Comparison of Methods to Model Suicide Rates by Country

Abstract

This paper is an applied statistical modeling experiment on the longitudinal suicide rate by each country in the world. We aim to apply different modeling approaches, including linear regression, linear mixed models (LMM) , generalized linear mixed models (GLMM) and generalized estimating equations(GEE) on the world suicide data. We examined both continuous target variable and binary target variable by transforming suicide rate into binary outcomes. We aim to study the relationship between the socioeconomic environment of a country and the suicide rate of its people. We also considered the influence of the characteristics of different sub-population in each country and the specific year when the suicide rate is measured. We measured the performance of different mothods and drew conclusions on the best performing modeling approach.

Keywords: Longitudianl study, LMM, GLMM, GEE, Suicide Rate

1 Introduction

Suicide is a tragic outcome of mental illness and a leading cause of death worldwide. It is both a complex and sensitive issue that researchers have studied for years in the hopes of reducing suicide rates worldwide. We are particularly interested in modeling how suicide rates vary by country over time.

Social and economic factors can influence a country's suicide rate [1]. There also may be crises specific to a country at a given time, such as viral outbreaks or wars, that cause widespread devastation, thereby leading to an increase in suicide rates. A recent study also found that sex is highly correlated with suicide rate, as men are more likely to commit suicide than men and have a higher mortality rate [4]. Lower GDP per capita is likely correlated with suicide rate, as people in low-income countries likely have less opportunities than those in high-income countries [3].

Other factors have been found to impact suicide on an individual level, including divorce, number of previous suicide attempts, and mental health disorders including bipolar disorder, depression, and schizophrenia [2]. However, most of these predictors can't be accounted for on the country-level, which is a limitation of our analysis.

In this paper, we propose modeling suicide rate by country as both a continuous and binary outcome. For each case, we propose a series of plausible models, including generalized estimating equations and generalized linear mixed models. We use different combinations of predictors including sex, year, population, GDP per capita, and generation to model suicide rates in different countries over time.

In the continuous setting, we will predict suicide rates using basic models like linear regression and also models geared toward dependent data like linear mixed models and generalized estimating equations. To model suicide rates as a binary outcome, we will

classify suicide rates as either high or low by taking the mean suicide rate over all countries over time and classifying rates above this value as high and rates below this value as low. In the binary outcome setting, we will use a mixed-effects logistic regression model and a conditional logistic regression model. We will compare these approaches by country over time and evaluate our results in terms of the relative standard errors of our estimates and interpretability.

2 Methods

We have included quite a few methods for comparison. Since we have both the continuous outcome suicide rates and the binary outcome which indicates whether the suicide rate is high or low, two sets of methods were attempted. Within each category, cross sectional and logitudinal methods were both considered. For the continuous case, we have used inear regression, gamma regression, including both a zero truncated and zero-inflated model, a linear mixed model, and a generalized estimating equation. For the binary case, we have included GLMM and the conditional logistic regression model. GLMM can model the probability of a given generation in a country having a suicide rate higher than the mean suicide rate in the dataset while accounting for country-specific random effects and multiple predictors. Conditional logistic regression, on the other hand, can be used to compare the odds of suicide between those who have died by suicide and those who have not, while adjusting for potential confounding factors.

2.1 Ordinary Least Squares

A linear regression model has been included to serve as a baseline model for the purposes of comparison. We define the following model:

$$\begin{aligned} y = \beta_0 + \beta_1 \times \text{year} + \beta_2 \times \text{age} + \beta_3 \times \text{sex} + \beta_4 \times \text{population} \\ + \beta_5 \times \text{GDP per capita} + \varepsilon \end{aligned}$$

In this model, y represents the suicide rate, each of the $\beta_0, \ldots, \beta_5$ represent the fitted regression coefficients for our predictors, including an intercept term, and ε is the model's prediction error.

2.2 Generalized Linear Model

Note that the distribution of suicide rates is right-skewed, as shown in Figure 1. Therefore, it may be appropriate to consider a gamma Generalized Linear Model (GLM) because this model is suitable for continuous, positive, right-skewed data. In particular, we consider a zero-inflated gamma GLM because we have a large amount of zero values in our dataset and the gamma distribution takes strictly positive values.

