Analysis the leading causes of death in Alberta*

Reproduce the comparison the Poisson with negative binomial models from 'Telling Stories with Data' (Alexander 2023).

Yunshu Zhang

March 17, 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

*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 heath 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 cause 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 analying 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 cause 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 statistical programming language `R` (R Core Team 2024), the library `tidyverse` (Wickham et al. 2019) for preparation and analysis of data, the `janitor` (Firke 2023), the `here` (Müller 2020) for read data, the `ggplot2` for generating figures, the dplyr (Wickham et al. 2023), and we run the model in R (R Core Team 2024) using the rstanarm package (Brilleman et al. 2018), the boom package (**boom?**), and the boom.Mixed package (Bolker and Robinson 2022). Since this paper is a reproduction of 'Telling Stories with Data' (Alexander 2023), some similar code were used.

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 dataset of other province since the representation of data in this dataset is clear and concise. There are some similar datasets, but the one we used is the more suitable for testing Poisson and the negative binomial models. 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.

The dataset consist first 30 leading causes of death from 2001 to 2022 and each cause has the corresponding number of deaths. In summary, there are 666 rows which contain four rows of descriptive words and two repeated ranking rows at row 34 and row 35. That is there are 664 variables in original dataset which is in form of csv document.

2.2 Data Visualization

Now, we can explore this dataset by using ggplot2 package (Wickham 2016) for generating plots, and the knitr package (Xie 2014) for presenting tables.

Table 1: A summary table of cleaned data

Year	Cause	Ranking	Deaths	Years
2001	All other forms of chronic …	1	1888	22
2001	Acute myocardial infarction	2	1330	22
2001	Malignant neoplasms of trac…	3	1095	22
2001	Other chronic obstructive p…	4	664	22
2001	Stroke, not specified as he…	5	663	22
2001	Atherosclerotic cardiovascu…	6	545	22
2001	Malignant neoplasm of breast	7	426	22
2001	Diabetes mellitus	8	397	22

Table 1: A summary table of cleaned data

Year	Cause	Ranking	Deaths	Years
2001	Other malignant neoplasms o...	9	389	16
2001	Malignant neoplasms of colon	10	386	22
2001	Congestive heart failure	11	338	22
2001	Malignant neoplasms of pros...	12	332	22
2001	Alzheimer's disease	13	329	22
2001	Organic dementia	14	280	22
2001	Atherosclerosis	15	253	11

Table 1 shows the detail of the cleaned dataset which includes 5 variables and 663 observations in total. The variables in the dataset include Year (in years), Cause, Ranking, Death, and Years (the frequency of each cause).

Table 2: A summary table of top fifteen causes of death

Year	Cause	Ranking	Deaths	Years
2022	Organic dementia	1	2,377	22
2022	All other forms of chronic ...	2	2,098	22
2022	Other ill-defined and unkno...	3	1,714	4
2022	COVID-19, virus identified	4	1,547	3
2022	Malignant neoplasms of trac...	5	1,523	22
2022	Acute myocardial infarction	6	1,240	22
2022	Accidental poisoning by and...	7	1,200	10
2022	Other chronic obstructive p...	8	1,183	22
2022	Diabetes mellitus	9	730	22
2022	Stroke, not specified as he...	10	650	22
2022	Atherosclerotic cardiovascu...	11	582	22
2022	Malignant neoplasm of breast	12	509	22
2022	Malignant neoplasms of pan...	13	488	10
2022	Malignant neoplasms of pros...	13	488	22
2022	Congestive heart failure	15	481	22

Table 2 shows top-15 causes of death in Alberta in 2022 and we found some interesting things such as some diseases just showed recently, for example the COVID-19.

