Mortality of Alberta: Study of causes of death after experiencing the Great Depression.[*]

Jierui Miao, Shenze Lu

17 March 2024

Abstract

The recession was horrific, with people facing all sorts of hidden dangers all the time, a collapsing healthcare system, and sky-high drug prices due to inflation. In this report, we choose 2010 as a proxy for the aftermath of the Great Recession and selected the top five causes of death for that year. Using Poisson and negative binomial models, we examine the impact of these causes of death on mortality over the past 20 years and use regression models to predict future trends. The findings suggest that surviving the Great Depression does not mean that chronic diseases, cardiovascular diseases, and tumors have become less harmful to humans. Governments should urge people to be proactive in preventing the onset of disease and raising awareness.

Contents

[*] Code and data supporting this analysis is available at: https://github.com/jerry-maker-765/Mortality--Alberta-during-economy-shut.git

Introduction

The dynamics of health status and causes of death in our society changed as the global economy and healthcare system faced unprecedented challenges, particularly during the latter part of the Great Depression. (Strumpf et al. 2017) The period not only marked a turning point in the global financial and economic landscape, but also had a profound impact on all aspects of society, including public health. Rising unemployment, falling incomes, and cutbacks in social services during the recession may have indirectly affected people's health status and lifestyles, thus having a significant impact on mortality.

This study will focus on the top five causes of death in Alberta in 2010, after the Great Recession, and we find that chronic heart disease dominated, with malignant neoplasms and acute myocardial infarction also contributing significantly to mortality during this period. By deeply analyzing the data from this particular period, we hope to better understand the specific mechanisms by which economic factors influence human health and mortality.

In conducting this study, we began by detailing the data sources and research methods. The dataset contains detailed cause-of-death records, which provided us with an opportunity to comprehensively analyze the impact of various mortality factors on population health. In the modeling section, we built linear regression models with these five major predictors of cause of death and used negative binomial and Poisson methods to study their impact on the number of deaths. Through our analysis, we found that among these five causes, chronic diseases have the greatest impact on the number of deaths, while Atherosclerotic cardiovascular disease seems to have the least impact.

We then illustrated the number of deaths caused by these causes each year from 2001 to 2022 with a line graph showing the change in the number of deaths caused by each cause. The images show that the number of deaths due to these diseases has not decreased over the 20-year period; in fact, the number of deaths due to chronic diseases and malignant neoplasms is on the rise, which is consistent with the conclusions we reached in the modeling section.

With this study, we hope to provide insight for policy makers to emphasize the importance of protecting and promoting public health during economic downturns and to reduce the negative impact of economic depression on population health. We also hope to provide an understanding of the relationship between economic factors and mortality, and also provide a basis for more robust health policies in the face of similar economic challenges in the future.

Data

Backgrounds

The open dataset used for this analysis is called "Leading Causes of Death" and is provided by the Government of Alberta. The dataset provides information on the 30 leading causes of death in Alberta and ranks these causes of death by total number of deaths. The causes of death used in the dataset are based on the International Classification of Diseases, 10th Edition. The dataset includes complete information from 2001 to 2022 and has been updated annually since 2001, with the most recent update occurring in September 2023. Thus, 30 leading causes of death and the number of deaths from those causes have been added to the dataset each year since 2001, and it is possible for a cause to appear in the rankings for several consecutive years. The variable "n" in the dataset will count the number of times each particular cause of death enters the ranking.

