Methodology-WriteUp

Sophia Li

2024-12-16

Methodology

Variable Selection

Based on the exploratory data analysis (EDA) graphs, the pre-selection of variables was guided by observed trends and personal preferences to focus on factors most relevant to the study objectives. Age was selected due to its clear association with death, as seen in the distribution where older age groups exhibited higher mortality rates. Gender was included because of notable differences in proportions of deaths between males and females. Education level was chosen as it revealed significant disparities, with lower education groups showing a higher proportion of deaths. Suicide methods, particularly pesticide use, hanging, and jumping, were retained due to their clear dominance in both frequency and association with fatality outcomes. Temporal variables, such as year and season, were included to explore potential patterns over time and seasonal variations, which might reflect environmental or social factors influencing fatality rates.

The urban versus rural classification was also considered, as it may reflect disparities in access to emergency care or resource availability. However, the exploratory data analysis (EDA) revealed no significant differences in survival outcomes between urban and rural areas. Consequently, the urban/rural variable has been excluded from the model. While the original paper that introduced the dataset does not provide a detailed explanation of the urban variable, based on my knowledge and observations, it is likely associated with the HuKou (registered residence location) system which is special to China, which may not necessarily influence fatality rates.

As mentioned above, I combined two variables(Died and Hospitalized) into one combined variable and categorize them into two levels - high fatality and medium fatality. I also generate EDA graphs to explore potential patterns and relationships between key predictors and the combined Outcome variable. The Age distribution indicates that high-fatality cases are concentrated among middle-aged individuals (40–60 years), with medium-fatality cases spanning a broader age range. Gender analysis reveals a higher proportion of high-fatality outcomes among males, while medium-fatality outcomes are more evenly distributed between genders. Education level shows a clear trend, where lower education levels, particularly among the illiterate and those with only primary education, are strongly associated with high-fatality cases. Suicide methods highlight the dominant role of pesticide ingestion in high-fatality outcomes, while methods like hanging and poisoning contribute more to medium-fatality cases. Yearly trends suggest relatively stable proportions of high-fatality outcomes from 2009 to 2011, whereas medium-fatality cases show a slight increase. Seasonal patterns indicate a slightly higher proportion of medium-fatality outcomes during spring and summer, which may reflect increased exposure to environmental or occupational stressors. Together, these findings highlight the importance of demographic, educational, temporal, and behavioral factors in understanding suicide fatality outcomes.

Model Assumption & Diagnostics

To analyze the combined variable as the response and include the selected predictors, I chose to use a multinomial logistic regression model. This approach also allows me to interpret the odds and identify which

variables, or specific categories within them, are statistically significant among those considered.

The multinomial logistic regression model built for this analysis largely satisfies the key model assumptions. Independence of observations is reasonable since each individual in the dataset is a separate suicide attempt, and no clustering within households or regions is explicitly indicated. No perfect multicollinearity is confirmed as the Variance Inflation Factor (VIF) values for all predictors are well below the accepted threshold, indicating no problematic correlation among predictors. Regarding the linearity of predictors with log-odds, Age, as a continuous predictor, appears suitable for the model; however, further diagnostics like logit plots may be needed to verify its linear relationship with the log-odds. The absence of outliers has not been explicitly tested yet, but influential observations could be evaluated through diagnostic tools such as Cook's distance to ensure model robustness. Additionally, the dataset meets the assumption of adequate sample size, as there are sufficient observations for each predictor category, reducing the risk of convergence issues or bias. Finally, the non-overlapping categories assumption for the response variable is satisfied since High_Fatality and Medium_Fatality are mutually exclusive and exhaustive, providing a clear and valid outcome structure. Overall, the model assumptions appear reasonably met, with minor areas for further verification.

Model Implementation

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{Education} + \beta_4 \text{method} + \beta_5 \text{Year} + \beta_6 \text{Season}$$

Here:

- $P(Y=1)$: Probability of the outcome being "High_Fatality"- immediate death or death afte treatment.
- $P(Y=0)$: Probability of the outcome being "Medium_Fatality" - survival after treatment.
- $\beta_0, \beta_1, \ldots, \beta_6$: Coefficients of the predictors.