Table 3: A summary table of leading causes of death

Year	Cause	Ranking	Deaths	Years
2001	All other forms of chronic ...	1	1888	22
2001	Acute myocardial infarction	2	1330	22
2001	Malignant neoplasms of trac...	3	1095	22
2001	Organic dementia	14	280	22
2002	All other forms of chronic ...	1	1847	22
2002	Acute myocardial infarction	2	1294	22
2002	Malignant neoplasms of trac...	3	1224	22
2002	Organic dementia	14	284	22
2003	All other forms of chronic ...	1	1749	22
2003	Malignant neoplasms of trac...	2	1257	22
2003	Acute myocardial infarction	3	1242	22
2003	Organic dementia	15	281	22
2004	All other forms of chronic ...	1	1739	22
2004	Acute myocardial infarction	2	1337	22
2004	Malignant neoplasms of trac...	3	1284	22

Table 3 shows some causes with variable 'Years' is equal to 22 since 2001. Firstly, we look at top-four causes of death in 2022. Since we do not want to consider some diseases get prevalent recently, these top-four causes of death have the variable 'Years' is equal to 22. Then we want to examine if there is a relationship between total deaths and each of this specific top-four causes since 2001.

Figure 1 shows the linear regression between different cause and time. Organic dementia and all other forms of chronic ischemic heart disease are top 1 and top2 cases of death in 2022. At the same time, malignant neoplasms of trachea, bronchus and lung and acute myocardial infarction are two causes of death at top 5 and top 6 in 2022. For the top 3 and top 4 cases of death in 2022, they do not satisfy the condition that the variable 'Years' is equal to the 22. Thus they were not in so-called top-four causes of death. When we observe these four causes, we can find that only organic dementia has showed surprising increase through 22 year. The initial deaths in 2001 was 280 and the latest death in 2022 was 2874. With the worsening aging issue, the number of deaths due to organic dementia has increased tenfold over the period of 22 years. This cause of death has also risen from the 14th position in 2001 to the current top position. In 2017, organic dementia got the top-one cause of death at the first time and then so far.

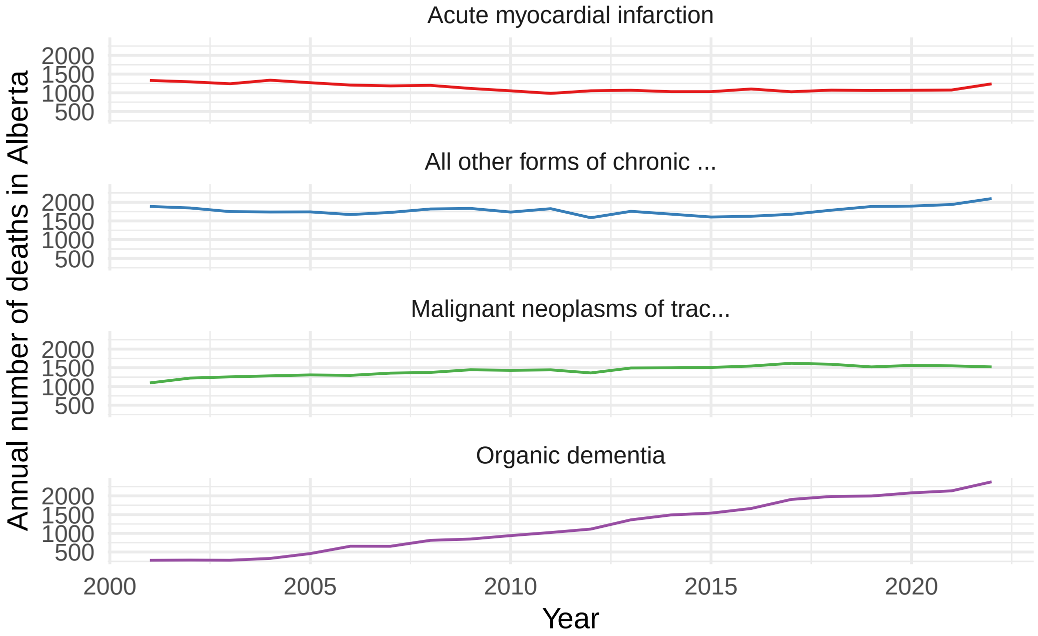

Figure 1: Annual number of deaths for the top-four causes in 2022, since 2001, for Alberta, Canada

Table 4: Summary statistics of the number of yearly deaths, by cause, in 2022, in Alberta, Canada

Min Deaths	Mean Deaths	Max Deaths	SD	Var	N
280	1383	2377	437	190696	88

Table 4 shows the summary of statistic of total deaths in Alberta. In this table, we notice that the mean, 437, is different to the variance, 190,696.