A zero-inflated gamma distribution models zero values as a binary random variable and models positive values using a gamma distribution. In other words, it is a mixture of a gamma and a degenerate distribution at zero.

Our GLM is defined as follows:

$$\begin{aligned} g(y) = \beta_0 + \beta_1 \times \text{year} + \beta_2 \times \text{age} + \beta_3 \times \text{sex} + \beta_4 \times \text{population} \\ + \beta_5 \times \text{GDP per capita} + \varepsilon, \end{aligned}$$

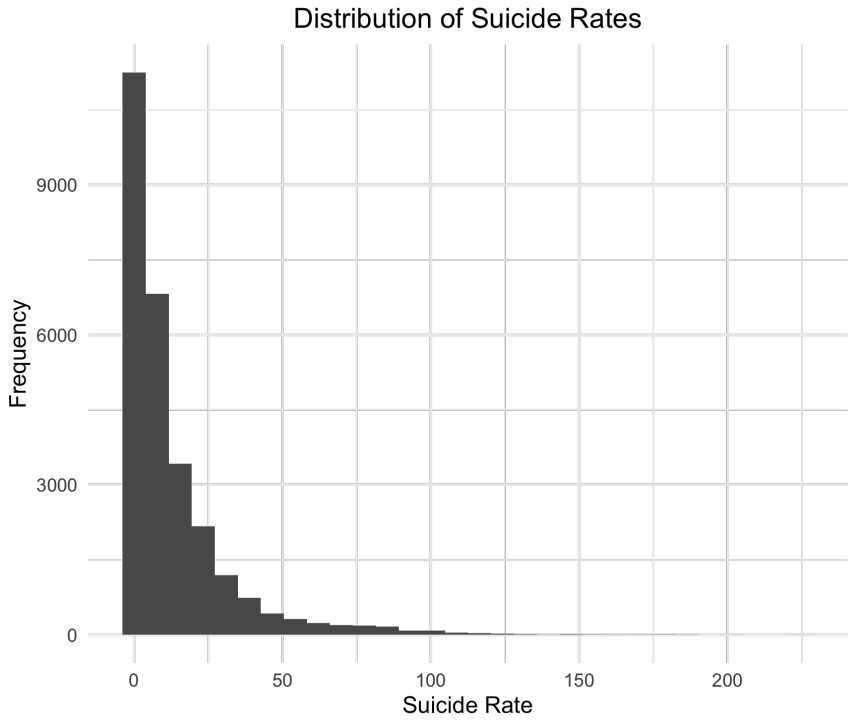

Figure 1: The distribution of suicide rates over all countries over time is strongly skewed to the right.

where the parameters and variables are defined the same as in Section 2.1, and we use the link function

$$g(\mu_i) = 1/\mu_i.$$

2.3 Linear Mixed Model

Linear mixed models (LMM) are an extension of linear models that allow for both fixed and random effects, which is particularly useful for modeling longitudinal data. This model is subject-specific, or conditional, meaning it yields parameters estimates that are conditional

on the cluster. In our dataset, we have longitudinal data by country, where the suicide rate in each country is measured for a range of years between 1980 and 2010, so our parameter estimates will be country-specific in this modeling framework.

We consider the following linear mixed model:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is a $\boldsymbol{N} \times \boldsymbol{1}$ column vector representing the response variable (suicide rate), $\boldsymbol{X}$ represents the $\boldsymbol{N} \times \boldsymbol{p}$ matrix of the $\boldsymbol{p}$ fixed effects, $\boldsymbol{\beta}$ is a $\boldsymbol{p} \times \boldsymbol{1}$ columns vector for the fixed-effects regression coefficients, $\boldsymbol{Z}$ represents the $\boldsymbol{N} \times \boldsymbol{qJ}$ design matrix for the random effects, $\boldsymbol{u}$ is a $\boldsymbol{qJ} \times \boldsymbol{1}$ column vector of the random effects, and $\boldsymbol{\varepsilon}$ represents the $\boldsymbol{N} \times \boldsymbol{1}$ vector of residuals.

