Decoding Mortality: Trends and Causes in Canada between 2001-2022*

A Bayesian Statistical Analysis Unveiling the Predominance of Malignant Neoplasms and Heart Diseases

Yuchao Niu

March 17, 2024

The study of mortality offers key public health insights into the fundamental causes and trends of death. This research investigates Canada's mortality trends from 2001 to 2022, employing Bayesian Poisson and negative binomial regression models to analyze data from the Canadian Vital Statistics Death Database. It reveals that malignant neoplasms and heart diseases significantly elevate mortality rates, surpassing the effects of unintentional accidents by five times and three times, respectively. Such findings highlight the critical necessity for targeted public health initiatives focusing on the prevention, early detection, and treatment of these conditions.

Table of contents

1 Introduction 2

2 Data 3
2.1 Data Source 3
2.2 Data Measurement 3
2.3 Data Characteristics 4
2.4 Data Visualization 4

3 Model 5
3.1 Model set-up 5
3.1.1 Model justification 6

[*] Code and data are available at: https://github.com/MelanieNiu/Canadian-Mortality.

4 Results 7

5 Discussion 9
5.1 Findings 9
5.2 Ethical implications 9
5.3 Weaknesses and next steps 10

Appendix 11

A Additional data details 11

B Model details 11
B.1 Posterior predictive check 11
B.2 Diagnostics 12

References 14

1 Introduction

Death is a dreaded yet unavoidable event in life. As Benjamin Franklin famously stated, "In this world, nothing can be said to be certain, except death and taxes." Population mortality has been meticulously recorded over the past century, enabling the study of epidemiological trends, such as those observed in the recent pandemic. Crucially, analyzing region-specific mortality and identifying the leading causes of death inform public health decisions and guide the effective allocation of government resources to prevent and treat diseases. The availability of vast data sources unveils patterns and valuable information regarding causes of death.

This study investigates mortality in Canada from 2001 to 2022, utilizing the Canadian Vital Statistics Death Database by Statistics Canada. The estimand of the study is the relative risk factors for death associated with each cause of death. We employed a Bayesian approach, using both the Poisson regression model and the Negative Binomial regression model, to analyze the top five leading causes of mortality in Canada. The Bayesian model applies Bayesian inference to continually update the probability of a hypothesis as more data becomes available. The Poisson regression model is traditionally favored for count data but assumes equal mean and variance, an assumption that does not always hold in mortality data due to over-dispersion. The Negative Binomial regression model addresses this by accounting for over-dispersion, offering a more flexible approach for such data. Through this study we also would like to compare the performance of the two models for the dataset we are interested in.

The top five leading causes of death identified are malignant neoplasm, diseases of the heart, chronic respiratory diseases, cerebrovascular disease, and unintentional accidents. Malignant neoplasm exhibits the strongest impact death counts, being 5 times more influential than accidents. Diseases of the heart account for 3 times more deaths than unintentional accidents.

The findings advocate for prioritizing public health resources towards the screening, prevention, and treatment of these conditions.

The remainder of this paper is organized as follows. Section 2 discusses the data and measurement methods used in this study. Section 3 presents the model utilized in the analysis. Section 4 presents the results, and Section 5 discusses the implications, limitations, and proposes future directions for research.

2 Data

2.1 Data Source

This study was inspired by the analysis of mortality in Alberta (Alexander 2023), focusing on the leading causes of death, and extends the analysis to Canada, exploring the leading causes of death nationwide. The dataset for this analysis is sourced from the open government portal of Statistics Canada, summarizing the causes of death and the number of individuals affected across all age groups from 2000 to 2022. This data originates from the Canadian Vital Statistics – Death Database (CVSD), an administrative survey that collects demographic and medical information, including the cause of death, from vital statistics registries in all provinces and territories, a practice ongoing since 1921.

2.2 Data Measurement

The target population of the Canadian Vital Statistics – Death database (CVSD) includes deaths occurring in Canada among both Canadian residents and non-residents. As the survey operates on a census with a cross-sectional design, it does not employ sampling methods.

Data on registered deaths is submitted to Statistics Canada by each provincial and territorial Vital Statistics Registry. The death registration form comprises personal details provided to the funeral director by someone knowledgeable about the deceased and a medical certificate of cause of death, completed by the last attending medical professional or a coroner in cases necessitating an inquest or inquiry.

