Evaluating the Effectiveness of Statistical Models in Analyzing Mortality in Canada*

A Practical Application of Negative Binomial Regression

Xiyou Wang, Yetao Guo

2024-03-17

This study explores the complex relationship between Canadian social trends, healthcare dynamics, and mortality rates. We compare the differences between Poisson and negative binomial distribution models using a comprehensive dataset and provide insights into the complexity of mortality. Our main findings reveal important associations between various social and healthcare factors and mortality rates, indicating that the negative binomial regression is more suitable for the data. Our research contributes to a deeper understanding of mortality rates and their influencing factors, laying a valuable foundation for future research and strategic initiatives aimed at improving public health outcomes.

Contents

*Code and data are available at: https://github.com/wxywxy666/Mortality-in-Canada.

1 Introduction

Studying public health mortality rates is of great significance to policy makers, healthcare professionals, and society as a whole. Mortality is an important indicator of social well-being and health system efficiency. It provides valuable insights for addressing the challenges faced by understanding the overall health situation, which in turn guides resource allocation and policy decisions.

For the purposes of this study, our primary estimate is mortality in Canada from 2000 to 2022, broken down by major cause for each year. In-depth analyses were conducted by fitting mathematical models, taking into account the impact of various social and health care factors on mortality. Through a comprehensive analysis of the mortality situation, we aim to reflect on overall social trends and health care dynamics in Canada. Our study provides a nuanced understanding of mortality, highlighting its complexity and the factors that influence it.

One important result of our analysis is that there is a complex relationship between mortality rate and various social and healthcare factors. This discovery is the focus of our report, providing valuable insights into the challenges and opportunities facing the Canadian healthcare system. By exploring the complexity of Canada's public health mortality rate, we hope to deepen our understanding of this key indicator and its impact on policy-making, healthcare services, and social well-being.

We use R(R Core Team 2024) for all data wrangling and analysis and R packages dplyr(Wickham et al. 2023) to clean the original data, ggplot2(Wickham 2016), knitr(Xie 2014), kableExtra(Zhu 2024) to produce the charts and modelsummary(Arel-Bundock 2022), rstanarm(Brilleman et al. 2018), broom.mixed(Bolker and Robinson 2022), loo(Yao et al. 2017) to fit the data into the model and generate relevant analysis charts.

2 Data

This section seeks to provide a comprehensive understanding of the dataset used in our analysis. The data offer a wider perspective allowing trends to be analyzed over time, including periods of COVID-19 outbreaks.

2.1 Source

Our study uses a dataset from Statistics Canada(Warin and Le Duc 2023) and focuses on mortality trends in Canada from 2000 to 2022, categorized by age, sex, and cause of death. This information is crucial for understanding public health trends and guiding policy decisions. Therefore, this dataset was selected due to its comprehensive coverage, high data quality and reliability.

Data from Statistics Canada is updated annually, and the specific data used in this article is the latest available as of 2022 . Raw data set presents data on the total number of deaths in Canada and ranks the leading causes of death, such as salmonella infections, shigellosis and amoebiasis, and tuberculosis, etc. Also included are maps of the age at time of death, the distribution of both genders, and partial places of residence. It is worth noting that The category "Age at time of death, all ages" includes the number of deaths for children aged under one year old, The deaths for which age is not stated are included in the "Age at the time of death, all ages" category but not distributed among age groups. All of the data is possessed and cleaned through R studio, a programming language for statistical computing and graphics.

2.1.1 Measurement

Using Poisson and negative binomial distributions as link functions, the data were fitted and the results of the two distributions were compared to select the best link function. The Poisson distribution is a simple way to predict events, but it expects the average number of events to be the same as the amount they vary, which isn't always true for actual data.The negative binomial distribution is a more flexible version that can deal with differences between the average number of events and how much they vary by including an extra detail, the dispersion parameter, that helps when the data is more spread out than the average.

2.2 Raw data

First, the original dataset file of "Leading causes of death", called "1310039401-eng", was downloaded from Statistics Canada and renamed as "raw_data". Five key variables are included. "Leading cause of death" in the table is defined as an illness or injury that triggers a sequence of events leading directly to death, or an accident or violent situation that results in a fatal injury. The underlying cause is selected from the conditions listed on the medical certificate of cause of death. “Characteristics” contains the “Rank of leading causes of death” which is based on the “Number of deaths”. “Age at time of death” is attained at the last birthday preceding death. “Reference period” contains specific numbers, from 2000 to 2022.

