A Statistical Analysis of Mortality in Ontario*

An Analysis of Poisson, Negative Binomial, and Gaussian Regression Models.

Nikhil Iyer

March 17, 2024

Mortality trends have been studied and monitored for centuries, in which they play a pivotal role in understanding the health dynamics of populations. This paper looks at mortality trends in Ontario from 2002 to 2022, using Generalized Linear Models. This study shows that mortality trends are nuanced and best modelled using a Negative Binomial model. The preferance for a Negative Binomial model in modelling mortality data in Ontario is evidence of the complex nature of mortality. This study highlights the importance of managing disparaties in the Ontario healthcare access and quality.

1 Introduction

In Ontario, Canada, the investigation of morality trends has attracted significant attention, driven by a commitment from government officials to make evidence-based policies. Morality trends are a vital metric in assessing the health of a population, and provides insight into the effectiveness of the healthcare systems. This paper models morality trends in Ontario from 2002 to 2022, using data (2) provided by (Statistics Canada 2023).

This study aims to fill the gap in understanding the underlying distribution of mortality patters in Ontario. Using mainly `R` (R Core Team 2023), `tidyverse` (Wickham et al. 2019), and `rstanarm` (Goodrich et al. 2024), and with heavy inspiration from TELLING STORIES WITH DATA (Alexander 2023) the data is analyzed using regression analysis to find the behavior of mortality trends.

The analysis revealed that mortality trends in Ontario are best represented using a Negative Binomial model which outlines the nuances and complex nature of mortality, while leaving space for outside factors that could affect the data. These findings can be used as indicators for understanding the health status of Ontario's population, which can be used to make evidence based decision that will positively impact the province.

*Code and data are available at: https://github.com/Niyer02/Mortality-in-Ontario

2 Data

The raw data was retrieved from Statistics Canada (Statistics Canada 2023). The data provided was a comprehensive table with many variables. However, for the purpose of this analysis, only the most prevalent features were retained. Cleaned data (Wickham et al. 2023) (Wickham, Hester, and Bryan 2024) was derived from the raw data, with features of interest being the cause of death (`Leading causes of death (ICD-10)`) and death count (`VALUE`). The data was extracted and cleaned such that each year has the same 5 causes of death (`Leading causes of death (ICD-10)`), which allowed for the fitting of the mode to the data set.

For a better understating, the data can be classified by year (`REF_DATE`) with 5 causes of death (`Leading causes of death (ICD-10)`) per year, in the range of 2002 to 2022. A time series of the variables can be seen in Figure 1.

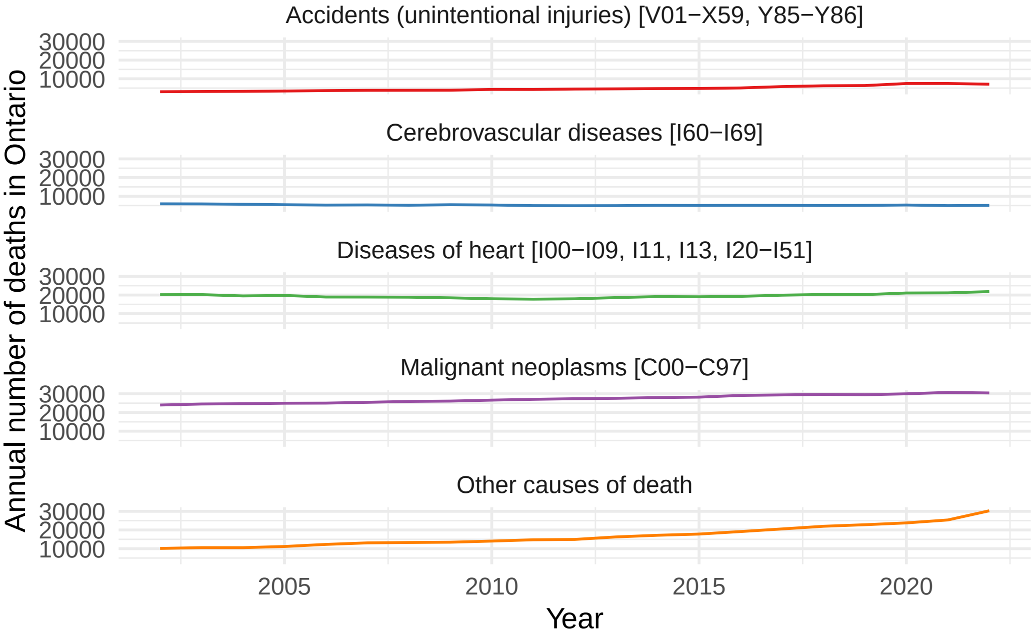

Figure 1: Time Series of Deaths in Ontario from 2002 to 2022

From Figure 1, we can see that most causes of death (`Leading causes of death (ICD-10)`) either remain stagnant or slightly increase in time. However `Other causes of death` sees a big increase in frequency. Alternatively, we can view the variables in a stacked bar chart (Wickham 2016) to get a better understanding of the trends over the time frame. The stacked bar chart with the variables of interest can be seen in Figure 2.

