Trends in Top 10 Causes of Death in Ontario*

Jacob Gilbert Liam Wall

March 16, 2024

Ontario, Canada, mortality data has been available on the internet for years and we have utilised data from 2000 to 2022 to predict the leading causes of death in 2023 to 2028. We examine the top ten leading causes of death and observe the trend in them over 23 years and train a model to use cause of death and the year as predictors for the number of yearly deaths for that cause. We find that year and cause prove to be very good predictors of total death for the years 2000 to 2022 and extends the trends observed into the future until 2028. However, this model does not take into account many of the external factors in medicine and society that are the true causes of these trends, and so we expect that this model explains the future of Ontario mortality if no intervention is made.

1 Introduction

In this paper, we attempt to analyze data on Ontario mortality from 2000 to 2022. In the analysis we create a model predicting the number of deaths per year for a certain cause. From this model we were able to create a prediction for the next five years of the top ten causes of death in Ontario and the corresponding predicted number of deaths. The top three leading causes of death from 2000 to 2022 are malignant neoplasms, diseases of the heart, and other causes of death not listed in the top 43 leading causes of death in Ontario. The model we create accurately reflects this data from 2000 to 2022 and does so to a high degree. The model, using posterior predicting, shows that the next five years will retain the same top three causes. In this paper we estimate that the top three leading causes of death will remain the top three causes, and the estimand is that a combination of the cause based on the year can predict the number of deaths in the near future.

The creation of this report was made possible by R Core Team (2023), Alexander (2023), Wickham et al. (2019) for data cleaning, Goodrich et al. (2022) for data modeling and posterior prediction checks, Wickham (2011) to test many aspects of the code used for the

*Code and data are available at: https://github.com/JfpGilbert0/Ontario-CA-Mortality.

analysis, Zhu (2024) for adjusting and displaying code, Chang (2015) to download the data, and Canada (2023) for making the data available.

2 Data

The data is this paper was retrieved from Canada (2023). The data set contains information on Ontario residents' mortality from 2000 to 2022. The data set lists information on the top 43 leading causes of death each year. It has values for rank of leading cause of death in Ontario, total number of deaths per cause, percentage of deaths pertaining to each cause compared to the yearly total, and age-specific mortality rate. It lists all this information for males, females, and both sexes combined individually as well as for each age group of around 5 years and a combined statistics for all ages. The data set also lists annual information like total deaths per cause per year and total deaths per year. For the purposes of our analysis we will only look at the top ten causes of death in Ontario per year **?@fig-all_values**.