Data cleaning will be performed after loading the data. Since the dataset contains one NA observation and three (blank) observations in 2014, 2015, and 2018, this may affect the process and results of this analysis. Therefore, we remove all observations with NA or (blank) from the dataset before conducting any analysis. The reason we do not use estimation or any other method to fix missing data is that the missing data appear in the "Reason" column, which is a "String" column. Therefore, instead of using estimation or any other method, we will remove these rows of missing data and there will be no waste of data since the missing data appears in the "String" column. After that, the pull function is used to extract only the names of these five causes of death. Finally, the original dataset, alberta_cod, is updated to retain those records whose cause-of-death names appear within the top five by means of the filter function and the %in% operator. After this processing, the dataset will contain only records related to the top five causes of death in Alberta in 2010

Our report's original dataset is sourced from the Open Government program of Alberta province(2024). And the R packages we used for programming this report by R language(R Core Team 2020) include following packages:tidyverse(Wickham et al. 2019), boot(Canty and Ripley 2022), broom.mixed(Bolker et al. 2021), collapse(Krantz 2021), dataverse(Leeper 2021), gutenbergr(Robinson 2021), janitor(Firke 2021), knitr(Xie 2021), marginaleffects(Arel-Bundock 2022a), modelsummary(Arel-Bundock 2022b), rstanarm(Team 2022), ggplot2(Wickham 2016), lubridate(Garrett Grolemund 2021), kableExtra(Zhu et al. 2024) and gridExtra(Auguie and Antonov 2017).

Variables

In table 1 the variables of the dataset are introduced.

Table 1: Types and descriptions for variables.

Column	Type	Description
Calendar year	num	A numeric variable. This variable represents the year in which the deaths occurred. It's used to track changes in death rates and causes over time.
Cause	str	A string variable. The cause variable would detail the reason for death, such as specific diseases or accidents. This information is crucial for understanding public health trends and for policy-making.
Ranking	num	A numeric variable. Ranking indicate the position of each cause of death in terms of its frequency compared to others within the same year. A lower ranking number(e.g., 1) would imply a higher frequency of deaths due to that cause.
Total deaths	num	A numeric variable. This variable represents the total number of deaths attributed to each cause in a given year. It's used to measure the impact of different causes of death on the population.
N	num	A numeric variable, counts the number of times for a specific cause of death enter the ranking, which is equivalent to the number of times for a specific cause of death appears in the dataset until the latest update.

Model

Table 2: Top 10 causes of death in Alberta in 2010.

Year	Cause	Ranking	Deaths	Years
2010	All other forms of chronic …	1	1,737	22
2010	Malignant neoplasms of trac…	2	1,431	22
2010	Acute myocardial infarction	3	1,053	22
2010	Organic dementia	4	939	22
2010	Atherosclerotic cardiovascu…	5	898	22
2010	Other chronic obstructive p…	6	826	22
2010	Stroke, not specified as he…	7	686	22
2010	Malignant neoplasms of colon	8	468	22

Year	Cause	Ranking	Deaths	Years
2010	Malignant neoplasm of breast	9	393	22
2010	Alzheimer's disease	10	390	22
2010	Diabetes mellitus	10	390	22

In reviewing the processed dataset, the leading causes of death for 2010 were extracted from the dataset, as shown in Table 2.

In this study, we used Poisson and negative binomial models to predict the total number of deaths caused by a specific cause of death, thus providing an in-depth analysis of the impact of various causes of death on the total number of deaths. This study focuses specifically on the top five causes of death as of 2010: all other forms of chronic ischemic heart disease; malignant neoplasms of the trachea, bronchus, and lungs; acute myocardial infarction; organic dementia; and atherosclerotic cardiovascular disease. By using a count data model, we will explore the extent to which different causes of death contribute to the number of deaths and compare which of the Poisson and negative binomial distributions is more appropriate for describing mortality data in Alberta. Given the characteristics of count data, the Poisson distribution, as the simplest regression model, is suitable for cases where the counts result in non-negative integers. (Consul and Famoye 1992) However, considering the over-dispersion that actual data may exhibit, the negative binomial model, which is an extension of the Poisson model to include a stochastic component that reflects uncertainty in event rates, (Gardner, Mulvey, and Shaw 1995) may be more appropriate for use in the analysis of the data in this study. With this comparison, this study aims to provide a more accurate model to predict the impact of different causes of death on the total number of deaths, thus providing a scientific basis for the formulation of targeted public health policies.