3 Model

Now we can use the `rstanarm` package (Brilleman et al. 2018) to fit two models: Poisson and the negative binomial models.

3.1 Model summary of Poisson and the negative binomial

Table 5 shows estimates for Poisson and the negative binomial models are similar. Thus we need to use other method to compare these two.

Table 5: Modeling the most prevalent cause of deaths in Alberta, 2001-2021

	Poisson	Negative binomial
(Intercept)	7.037	7.039
		(0.079)
causeAll other forms of chronic ...	0.446	0.446
		(0.113)
causeMalignant neoplasms of trac...	0.224	0.221
		(0.110)
causeOrganic dementia	0.046	0.043
		(0.111)
Num.Obs.	88	88
Log.Lik.	−5265.236	−661.564
ELPD	−5445.2	−665.4
ELPD s.e.	1201.2	8.5
LOOIC	10 890.3	1330.8
LOOIC s.e.	2402.4	17.0
WAIC	11 036.1	1330.7
RMSE	353.43	353.44

3.2 Model equation

Firstly, we can get the Poisson probability mass function which is same as the negative binomial is (Pitman 1993)

$$P(X = k) = e^{-\lambda}\lambda^k/k!, fork = 0, 1, 2, 3, ...$$

Following an exploratory analysis of the dataset, the ultimate model is presented below.

$$y_i|\lambda_i \sim Poissonlog_e(\lambda_i) = \beta_0 + \beta_1 \times cause_i$$

where:

- $y_i|\lambda_i$ is the number of deaths for different causes.
- λ_i is the intercept which means the basic mortality without any causes we control.
- β_0 is the intercept or constant term, which represents the expected value of basic mortality when there is no active cause that we interested.
- β_1 is the coefficient or the estimated change in mortality for a change in the cause.
- $cause_i$ is the cause of the death.

In specific, we have two equations of mortality for two models:

Firstly, since

$$y_i = e^{log_e(\lambda_i)}$$

We can have two equations such that

$$y_i = e^{log_e(\lambda_i)} = e^{7.037+0.446I_1+0.224I_2+0.046I_3}, for\ Poisson\ Model y_i = e^{log_e(\lambda_i)} = e^{7.039+0.446I_1+0.221I_2+0.043I_3}, for$$

where

- I_1 means the indicator function such that if this cause happened, that is all other forms of chronic ischemic heart disease, then we get the function value of 1.Otherwise, we get the function value of 0
- I_2 represents the indicator function such that if this cause happened, that is malignant neoplasms of trachea, bronchus and lung, then we get the function value of 1.Otherwise, we get the function value of 0
- I_3 is the indicator function such that if this cause happened, that is organic dementia, then we get the function value of 1.Otherwise, we get the function value of 0

4 Results

4.1 Comparing posterior predictive checks for Poissson model and Negative binomial model

Since two equations have similar estimates in each row and from 'Telling Stories with Data' (Alexander 2023), we can use function `pp_check` in the `rstanarm` package (Brilleman et al. 2018) to compare these two models. This involves comparing the observed outcome variable with simulations generated from the posterior distribution.

Figure 2 and Figure 3 are fitted plots of the Poisson model and Negative binomial model, where the dark and bold lines indicate that the negative binomial approach is a better choice for our circumstance.

Lastly, from 'Telling Stories with Data' (Alexander 2023), we can assess and compare models using the leave-one-out (LOO) cross-validation (CV) resampling method. This approach is a form of cross-validation where each fold consists of only one observation.