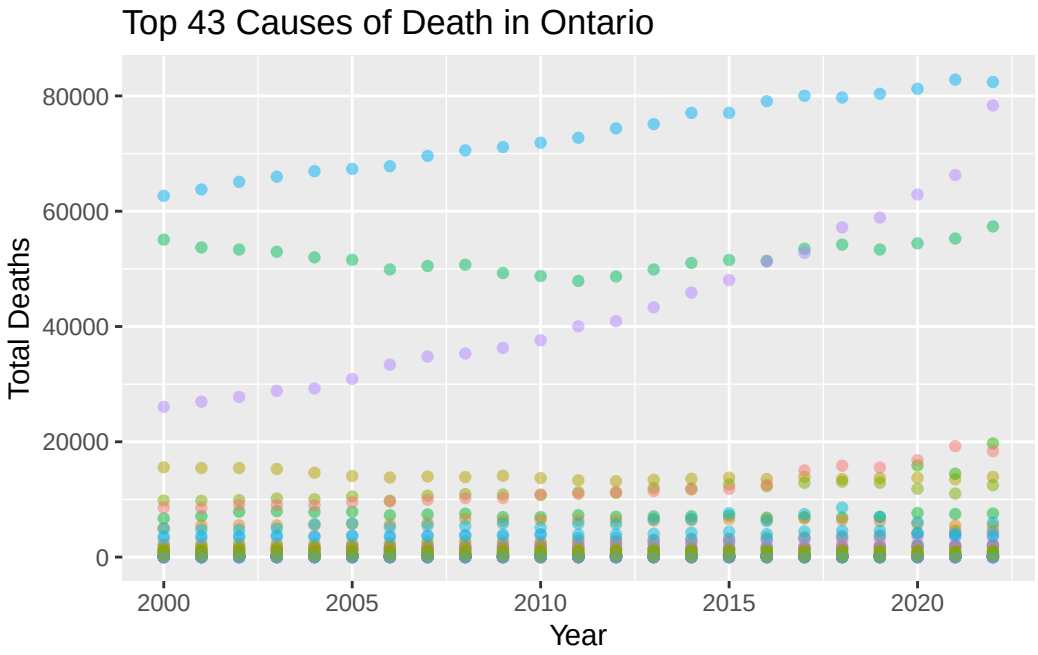

Figure 1: Visual of each value in the dataset, the number of deaths per year per cause. Each color corresponds to a cause

For this research we look at the data pertaining to both sexes and all age groups, which are the cumulative results from the individual information from male and female and from all the age groups combined. The purpose of this paper is not to define age specific or gender specific trends in the mortality of Ontario residents, but rather the overall trend of the leading causes

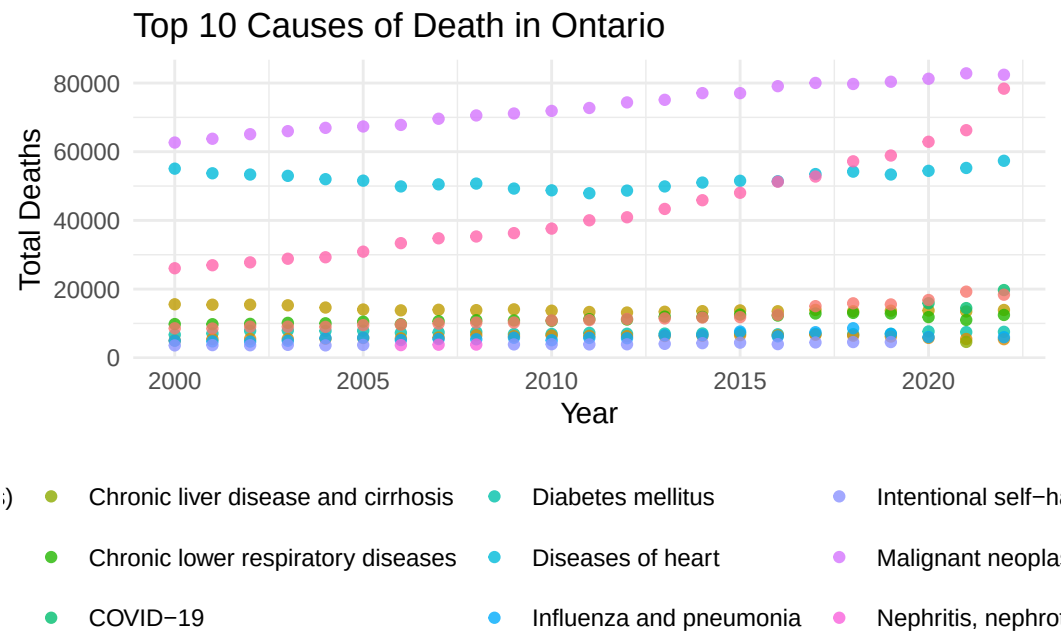

Figure 2: Visual of each value in the dataset, the number of deaths per year per cause. Each color corresponds to a cause

of death and predictions for the future. In this way way we can use the combined data from these groups and still achieve meaningful results in our analysis **?@tbl-header-1**.

The original data set from https://www150.statcan.gc.ca/ contains 407,334 observations of 18 different variables. In our analysis, using the combined data of gender and age, we look at 230 observations of 5 variables: year, cause, total deaths per cause, yearly rank of leading cause of death, and total deaths that year. There are 230 observations for the top ten causes of death over 23 years, 2000 to 2022. In the 23 years, there has only been 13 different causes ranked as the top ten leading causes of death **?@tbl-header-2**. The yearly deaths for each of the top ten causes range from 3,606 to 82,822. The annual deaths in Ontario from 2000 to 2022 range from 218,062 to 334081.