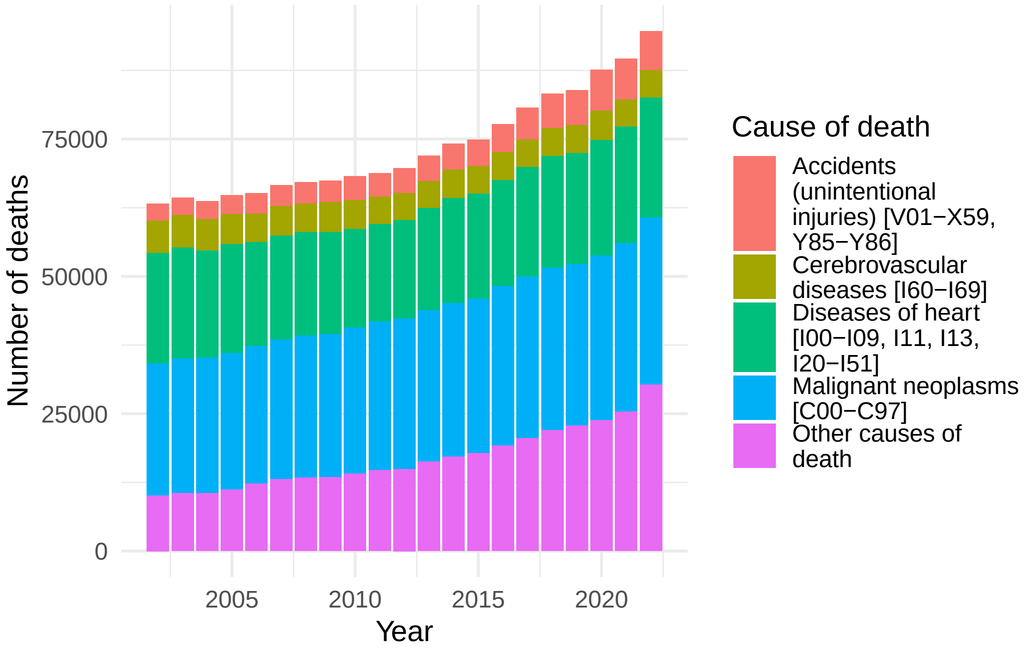

Figure 2: Stacked Bar Chart of Deaths in Ontario from 2002 to 2022

From Figure 2, it is more evident that the number of deaths has increased greatly, and all causes of death (`Leading causes of death (ICD-10)`) follow an exponential growth. Visualizing the data was important, as it allowed for a better understanding and thus a better model to be fit.

3 Model

The goal of our modelling strategy is analyze the data (2), and to identify patterns and relationships. We showcase the applications of Poisson, Negative Binomial, and Gaussian regression models, and determine the model that captures the underlying distribution the most effectively.

3.1 Model set-up

Define y_i as the number of deaths. Then μ_i represent the mean of the negative binomial distribution for observation i.

$$y_i \sim \text{NegBinomial}(\mu_i, \phi) \tag{1}$$

$$\mu_i = e^{\eta_i} \tag{2}$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \tag{3}$$

Where β_0 is the intercept, $\beta_1 x_{i1} + ... + \beta_p x_{ip}$ are the coefficients, and ϕ is the dispersion parameter of the Negative Binomial distribution. We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2024). We use the default priors from `rstanarm`. The other models tested did not achieve the same accuracy as the Negative Binomial model, as such they will not be defined in this paper.

