Evaluating the Effectiveness of Statistical Models in Predicting Mortality RatesAn Analysis of Negative Binomial Distribution Model in Canada*

An Analysis of Negative Binomial Distribution Model

Xiyou Wang, Yetao Guo

2024-03-14

This study explores the complex relationship between Canadian social trends, healthcare dynamics, and mortality rates. We compared the goodness of fit between the negative binomial distribution model and the Poisson distribution model using a comprehensive dataset, and conducted in-depth research on the complexity of mortality rates. Our main findings reveal important associations between various social and healthcare factors and mortality rates, indicating that the negative binomial distribution model 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

The field of public health mortality is of great significance to policy makers, healthcare professionals, and society as a whole. Mortality rate is an important indicator for measuring social well-being and health system efficiency. They provide valuable insights into the challenges faced by understanding the overall health situation, thereby providing guidance for resource allocation and policy decision-making.

In this study, we delved into the complexity of mortality rates, particularly in the context of Canada. Through a comprehensive analysis of the mortality rate situation, we aim to reflect on Canada's overall social trends and healthcare dynamics. Our study provides a subtle understanding of mortality, highlighting its complexity and influencing factors.

In this study, our primary estimand is the mortality in Canada from 2000 to 2022, by top cause each year. We perform in-depth analysis by fitting mathematical models, taking into account the impact of various social and health care factors on mortality.

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.

We have provided more details in this introduction than in the abstract, but we will not disclose the full content of our research findings. On the contrary, we have provided a high-level overview of our research findings, laying the foundation for subsequent chapters of this article.

Looking ahead, we discussed the next steps of our research and outlined the direction for future exploration. Finally, we conclude the introduction with a brief paragraph, emphasizing the structure of the paper and guiding readers to understand the following chapters and their respective contributions to the overall narrative.

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.

2 Data

This section seeks to provide a comprehensive understanding of the dataset used in our analysis. The dataset captures the mortality rates across various demographics in Canada from 2000 to 2022. The data offer a wider perspective allowing trends to be analyzed over time, including periods of COVID-19 outbreaks.

2.1 Raw data

First, the original dataset file of "Leading causes of death", named "1310039401-eng", was downloaded from Statistics Canada(https://www.statcan.gc.ca). Five key variables are included. "Leading causes of death" tabulated is the underlying cause of death. This is defined as the disease or injury which initiated the train of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury. 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. This article mainly studies the statistical model of death toll, so only the data of all ages and both sexes has been downloaded.

2.2 Data cleaning

In order to understand the data more intuitively, data cleansing is necessary. A detailed description of each variable explains how these variables are critical for understanding public health trends and policy implications.

table1 showcases the first ten rows of Leading Causes of Death in Canada for the Year 2022.To further understand the table,"cause" refers to the event that led to the deaths recorded in the dataset. Each cause is identified by its medical name."deaths" represents the total number of people who passed away due to each listed cause. Years represent……..It should be noted that since the COVID-19 outbreak began in 2020, the data for 'years' is limited to only three years

2.2.1 Preview

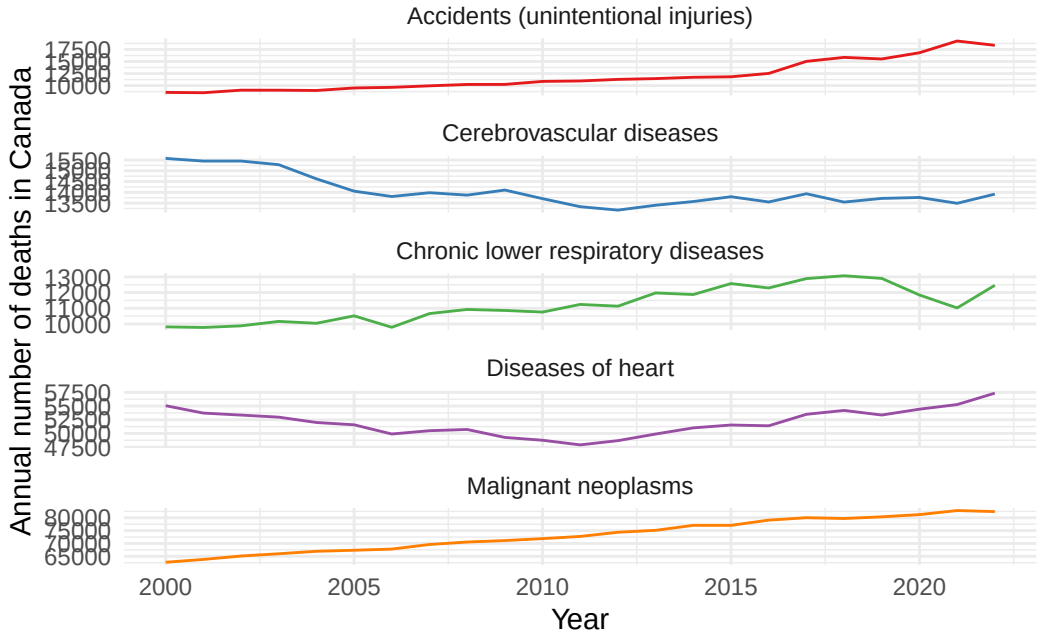
Year	Cause	Deaths	Ranking	Years
2022	Malignant neoplasms	82,412	1	23
2022	Diseases of heart	57,357	2	23
2022	COVID-19	19,716	3	3
2022	Accidents (unintentional injuries)	18,365	4	23

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

table1 showcases the first ten rows of Leading Causes of Death in Canada for the Year 2022. To further understand the table, "Cause" refers to the event that led to the deaths recorded in the dataset. Each cause is identified by its medical name."Deaths" represents the total number of people who passed away due to each listed cause. "Years" represent……..It should be noted that since the COVID-19 outbreak began in 2020, the data for 'years' is limited to only three years

2.2.2 Visualization

Since the entire data set is so large, a simple approach would be to focus on the top five causes of death. Of course, these five causes of death must occur every year from 2000 to 2022, which means that its "Years" value is 23.



3 Result

This table 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:

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

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.

3.2 Estimate

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

3.3 Posterior predictive check

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).

3.4 Resampling

Following this logic, LOOCV(Leave-one-out cross-validation) will not be trustable 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.

The information provided in the table allows us to compare the relative performance of the negative binomial distribution model and the Poisson distribution 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 distribution model has higher "elpd_diff" and "elpd_kfold" values and also has a smaller standard deviation. This shows that the negative binomial distribution model has better predictive performance than the Poisson distribution model in cross-validation. Therefore, we can infer that the negative binomial distribution 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. ## Finding The main focus of this document revolves around a comparison of the performance of two statistical models - the negative binomial distribution model and the Poisson distribution model - in predicting mortality in Canada. Through detailed data analysis, we find that the negative binomial distribution model outperforms the Poisson distribution 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 distribution model has a smaller standard deviation, suggesting that its predictions are more stable. This finding provides a new perspective that negative binomial distribution models may be more appropriate for describing and predicting mortality in Canada.

4.1 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. The superiority of the negative binomial distribution model 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. In terms of insight, the reason why the negative binomial distribution 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.

4.2 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. 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 optimise the negative binomial distribution model, or find other more suitable models to improve the prediction accuracy and stability.

4.3 Weakness and Future Research Directions

Despite the superiority of the negative binomial distribution model in predicting mortality in Alberta, 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.

5 Conclusion