In our scenario, $\boldsymbol{N} = \boldsymbol{90}$ countries were included in the study. The response variable $\boldsymbol{Y}$ is a continuous variable representing suicide rate. We have $\boldsymbol{5}$ fixed effects predictors: year, generation, sex, population, and GDP per capita. We also have a random intercept for each of the $\boldsymbol{J} = \boldsymbol{90}$ countries in our dataset.

2.4 Generalized Estimating Equation

Generalized Estimating Equations (GEE) are a way of modeling longitudinal, and generally non-normal, data. On a theoretical level, this method involves solving a set of equations to obtain estimates of model parameters. This is a marginal model that seeks to model a population average, namely the overall suicide rate over time. However, GEE can accommodate for repeated measures, and it's a common alternative to likelihood-based generalized linear mixed models (GLMM).

This modeling approach allows us to model the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ while treating

the correlation as nuisance parameters. Note that we assume marginal mean and variance such that

$$\mathbf{E}(\boldsymbol{Y_{ij}}) = \boldsymbol{\mu_{ij}}, \text{ Var}\{\boldsymbol{Y_{ij}}\} = \boldsymbol{\phi a_{ij}^{-1} v(\mu_{ij})},$$

and we assume the mean model:

$$\boldsymbol{g(\mu_{ij}) = X_{ij}^T \beta.}$$

Then, our generalized estimating equations are:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{D_i^T V_i^{-1} (Y_i - \mu_i) = 0},$$

where $\boldsymbol{D_i} = \frac{\boldsymbol{\partial \mu_i}}{\boldsymbol{\partial \beta^T}}$ is $\boldsymbol{n_i \times p}$, $\boldsymbol{(Y_i - \mu_i)}$ is $\boldsymbol{n_i \times 1}$, and $\boldsymbol{V_i}$ is an $\boldsymbol{n_i \times n_i}$ working covariance matrix.

2.5 Generalized Linear Mixed Model

In the context of analyzing suicide rates across countries, the suicide rate can be treated as a dichotomous variable, representing either a high or low rate. One approach to categorizing a continuous variable into two groups is to use a cut-off value based on the mean. This method is commonly employed across various fields and allows for comparisons across different studies. Therefore, we propose to use the mean suicide rate as the cut-off value in our analysis. Specifically, we consider a country to have a high suicide rate if its suicide rate is above the mean, and a low suicide rate if its suicide rate is below the mean.

To model the dichotomous suicide rate outcome, we employ the Generalized Linear Mixed Model (GLMM), which is a statistical model suitable for non-normal and non-continuous response variables. GLMMs combine the benefits of GLMs, which can handle

non-normal data, and MEMs, which can model the effects of fixed and random factors on the response variable. We use the GLMM framework to analyze the suicide rate data and incorporate random effects to account for the correlation between observations from the same country. The proposed model is as follows

$$\begin{aligned}\textbf{logit}(\boldsymbol{p}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \times \text{year} + \boldsymbol{\beta_2} \times \text{age} + \boldsymbol{\beta_3} \times \text{sex} + \boldsymbol{\beta_4} \times \text{population} \\ + \boldsymbol{\beta_5} \times \text{GDP-per-capita} + \boldsymbol{Zb} + \boldsymbol{\varepsilon},\end{aligned}$$

where $\boldsymbol{p} = \boldsymbol{E}(\boldsymbol{Y} > \hat{\boldsymbol{y}})$ is the probability that a given generation in a country will have a suicide rate higher than the mean suicide rate. $\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \ldots, \boldsymbol{\beta_5})$ represents the fitted regression coefficients for our predictors, including six predictors: year, age, sex, population, and GDP-per-capita. $\boldsymbol{Z}$ is the design matrix for the random effects and $\boldsymbol{b}$ represents the random effects, and $\boldsymbol{\varepsilon}$ is the model's prediction error.

2.6 Conditional Logistic Regression

Conditional logistic regression is an extension of logistic regression that allows one to take into account stratification and matching. Observational studies use stratification or matching as a way to control for confounding[5].