3.1.1 Model justification

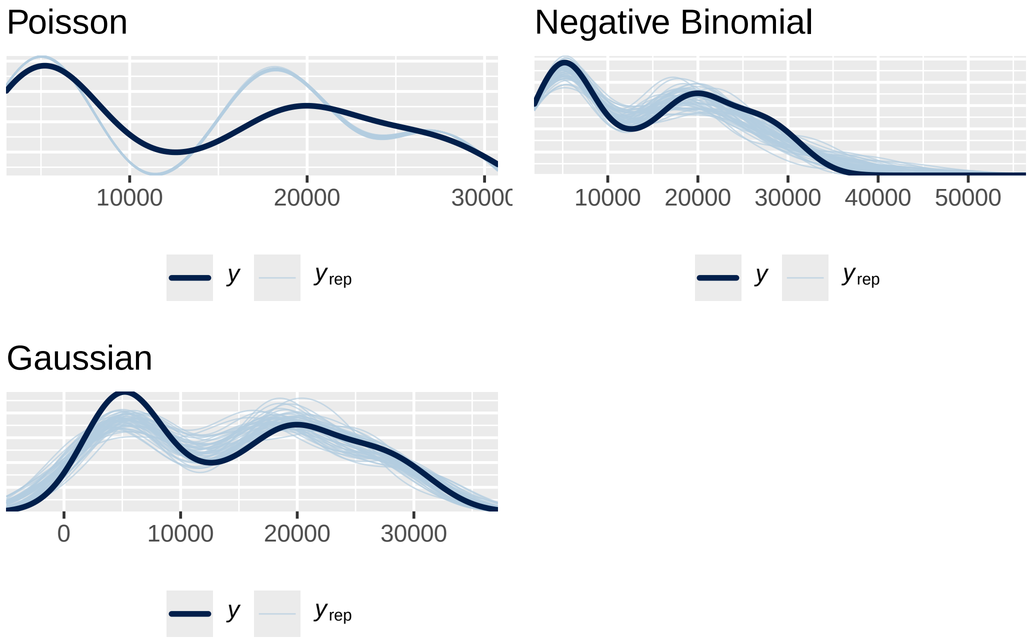

Figure 3: Poisson, Negative Binomial, and Gaussian Regression Models Fitted on the Data

From Figure 3, we can see that the Negative Binomial model fits the best, however it is very close to the Gaussian model. To further confirm this we can further compare the models using the re-sampling method leave-one-out(LOO) cross-validation (CV) seen in Table 1.

Table 1

```
                  elpd_diff se_diff
model_neg_binomial      0.0       0.0
model_gaussian        -18.4      11.2
model_poisson      -24246.1    5629.2
```

From the leave-one-out(LOO) cross-validation (CV) re-sampling method, we can confirm that the Negative Binomial model was the best choice, because ELPD is larger.

4 Results

The result of this analysis is that the Negative Binomial model is the optimal choice for modelling mortality trends in Ontario. This suggests that the Negative Binomial distribution captures the variability and complexities in the mortality data. This is likely because the Negative Binomial model accounts for over dispersion among other factors within the data.

The ability of the model to accurately model the trend enables officials and authorities to gain deeper insights, and to make a more educated decision based on the evidence. Overall, officials can use the Negative Binomial model, and leverage its predictive power to implement practices and make changes to better the health of Ontario's population.

5 Discussion

5.1 About the Findings

By selecting a model that accurately fits the mortality data (2), officials gain a powerful tool in analyzing and predicting the health of Ontario's population. Leveraging this tool can be used for targeted developments and interventions tailored to the population. Insights from the model can be used to allocate funds and resources, and find discrepancies in the healthcare system.

It can be seen from the model summary in the appendix (6) that some features (Causes of Death) have a higher overall presence than others. Officials could use this information to allocate resources towards the features that have the highest mortality rate. Furthermore, the Negative Binomial model can be used to forecast future mortality trends, which can be used to proactively plan and adapt the healthcare system wherever possible.

5.2 Weaknesses and next steps

Weaknesses in this study are most prevalent in the model (3) section. Models were restricted to Generalized Linear Models (GLM's), and although they accurately captured the underlying distribution, a different type of model could have done a better job. Additionally, access to more data that covers a wider time rane would yield a better model that could be used for more general inference.

6 Appendix

6.1 Negative Binomial Summary for one Feature

The `model summary` (Arel-Bundock 2022) library was used to display the summary statistics.

Table 2

```
                                                                           mean
(Intercept)                                                          8.47398895
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 0.09109709
                                                                            mcse
(Intercept)                                                          0.001230250
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 0.001561911
                                                                             sd
(Intercept)                                                          0.05270501
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 0.07388602
                                                                               10%
(Intercept)                                                           8.4072424923
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] -0.0005822092
                                                                            50%
(Intercept)                                                          8.47394887
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 0.09041986
                                                                           90%
(Intercept)                                                          8.5426248
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 0.1856669
                                                                     n_eff
(Intercept)                                                           1835
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69]  2238
                                                                         Rhat
(Intercept)                                                          1.000724
`Leading causes of death (ICD-10)`Cerebrovascular diseases [I60-I69] 1.000265
```

References

Alexander, Rohan. 2023. *TELLING STORIES WITH DATA with Applications in r*. Boca Raton: CRC Press.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

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

Statistics Canada. 2023. "Leading causes of death, total population (age standardization using 2011 population)." https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310080101.

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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.