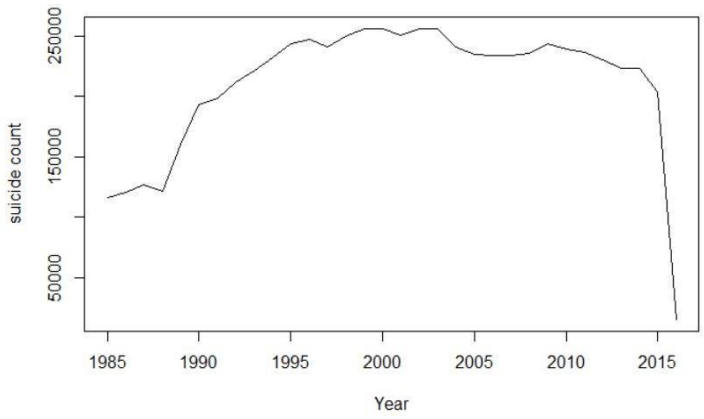
Stat 431 Final Report

Conrad Manaugh, Holiday Tang, Bradley Gibbons

Abstract

For our project we chose to analyze a dataset from Kaggle on suicide rates from the years 1985 to 2016[1]. The data included 27,820 observations among 12 different variables (details listed below). It is important to note that less than half of observations include HDI for the country at a given year, but the rest do not. Additionally, we came across values of zero under the suicide_no variable; we were not able to determine whether these were missing values or the true value of the variable, as some places with high population have 0 suicides. Regardless for our analysis we treat these as 0 suicides and not an NA observation. Per the description, "The dataset was compiled from four separate datasets--linked by time and place--built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum."[2] Therefore the purpose of this dataset is to understand what contributes to suicide. Hopefully by understanding what can cause suicide people in the future can avoid or limit the effects. The plot to the right shows a significant drop in suicide counts for the year 2016. This indicates there is most likely incomplete data so we excluded 2016 from the analysis.



[1] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
[2] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Variables

1. *Country*: Categorical variable for each country
2. *Year*: 1987 until 2015
3. *Sex*: Categorical, Male or Female (2 levels)
4. *Age*: Categorical, numeric age above 15 divided into intervals (6 levels)
5. *Suicide_no*: Integer, total number of suicides for this group of people
6. *Population*: Integer, population of this group
7. *Suicides/100k*: Numeric, number of suicides per 100,000 people
8. *Country-Year*: Categorical, indicator of both country and year
9. *HDI*: Numeric (between 0 and 1), Human Development Index; more than half are NA
10. *GDP_per_year*: Numeric, a country's GDP at this year
11. *GDP_per_capita*: Numeric, a country's GDP/total_population, for each year
12. *Generation*: Categorical, the generation a person who committed suicide is apart of

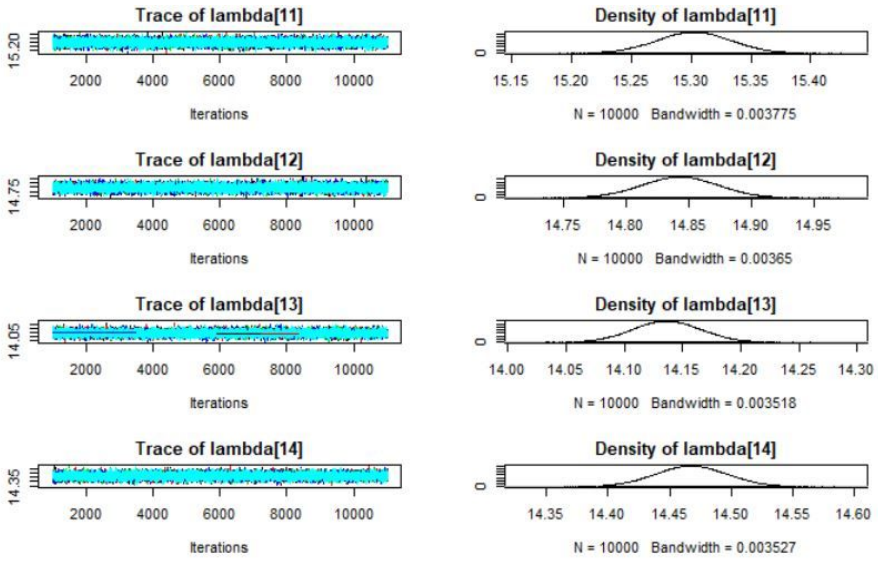
Overall Suicide Rate by Year

The first topic of interest in the present project is suicide rates across years from 1985 to 2015. By modeling a poisson rate model $Y_i|\lambda_i \sim indep.\ Poisson\ (\lambda_i T_i)$, with Y_i being number of suicides, λ_i being suicides per 100,000 population, T_i being the total $population/100,000$, and the index i denoting the $year - 1984$, that is $i = 1$ corresponds to the data of the year 1985.

