	0.0634	(0.0628, 0.0640)
15-24 YO	1.0110	(1.0066, 1.0155)
25-34 YO	1.4492	(1.4449, 1.4535)
35-54 YO	1.8218	(1.8173, 1.8264)
55-74 YO	2.0229	(2.0133, 2.0339)
75+ YO	2.9202	(2.9027, 2.9409)
G.I. Generation	1.0260	(1.0170, 1.0338)
Silent	0.8785	(0.8729, 0.8833)
Boomers	0.9074	(0.9038, 0.9111)
Generation X	0.9317	(0.9279, 0.9355)
Millenials	1.0070	(1.0002, 1.0131)
Generation Z	1.3031	(1.2798, 1.3246)

Table 1: Exponential form of fixed effects for GDP, sex, age group and generation.

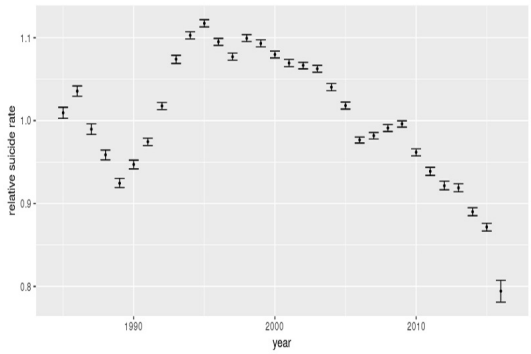

Figure 3: Year-specific random effects in exponential form.

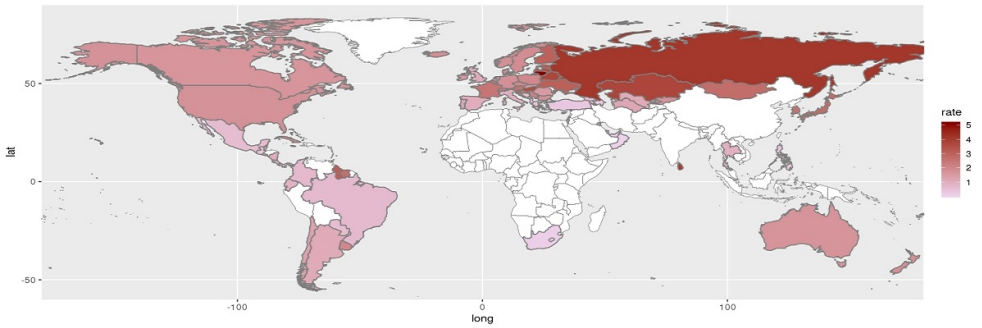

Figure 4: Country-specific random effects in exponential form.

We then discuss the relation between suicide rates and country, year. The year, country random effects are displayed in Figure 3, 4 respectively. We can see there is a significant "peak" of year-specific random effects in the late 1990s, but we don't know how to interpret this. The most influential crisis in social economical history at that time may be the asian finacial crisis, however, few asian countries are recorded in our dataset. For country random effects, we see a consistent pattern as in the EDA part: Former Suviet Union countries and eastern European countries suffer from relative high suicide rate than most other places in the world. It's surprising that Sri Lanka also has a very high suicide rates, and we don't have much geological information for Sri Lanka as India's data is not available here. Our country random effects do capture the spatial pattern we have seen, but it's unclear whether this spatial pattern is an inherent feature of underlying true distribution, or is imposed by our CAR prior. This will be addressed by comparing other two models in the next section.