The measurements recorded in the original data set are sourced from hospitals and clinics from around Ontario. The only inputted data to these data sets are the number of deaths per cause. These are annually updated statistics and originally inputted as either male or female statistics and in a certain age group. The resulting cumulative data of both sexes and all age groups is derived from this original data. As both sexes is simply the sum of male and female and all age groups is again the sum of all age groups. The mortality rate is also an annually updated statistics per cause and calculated based on information on the total number of contractions of a disease (not included in thsi data). The total number of deaths, rank of leading cause of death, and percentage of deaths per each cause and for the total of the year are calculated from the originally inputted male and female data for all ages.

3 Model

The model used to simulate, analyse, and predict the next five years of mortality data for the top ten causes of death in Ontario was produced using the rstanarm package and the stan_glm function. Taking advice from Alexander (2023) in his book Telling Stories With Data, we set out to fit our data to a negative binomial distribution. We regressed total deaths for each cause against cause of death dependent on the year. We used the negative binomial

Table 1: Visual of the original dataset and the top 13 leading causes of death from 2000 to 2022.

Year	Leading Cause of Death	Deaths	Rank	Total Annual Deaths
2000	Malignant neoplasms [C00-C97]	62672	1	218062
2000	Diabetes mellitus [E10-E14]	6714	7	218062
2000	Alzheimer's disease [G30]	5007	8	218062
2000	Diseases of heart [I00-I09, I11, I13, I20-I51]	55070	2	218062
2000	Cerebrovascular diseases [I60-I69]	15576	4	218062
2000	Influenza and pneumonia [J09-J18]	4966	9	218062

Leading Causes of Death	Average Rank	Number of Years Present
Malignant neoplasms	1.000000	23
Diseases of heart	2.217391	23
Other causes of death	2.782609	23
Cerebrovascular diseases	4.391304	23
COVID-19	4.666667	3
Accidents (unintentional injuries)	5.391304	23
Chronic lower respiratory diseases	5.521739	23
Diabetes mellitus	7.260870	23
Influenza and pneumonia	8.500000	22
Alzheimer's disease	8.565217	23
Intentional self-harm (suicide)	10.000000	17
Nephritis, nephrotic syndrome and nephrosis	10.000000	3
Chronic liver disease and cirrhosis	10.000000	1

which was predicted by our model. We can see this distribution is very normal looking with a clear mean of 0. The outcome of this model is discussed further in the discussion and results sections.

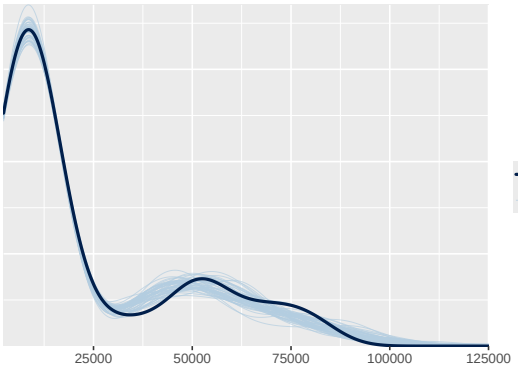

Figure 3: Modeling the most prevalent cause of deaths in Ontario, 2000 - 2022.

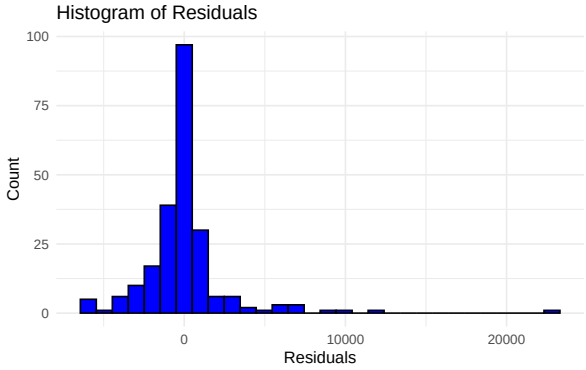

Figure 4: Modeling the most prevalent cause of deaths in Ontario, 2000 - 2022.