If there are s strata (matched sets) and p independent variables (x's), the conditional logistic regression model is

$$\text{logit}(\boldsymbol{p}) = \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2 z_2} + \cdots + \boldsymbol{\alpha_s z_s} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_p x_p}$$

where the z's are binary indicator variables for each strata (note that there are only s – 1 z variables needed), the $\boldsymbol{\alpha}$'s are the regression coefficients associated with the stratum indicator variables, the x's are the covariates, and the $\boldsymbol{\beta}$'s are the population regression coefficients to be estimated.

The conditional likelihood approach eliminates nuisance parameters $\boldsymbol{\alpha}$'s by conditioning on their sufficient statistics. And the parameter $\boldsymbol{\beta}$'s can be used to analyze the odds ratios when adjusting for the others covariates.

For our data, we fit the conditional logistic regression model

$$\begin{aligned}\text{logit}(\boldsymbol{p}) =& \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} \times \text{country}_\mathbf{2} + \cdots + \boldsymbol{\beta_1} \times \text{year} + \boldsymbol{\beta_2} \times \text{sex} + \boldsymbol{\beta_3} \times \text{population} \\ &+ \boldsymbol{\beta_4} \times \text{GDP} + \boldsymbol{\beta_{51}} \times \text{generation}_\mathbf{1} + \cdots + \boldsymbol{\beta_{55}} \times \text{generation}_\mathbf{5}\end{aligned}$$

where $\boldsymbol{p} = \boldsymbol{E(Y > \bar{y})}$ is the probability that a given generation in a country will have a suicide rate higher than the mean suicide rate in our dataset. $\text{country}_{\boldsymbol{i}}$ is the indicator variable of the given group of people from the ith country. The conditional logistic regression algorithm treats $\boldsymbol{\alpha}$'s as nuisance parameters and only estimates $\boldsymbol{\beta}$'s.

3 Results

We will first discuss our baseline, non-longitudinal models. In Figure 2, we show the comparison of coefficient estimates along with their respective 95% confidence intervals for both a standard linear regression model and a zero-inflated gamma GLM. Notice that linear regression and the zero-inflated gamma GLM yield extremely different coefficient estimates, and the standard errors of coefficient estimates produced by the GLM gamma model tend to be much lower than those from linear regression. The coefficient estimates for LMM and GEE model are similar for several variables, while the standard deviation for parameter estimates of the LMM model are smaller than those for the GEE model. When we measure the model performance by AIC, LMM has the best performance among the linear regression, linear mixed models and zero-inflated gamma GLM. Since we can't measure the model performance of GEE with AIC, we instead used QIC (quasilikelihood under the

independence model criterion) and showed that GEE with independence correlation has slightly better performance than GEE with exchangeable correlation.

	Linear Regression		GLM Gamma	
	coef	se(coef)	coef	se(coef)
year	2.480e-01	1.390e-02	-8.477e-04	4.896e-05
sex:male	1.501e+01	1.952e-01	-9.328e-02	1.381e-03
population	1.510e-07	2.519e-08	-1.176e-10	8.904e-11
GDP_capita	1.466e-05	5.527e-06	1.971e-09	1.801e-08
generation:Silent	-2.245e+01	3.710e-01	6.269e-01	1.940e-02
generation:Boomers	-3.949e-01	3.211e-01	4.958e-01	1.770e-02
generation:Generation.X	-2.158e+00	2.776e-01	3.170e-01	1.214e-02
generation:Millenials	1.566e+00	2.321e-01	1.435e-01	6.261e-03
generation:Generation.Z	3.174e-02	2.175e-01	4.447e-02	2.286e-03

Table 1: Parameter Estimates of LM and GLM Gamma

To assess the model when we have binary outcomes, we convert the suicide rate into a binary variable using the mean of our observed suicide rate as the threshold, that is, we categorized our data into two categories: above-average and below-average.

We fit the conditional logistic regression and GLMM using the same model with a country-specific random effect and the estimates from both models are shown in Table 3. As we can see, the two models give similar results. Also, the parameters for the population of each country and every category of generation are not significant based on both models.

