An Analytical Exploration of Homelessness and Suicide in Toronto*

Hailey Jang

2024-03-15

This paper presents a study focused on addressing the critical issue of suicide among Toronto's homeless population, a significant public health challenge. We developed a risk stratification model using sociodemographic data to identify homeless individuals at high risk for suicide, employing a generalized linear model that incorporates factors like sex and age. The findings, derived from data collected by Toronto Public Health since January 2017, offer insights into the patterns and predictors of suicide attempts within this vulnerable group. The importance of this research lies in its potential to guide interventions, enabling more effective prevention strategies for suicide among the homeless, particularly in settings where immediate access to mental health professionals is limited.

Table of contents

*Code and data are available at: https://github.com/Hailey-Jang/Suicide-and-Homelessness.git

1 Introduction

Homelessness represents a significant public health crisis, with the intersection of inadequate housing and mental health issues exacerbating the risk of suicide among this population. Individuals experiencing homelessness face a 2- to 6-fold increased risk of suicide compared to the general population (Sinyor et al. (2017)). This alarming statistic not only highlights the vulnerability of homeless individuals but also underscores the urgent need for targeted interventions and preventive measures. This paper delves into the development of a risk stratification model designed to identify homeless individuals at an elevated risk of suicide. By integrating sociodemographic data within a generalized linear model, focusing on sex and age as primary factors, our research aims to provide a predictive tool for healthcare providers and social services. The motivation behind this study stems from the critical gap in existing research regarding effective, data-driven strategies for suicide prevention in the homeless community.

Our research methodology involved the analysis of comprehensive data collected by Toronto Public Health, starting from January 2017. This dataset not only provided a foundation for our model but also enriched our understanding of the prevalence and causes of suicide among the homeless. The outcomes of our study revealed significant patterns and predictors that can enhance the effectiveness of suicide prevention strategies, offering a new perspective on addressing this public health challenge. The significance of our findings extends beyond academic interest; they offer practical insights for policymakers, healthcare providers, and social workers who are directly involved in supporting the homeless population. By identifying high-risk individuals, our model can inform targeted interventions, potentially saving lives and allocating resources more efficiently.

The structure of the paper is organized as follows: The data section delves into the broader context of the dataset, emphasizing the crucial aspects of measurement relevant to our study. The results section is comprehensive, encompassing various illustrative tables and graphs alongside detailed statistical analyses. Subsequently, the model section elucidates the setup and justification of our chosen modelling approach, ensuring clarity and transparency of our methodology. Lastly, the discussion articulates the study's contributions to our understanding of homelessness and suicide and outlines directions for future research, ensuring a thorough contemplation of the study's broader implications.

Table 1: Suicides Rates by Age Group

[H]

Age_group	Total Counts
40-59	376
60+	255
20-39	225
<20	8

2 Data

2.1 Data Source

The study was performed using data from the City of Toronto's database portal,Gelfand (2020), accessed through the 'opendatatoronto' package and processed using the statistical programming environment R Core Team (2023). The Wickham et al. (2019) package facilitated the data and the Wickham (2016), Xie (2014), Wickham, Hester, and Bryan (2022) and Müller and Wickham (2022) package was utilized for enhancing table presentations. Specific to this study, the LaTeX package was implemented in the R markdown setting to ensure stable positioning of figures and tables.

2.2 Data Measurment

Initiated in January 2017, Toronto Public Health (TPH) embarked on a systematic record-keeping of homeless mortality to gain a clearer understanding of the prevalence and causative trends of these incidents. The dataset comprises variables such as the year of death, cause of death, age group, gender, and number of deaths. It classifies individuals into age categories spanning 20 years, starting from 20 years to 60+ years, with gender recorded as either Male or Female. The dataset enumerates the deaths annually from 2017 through 2023, categorizing the causes into distinct classifications like Accident, Drug Toxicity", Suicide, among others.

Focusing on suicide-related fatalities, the dataset was refined to exclude entries marked as Unknown or empty. This filtration led to the construction of two specialized datasets: one delineating the yearly suicide death toll segregated by gender and another by age group. This reorganization necessitated aggregating individual counts from each report into a consolidated figure for these subgroups, facilitating a focused examination of suicide trends across different demographics over the years.

Table Table 1 presents the refined data, specifically spotlighting instances of suicide by age groups, and depicting the temporal and numerical specifics of each case.