Table 3: Comparison of Poisson and negative binomial results

	Poisson	Negative binomial
(Intercept)	7.038	7.040
		(0.072)
causeAll other forms of chronic ...	0.446	0.444
		(0.100)
causeAtherosclerotic cardiovascu...	−0.436	−0.437
		(0.100)
causeMalignant neoplasms of trac...	0.223	0.222
		(0.101)
causeOrganic dementia	0.045	0.046
		(0.097)
Num.Obs.	110	110
Log.Lik.	−5571.913	−805.093
ELPD	−5741.1	−809.3
ELPD s.e.	1213.1	11.1
LOOIC	11 482.2	1618.7
LOOIC s.e.	2426.3	22.1
WAIC	11 671.4	1618.6
RMSE	320.44	320.45

Poisson model

$$log(E(y)) = 7.038 + 0.446x_1 - 0.436x_2 + 0.223x_3 + 0.045x_4$$

{#eq-equation1}

The Poisson model analyzes the top five causes of death in 2010 and how they affect the expected total number of deaths, denoted as E(y). The intercept corresponds to the impact of Acute Myocardial Infarction. Variables x1 to x4 represent the effects of four different causes of death, with their coefficients indicating the direction of impact (positive values indicate a positive effect, negative values indicate a negative effect): - x1 represents the impact of "All other forms of chronic ischemic heart disease." - x2 represents the impact of "Atherosclerotic cardiovascular disease," with a negative coefficient, indicating that this cause is negatively correlated with the number of deaths. - x3 represents the impact of "Malignant neoplasms of trachea, bronchus, and lung." - x4 represents the impact of "Organic dementia."

This model indicates that x1 (other forms of chronic ischemic heart disease) has a positive impact on the number of deaths, meaning it increases the total number of deaths. On the other hand, x2 (atherosclerotic cardiovascular disease) has a negative impact on the number of deaths, implying that it reduces the total number of deaths. The positive coefficients for x3 and x4 demonstrate that they also contribute to an increase in the total number of deaths, although the magnitude of their impact varies.

$$E(y) = e^{7.038+0.446x_1-0.436x_2+0.223x_3+0.045x_4}$$

{#eq-equation2}

This equation is derived from the exponential transformation of the Poisson regression model, providing a more intuitive way to understand how various factors affect the expected value of the number of deaths, denoted as E(y). For instance, if the value of a factor increases and its corresponding coefficient is positive, then E(y), the expected total number of deaths, will increase. Conversely, if a factor's coefficient is negative, then E(y) will decrease. Such a model allows us to quantify the impact of specific health conditions on the total number of deaths, offering a scientific basis for the formulation of public health strategies and the allocation of resources.

negative binominal model

Considering the possibility of overdispersion after completing the Poisson model analysis, we decided to conduct an analysis using the negative binomial regression model next. This approach is often taken when the variance in the data significantly exceeds the mean, which can happen in count data like the number of deaths. The negative binomial regression can handle overdispersion by introducing an extra parameter to account for the variability, making it a suitable alternative for more accurately modeling the data under such conditions.