The data collected includes age, sex, marital status, residence and birthplace of the deceased, date of death, underlying cause of death according to the World Health Organization's International Statistical Classification of Diseases and Related Health Problems (ICD), province or territory of death occurrence, place of accident for non-transport accidental deaths, and autopsy details.

Statistics Canada conducts routine quality checks to ensure an error rate below 3%. However, potential biases may arise from the classification of death causes, adherence to the ICD, inclusion of resident data, and data transmission methods. Specifically, the shift in data collection for Canadian residents dying in the United States post-2009, the unavailability of Yukon data post-2017, and the transition to electronic data transmission via the National Routing System (NRS) could introduce biases. These factors may affect the survey's representativeness and accuracy, which are crucial for researchers to consider.

2.3 Data Characteristics

The original dataset obtained from the Statistics Canada open government portal includes 2,268 observations across 18 variables, offering detailed demographic and geographical information on registered deaths. I aim to focus on five specific variables: 'Reference Period', 'Leading Causes of Death (ICD-10)', 'Characteristics', 'Ranking', and 'Number of Deaths'. The 'Reference Period' denotes the year of death registration (2000-2022), while 'Leading Causes of Death' categorizes the cause according to the International Statistical Classification of Diseases and Related Health Problems (ICD-10). 'Characteristics' encompasses both the 'Number of Deaths' and their 'Ranking', indicating either the total deaths or the cause's rank in a given year, respectively. A sample of the cleaned data can be found in Section A.

R (R Core Team 2023) was the language and environment used for the bulk of this analysis, alongside the tidyverse (Wickham et al. 2019), arrow(Richardson et al. 2024), knitr(Xie 2014) , ggplot2(Wickham 2016), broom(Bolker and Robinson 2022), dplyr(Wickham et al. 2023) and have been used in data downloading, cleaning and visualization.

2.4 Data Visualization

Table 1 highlights the top ten causes of death in 2022, with counts reflecting their frequency over 22 years. Except for COVID-19, all other causes have consistently been among the top ten annually. We selected the top five for detailed examination.

Table 1: Annual number of deaths for the top-five causes in 2021, since 2001, for Canada

Year	Cause	Death	Ranking	Counts
2021	Malignant neoplasms	82,822	1	22
2021	Diseases of heart	55,271	2	22
2021	Accidents (unintentional injuries)	19,257	3	22
2021	COVID-19	14,466	4	3
2021	Cerebrovascular diseases	13,491	5	22
2021	Chronic lower respiratory diseases	11,018	6	22
2021	Diabetes mellitus	7,472	7	22
2021	Alzheimer's disease	5,471	8	22
2021	Chronic liver disease and cirrhosis	4,617	9	7
2021	Influenza and pneumonia	4,115	10	22

Figure 1 depicts the annual death trends for these top five causes. There has been a modest increase in deaths from Accidents (unintentional injuries) over two decades. Deaths due to Malignant Neoplasms have risen from 60,000 in the early 2000s to nearly 80,000 in 2023, while Cerebrovascular Diseases, Chronic Lower Respiratory Diseases, and Diseases of the Heart have maintained steady annual deaths at approximately 25,000, 20,000, and 60,000, respectively.