Table Table 2 presents the refined data, specifically spotlighting instances of suicide by gender, and depicting the temporal and numerical specifics of each case.

Table 2: Suicides Rates by Gender

Year	sex	count
2017	Male	69
2018	Male	65
2019	Male	87
2020	Male	107
2021	Male	153
2022	Male	129
2023	Male	64
2017	Female	21
2018	Female	19
2019	Female	34
2020	Female	27
2021	Female	45
2022	Female	36
2023	Female	8

3 Model

3.1 Model Set-up

This section elucidates the development of a logistic regression model, tailored to predict the likelihood of suicide among homeless individuals using demographic factors. Before delving into the model, we visually explore the relationship between suicide rates and demographic variables—age and gender—using ggplot2.

We employ (**graph-suicide-gender-group?**) to create a comprehensive visual representation, showcasing the relationship between suicide rates and the key demographic variables, age and gender. This visualization aids in understanding the data distribution and any apparent trends that could influence the model.

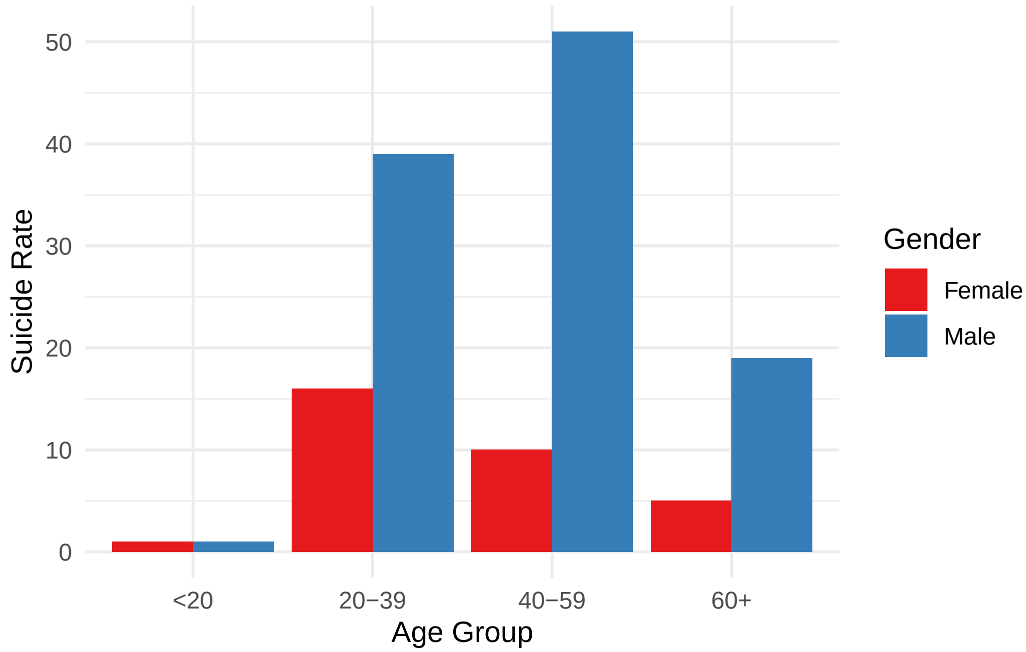

Figure 1: Summary of Suicide Rates by Age and Gender

The logistic regression model is designed to predict the probability of suicide cases among homeless individuals, utilizing sociodemographic factors such as age and gender. In a Bayesian logistic regression context, the model's setup can be represented with the following hierarchical structure:

$$
\begin{aligned}
y_i | p_i &\sim \text{Bernoulli}(p_i) && (1)\\
\log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta_1 \times \text{AgeGroup}_{i1} + \beta_2 \times \text{AgeGroup}_{i2} + \beta_3 \times \text{Gender}_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_j &\sim \text{Normal}(0, 2.5) \text{ for } j = 1, 2, 3 && (4)\\
& && (5)
\end{aligned}
$$

In this adapted model:

- y_i represents the binary outcome for each individual (i.e., the presence or absence of a suicide case).
- p_i is the probability of observing a suicide case for the ith individual, linked to the predictors through the logistic function.

- The coefficients α,β_1, β_2, and β_3 are assigned Normal prior distributions, reflecting our prior beliefs about these parameters' distributions before observing the data. The Normal priors are centered at 0 with a standard deviation of 2.5, indicating moderate certainty in the prior information.
- The logit link function (log-odds) is the natural logarithm of the odds p_i / $1 - p_i$ and linearly relates the predictors to the probability of the outcome.