Table 6 shows the result of comparison. Since the Negative binomial has higher value in ELPD than Poisson has, the Negative binomial model is our optimal model.

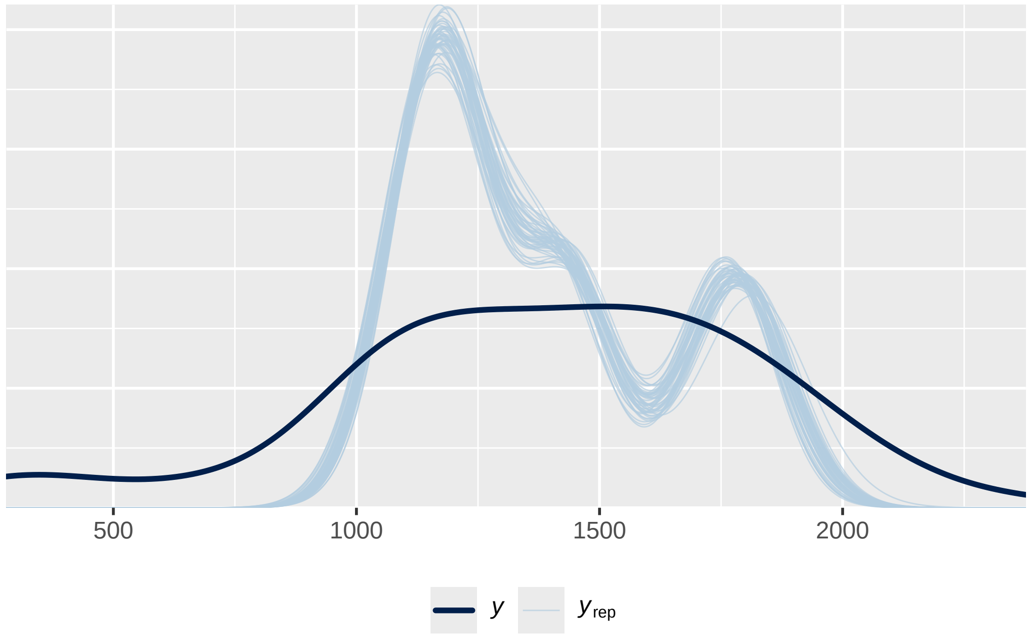

Figure 2: Poisson model

Table 6: Compared data frame

```
                    elpd_diff se_diff
alberta_neg_binomial     0.0       0.0
alberta_poisson      -4779.8    1193.0
```

5 Discussion

5.1 General prosess of this paper

In section three, namely the model section, we firstly examine the formula of Poisson regression. Based on the result from the data visualization, we can know that the SD and Variance of total death we want to do research is not equal. This means Poisson may not the best choice for this situation. Therefore we introduce the negative binomial regression. In the next section, we used two models to fit same dataset. At the beginning of the thing, (**table-model-1?**) showed almost identical statistic. After applying posterior predictive checks and the leave-one-out (LOO) cross-validation (CV) resampling method, we get Figure 2, Figure 3, and Table 6. From two fit lines, we can clear infer that the negative binomial model is our optimal solution for this dataset.