5 Comparison: Strength of Spatial Dependency

Our proposed model in section 3 incorporate the spatial structure among countries by a CAR prior with moderate prior on precision τ. It would be interesting to consider other two variants: one has no spatial dependency, and the other has a strong spatial prior. For the spatial-independent model, we assume country random effects u_j comes from i.i.d. Normal distribution:

$$u_j \sim \mathrm{N}(0, 1/\tau_s), \quad \tau_s \sim \mathrm{Gamma}(a, b)$$

with $a = 5, b = 10$. For the strongly spatial-dependent model, we still assume a CAR prior on country random effects u_j the same as in 1, but has a very strong gamma prior on τ:

$$\tau \sim \mathrm{Gamma}(10^8, 10^4).$$

We do the same Gibbs-Metropolis Hasting framework as before. The posterior mean of country random effects are shown in Figure 5. Even with absurdly strong spatial dependency, we still find former Suviet Union and eastern European countries to have higher suicide rates. When no spatial dependency is assumed, we get nearly the same results as in section 4. In that sense, we can claim the spatial dependency structure is an inherent feature of underlying true model, and insensitive to prior settings as long as prior is not too powerful.

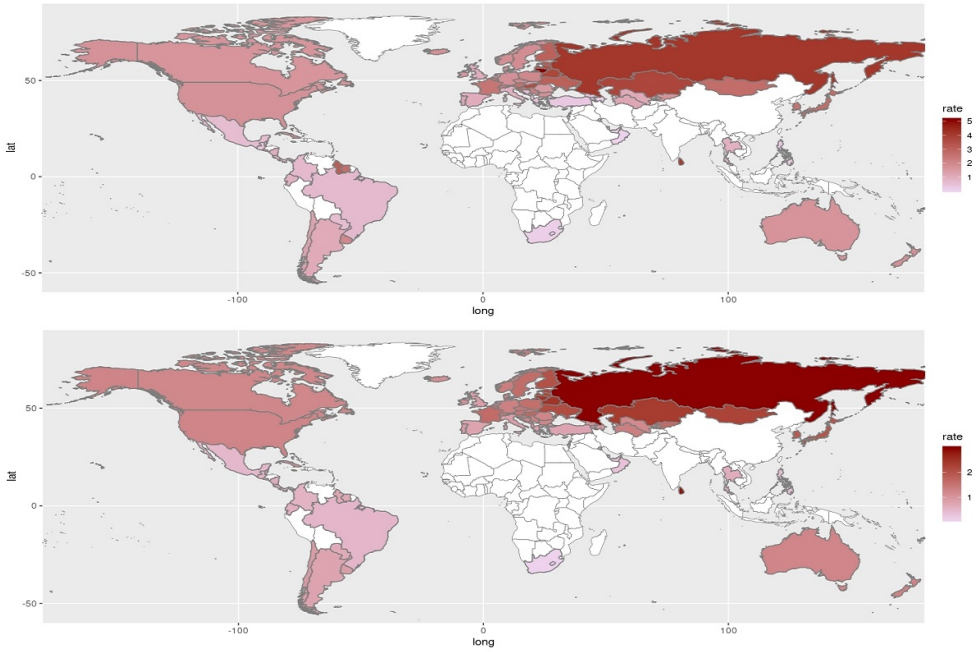

Figure 5: Country specific random effects. Top: Spatial independent model; Bottom: Strongly spatial dependent model.

6 Conclusion and Discussion

We discovered spatial pattern in suicide rates of different countries, and employed a Conditional AutoRegression to model this pattern. Comparison among EDA and different variant models confirms that our model captures the major spatial features and is insensitive to prior choice. We conclude that former Suviet Union countries and eastern European countries, as well as Sri Lanka, have a relatively high suicide rate than other countries. For year random effects, we find a relative high suicide rate is late 1990s, which we can't explain right now. For fixed effects, male and older people tend to have a higher suicide rates than female and younger people; nearly no effects for GDP is discovered, and people of later generations, especially generation Z, tend to have a higher suicide rate than other generations after accounting for correlation between generations and age.

For future research, it may be helpful to collect more social - economical covariates which have hugh influence on one's life, like unemployment rate and gini index of a subgroup of people. Whether they have perticipated in a war may also be important. The human development index should be better collected to serve as a reliable covariate. In that sense we may not only discover which country has higher suicide rate, but also the reason behind these country random effects.