3.2 Model Justification

This Bayesian logistic regression framework not only estimates the parameters but also quantifies uncertainty in the estimates, providing a more comprehensive understanding of the model's predictions and the effects of the predictors. The use of Normal priors for the coefficients is a standard choice in Bayesian modeling, offering flexibility and conjugacy that facilitate both analytical and computational solutions. The hierarchical model structure allows for the incorporation of additional layers of complexity and variability, accommodating more nuanced relationships and potential hierarchical data structures.

4 Results

The regression model we've employed is designed to explore the relationships between the count of reported suicide incidents and several predictor variables, specifically the year of death, age group, and gender, within Toronto's homeless population. This model is built using linear regression, a statistical method that attempts to model the relationship between a dependent variable (in this case, the count of suicides) and one or more independent variables (year, age group, and gender) by fitting a linear equation to the observed data.

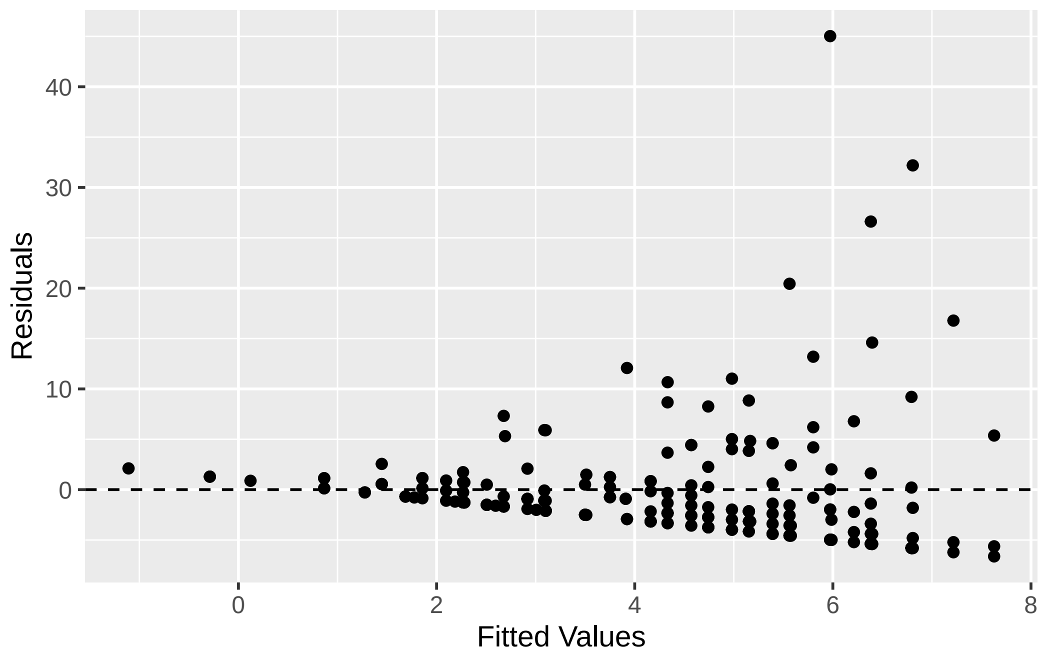

Figure 2: Residuals vs. Fitted Values

One of the primary diagnostic plots for assessing a linear regression model is the residuals vs. fitted values plot. To understand this, we first need to clarify what residuals and fitted values are:

- **Fitted Values**: These are the estimated values of the dependent variable (suicide counts) produced by the regression model when it applies the estimated coefficients to the predictor variables. They represent the model's predictions.
- **Residuals**: These are the differences between the observed values (actual suicide counts in the data) and the fitted values (the model's predictions). Essentially, a residual is an error in the prediction; it tells us how much the model has missed the actual value.