Regarding λ_i, it is modeled as $\lambda_i \mid \alpha, \beta \sim i.i.d.\ gamma(\alpha, \beta)$.Instead of specifying fixed values for α, β, we applied a hierarchical model through defining hyperparameters $\alpha,\ \beta \sim indep.\ exp(1)$. The model we use is similar to the model used in the airliner fatalities example 9.2.

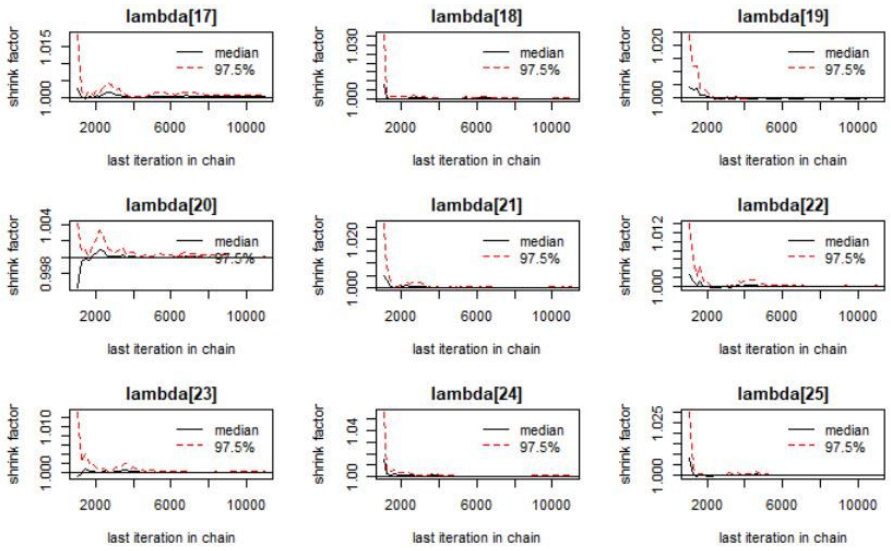
To achieve a suitable data structure, the variables "*suicide_no*", and "*population*" are grouped by year to obtain the total amount of suicides for each year and the total population for each year. In the model, the population variable is divided by 100,000 to cater to the interest for λ_i to be suicides per 100,000 population.

Note these variables do not comprehensively represent the world data since the dataset did not capture every country.



The values initialized for the model are α, β. 5 chains of initializations are utilized with the following values for (α, β): (1,1), (0.01,0.01), (100, 100), (0.01, 100), (100, 0.01). With the interest to inspect $\alpha, \beta, \lambda_i, mean(\lambda), var(\lambda)$, and a new lambda as an expectation for a new year--assuming the data is representative. 11000 initial iterations are run, with the first 1000 burn-in performed automatically, and following is a portion of the trace plot:
As we can see, the trace plots are stable around a certain value, and the density plots look reasonable, and this is the same across all λ_i, including the mean and variance of λ,

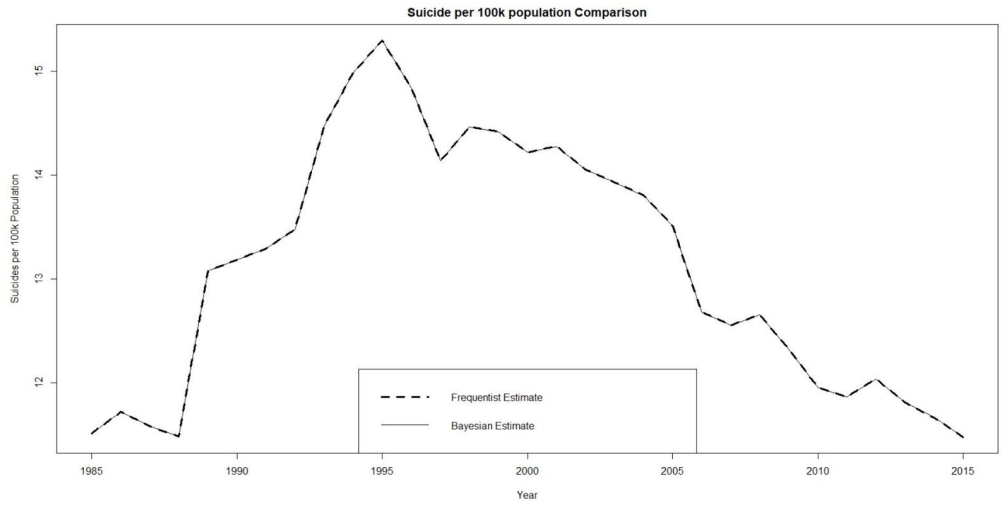
Since the time range of interest is 1985 ~ 2015, there will be 32 lambda values in toal. There are 5 chains in total. Following is one of the many trace plots from this model.
As illustrated, the trace plot shows convergence of the parameters of interest, and this is the same for all other parameters of interests. One other convergence diagnostic plot is the Gelman Diagnostic plot, here is one examined example:



