From the World Health Organization data, on average, 800 000 people die due to suicide every year. Suicide is the second leading cause of death among 15–29-year-olds. Suicide does not just occur in high-income countries but is a global phenomenon in all regions of the world. Yet, only 38 countries report having a national strategy for suicide prevention. The Eastern Europe and East Asia regions have the highest suicide rate worldwide.

To study our research question, we use a database called Suicide Rates Overview 1985 to 2016 from the WHO obtained from Kaggle. The data compares socio-economic information with suicide rates by year and country.

There are 12 variables in the raw data: country, year, sex, age, suicide_no, population, suicides/ 100k pop, county-year, HDI for year, gdp_for_year ($), gdp_per_capita ($), and generation. The variable **year** ranges from 1985 to 2016. A**ge** includes five groups: 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years. **Suicides/ 100k pop** is the standardized suicide rate. **Country-year** is the combination of country and year in a format like "Albania1985". **HDI for year** is the Human Development Index for the year. It is a statistical composite index of life expectancy, education, and per capita income indicators. **Generation** includes G.I. Generation, Silent, Boomers, Generation X, Millennials, and Generation Z and is based on age group.

This paper aims at discovering the important candidates and measuring their impacts for suicide count in order to make suggestions on suicide prevention in the future.

Before doing the prediction, we do the following literature review: There is a strong negative correlation between suicide rates and GPD per capita. Males were between 1.7 and 2.1 times more likely to commit suicide than females in 2015. Suicide rates have increased by 60% globally in the last 45 years.

Based on that, we predict that higher GDP is associated with lower suicide count. We predict that male will commit suicide of a rate triple that of females. We predict that there is an upward trend for suicide rate over time, meaning that for every additional year, suicide rate increases.

Before fitting a model to the data, we can plot the response against the predictors to explore the data. We begin by plotting a box plot of suicide rate per 100 thousand people by sex. From the graph, males typically have larger values of suicide rate than females. Both males and females have longer upper whiskers than lower whiskers, meaning that both sexes have suicide rates more varied among the most positive quartile group.

Next, we construct box plots for suicide rate per 100 thousand people by age and generation. In general, higher age is associated with higher suicide rates. People older than 75 have the most varied suicide rate in the second to third quartile and in the upper whisker among the other age groups. They also have the highest amount of extremely large suicide rates. Compared to younger age groups, this group is possibly suffering from more physical illness and psychological obstacles, which may give them greater motivation to commit suicide. The second most varied is the 35 to 54 years group.

Considering the generation plot, older generations are associated with larger suicide rates.

We then plot time series of suicide rate per 100 thousand people for various countries, and here we only show four as an example. 11 counties including Brazil have an upward trend. 11 countries including Finland have a downward trend. 5 countries including the United States have a downward then upward trend. 8 countries including Japan have an upward then downward trend. Other countries don't have an obvious trend in time.

Next, we plot suicide rate per 100 thousand people by GDP. Since the scale of GDP is not friendly to display, we consider the log transformation. We now see perhaps a weak linear relationship between suicide rate and log (GDP). This makes sense because different countries have different patterns of suicide rate over years, altogether, the suicide rate may not have a strong relationship with GDP.

Then we begin building a model to our data. A typical linear model is not appropriate for the suicide data because: First, the suicide data are count data. Therefore, the response variable is not continuous. Second, the variance of the suicide data is unlikely to be constant, as count data typically has larger variability for larger counts. They both violate the assumptions for a linear model.

However, generalized linear models cover both situations by allowing for response variables that have arbitrary distributions other than simply normal distributions, and for the link function to vary linearly with the predicted values. Therefore, the generalized linear model for Poisson data is suitable for the suicide data.

There are three components of a GLM: the random component, the linear predictor, and the link function. The random component is a response variable y with independent observations $y_1, y_2, \dots, y_n$. The random component must also be a member of an exponential family such as the Poisson distribution.

The link function transforms the mean of the response and connects it to the linear predictors.

The suicide data has a Poisson response, *where* n_i is population size, and λ_i is suicide rate. $n_i\lambda_i$ equals the mean suicide count for observation i. For Poisson distribution the link function is often log. We connect the suicide rate to the linear predictors: $\log\left(\frac{E(count)}{population}\right) = X\beta \rightarrow \log(count) = \log(population) + X\beta$. In order to get the suicide rate, we adjust for the population size by adding the term offset(log(population)) in the implementation regression in R.

Initially there were 12 variables in the raw data. We drop HDI for it is not applicable for 2/3 of the countries. We are examining the suicide count as a count number, so, suicides/ 100k is dropped. For the same reason, we keep gdp_for_year ($) instead of gdp_per_capita ($). We drop the country-year variable and keep country and year. We drop both generation and age because they are categorical regressors. We use an interaction variable country*year to examine the additional year effect on each country, and it automatically includes two variables: country and year with respect to model hierarchy. We use AIC and BIC to check for different combination of interaction variables and find the model with sex, country * year, and log (GDP) the best. We also test for the significance of the regression coefficients by comparing this full model with sex, country, year, country * year, log (GDP), to different reduced models using the Analysis of Deviance method which give us the same result.

GLM makes several assumptions. The data must be independent. The response must be a member of an exponential family distribution with mean μ_i. There should be no existence of high leverage outliers to influence the fitness of the model to the data. The suicide count data is from a Poisson distribution, which belongs to the exponential family of distributions. So, the 2nd assumption is checked. For the rest of the assumptions we need to check them in the following sessions one by one.