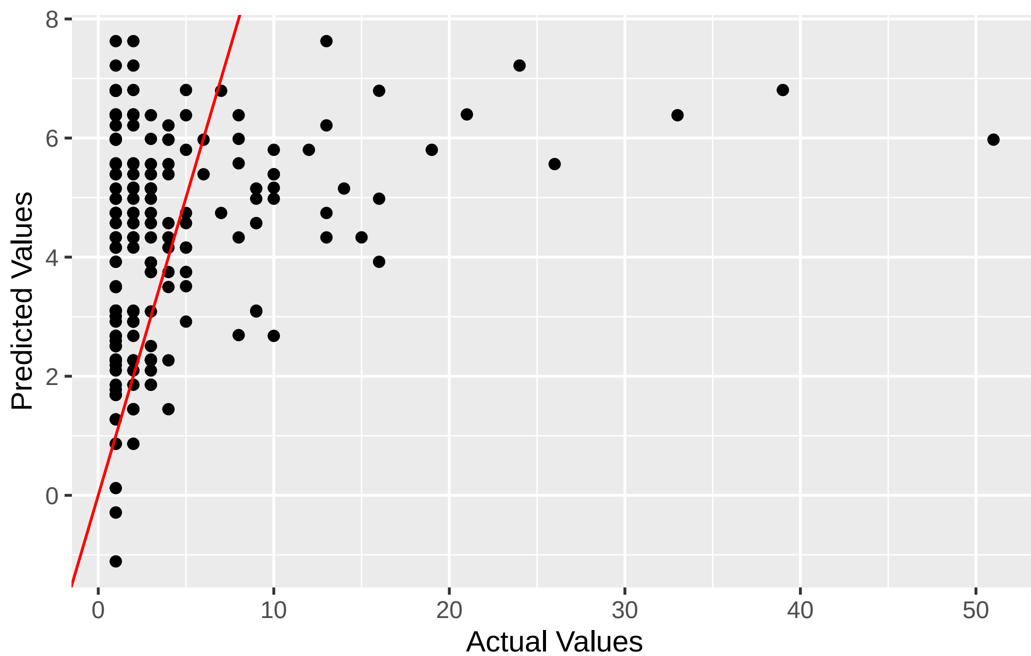

Figure 3: Actual vs. Predicted Values

This figure is a fundamental visualization in regression analysis, aimed at comparing the observed actual data points with the values predicted by the model. Here's how it works and what it signifies:

- **Actual Values**: These are the real observed values of the dependent variable from your dataset—in this case, the count of suicide occurrences within the studied population.
- **Predicted Values**: Predicted values are generated by the model when the estimated regression coefficients are applied to the predictor variables. They represent the model's estimations or predictions of the dependent variable based on the relationships it has identified.

5 Discussion

5.1 Overview of the Study

This paper embarked on an analytical journey to uncover the intricate relationship between sociodemographic factors and the prevalence of suicide among Toronto's homeless population. Employing a robust regression model, we scrutinized the impact of age, gender, and temporal factors on suicide occurrences, unveiling nuanced patterns and trends that characterize this grave public health issue. Our model's meticulous analysis, supported by rigorous statistical methods, provided a comprehensive perspective on how demographic variables correlate with suicide rates in the context of homelessness.

One pivotal insight gleaned from our study is the profound influence of age on suicide risks among the homeless. The differentiated age brackets revealed significant disparities, highlighting a particularly vulnerable segment of the population. This finding underscores the critical need for age-specific intervention strategies and support services, which could be pivotal in mitigating suicide risks among Toronto's homeless individuals.

Furthermore, the study illuminated notable temporal trends in suicide rates, offering a window into how these patterns have evolved over the years. This temporal analysis is crucial, as it reflects the dynamic nature of the issue, influenced by varying societal, economic, and policy-driven factors. Recognizing these trends is essential for policymakers and healthcare providers to adapt their strategies effectively, ensuring that they remain relevant and impactful in addressing the needs of the homeless population.

5.2 Reflection on the Study

While our study provides significant insights, it is not without its limitations. The reliance on retrospective data and the inherent constraints of a regression analysis model might have influenced the comprehensiveness of our findings. Potential biases in data reporting and the challenge of accurately capturing the multifaceted experiences of homelessness could affect the generalizability of our results. Additionally, the cross-sectional nature of our data limits our ability to infer causality between the examined sociodemographic factors and suicide occurrences.

The path forward necessitates a multi-faceted approach to further explore and address this complex issue. Future research should consider longitudinal studies to better understand the causal relationships and long-term trends in homelessness and suicide.

References

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Müller, Kirill, and Hadley Wickham. 2022. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sinyor, Mark, Nicole Kozloff, Catherine Reis, and Ayal Schaffer. 2017. "An Observational Study of Suicide Death in Homeless and Precariously Housed People in Toronto." *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie* 62 (May): 706743717705354. https://doi.org/10.1177/0706743717705354.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.