

Analysis the leading causes of death in Alberta*

Comparison the Poisson with negative binomial models

Yunshu Zhang

March 16, 2024

This paper uses a dataset from the government of Alberta to fit Poisson and negative binomial models. When we focused on the top-fifteen causes of death in 2022, the result showed the negative binomial model is a better fit than the Poisson. This analysis revealed the fact that the negative binomial model may fit better than Poisson model in some reality circumstances. In addition, from the data about the cause of death, we can find the most widespread causes of death in Alberta. These insights can guide public health scientists and the policymaker in publishing healthy handbook or guidelines to decrease the mortality.

Table of contents

1	Introduction	2
2	Data	2
2.1	Data Description and Methodology	2
3	Model	3
4	Results	3
5	Discussion	3
6	Conclusion	3
	References	4

*Code and data are available at: https://github.com/Yunshu921/mortality_in_Alberta.git.

1 Introduction

Due to the development of economy, the progress of medical and the electronization of life, our life expectancy has remarkably increased. We can perceive this point by the higher risk of developing chronic diseases compared to the past Diaconu et al. (2016). An aging population is a problem endemic to the Western and industrialized countries. Therefore, by analyzing data about causes of death, we can provide insights to improve the old-age survival. More importantly, normal people can take the information as one's health guideline. To be more specifically, we can take aged parents to do health check for checking specific and prevalent diseases.

In this paper, we will firstly examine the relationship between the time and the number of death for several leading causes of death by using linear regression analysis. Then we will utilize data from leading causes of death in 2022 to fit two models which are Poisson model and the negative binomial model. The estimand is which model fit better in this situation. The result of analysis showed the negative binomial model fit better than Poisson model.

After reviewing literature, they have pointed out the difference between Poisson regression and the negative binomial regression. Thus there is a gap about a specific example which shows advantages of the negative binomial model overweight that of Poisson model. Furthermore, during analyzing the data, some interesting findings can assist healthcare professionals and public health related department to allocate resources effectively.

In this paper, there are 4 sections, excludes the introduction. In the first section, we review the source of data from the government of Alberta, the advantages and disadvantages of data, methodologies that follow it, and data terminology. In addition, we have some plots to show the distribution of the time and the number of death for several leading causes of death. For the second section, we run two linear regression models and explain each variable in detail. In next section, we will display the results by using tables and plots. In the final section, we discuss our results and point out some weaknesses. This paper is carried out using R (R Core Team (2024)), the library Tidyverse (Wickham et al. (2019)), the janitor (Firke (2023)), the here (Müller (2020)) and were used in this paper.

2 Data

2.1 Data Description and Methodology

The dataset used in this paper is from the open data of government of Alberta and can be freely downloaded at their website ((2022)). This dataset was created at 2015 May 13 and last modification was done at 2023 Sept 22. The update frequency is annual and the publisher is Service Alberta. We use this dataset not another province since the representation of data in this dataset is clear and concise. At the same time, the government of Alberta

has collected data regarding economy and finance, society and communities, employment and labour, environment, health and wellness, government, agriculture and other varied topics. The government of Alberta keeps records on a weekly, monthly, and yearly data.

3 Model

4 Results

5 Discussion

6 Conclusion

References

2022. *Alberta.ca*. <https://open.alberta.ca/dataset/03339dc5-fb51-4552-97c7-853688fc428d/resource/3e241965-fee3-400e-9652-07cfbf0c0bda/download/deaths-leading-causes.csv>.
- Diaconu, Viorela, Nadine Ouellette, Carlo G. Camarda, and Robert Bourbeau. 2016. “Insight on ‘Typical’ Longevity: An Analysis of the Modal Lifespan by Leading Causes of Death in Canada.” *Demographic Research* 35 (Vol. 35, JULY - DECEMBER 2016): 471–504. <https://doi.org/https://doi.org/10.4054/demres.2016.35.17>.
- Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://github.com/sfirke/janitor>.
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://here.r-lib.org/>.
- R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.