2.3 Data cleaning

In order to understand the data more intuitively, data cleansing is necessary. The other reason is that the original data file contains many irrelevant instructions and variables. The first step is to delete them directly and to rename “Leading causes of death” as “cause”. And then, replace the original classification method based on the cause of death with year, named “year”. Next, add two variables, “ranking” represents the ranking of the number of people for this cause in the same year, and “years” represents the number of times this cause appears from 2000 to 2022. Finally, sort them with “year” as the main one, and then sort them in the reverse order of “total_death” according to this cause in this year, that is, in the forward order of “ranking”.

2.3.1 Preview

Table 1: Modeling the most prevalent cause of deaths in Canada, 2000-2022

Year	Cause	Deaths	Ranking	Years
2022	Malignant neoplasms	82,412	1	23
2022	Diseases of heart	57,357	2	23
2022	COVID-19	19,716	3	3
2022	Accidents (unintentional injuries)	18,365	4	23
2022	Cerebrovascular diseases	13,915	5	23
2022	Chronic lower respiratory diseases	12,462	6	23
2022	Diabetes mellitus	7,557	7	23
2022	Influenza and pneumonia	5,985	8	23
2022	Alzheimer’s disease	5,413	9	23
2022	Chronic liver disease and cirrhosis	4,530	10	23

Preview the processed clean data first, which is very helpful for understanding the overall data. Table 1 shows the top-ten causes in 2022, from which some rough findings can be obtained. We previously predicted that the number of occurrences of most causes of death would be 23, which means they would occur every year, but the actual situation is that except for COVID-19, the rest are 23, and COVID-19 is 3, because COVID-19 suddenly appeared in 2020. Does this mean that Canada’s medical and health system has not undergone significant changes? The same diseases are still plaguing the people in 23 years.

2.3.2 Trend

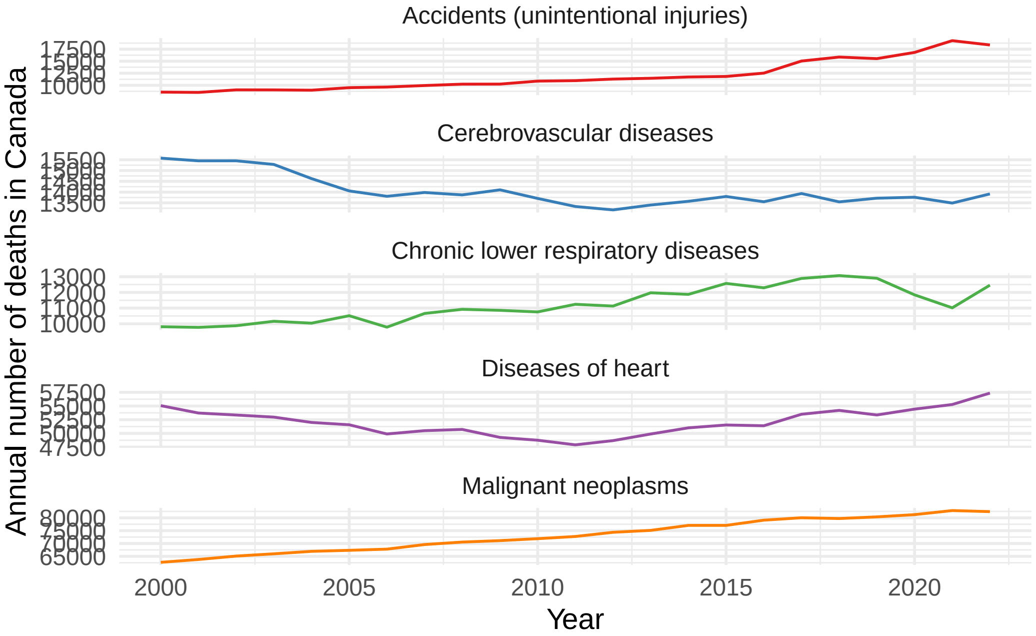

Figure 1: Annual number of deaths for the top-five causes in 2022, since 2001, for Canada

Figure 1 shows how many people died from different causes each year in Canada from 2000 to around 2022. The vertical axis represents the annual number of deaths. The horizontal axis represents time, from the year 2000 to about 2024. Different colored lines for each cause show whether the number of deaths is going up, going down, or staying about the same over time. The red line has some fluctuations, but the overall trend is upward, indicating that deaths from accidents have been rising. The blue line is mostly flat with a bit of a drop, which could mean fewer people are dying from strokes and related diseases. This suggests improvements in healthcare or prevention measures for these conditions. The green line climbs up slowly, indicating more deaths from diseases like emphysema. The purple line indicates The number of deaths from heart disease is the highest among the listed causes and shows a slight upward trend over the two decades, with some fluctuations but no significant increase or decrease. Orange Line: The trend for cancer-related deaths is gradually increasing, indicating a growing number of deaths due to malignant neoplasms over the years.