4 Results

From the analysis of the mortality data in Ontario of this century, we can see that there are three longstanding leading causes of death whereas the third through tenth leading causes of death are relatively similar and do not undergo large changes relative to total deaths per year in Ontario. We can see that in 2017 diseases of the heart was surpassed by other causes of death for the number two ranking of leading cause of death. Besides that there has been no change of rank for the top three causes of death since 2000 and malignant neoplasms has remained at rank number one Figure 5. As for the fourth through tenth leading causes, there has been many changes in rank however the amount of deaths for each of these causes does not 20,000, and rarely surpasses 15,000. In this way a small change in the amount of deaths per cause could lead to a change of rank much easier than in the case of the top three leading causes.

In the case of the fourth through tenth leading causes, there is one outlier that has been trending upwardly in recent years: accidents and unintentional injuries. We can see this as the orange points in Figure 6. In 2000 this was ranked sixth with only 8,589 deaths to rank four in 2021 with 19,257 deaths.

Using the model we built, having regressed cause and year together to predict the total number of deaths per cause each year, we can use the posterior predictors to predict value for the future. We did this for the next 5 years after 2022. We can see the visual extension of the predicted data with the observed data from 200 to 2022. In Figure 7 we can see the predicted values on the right.

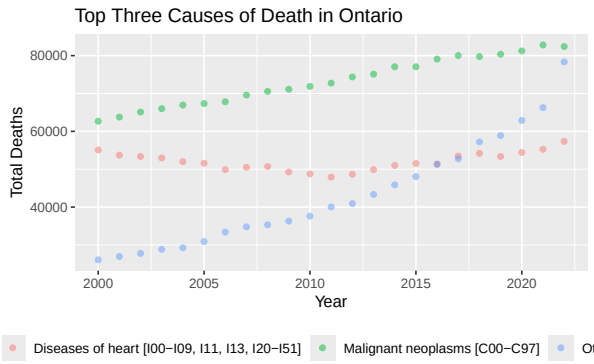

Figure 5: Visual of the trends in total deaths from the top three causes versus the next seven.

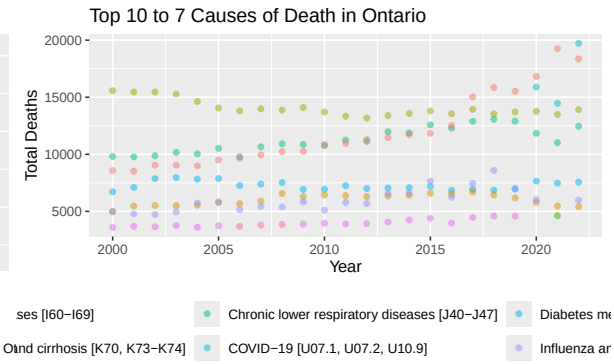

Figure 6: Visual of the trends in total deaths from the top three causes versus the next seven.