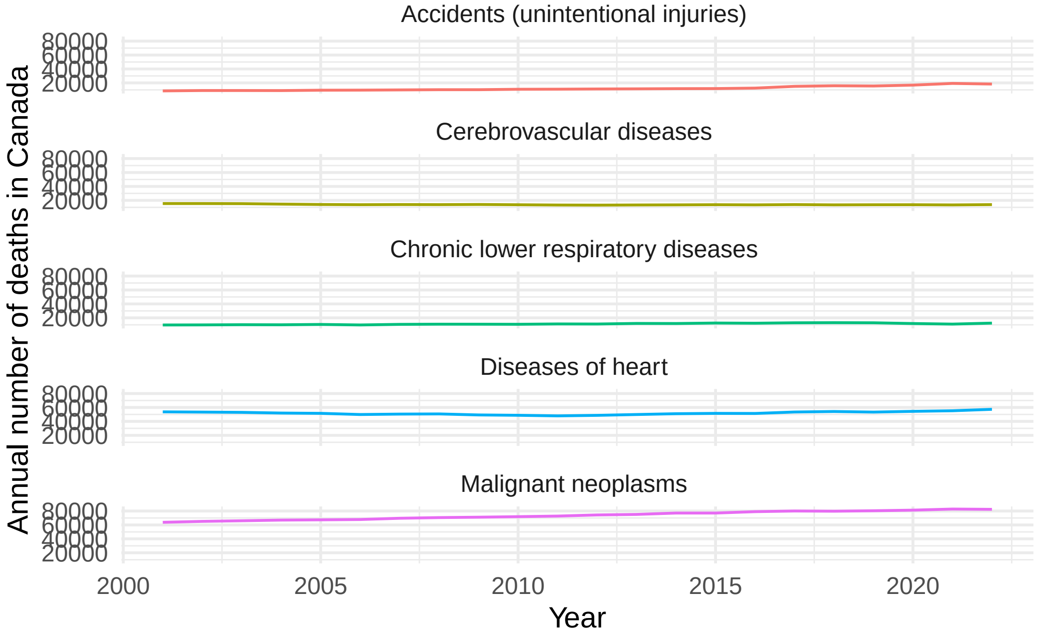


Figure 1: Annual number of deaths for the top-five causes in 2021, since 2001, for Canada

3 Model

The goal of our modeling strategy is to analyze mortality data for the leading causes of death in Canada over the past two decades, aiming to understand the relationship between disease and the number of deaths. We employ a Bayesian approach, assigning prior distributions to parameters based on prior knowledge and specifying a likelihood function based on the observed data. By applying Bayes' theorem, we update our prior beliefs to form posterior distributions of the parameters. Markov Chain Monte Carlo (MCMC) methods are utilized to approximate these posterior distributions, allowing us to integrate prior knowledge and interpret our findings within a probabilistic framework.

3.1 Model set-up

In our Poisson model, define y_i as the number of deaths in a year due to a leading cause of death. Then λ_i is the average rate of deaths due to this cause per year. λ_i is linked to the predictor x_i for the ith observation by a log link function where β_0 is the intercept and β_1 is the coefficient. We specifiy prior distributions for the parameters β_0 and β_1 in Bayesian analysis. We choose β_0 and β_1 to follow a normal distribution with mean 0 and conservative standard deviation of 2.5.

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 X_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)
\end{aligned}
$$

In our negative binomial model, similarly define y_i as the number of deaths due to a leading cause of death in a year. Then μ_i is the average count of deaths in a year. μ_i is linked to the predictor cause of death x_i by a log link function with the intercept β_0 and the coefficient β_i. ϕ_i is the dispersion of the distribution measuring the extent of deviation from the count expected under a Poisson distribution. We specifiy the prior distributions for β_0 and β_1 to follow a normal distribution with mean 0 and conservative standard deviation of 2.5. We specify the prior of ϕ_i to follow a gamma distribution.

$$
\begin{aligned}
y_i &\sim \text{NegativeBinomial}(\mu, \phi) && (5)\\
\log(\mu_i) &= \beta_0 + \beta_1 X_i && (6)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (7)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (8)\\
\phi &\sim \text{Gamma}(2, 0.1) && (9)
\end{aligned}
$$

3.1.1 Model justification

Since the residual of the number of deaths by a leading cause in a year does not follow a normal distribution, we consider using a generalized linear regression model. Since mortality occurrences are discrete events, we consider the Poisson regression model and its variant, negative binomial regression model, both of which relate explanatory variables to dependent variables representing counts of events. Table 2 shows that the mean, 32,603, is different from the variance, 673,243,267. This indicates dispersion in the data, therefore negative binomial regression model may be more suitable.

For model justification, we acknowledge that mortality occurrences are discrete events, making the Poisson and negative binomial regression models suitable for relating explanatory variables to event counts. The negative binomial model is particularly chosen for its capacity to handle

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

	Min	Mean	Max	SD	Var	N
Death	8521	32 603	82 822	25 947	673 243 267	110

overdispersion, a common feature in mortality data, under the assumption that the model's residuals are uncorrelated.

We model the incidences of death by one of the leading cause of deaths using a Poisson regression model and a negative Binomial regression model. Background details and diagnostics are included in Section B.

R packages rstanarm(Goodrich et al. 2022) and modelsummary (Arel-Bundock 2022) are used to build and analyze the models. We use the default priors from `rstanarm`.

4 Results

Table 3 presents coefficients from both Poisson and negative binomial regression models for various causes of death. The intercept, representing the baseline of Accidents (unintentional injuries), indicates the log count of deaths when no other specific cause is considered. Coefficients for causes such as cerebrovascular diseases, chronic lower respiratory diseases, diseases of the heart, and malignant neoplasms reflect log-relative differences in death counts relative to Accidents. Positive coefficients indicate a positive association with the number of deaths.