The exponential of the parameters can be interpreted as odds ratios, and the confidence intervals can be converted to the exponential scale and give a confidence interval for the odds ratios as shown in Table 3. The odds of having an above-average suicide rate is about

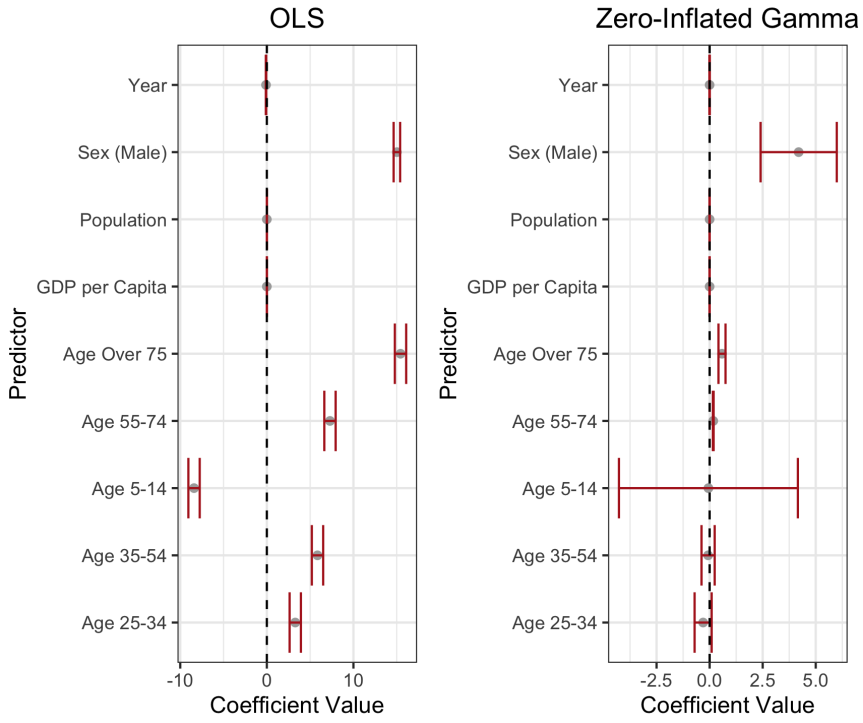

Figure 2: We can compare the $\boldsymbol{\beta}$ coefficient estimates for a basic linear regression model and a zero-inflated gamma GLM.

1.077 times the odds of a year ago for a given generation of people of the same sex in a given country with the population and GDP per capita fixed. Also, the odds of having an above-average suicide rate for males is about 64 times for females of the generation in the same year in a given country with the population and GDP per capita fixed.

	LMM		GEE Independent	
	coef	se(coef)	coef	se(coef)
year	2.715e-01	1.448e-02	2.480e-01	5.866e-02
sex:male	1.500e+01	1.630e-01	1.501e+01	1.131e+00
population	-2.020e-08	4.583e-08	1.510e-07	2.036e-07
GDP_capita	-8.418e-05	1.113e-05	1.466e-05	3.616e-05
generation:Silent	-2.217e+01	3.108e-01	-2.245e+01	1.975e+00
generation:Boomers	-2.979e-01	2.725e-01	-3.949e-01	9.076e-01
generation:Generation.X	-1.886e+00	2.322e-01	-2.158e+00	3.857e-01
generation:Millenials	1.717e+00	1.944e-01	1.566e+00	3.255e-01
generation:Generation.Z	1.889e-03	1.824e-01	3.174e-02	2.301e-01

Table 2: Parameter Estimates of LMM and GEE Independent Correlation

4 Conclusion and Discussion

Our study has contributed to the existing body of research on longitudinal suicide rates across the world. We employed various modeling approaches, such as linear regression, linear mixed models, generalized linear mixed models, generalized estimating equations, and conditional logistic regression to investigate the relationship between suicide rates and socioeconomic factors. Our results suggest that different models can effectively predict continuous and binary suicide rates. The zero-inflated gamma GLM outperformed the linear regression model for predicting continuous suicide rates, while both conditional logistic regression and GLMM models were effective in analyzing binary suicide rate outcomes.