$$log(E(y)) = 7.04 + 0.444x_1 - 0.437x_2 + 0.222x_3 + 0.046x_4$$

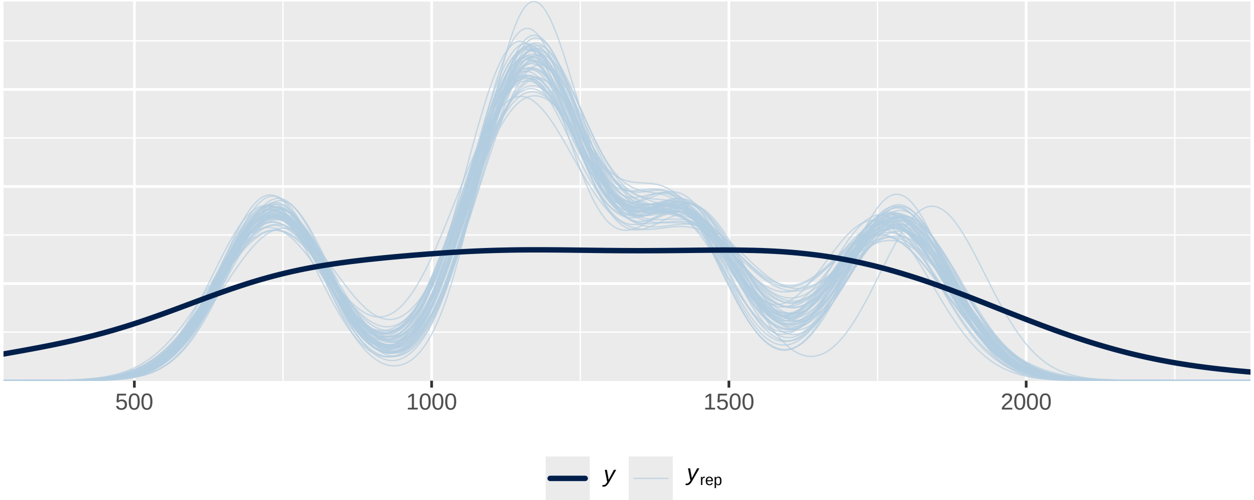
{#eq-equation3}

$$E(y) = e^{7.04+0.444x_1-0.437x_2+0.222x_3+0.046x_4}$$

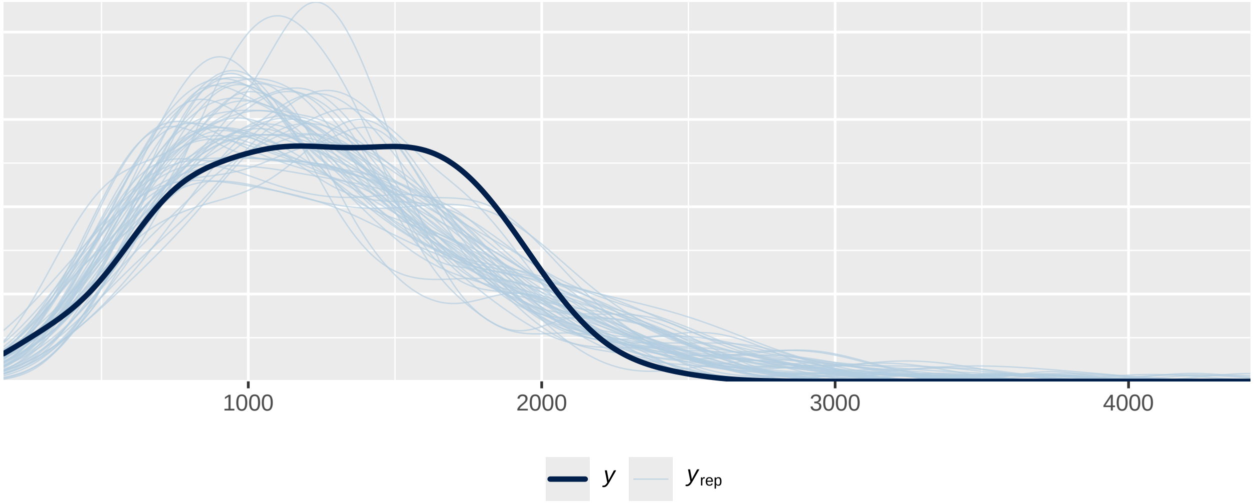
{#eq-equation4}

Poisson vs Negative Binomial model

Poisson Model Check for Alberta Cause of Death



Negative Binomial Model Check for Alberta Cause of Death



We created graphs for both the Poisson model and the negative binomial model to observe if these models are suitable for estimation.

The first graph shows a distribution with clear peaks and troughs. The predictions generated by the model align well with the actual data in some areas but show significant deviations in others. This discrepancy might indicate that the model has not captured all the variations in the data, especially at the peaks and troughs.