Table 3: Modeling the most prevalent cause of deaths in Canada, 2001-2022

Model Comparison: Poisson vs. Negative Binomial

	Statistics		
Term	Estimate	Standard Error	Model
(Intercept)	9.403	0.002	Poisson
CauseCerebrovascular diseases	0.142	0.003	Poisson
CauseChronic lower respiratory diseases	0.320	0.005	Poisson
CauseDiseases of heart	1.454	0.002	Poisson
CauseMalignant neoplasms	1.805	0.002	Poisson
(Intercept)	9.404	0.042	Negative Binomial
CauseCerebrovascular diseases	0.141	0.061	Negative Binomial
CauseChronic lower respiratory diseases	0.322	0.126	Negative Binomial
CauseDiseases of heart	1.451	0.061	Negative Binomial
CauseMalignant neoplasms	1.804	0.060	Negative Binomial

For instance, Table 3 shows that diseases of the heart have a coefficient of 1.454 in both models, suggesting the expected death count from this cause is approximately 4.28 times higher than from Accidents (unintentional injuries), after exponentiating the coefficient. Similarly, malignant neoplasms have a coefficient of approximately 1.805, indicating an expected count more than 5.08 times higher than Accidents. Chronic lower respiratory diseases show a modest increase in mortality compared to Accidents, with a coefficient of 0.320 in the Poisson model and 0.322 in the negative binomial model, indicating about 0.38 times more deaths than Accidents. Lastly, Cerebrovascular diseases has a coefficient of 0.141, showing 0.15 times more death counts compare to Accidents although the impact is also modest.

Table 4 displayed the values of the parameters in both models.The LOOIC value (Leave-One-Out Cross-Validation Information Criterion) is useful in comparing model fitness. A lower LOOIC value suggests a model that is expected to make more accurate predictions for new, unseen data. This is because LOOIC penalizes models for having too many parameters (complexity), helping to avoid overfitting. The Poisson model has a LOOIC value of 33604.96 whereas the negative binomial model has a LOOIC value of 1787.36 much lower than the Poisson model. This indicates a better fit of the data by the negative binomial model.

Table 4: Model summary of Poisson model vs. negative binomial model

Model Summary: Poisson vs. Negative Binomial

Metric	Poisson	Negative Binomial
LOOIC	$33,604.96$	$1,787.36$
ELPD	$-16,802.48$	-893.68
ELPD SE	290.89	0.82

Both models yield similar coefficients, suggesting they capture a similar relationship between causes and death counts. The negative binomial model, with a log likelihood of -1050.577, fits the data better than the Poisson model, which has a log likelihood of -16957.794. The negative binomial model's better fit is further supported by a higher ELPD value compared to the Poisson model as shown in Table 5. Posterior predictive checks, trace plots, and Rhat plots assess MCMC convergence and the reliability of the Bayesian analysis, with details in Section B.

Table 5: Cross validation comparision between Poisson model and negative binomial model

Cross-Validation Comparison

Poisson model vs. Negative Binomial model

elpd_diff	se_diff	elpd_kfold	se_elpd_kfold	p_kfold	se_p_kfold
0.00	0.00	-895.2986	7.657298	4.856994	0.5902058
-16538.21	3011.67	-17433.5116	3015.707031	1784.512213	545.2201742

5 Discussion

5.1 Findings

Our analysis over the past 20 years reveals that the expected count of deaths from malignant neoplasm, the leading cause of death, is approximately 6.08 times higher than that from accidents (unintentional). Diseases of the heart, ranking second, account for 3.28 times more deaths than accidents. This contrasts with global statistics from the WHO, highlighting the significance of allocating resources towards cancer prevention and treatment in Canada (Organization 2020). The analysis of Alberta's data shows similarities with national trends, with the notable exception of organic dementia replacing accidents due to its correlation with aging, despite Alberta having the lowest percentage of senior population at 15.2% by 2023 (Denton and Spencer 2019). This calls for further investigation into Alberta's health indicators and lifestyle factors.

The negative binomial regression model is well-suited for datasets exhibiting overdispersion, which is evidenced in our dataset by various model parameters. This overdispersion signifies that the variance of the dependent variable is larger than would be expected under a Poisson distribution, thus making the negative binomial model a more appropriate choice due to its ability to account for this variability.