3 Result

Table 2: Summary statistics of the number of yearly deaths, by cause, in Canada

Min	Mean	Max	SD	Var	N
8521	32504.35	82822	25849.99	668221779	115

Table 2 tells us the summary statistics of the number of yearly deaths by cause in Canada. The number of observations in the data is 115. The minimum value in the data set is 8521 and the maximum value is 82822.SD refers to Stands for Standard Deviation, a measure of how spread out numbers are in the set is 25849.99. Var refers to variance, indicating how much the numbers vary from the average, and it's 668221779.

3.1 Model fitting

The Poisson distribution describes the probability distribution of the number of events that occur within a fixed unit of time or space. Its probability mass function is:

$$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$$

The Poisson distribution is usually more concise and convenient, can fit a wider range of models and is easier to interpret. Nevertheless, Poisson distribution assumes that the mean(λ) and variance(λ) are equal.

Compared with the Poisson distribution, the calculation and derivation of the negative binomial distribution are more complicated, and the parameter estimation of the model may be more difficult, especially when the amount of data is small or the model does not fit well. Its probability mass function is:

$$P(X=k)=\binom{k+r-1}{k}(1-p)^k p^r$$

However, the negative binomial distribution is a generalization of the Poisson distribution that allows for differences between mean($\frac{r(1-p)}{p}$) and variance($\frac{r(1-p)}{p^2}$). It does this by processing an extra parameter: the dispersion parameter(r), which is the number of successes that need to be observed before a number(k) of observed failures.

When means and variances are not equal, especially when overdispersion occurs, using the negative binomial distribution instead of the Poisson distribution can provide a more accurate statistical model that better fits the data. For the sake of rigor, this study still completely compares the two.

Table 3: Modeling the most prevalent cause of deaths in Canada, 2001-2020

	Poisson	Negative binomial
(Intercept)	9.390	9.389
		(0.038)
causeCerebrovascular diseases	0.160	0.161
		(0.053)
causeChronic lower respiratory diseases	−0.063	−0.062
		(0.055)
causeDiseases of heart	1.469	1.469
		(0.055)
causeMalignant neoplasms	1.812	1.813
		(0.054)
Num.Obs.	115	115
Log.Lik.	−18 651.498	−1098.843
ELPD	−19 266.9	−1102.0
ELPD s.e.	3030.0	9.7
LOOIC	38 533.9	2204.0
LOOIC s.e.	6060.0	19.4
WAIC	39 297.8	2204.0
RMSE	3355.03	3355.08

3.2 Estimate

Implementing Poisson and negative binomial regression, simultaneously. Model the most prevalent cause of deaths. According to the Table 3, the estimates are similar, so it is necessary to use a posterior predictive check in addition.

3.3 Posterior predictive check

Figure 2, Figure 3 indicate that negative binomial approach is a better choice for this circumstance. But it's too early to draw conclusions from the figures alone.

There are many observations with a pareto_k > 0.7. Large Pareto's K means the model posterior would be too different if one data point is being removed. This suggests that the model is not capturing the data well (i.e. some data points with high K are highly influential and not being considered by the model).