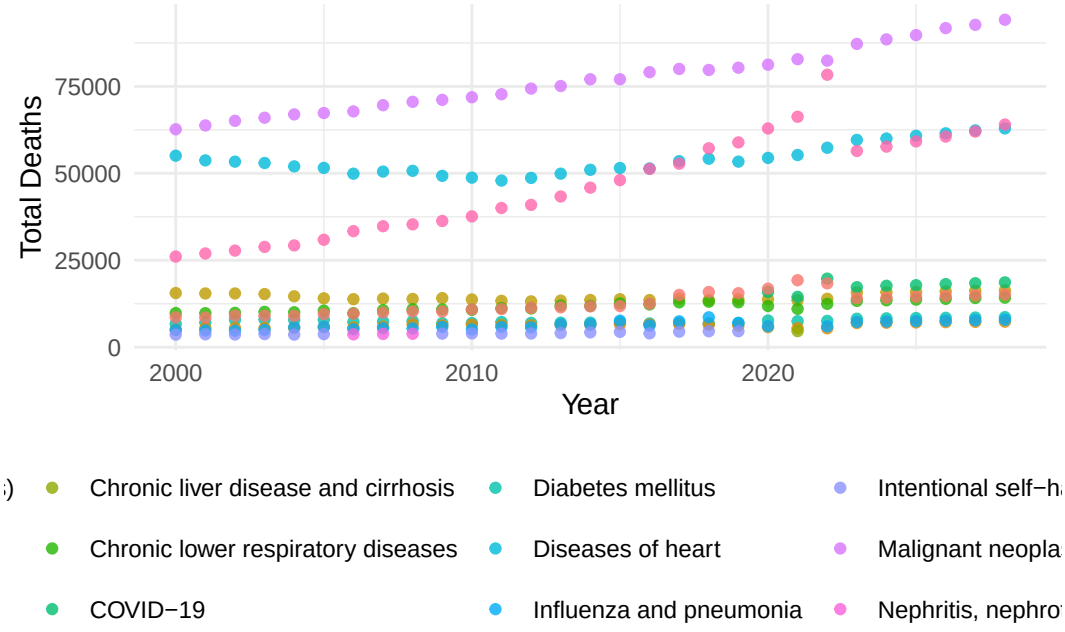

Figure 7: A visual extension of the observed data with predictions from the neagtive binomial model from 2023 to 2028.

5 Discussion

We can see in Figure 8 the total deaths in Ontario have increased greatly, and in recent years has had a somewhat exponential increase. Our model continues this trend and we can see the deaths continue to rise in Figure 9, however, it seems that our model is mistaken in that in 2023 it predicts less than 290,000 deaths whereas in 2022 we observed over 330,000 deaths. Further, when we look at Figure 7, we see again see a similar extension by the model in a linear fashion. It predicts the deaths for each cause to increase in the next years, which is almost certainly an accurate prediction, however in a completely linear way. Naturally, this is reflected in the total deaths and we begin to learn of some of the flaws of this model.

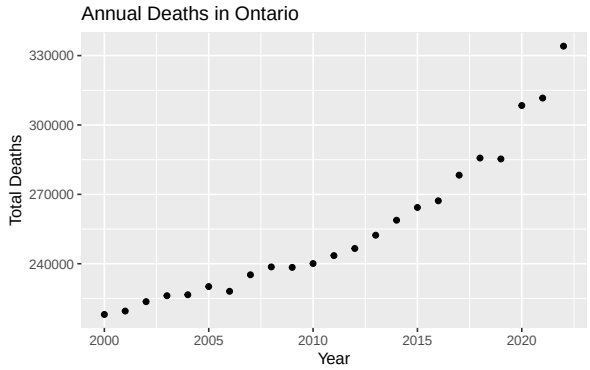

Figure 8: Visual of the total deaths in Ontairo from 2000 - 2022.

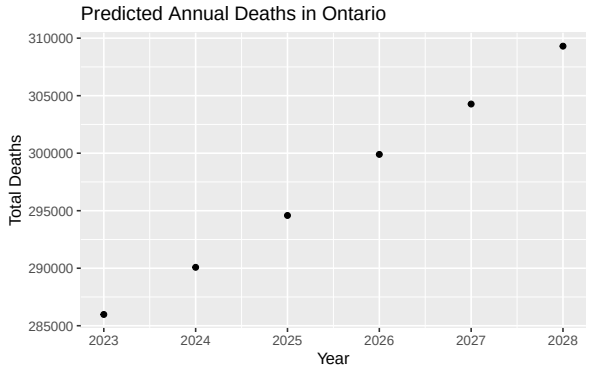

Figure 9: Visual of the total deaths in Ontairo from 2000 - 2022.