We drop 8 countries that have perfect multicollinearity problem between the country*year variable and year variable. They all have less than 10 years' data for suicide count. The correlation between GDP and year is small. There is no serious collinearity problem between them.

From the marginal model plot, neither the data-driven nor the model-driven smooths fit the data very well, meaning that we are doing a bad job on describing the relationship between suicide count and the regressors year and log (GDP).

The added variable plot indicates that the negative relationship between suicide count and the variables log (GDP) and year is weak.

The residuals versus fitted plot has a non-constant thickness. This makes sense because for a Poisson model, the mean of response equals to its variance, and variance increases as the mean increases.

Both the Successive Residual plot and the Durbin-Watson test yields similar results that there is very strong evidence of positive serial autocorrelation in our residuals. So, the 1st assumption for GLM is violated, the data may not be independent in time. Since the suicide count data is a time-ordered data that measured yearly from 1985 to 2016, the previous year observation can have an influence on the current observation, which may be the cause of the autocorrelation in our residuals.

From the influence plot, all these five observations have unusually large residuals and hat values. Netherlands1990 male has large hat-value, which means that it has large leverage. Kazakhstan1993 male has large studentized residual, so it could be an outlier. All five observations are candidates of being influential observations. So, the 3rd assumption for GLM may not hold as there are some high leverage outliers that influence the fitness of the model to the data. After dropping the observations with those five countries, the coefficient for sex decrease by 7%;

year decrease by 25%, and GDP decreased by 32%, meaning that these five countries are influential observations. However, as part of the population of interest, they are kept in the model.

So, this is our model. To make it easier to interpret, we raise both sides of the equation to the power of e.

With other factors being constant, we estimate that males have 3.56 times of suicide count of females. The result coincides with our prediction that males will commit suicide of a rate triple that of females. From the WHO, females tend to have higher rates of reported nonfatal suicidal behaviors, while males have a much higher rate of completed suicides, which may explain why males have two times higher suicide count than females.

With other factors being constant, a 1% increase in GDP is associated with a 0.241 % decrease in suicides than the previous year. This is a very small scale. As showed previously, different countries have different patterns of suicide rate over years, altogether, the suicide count may not have a strong relationship with GDP. The result coincides with our prediction that higher GDP is associated with lower suicide rate. This makes sense as a higher GDP country is usually associated with better health care and mental care systems. Suicide also results from many sociocultural factors such as unemployment, which could also explain their negative association.

Since Albania is the reference level country by default, when we are examining Albania, the value of country equals 0, otherwise country equals 1. To analyze the regressor year, with other factors being constant, we estimate that for Albania, each additional year is associated with 1.03 times more suicides than the previous year. For other countries, each additional year is associated with an extra $e^{\widehat{\beta_2}} * e^{\widehat{\beta}_{country,year}*country*year}$ more suicides than the previous year. For each additional year, in comparison with Albania, a country is associated with $e^{\widehat{\beta}_{country,year}*country*year}$ times additional effect of change in expected suicide count.

For instance, for Finland, each additional year is associated with 1.09 times more suicides than the previous year, meaning a 9% increase in suicide count. In comparison to Albania, for Finland, each additional year is associated with 1.05 times additional effect of change in expected suicide count, meaning that a 5% increase in suicide count related to Albania. The result coincides with our prediction that suicide count increase by year. These interpretations assume we can hold other variables constant, which is not the case.

We also interpret Denmark males and United States females in year 2015 as examples. On average, we estimate that Denmark males in year 2015 have 446 suicide count, or 16 suicides per 10,000 people. The United States females in year 2015 have 8,757 suicide count, or 6 suicides per 10,000 people.

Now, we are all familiar with the Fight Club. Different from the Fight Club, in the United Kingdom, there is a club called Andy's Man Club. It aims at encouraging males to start taking about themselves, their feelings and difficulties, which is usually what males fail to do. In the club, it's okay to talk.

What can we learn from Andy's Man Club? According to the WHO, clinical depression, substance abuse and severe physical disease or infirmity are major causes of committing suicides. To address the suicide problem, we should focus on prevention, diagnosis, and treatment on committing suicides. Prevention involves restriction of access to common methods of suicide. Back to the age group boxplot. From the WHO, for the 15-24 years old, school-based interventions have been demonstrated to reduce the risk of suicide. For the 35 to 54 years old, adequate treatment of depression, alcohol and substance abuse can also help reducing their suicide rates. For people above 75 years old, more efforts surrounding community-based care and educational programs for primary health care providers on the identification and treatment of late-life depression may also help lowering the suicide rates.

The first improvement that can be made in our model is that we only have 5 predictors: sex, year, log (GDP), country, country*year. Among them we only have two numeric predictors: log (GDP) and year. There may be omitted variable bias existing, which may influence the accuracy of the coefficients and dilute the variances of the coefficients. Future model building may consider including more numeric predictors such as Gini Coefficient and Engel's Coefficient which measures the proportion of income spent on food or even HDI. As long as the predictor is applicable to most of the observations, otherwise it is not representative for the data.

The second improvement is that residuals are autocorrelated in time. Since suicide count is measured yearly, previous observation can influence current observation in a time series. Adding more valid numeric predictors might help with the problem but the problem is hard to be eliminate from a time series data.

The third improvement is the weak relationship between suicide count and regressors including sex, year, and log (GDP). Suicide count is subject to macro-economic trends, sociocultural condition and many other factors that are different across countries. Also, different countries have different patterns of suicide count over years. Altogether, suicide count may not have a strong relationship with GDP globally. Suicide count may not have a strong relationship with sex, or year globally, either due to the similar reasons.