5.2 Ethical implications

The study ensures confidentiality and respectful data handling, particularly when dealing with rare causes of death. It raises concerns about potential disparities in resource access and safety, urging further research into occupational and transportation inequities. Sampling biases were minimized, but differential data access could lead to underrepresentation, notably for Yukon post-2017, potentially affecting policy benefits (Statistics Canada 2023)

5.3 Weaknesses and next steps

While the Poisson and negative binomial regression models demonstrate good fit, they assume independence of death counts, which may not reflect reality due to clustering by geographic and demographic factors. Additionally, we only considered one variable the leading cause of death, without considering rarer causes or potential confounders like socioeconomic factors. This constrains the ability of the model to capture the nuances of mortality and its drivers.

Future studies could aim to address these weaknesses by incorporating spatial and demographic variables to account for clustering effects.For instance, additional provinces could be included in the investigation to compare and elucidate factors impacting regional mortality. Future studies can broaden the scope to include more socioeconomical variables, such as diet, lifestyles, environmental pollutants and their interactions. Canada is also home to omany centenarians in the world (Statistics Canada 2023); conversely, we could compare mortality and the phenomenon of longevity. Longitudinal studies could offer insights into mortality trends and the long-term effects of environmental factors like climate change.

Appendix

A Additional data details

The raw dataset was downloaded from Statistics Canada open data portal and cleaned by selecting the variables of interest "Year", "Causes of Death", "Characteristics" and "Value". The "Characteristics" variable is converted into a new variable "Death" indicating the number of death attributed to the cause and a new "Rank" variable showing the rank of the cause in a given year. Table 6 presents a sample of the cleaned dataset.

Table 6: A sample of the cleaned dataset of number of death attributed to causes of death in Canada, 2001-2022

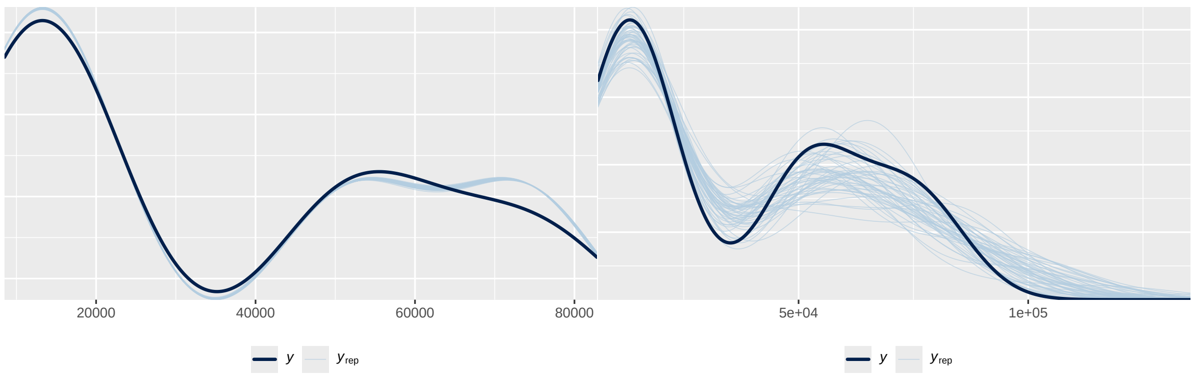
Year	Cause	Death	Ranking
2003	Diseases of heart	52974	2
2020	Salmonella infections	9	41
2005	Operations of war and their sequelae	0	45
2004	Certain conditions originating in the perinatal period	1024	19
2006	Certain conditions originating in the perinatal period	1013	18

B Model details

B.1 Posterior predictive check

In Figure 2a we implement a posterior predictive check for the Poisson regression model described in Section 4. This shows predictions generated from the posterior distribution of the model parameters (represented by the light blue lines) align reasonably well with the actual data (represented by the dark blue line) with minor misalignment. This suggests the model represent the data well.

In Figure 2b we implement a posterior predictive check for the negative binomial regression model described in Section 4 This shows predictions generated from the posterior distribution of the model parameters (represented by the light blue lines) align generally well with the actual data (represented by the dark blue line). This suggests the model represents the data well.