Our study revealed that sex and year were significant predictors of suicide rates, whereas population and GDP per capita did not significantly impact suicide rates. Males were found

	GEE Exchangeable	
	coef	se(coef)
year	2.71e-01	4.34e-02
sex:male	1.50e+01	1.13e+00
population	-1.88e-08	1.49e-07
GDP_capita	-8.34e-05	2.16e-05
generation:Silent	-2.22e+01	1.96e+00
generation:Boomers	-2.97e-01	9.34e-01
generation:Generation.X	-1.89e+00	3.99e-01
generation:Millenials	1.72e+00	3.18e-01
generation:Generation.Z	2.42e-03	2.02e-01

Table 3: Parameter Estimates of GEE Exchangeable

	LM	LMM	GLM Gamma
AIC	230372	157715	220985

Table 4: AIC of LM, LMM and GLM Gamma

to have higher odds of above-average suicide rates than females, with a slight increase over time. Our findings highlight the importance of considering sub-population characteristics, such as gender, and the year in which suicide rates were measured when designing targeted suicide prevention interventions. Moreover, targeting socioeconomic factors such as GDP per capita and population may help reduce suicide rates.

Our study has significant implications for public health and suicide prevention efforts worldwide. Suicide is a complex phenomenon influenced by individual, societal, and cultural factors. Our findings underscore the need for tailored suicide prevention efforts that

	GEE Independent	GEE Exchangeable
QIC	7157371	7263995

Table 5: QIC of GEE Independent and GEE Exchangeable

	Conditional Logistic Regression			GLMM		
	coef	exp(coef)	se(coef)	coef	exp(coef)	se(coef)
year	0.074	1.077	0.004	0.073	1.076	0.004
sex:male	4.163	64.279	0.060	4.148	63.304	0.060
population	-0.000	1.000	0.000	-0.000	1.000	0.000
GDP_capita	-0.000	1.000	0.000	-0.000	1.000	0.000
generation:Silent	-16.130	0.000	401.309	-15.938	0.000	210.421
generation:Boomers	-10.697	0.000	366.343	-10.540	0.000	192.087
generation:Generation.X	-6.796	0.001	250.260	-6.692	0.001	131.221
generation:Millenials	-2.917	0.054	126.905	-2.866	0.057	66.541
generation:Generation.Z	-0.983	0.374	42.302	-0.964	0.381	22.180

Table 6: Parameter Estimates of Conditional Logistic Regression and GLMM

consider the unique risk factors and characteristics of different sub-populations. Future research should focus on identifying effective suicide prevention interventions that are adapted to specific populations and socio-economic contexts. Additionally, longitudinal studies that investigate changes in suicide rates over time could provide further insight into the complex nature of suicide and inform the development of more effective prevention strategies.

In summary, our study has provided valuable insights into the relationship between suicide rates, socioeconomic factors, and sub-population characteristics. We hope that our findings will contribute to global efforts to reduce the burden of suicide and improve mental

	Conditional Logistic Regression	GLMM
	OR (95% CI)	OR (95% CI)
year	1.077 (1.068, 1.085)	1.076 (1.067, 1.084)
sex:male	64.279 (57.134, 72.284)	63.304 (56.284, 71.207)

Table 7: Odds Ratio Estimates by Conditional Logistic Regression and GLMM

health outcomes for all.

References

[1] David A Brent and J John Mann. Family genetic studies, suicide, and suicidal behavior. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 133, pages 13–24. Wiley Online Library, 2005.

[2] Rise B Goldstein, Donald W Black, Amelia Nasrallah, and George Winokur. The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Archives of general psychiatry*, 48(5):418–422, 1991.

[3] Anders Niméus, Lil Träskman-Bendz, and Margot Alsén. Hopelessness and suicidal behavior. *Journal of affective disorders*, 42(2-3):137–144, 1997.

[4] Katarina Skogman, Margot Alsén, and Agneta Öjehagen. Sex differences in risk factors for suicide after attempted suicide: a follow-up study of 1052 suicide attempters. *Social psychiatry and psychiatric epidemiology*, 39:113–120, 2004.

[5] Wikipedia contributors. Conditional logistic regression — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Conditional_logistic_regression&oldid=1122153789`, 2022. [Online; accessed 13-March-2023].