The second graph presents a relatively smooth and continuous distribution, where the model's predictions (in light blue) closely follow the actual observations (in dark blue). This typically indicates that the model's predictions are more consistent with the actual data.

It is evident that the precision of the Poisson model's predictions is not as high as that of the negative binomial model. To further confirm our observation, we can use the resampling method Leave-One-Out (LOO) Cross-Validation (CV) to compare the models.

Table 4: ELPD and SE difference in Negative Binomial and Poisson models.

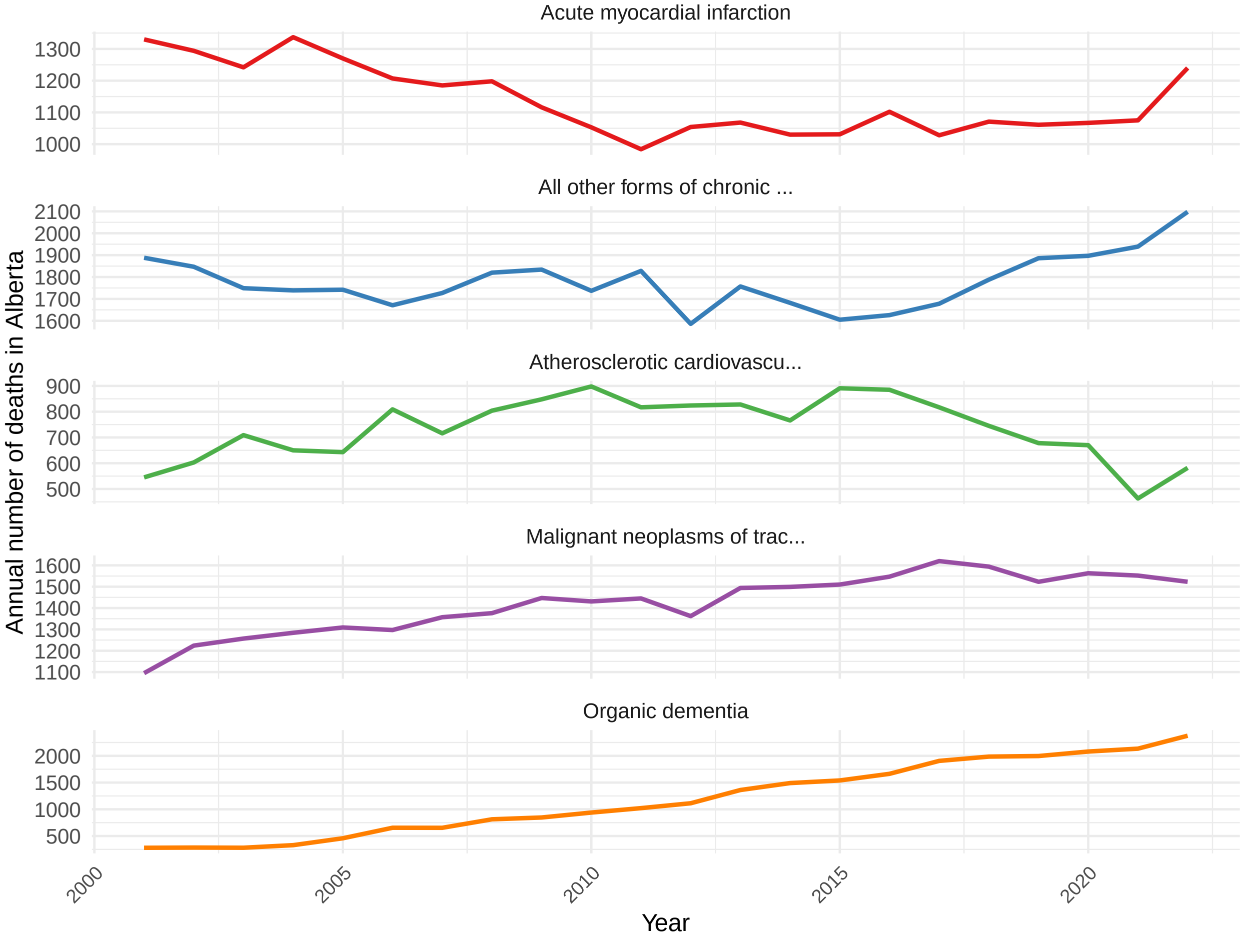
	ELPD difference	SE difference
Alberta CoD under Neg. Binomial	0.000	0.000
Alberta CoD under Poisson	-4931.767	1202.595