Before considering these flaws it is important to note that this model does fit the observed very well. The results of the model show it converges, that is the that values are all one and the as the model increases in sample size, the error will be contained within an interval. Further, using posterior predictive checking **?@fig-model-three** and the "Leave-One-Out" re-sampling method (code for the LOO re-sampling method is available on the Github repo at https://github.com/JfpGilbert0/Ontario-CA-Mortality/tree/main and was not small enough to be in this paper) on the negative binomial model compared to a Poisson fitted model shows the negative binomial approach to be far more accurate to the data. We can also see through posterior predictive checking and looking at the residuals that there is very little error when using year and cause to predict the number of deaths.

Having seen that model fits very well for the observed data and that in the future it predicts very linear changes, one may conclude that the real predictors are not year and cause of death but rather much more specific and less quantifiable factors. Year and cause of death are very good predictors when we have the observed data, however we cannot get mortality data for the future and so the model can only replicate trends it has seen in the data from 2000 to 2022. It is replicating the almost linear gradual change in the values of our data we can see in Figure 2. What the model does not pick up on well enough is that there is exponential

growth of the total deaths per cause, and total deaths in Ontario over the recent past, more specifically since 2017.

The reasons this model is not perfect are many and we can start discussing some of the more obvious ones. First, since 2020 COVID-19 has been prevalent and its effect is not small. In it's first year, it was the fifth ranked leading cause of death in Ontario and in 2022 it became the fourth ranked leading cause. The model predicts these deaths will rise although not enough in the next five years to surpass diseases of the heart at rank three. In the past years, COVID-19 has proved quite sporadic with spikes in summer and new strains always potentially coming. In the future, this model does not account for the effect and nature of COVID-19 and the possibility of it disappearing or becoing more common.

Second, medical intervention and innovation ais always happening and the most money for innovation goes to the leading causes of death. This is obvious because if so many people are dying from malignant neoplasms and diseases of the heart it is only natural that such a large part of the medical industry strives to improve this. The model also does nto account for any kind of medical intervention in the future. It is plausible to assume that there will be significant medical advancements, enough to curb some of the leading causes of death.

Third, the method od reporting these deaths and the leading cause of death is always improving and changing. 20 years ago, we may not have recorded all the deaths as accurately as we would today. Perhaps someone who died of a certain disease was placed into the 'other causes' category when today they could have been placed in 'malignant neoplasms' because of the advancements of the medical industry.

Lastly, this model is not a perfect reflection of the future because over the 20 years, the trends are more linear than anything else. We have seen, especially in Figure 7 that our model extends the data in a very linear fashion, and it does line up with the general data. However, we can see from 2017 to 2022 some causes have taken a more exponential growth, particularly 'other causes of death' and 'accidental deaths.' If the model had been more specific in targeting theses trends, perhaps we would see a graph that flows better, however there are things we can still learn from this model. We learned that although this model takes into account 23 years of mortality data, it is the past few years since 2017 that seem to reflect the trend we should expect. We also learned that even with any kind of exponential growth or linear growth, the top three causes of death will remain the top three causes of death for the next five years. The gap from the top three to next seven causes is more than double the number of deaths of the fourth ranked leading cause of death.

In conclusion, we produced a model to predict the leading causes of death over the next five years, and what we have learned is that there has to be many more factors involved. Simply the year and leading cause of death is not enough to accurately predict the next five years, let alone the smaller trends in the 4th through 10th leading causes of death. We can howveer, learn from the model and reinforce that the top three leading causes of death today, will remain so in the future as their gap is so large to the causes below. These three causes are malignant neoplasms, diseases of the hearts, and other causes of death not listed in the current

top 43 causes of death in Ontario. It is important we take into account modern advancements and data reporting differences from twenty years ago. The method of reporting and medical innovations are probably the most effective factors in predicting the number of deaths. Despite the obvious errors visible in the prediction of the next five years, the model is shown to have fit the observed data very well.