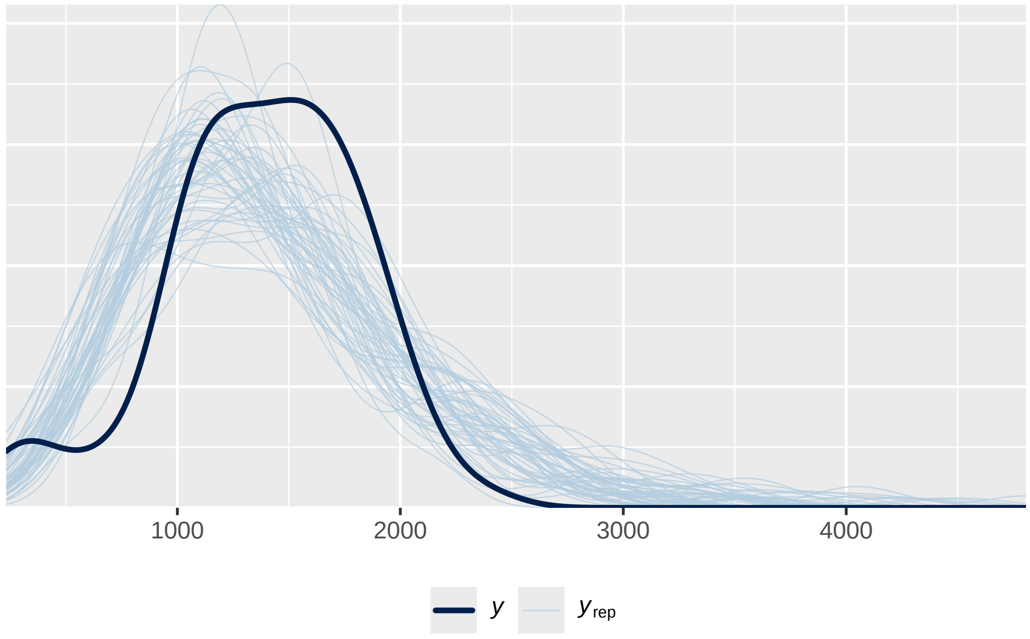

Figure 3: Negative binomial model.

5.2 Organic dementia has taking top one cause of death

From the Figure 1, we can find that the cause of death: organic dementia has irresistibly risen to the top cause of death. After exploring the literature around the word, we find two main reason behind this. The first reason is the spectacular longevity gains. As people's lifespans extend, the likelihood of developing senile dementia also increases. And I think may not the most important reason is that dementia can easily be triggered by other illnesses. For example, from the article published at The Lancet Neurology (Leys et al. 2005), having a stroke will increase the risk of dementia greatly.

5.3 Some inspiration about the leading cause of death

In the real life, data analysis about leading cause of death helps reveals the secret of different disease and shows the truth gradually. From the article: Insight on 'typical' longevity: An analysis of the modal lifespan by leading causes of death in Canada (Diaconu et al. 2016), the greatest gap in modal age-at-death estimates is seen between lung cancer and cardiovascular diseases. When we look at the long time range, from 1974 to 2011, for male, there was a 10 year gap between lung cancer and cardiovascular diseases. For female, there was a nearly 15 years gap between lung cancer and cardiovascular diseases. These two gaps maintained for almost 40 years. Although the modal age is postpones, the difference of two cause of death doesn't change. From this point, doctors in hospital can make more targeted judgments about different diseases.

5.4 Weaknesses

During constructing the model, we can find the Table 5 only has three predictors not four. In this part, we need to examine the principle behind the code. I think give a more detail explanation should be a better solution. In addition, for the overall logic of this paper, we need do more work. For instance, we should use concise and simple sentence to interpret relative complex definitions.

5.5 Improvement

We are supposed to do more examples related to comparing different models. Because in this way we can learn thing better. At the same time, we need read and pracice all the time to keep our mind clear and positive. For this paper, I think the detail of processing data and the wording can be improved.

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.

Alexander, Rohan. 2023. "Telling Stories with Data." *Tellingstorieswithdata.com*. https://tellingstorieswithdata.com.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://github.com/bbolker/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/.

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.

Leys, Didier, Hilde Hénon, Marie-Anne Mackowiak-Cordoliani, and Florence Pasquier. 2005. "Poststroke Dementia." *The Lancet Neurology* 4 (11): 752–59. https://doi.org/https://doi.org/10.1016/s1474-4422(05)70221-0.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Pitman, Jim. 1993. "Repeated Trials and Sampling." *Springer eBooks*, January, 79–137. https://doi.org/https://doi.org/10.1007/978-1-4612-4374-8_2.

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. 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.

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.