In the LOO comparison results, the LOOIC difference for the cause_of_death_alberta_neg_binomial model is 0.0, while for the cause_of_death_alberta_poisson model, it is -4954.0. This negative value indicates that the predictive performance of the Poisson model is significantly worse than that of the negative binomial model. The second column shows the change in the effective sample size (displayed here as 1208.0), which is an estimate of the uncertainty in the model comparison.

These results indicate that among the two models, the negative binomial regression model (cause_of_death_alberta_neg_bin performs better under Leave-One-Out cross-validation, meaning it has better predictive performance for the data. This further supports the previous observation that the negative binomial model is more suitable for handling overdispersed count data.

result

Figure 2: Annual number of deaths in Alberta for the top−five causes in 2010, since 200



Using line graphs, we looked at trends in the number of deaths from these five causes of death from 2001 to 2022. In Alberta, trends over time in the annual number of deaths from the five leading causes of death show different patterns. Annual deaths from acute myocardial infarction and all other forms of chronic ischemic heart disease have remained relatively stable, with no significant upward or downward trends. In contrast, the number of deaths from atherosclerotic cardiovascular disease has declined slightly. On the other hand, the number of deaths from malignant tumors of the trachea, bronchus and lungs rose slightly, while the number of deaths from organic dementia showed a more significant increasing trend. Such changes may reflect the aging of the population and changes in the health care field.

Table 5: Comparison of Poisson and negative binomial estimates

Cause	Poisson	Negative.Binomial
chronic ischemic heart disease	1779.3439	1779.3439
Atherosclerotic cardiovascular disease	736.5669	737.3038
Malignant neoplasms of trachea, bronchus and lungr	1423.6795	1425.1039
Organic dementia	1191.5378	1195.1178
Acute myocardial infarction	1139.1071	1141.3876

Through the models we have developed, we can simulate the number of deaths and ultimately make a comparison. The comparison shows that the negative binomial model's results are closer to the actual situation for deaths caused by “Malignant neoplasms of trachea, bronchus, and lung” (1425 vs. 1431). The Poisson model’s predictions for deaths caused by “Organic dementia” and “Atherosclerotic cardiovascular disease” are closer to the real situation.

This conclusion might seem to conflict with the earlier observation that the negative binomial model fits the real situation more closely. However, this does not imply that the negative binomial model is less accurate. Earlier studies have shown that the negative binomial model has better predictive power. But under certain conditions (i.e., when the data is not overly dispersed), the Poisson model might be sufficient. In most real-world situations, especially when data exhibits overdispersion, the negative binomial model often provides more accurate predictions. The negative binomial model is better suited to adapt to different data fluctuations. This might explain why the Poisson model appears to perform better in simulating data in this context.

Discussion

From the analysis of trends in mortality rates in Alberta from 2000 to 2020 reveals distinct patterns in public health outcomes. Specifically, the data demonstrates a stable decreasing trend in deaths attributed to cardiovascular diseases, which could be caused by advancements in medical technology, enhanced emergency response protocols, and improved management of risk factors such as hypertension and high cholesterol. Public education on heart health appears to have contributed positively to these outcomes. In contrast, a marked increase in mortality rates due to organic dementia points towards demographic shifts towards an older population and advancements in diagnostic techniques. Additionally, an uptick in deaths from malignant respiratory system neoplasms underscores the potential impacts of long-standing smoking habits and environmental factors such as industrial pollution, coupled with gaps in screening processes. These findings highlight the successes in combating cardiovascular diseases through innovation and public health initiatives, while also stressing the need for focused efforts on dementia and certain cancers, with an emphasis on prevention, early detection, and novel treatment approaches.