6 Appendix

6.1 More on the Model

In this section I will show that we created two models, on ewith a negative binomial distribution and one with a poisson distribution. The negative binomial fits very well as can be seen in the posterior predictive and residuals graphs **?@fig-model-two**. We also checked the robustness of the models by comparing them in a 'Leave-One-Out' test.

Lastly, we can see using the posterior predictive checking **?@fig-model-three** and the Leave-One-Out re-sampling method that the negative binomial distribution fits the observed data much better than a poisson distribution. (The LOO sampling method code is available on the Github repo available at https://github.com/JfpGilbert0/Ontario-CA-Mortality/tree/main)

```
SAMPLING FOR MODEL 'count' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.000105 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 1.05 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 0.26 seconds (Warm-up)
Chain 1:                0.324 seconds (Sampling)
Chain 1:                0.584 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'count' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 2.6e-05 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.26 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2:
Chain 2:  Elapsed Time: 0.295 seconds (Warm-up)
Chain 2:                0.258 seconds (Sampling)
Chain 2:                0.553 seconds (Total)
Chain 2:

SAMPLING FOR MODEL 'count' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 3e-05 seconds
```

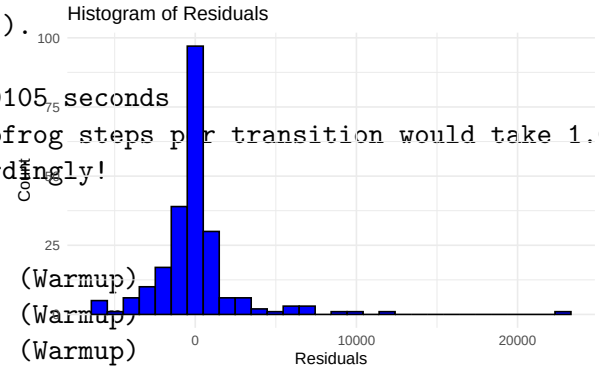

Figure 10: Residuals of the negative binomial distribution are plottedon the left. Residuals of the poisson fitted distribution are plottedon the right.

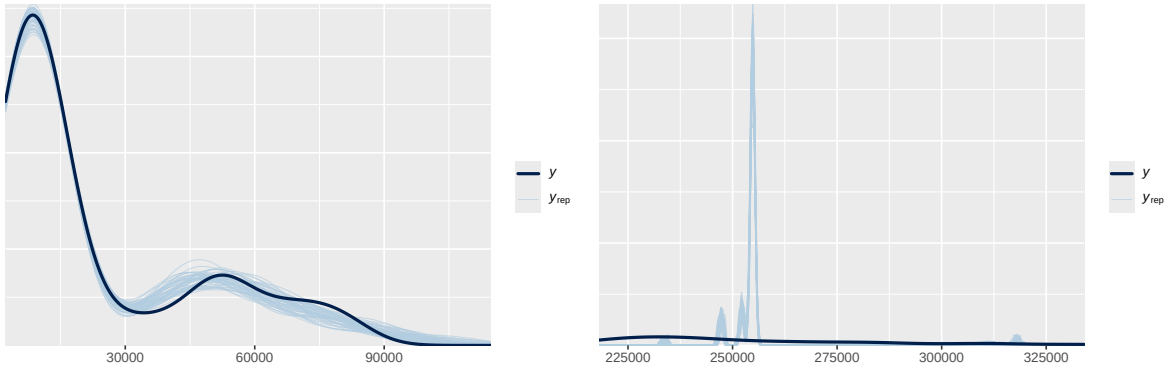

Figure 12: Posterior prediction check for the negative binomial (left) and poisson (right) models.

Figure 13: Posterior prediction check for the negative binomial (left) and poisson (right) models.

References

Alexander, Rohan. 2023. *Telling Stroies with Data*. https://tellingstorieswithdata.com/13-ijaglm.html#negative-binomial-regression.

Canada, atistics. 2023. *Mortality Rates, by Age Group*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310071001.

Chang, Winston. 2015. *Downloader: Download Files over HTTP and HTTPS*. https://CRAN.R-project.org/package=downloader.

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

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

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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.

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