As illustrated, all the gelmen statistics stably converge to 1, this trend is shown for all other parameters of interest, which provide evidence for convergence. Last but not least, most Gelman Statistics are 1 at its upper and lower bound of the confidence intervals for the first 10000 iterations after 1000 burn-ins. After 20000 more iterations, all Gelman Statistics turn out to be 1 at its upper and lower and lower bound of its confidence intervals, providing strong evidence of convergence.

It is important to point out that due to the number of parameters, including all diagnostic plots will be messy. Therefore, for further models, all diagnostic procedures for convergence are the exact sample replicate from the procedure mentioned in the previous paragraph, and all results provide evidence for convergence. If doubtful, please download data and run our codes for confirmation.

Regarding the results, number of suicides per 100,000 population shows an upward trend before 1995, and a downward trend after 1995, following is a graph for the mean of λ_i over the years. The frequentist estimate for the number of suicides per 100,000 population is also calculated, and these two measures are compared visually as follow:



As illustrated, the frequentist estimates almost identically match with the means of the bayesian estimates.

Regarding more details on the suicide rates across different years, following is a table of the confidence intervals for each lambda:

	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
λ_i	11.44, 11.57	11.65, 11.78	11.51, 11.64	11.42, 11.54	13.01, 13.14	13.12, 13.24	13.23, 13.34	13.41, 13.53	14.41, 14.53	14.92, 15.05

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
15.24, 15.36	14.78, 14.90	14.07, 14.19	14.41, 14.52	14.36, 14.47	14.16, 14.27	14.22, 14.33	14.00, 14.10	13.87, 13.98	13.74, 13.86	13.45, 13.56

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
12.62, 12.72	12.50, 12.60	12.60, 12.71	12.27, 12.37	11.90, 11.99	11.81, 11.91	11.98, 12.08	11.75, 11,85	11.61, 11.71	11.43, 11.53

As illustrated, all the confidence intervals are very small in range.

Comparing the Monte Carlo Standard Errors for each year to the standard deviations we see that for each year the standard deviation is more than twenty times the Monte Carlo standard error. This result indicates that the error is small for all years, and our analysis is not invalid because of hidden unseen errors.

	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
MCE	.00015	.00015	.00014	.00014	.00014	.00013	.00013	.00013	.00013	.00013

1995	1996	1997	1998	1999	2000	2001	2002	1003	2004	2005
.00013	.00013	.00012	.00012	.00012	.00012	.00012	.00012	.00012	.00012	.00012

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
.00011	.00011	.00011	.00011	.00011	.00011	.00011	.00011	.00011	.00011

	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
SD	.033	.033	.032	.033	.032	.029	.029	.029	.030	.031

1995	1996	1997	1998	1999	2000	2001	2002	1003	2004	2005
.030	.029	.028	.028	.028	.028	.028	.027	.027	.028	.027

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
.026	.026	.026	.025	.024	.024	.025	.025	.024	.025

In conclusion, our analysis shown that the suicide rate from 1985 to 2015 demonstrated the following trend: before 1995, suicide rates grew as years passed, and after 1995, the suicide rates declined as years passed.

Group Differences in Suicide Rate

Moving on, we are also interested in discovering how suicide rate differs among groups of people. In order to analyze this we will use a similar model to before, as we assume suicides happen rarely and are independent of each other. Different from before is that the data we entered will be sunsetted by gender and age bracket, so we can determine the difference between the many categories. For the priors we must specify alpha and beta, and we use 5

chains as these were enough to allow the Gelman Diagnostics to show convergence. Convergence occurs in the same manner as shown before, so will not be graphically displayed.

For the initialization, as well as the model, we use the same model from our prior section, similar to the class example ex9.2 model1 about airlines. The only difference between the models is we divide $t[i]$ by 100,000. In order to compare the average suicide rate we run the same model 12 times. This is because we have 6 age brackets and 2 genders. If we provide the data for the amount of suicides per 100,000 from these categories for each year we get 12 two column data frames containing amount of suicides and population for each year given the category. We run the same model on each subset and get the average lambda for each category, allowing us to compare and contrast the age brackets and genders.