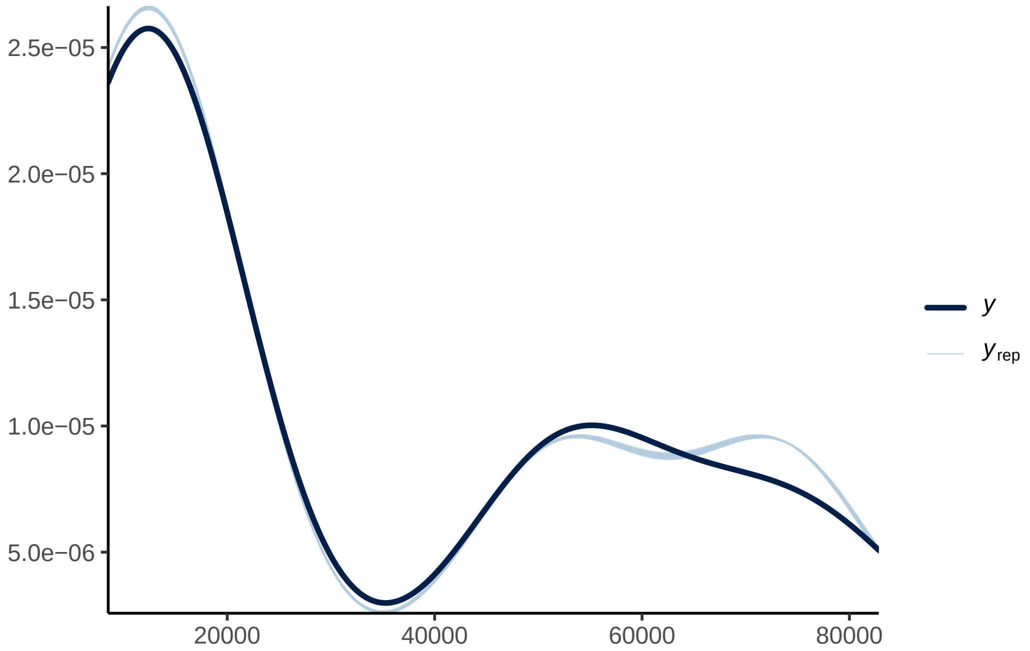

Figure 2: (a) Poisson model

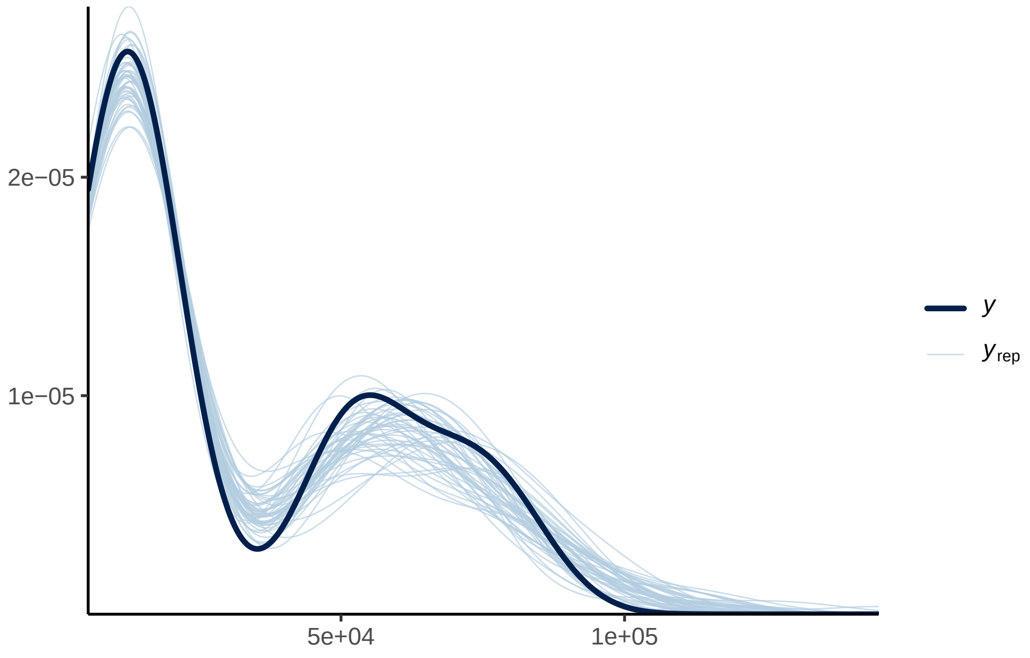

Figure 3: (b) Negative binomial model

3.4 Resampling

Following this logic, LOOCV(Leave-one-out cross-validation) will not be trustworthy anymore if most of the Pareto's K is higher than 0.7, although it can usually get a more accurate prediction. With this many problematic observations, K-fold cross-validation with argument 'K=10' to perform 10-fold cross-validation, that the data used for each training becomes closer to the entire data set, and the deviation decreases.

It is less computationally intensive, and more efficient when processing large data sets. So K-fold will be a more suitable resampling method in this study.