what is more,our study examined the effectiveness of Poisson and Negative Binomial Regression models in estimating the expected total deaths (E(y)) from the top five causes of death before 2010. Through thorough model evaluation, including graphical analysis and Leave-One-Out Cross-Validation (LOO CV), our findings highlight significant differences in model performance and applicability.

The Poisson Regression Model is known for its simplicity and the assumption that the mean and variance are equal, providing a basic understanding of how various causes of death impact expected mortality numbers. However, this model shows limitations in capturing the complexity of the data, especially in cases of overdispersion (where variance exceeds the mean). This is evident from the observed differences between model predictions and actual data, particularly at points of fluctuation like peaks and troughs.

In contrast, the Negative Binomial Regression Model offers a more robust alternative by incorporating an additional parameter to account for variance, addressing the challenge of overdispersion. Comparative analysis shows that the Negative Binomial model has superior predictive performance, with predictions more consistent with actual observations and significantly better LOO CV scores. The effectiveness of this model is demonstrated by its ability to accurately simulate the number of deaths from specific causes, such as “malignant neoplasms of the trachea, bronchus, and lung,” showing its adaptability to data fluctuations.

Also, while the Poisson model produced closer predictions in some instances, such as deaths caused by “organic dementias” and “atherosclerotic cardiovascular diseases,” this did not detract from the overall superior performance of the Negative Binomial model in handling overdispersed count data. Our results emphasize the importance of model selection in epidemiological research, advocating for the use of the Negative Binomial model in cases of overdispersion to ensure more accurate and reliable estimation of mortality numbers.

This comparative analysis underscores the urgent need for researchers to carefully consider the characteristics of their data and the fundamental assumptions of their statistical models. By doing so, we can enhance the accuracy of public health predictions and interventions, ultimately contributing to more effective health resource management and better health outcomes.

references

2024. *Alberta.ca*. https://www.alberta.ca/open-government-program.

Arel-Bundock, Vincent. 2022a. *Marginal Effects for Regression Models*. https://CRAN.R-project.org/package=marginaleffects.

———. 2022b. *Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Auguie, Baptiste, and Anton Antonov. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://cran.r-project.org/package=gridExtra.

Bolker, Ben et al. 2021. *Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Canty, Angelo, and Brian D. Ripley. 2022. *Bootstrap Functions (Originally by Angelo Canty for s)*. https://CRAN.R-project.org/package=boot.

Consul, PoC, and Felix Famoye. 1992. "Generalized Poisson Regression Model." *Communications in Statistics-Theory and Methods* 21 (1): 89–109.

Firke, Sam. 2021. *Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gardner, William, Edward P Mulvey, and Esther C Shaw. 1995. "Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models." *Psychological Bulletin* 118 (3): 392.

Garrett Grolemund, Hadley Wickham. 2021. *Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Krantz, Sebastian. 2021. *Advanced and Fast Data Transformation*. https://CRAN.R-project.org/package=collapse.

Leeper, Thomas. 2021. *Client for Dataverse 4 Repositories*. https://CRAN.R-project.org/package=dataverse.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2021. *Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Strumpf, Erin C., Thomas J. Charters, Sam Harper, and Arijit Nandi. 2017. "Did the Great Recession Affect Mortality Rates in the Metropolitan United States? Effects on Mortality by Age, Gender and Cause of Death." *Social Science & Medicine* 189: 11–16. https://doi.org/https://doi.org/10.1016/j.socscimed.2017.07.016.

Team, Stan Development. 2022. *Bayesian Applied Regression Modeling via Stan*. https://CRAN.R-project.org/package=rstanarm.

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.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

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