(a) Poisson model

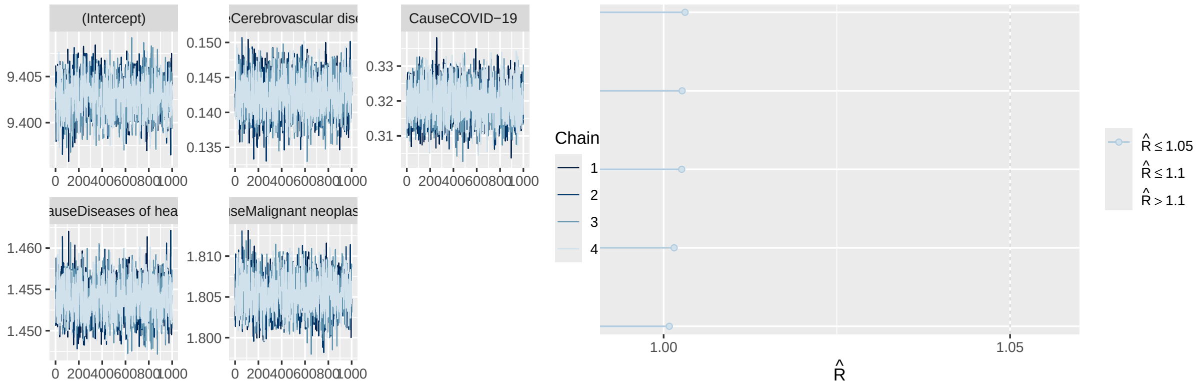
(b) Negative binomial model

Figure 2: Comparing posterior prediction checks for Poisson and negative binomial models

B.2 Diagnostics

Figure 3a is a trace plot for the Poisson regression model. It shows a horizontal, dense band of samples without any systematic patterns, drifts, or long periods of stagnation. This pattern suggests that the chain is mixing well and sampling efficiently from the posterior distribution. This suggests the Poisson regression model is suitable.

Figure 3b is a Rhat plot. It shows an Rhat value falling in the range 1-1.1. This suggests that the chains have converged to the target distribution in the Poisson model.

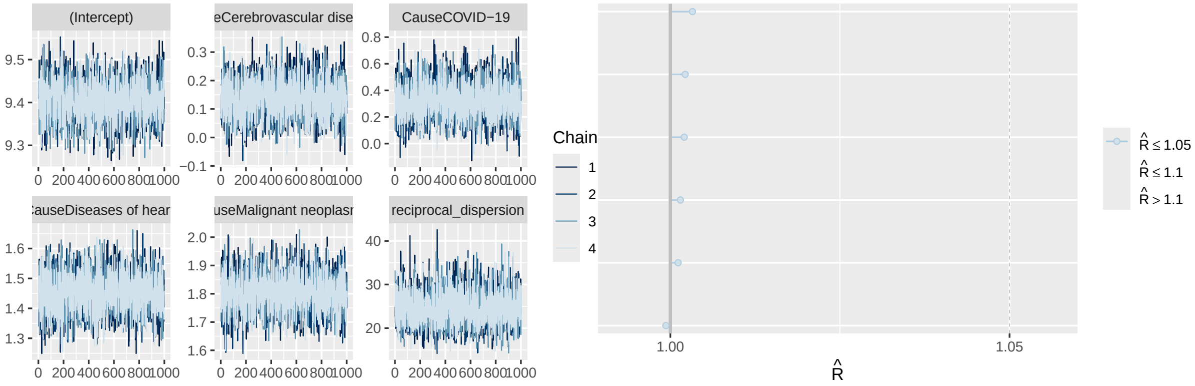


(a) Trace plot

(b) Rhat plot

Figure 3: Checking the convergence of the MCMC algorithm

Figure 4a is a trace plot for the negative binomial regression model. It also shows a horizontal, dense band of samples without abnormalities suggesting the chain is mixing well and sampling efficiently from the posterior distribution. This suggests the negative binomial model is suitable.



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

Figure 4b is a Rhat plot for the negative binomial regression model. It shows an Rhat value close to 1. This again suggests that the chains have converged to the target distribution in the negative regression model. The Rhat value is closer to 1 compared to in the Poisson model showing that the negative binomial regression model fits the data better.

References

Alexander, Rohan. 2023. "13 Generalized Linear Models: Multilevel Modeling." Telling Stories with Data. https://tellingstorieswithdata.com/13-ijaglm.html#multilevel-modeling.

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.

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

Denton, Frank T., and Byron G. Spencer. 2019. "Effects of Population Aging on Gross Domestic Product Per Capita in the Canadian Provinces: Could Productivity Growth Provide an Offset?" *Canadian Public Policy* 45 (1): 16–31. https://doi.org/10.3138/cpp.2018-003.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Organization, World Health. 2020. "The Top 10 Causes of Death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Statistics Canada. 2023. "Leading causes of death, total population, by age group." https://doi.org/10.25318/1310039401-eng.

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.