Table 4: K-fold cross-validation summary

	elpd_diff	**se_diff**	**elpd_kfold**	**se_elpd_kfold**	**p_kfold**	**se_p_kfold**
neg_binomial	0.00	0.000	-1104.82	9.96637	6.224038	1.472227
poisson	-20478.88	3829.983	-21583.70	3836.61495	3530.742609	1130.862683

The information provided in the Table 4 allows us to compare the relative performance of the negative binomial model and the Poisson model. The result mainly relies on the "elpd_diff" and "elpd_kfold" values of the two models, namely "expected log pointwise predictive density difference" and "ELPD for K-fold cross-validation". The higher these two values are, the better the model's predictive performance is.

From table, we can see that the negative binomial model has higher "elpd_diff" and "elpd_kfold" values and also has a smaller standard deviation. This shows that the negative binomial model has better predictive performance than the Poisson model in cross-validation. Therefore, we can infer that the negative binomial model is a better fit for the overall number of deaths by cause, in Canada, from 2000 to 2022.

4 Discussion

Plots of time-series changes in the number of deaths in Canada due to different factors are provided in the visualisation results of this study. The excellence of two forecasting models for mortality is analysed in comparison.

4.1 Finding

The main focus of this document revolves around a comparison of the performance of two statistical models - the negative binomial model and the Poisson model - in analyzing mortality in Canada. Through detailed data analysis, we find that the negative binomial model outperforms the Poisson model in terms of both the expected log-point-by-point predicted density difference(elpd_diff) and the K-fold cross-validated ELPD value(elpd_kfold). In addition, the negative binomial model has a smaller standard deviation, suggesting that its predictions are more stable. This finding provides a new perspective that negative binomial regression may be more appropriate for describing and predicting mortality in Canada.

4.2 Economic Impact Insights

From the perspective of economic impact, accurate mortality prediction is an important reference value for a number of industries, including insurance, pension, and healthcare(Or 2001). The superiority of the negative binomial regression means that we can predict future mortality rates more accurately, thus helping these industries to make more rational economic decisions, such as setting insurance premium rates and planning medical resources. In addition, for governments, accurate mortality prediction can also help to formulate more effective social security policies and reduce financial pressure(Whitehouse and Zaidi 2008). In terms of insight, the reason why the negative binomial model can achieve better prediction results may be related to its ability to better handle the discrete and over-discrete nature of the data. This suggests that we need to pay more attention to the characteristics of the data and choose more appropriate statistical models when dealing with similar problems.

4.3 Societal and Technological Influences

In terms of social impact, accurate mortality forecasts help the public to better understand the trends in demographic and health conditions, and thus make more informed life decisions. For example, the public can adjust their health management and retirement planning based on the prediction results(Holt-Lunstad and Smith 2012). In terms of technical implications, the findings in this paper provide new ideas for the application of statistical models in the field of mortality prediction. In the future, we can further explore and optimize the negative binomial distribution model, or find other more suitable models to improve the prediction accuracy and stability.

4.4 Weakness and Future Research Directions

Despite the superiority of the negative binomial regression in predicting mortality in Canada, there are still some shortcomings. For example, only two models were compared in this paper and other possible models were not considered; furthermore, the parameters of the models were not analysed in detail to identify the key factors affecting the prediction results. Future research directions can be developed in the following aspects: firstly, other possible statistical models can be further explored to find a more suitable model for predicting mortality in the Canada; secondly, an in-depth analysis of the model's parameters can be carried out to reveal the specific factors affecting the prediction effect; and lastly, attempts can be made to incorporate more factors into the model in order to improve the prediction accuracy and the scope of application. In summary, the findings of this paper are of great guiding significance for the modelling of mortality in the Canada, and also provide us with an opportunity to explore in depth the application of statistical models in the field of mortality prediction.

Reference

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Holt-Lunstad, Julianne, and Timothy B Smith. 2012. "Social Relationships and Mortality." *Social and Personality Psychology Compass* 6 (1): 41–53.

Or, Zeynep. 2001. "Exploring the Effects of Health Care on Mortality Across OECD Countries."

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Warin, Thierry, and Romain Le Duc. 2023. *statcanR: Client for Statistics Canada's Open Economic Data*. https://CRAN.R-project.org/package=statcanR.

Whitehouse, Edward, and Asghar Zaidi. 2008. "Socio-Economic Differences in Mortality: Implications for Pensions Policy."

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

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

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.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2017. "Using Stacking to Average Bayesian Predictive Distributions." *Bayesian Analysis*. https://doi.org/10.1214/17-BA1091.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.