Running the same model (as previously mentioned), multiple times on the 12 subsets we see clear convergence. This observation is backed up by the Gelman Diagnostics which show 1 for each model indicating convergence.

Using the quantiles obtained from the summary found in the appendix we see the difference between rates for those of different ages and sexes. These findings are summarized in the table below.

.95 λ_{mean} credible	Age 5-14	Age 15-24	Age 25-34	Age 35-54	Age 55-74	Age 75+
Male	[.73, .97]	[13.01, 15.82]	[19.52, 23.74]	[24.64, 30.24]	[27.82, 33.78]	[42.59, 52.20]
Female	[.37, .53]	[3.77, 4.61]	[4.57, 5.56]	[6.43,7.84]	[8.4, 10.5]	[12.25, 15.46]

Looking at the summary table above we see clear and obvious trends for the rate of suicide of each group. Within gender it looks like rate of suicide is increasing, as each age bracket has a higher rate of suicide then the previous bracket. Perhaps this is because quality of living goes down as we age, and the amount of traumatic events one goes through rises as we age. It makes sense for the young to not commit suicide as they do not have an understanding of the world good or bad.

Even more surprising, then, is that the increase in suicide rate as a person gets older is far larger for men compared to women. In every age bracket we see that the rate of suicide for men is twice--or even more than twice the rate of suicide for women. Such a drastic difference is truly substantial, so we can say with good confidence that men commit suicide at higher rates

than women. This could be because in parts of the world men have more demanding jobs physically, or face more social pressures to be the leader of the family unit.

We can compare our bayesian results with frequentist estimates. A frequentist estimate of the mean of the rate of suicides can be calculated by simply dividing the total suicides by the total population and taking the mean for all the years. We see a table of these results below, pulled from R commands in the appendix.

Frequentist λ_{mean}	Age 5-14	Age 15-24	Age 25-34	Age 35-54	Age 55-74	Age 75+
Male	0.81	14.32	21.5	27.27	30.62	47.14
Female	0.41	4.13	5.01	7.07	9.35	13.74

As we can see when comparing the frequentist mean estimates (which is just the true observed rate) with the bayesian credible intervals, all of the frequentists means are inside the credible intervals. This result gives good evidence that our bayesian methodology is correct, as it matches the frequentist results.

The last thing worth mentioning is that the Monte Carlo Time-Series SE is pretty low for all models. It barely ever goes up to 0.01, and for the young age brackets it is even below 0.0001. We can also determine that the monte carlo error rises as the rate of suicide rises, so even though some errors are bigger than others they are all very small in comparison to the statistic that we measured. For all Monte Carlo Errors shown the Standard Deviation is over twenty times its Monte Carlo Error, so they are all sufficiently small, refer to the SD table to compare.

Time-Series se	Age 5-14	Age 15-24	Age 25-34	Age 35-54	Age 55-74	Age 75+
Male	.00026	.00317	.00479	.00631	.00676	.01095
Female	.00017	.00096	.00112	.00160	.00233	.00375

SD	Age 5-14	Age 15-24	Age 25-34	Age 35-54	Age 55-74	Age 75+
Male	.0596	.7127	1.0717	1.419	1.5192	2.4446

Female	.0397	.2188	.2514	.3575	.5197	.8173

Summary

To sum up, the analysis were interested in modeling the rate of suicide, to be specific, modeling the number of suicides per 100,000 populations. Our major area of interest include, suicide rate over the years, and across different categorical variables.

Regarding the changes over the year, we found an upward trend before 1995, indicating more people tend to commit suicides during that range of time, and an downward trend afterward, indicating decline in suicide attempts.

When comparing the rate of Suicides for different types of people based on sex and age we see a clear trend. These trends exist for both frequentist and bayesian implementations. As people age they become more likely to commit suicide. Also males are much more likely to commit suicide than females. Men commit suicide at over 3 times the rate compared to women consistently, except at ages before puberty where they only have double the suicide rate of women.

Regarding potential improvements if the team has more time. Given more time we could compare how each categories suicides change overtime. We know the trend changes around year 1995 for the total suicides, but perhaps different age brackets or genders experience the trend change at different times, or different intensities. We could attempt to compare categories such as country, and continents. Also, other numeric variables in the data include the human development index, and GDP-related measures. It will be interesting if we can formulate an model to investigate the suicide rate changes in a regression framework to understand how the numeric variables contribute to the changes.

Appendix: