Top 10 Mortality Causes for Republic of Korea of the years 2000, 2010, 2015, 2019*

Poisson and Negative Bionimal Modelling of Annual Death Number and Cause

Hyuk Jang

March 15, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/anggimude/Top-Mortality-Causes-of-South-Korea

1 Introduction

The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Raw Data

The data used in this paper is derived from WHO(WHO) and was downloaded from the WHO Mortality Database. WHO(WHO) provides data for years country-level Global Health Estimates(GHE2019) for the years 2000-2019. Because the years that have estimates of a list provided of the cause of death categories in terms of International Classification of Diseases, Tenth Revision(ICD-10) in terms of a summary table of number of deaths by cause, age, and sex for WHO(WHO) member states for the years are 2000, 2010, 2015, 2019; the data for these four years are cleaned and analyzed for this paper. The analysis of deaths by cause of the raw data is executed for the age groups from 5 to 85+. This paper looks into the data for Republic of Korea as the author is South Korean but also because WHO methods and data sources(citation) certifies a high quality of data. The raw data includes columns such as code, cause, IS03, year, sex, age group, population, deaths, death rate per 100000 population, DALY, DALY rate per 100000 population.

The cleaning, testing, and modelling of the data for this paper was done through R (R Core Team 2023) with the aid of the following packages: tidyverse (**citetidyverse?**), dplyr (**citedplyr?**), rstanarm (**citerstanarm?**), ggplot2 (**citeggplot2?**), modelsummary (**citemodelsummary?**), bayesplot (**citebayesplot?**), parameters (**citeparameters?**), broom (**citebroom?**), kableExtra (**citekableExtra?**), gt (**citegt?**), readr (**citereadr?**), broom.mixed (**citebroommixed?**).

2.2 Cleaned Data

The data that is needed for this paper is year, cause, and number of deaths, so the raw data is cleaned to contain only the three columns we need. Because we are looking into the top 10 mortality in this paper, we rank the causes based on its number of deaths and merge it into Table 1. Now that the cleaning is done we can see the top 10 causes of deaths for the years 2000, 2010, 2015, and 2019, however, this isn't enough because to make a graph and create a poisson and negative binomial model, we must find the causes that are common in all of the years. Table 2 represents the table in which only the causes that appear among all the years descending order of ranking. Now we can recognize the six main causes of death in South Korea is stroke, Ischaemic Heart Disease, Stomach Cancer, Trachea Bronchus Lung Cancer, Liver Cancer, and Self Harm. Looking at Figure 1, we can see some interesting trends. For example, there is a plunge in the annual number of deaths from stroke over the 9 year

span. On the other hand, there has been increases for the causes of Ischaemic heart disease and Trachea, Bronchus, Lung cancer. There is a slight decrease in stomach cancer while liver cancer didn't fluctuate as much. Self harm is interesting because Korea is known to have the highest suicide rates out of all the OECD countries(Citation), and we see an increase in the number deaths from self harm has increased rapidly from 2000 to 2010 and a small decrease from 2010 to 2015. It seems like the death numbers from self harm has plateaued with a very slow rate of increase. In general, the top 10 most common causes of deaths are stroke, heart disease, stomach cancer, lung cancer, road injury, liver cancer, diabetes, Cirrhosis of the liver(liver damage), and self harm.

Table 1: Top 10 Mortality Rates of South Korea

Year	Cause	Deaths	Death Rate	Ranking
2000	Stroke	44109	93	1
2000	Ischaemic heart disease	18837	39	2
2000	Stomach cancer	13205	27	3
2000	Trachea, bronchus, lung cancers	12879	27	4
2000	Road injury	12141	25	5
2000	Liver cancer	10893	22	6
2000	Diabetes mellitus	10414	21	7
2000	Cirrhosis of the liver	9968	21	8
2000	Self-harm	6860	14	9
2000	Chronic obstructive pulmonary disease	6783	14	10
2010	Stroke	31934	64	1
2010	Ischaemic heart disease	23696	47	2
2010	Trachea, bronchus, lung cancers	17121	34	3
2010	Self-harm	16852	34	4
2010	Liver cancer	12204	24	5
2010	Stomach cancer	11507	23	6
2010	Diabetes mellitus	9225	18	7
2010	Lower respiratory infections	9013	18	8
2010	Colon and rectum cancers	8886	17	9
2010	Chronic obstructive pulmonary disease	7904	15	10
2015	Stroke	28655	56	1
2015	Ischaemic heart disease	27336	53	2
2015	Trachea, bronchus, lung cancers	18806	37	3
2015	Lower respiratory infections	17164	33	4
2015	Self-harm	14255	28	5
2015	Liver cancer	12217	24	6
2015	Alzheimer disease and other dementias	11164	21	7
2015	Stomach cancer	9534	18	8
2015	Colon and rectum cancers	9405	18	9

Year	Cause	Deaths	Death Rate	Ranking
2015	Kidney diseases	9188	18	10
2019	Ischaemic heart disease	28042	54	1
2019	Lower respiratory infections	26649	52	2
2019	Stroke	25596	49	3
2019	Trachea, bronchus, lung cancers	20293	39	4
2019	Self-harm	14635	28	5
2019	Alzheimer disease and other dementias	12144	23	6
2019	Liver cancer	11589	22	7
2019	Colon and rectum cancers	10180	19	8
2019	Kidney diseases	10107	19	9
2019	Stomach cancer	8624	16	10

Table 2: Common Mortality Causes of All Four Years 2000, 2010, 2015, 2019

Year	Cause	Deaths
2000	Stroke	44109
2000	Ischaemic heart disease	18837
2000	Stomach cancer	13205
2000	Trachea, bronchus, lung cancers	12879
2000	Liver cancer	10893
2000	Self-harm	6860
2010	Stroke	31934
2010	Ischaemic heart disease	23696
2010	Trachea, bronchus, lung cancers	17121
2010	Self-harm	16852
2010	Liver cancer	12204
2010	Stomach cancer	11507
2015	Stroke	28655
2015	Ischaemic heart disease	27336
2015	Trachea, bronchus, lung cancers	18806
2015	Self-harm	14255
2015	Liver cancer	12217
2015	Stomach cancer	9534
2019	Ischaemic heart disease	28042
2019	Stroke	25596
2019	Trachea, bronchus, lung cancers	20293
2019	Self-harm	14635
2019	Liver cancer	11589
2019	Stomach cancer	8624

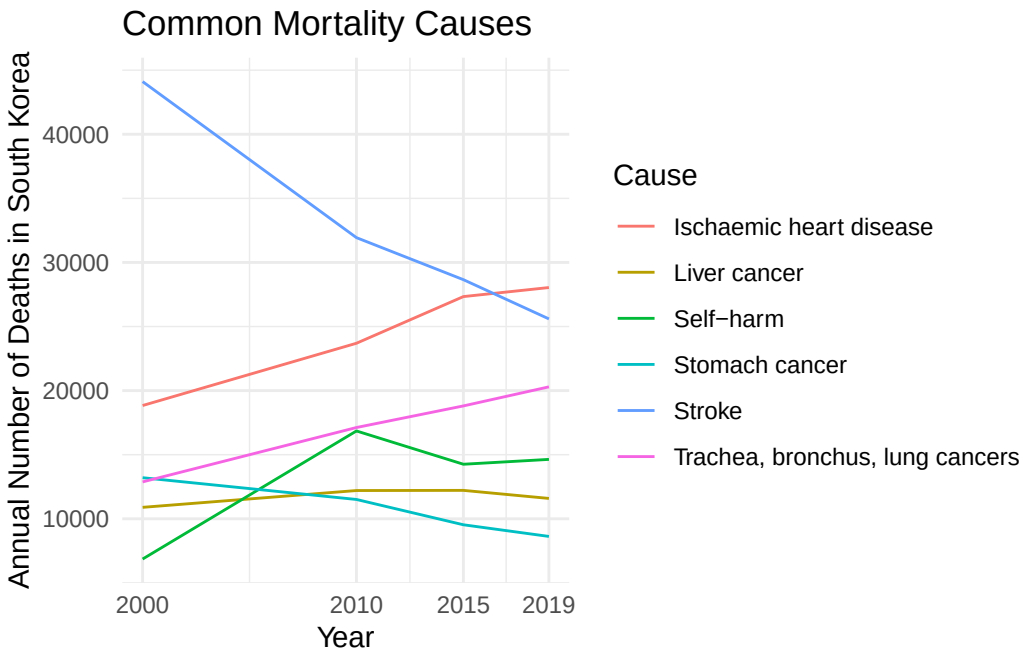


Figure 1: Line Graph of Common Mortality Causes of South Korea

Table 3: Summary statistics of the number of yearly deaths, by cause, in South Korea

	Min	Mean	Max	SD	Var	N
Deaths	6860	18 320	44 109	8927	79 687 233	24

2.3 Basic Summary Statistics

Table 3 is a representation of the Table 2 showing its minimum, mean, maximum, standard deviation, variance, and sample size. The summary statistics shows that the mean of the number of deaths are 18320, with a minimum of 6860, and a maximum of 44109 for a sample size of 24 as the data selected is from Table 2. The standard deviation and variance may be abnormally high because the range of the data is large, in other words, because the mean is 18320, and the maximum is 44109, there is likely a few outliers in the data creating such a high standard deviation and variance.

3 Model

The goal the Bayesian model is to incorporate prior knowledge such as previous studies or analysis into the choice of model. In this paper we use poisson and negative binomial model

because both of these models is often used when there occurs a certain number of events in a certain time period or intervals. The poisson distribution is efficient when used for situations where events occur independently over intervals of time. The data represents the cause and the annual number of death for the selected years, the poisson distribution is a measure that is appropriate. However, the poisson distribution is not the perfect fit because it assumes that the mean and the variance are the same whilst Table 3 shows us there is a significant difference in the mean and standard deviation. Negative binomial distribution is a perfect for situations like this. The negative binomial distribution accounts for additional variability when the data shows evidence of over dispersion. In addition, the negative binomial distribution has an extra parameter that considers the variance being larger than the mean. Thus, executing both models and comparing the results will determine the goodness of the fit.

3.1 Model set-up

The two models that are used in this paper are poisson and negative binomial which are both run in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. Both models define y_i as the number of deaths to a specific cause. For the poisson model it is followed by the mean parameter γ_i. The mean parameter γ_i is modeled as an exponential of a linear combination of the intercept α_i, regression coefficient β_i where i is the predictor, and an extra coefficient γ_i. All three α_i, β_i, and γ_i follows a normal distribution with mean 0 and standard deviation of 2.5 as the prior. The negative Binomial model has a mean parameter μ_i and a dispersion parameter ϕ. μ_i is modeled as the exponential of a lienar combination of an intercept α_i, coefficients β_i which represents the effects of each cause on the number of deaths, and γ_i will account for any extra variability in our scenario. The intercept α_i and coefficient of cause β_i follow a normal distribution with mean 0 and a standard deviation of 2.5. The dispersion parameter ϕ follows an exponential distribution with a rate parameter of 1. In summary, both poisson and negative binomial models work towards predicting the number of deaths based on the different causes assuming each cause has a coefficient affecting the number of deaths.

$$y_i \sim \text{Poisson}(\lambda_i) \quad (1) \tag{1}$$
$$\lambda_i = \exp(\alpha + \beta_i + \gamma_i) \quad (2) \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \quad (3) \tag{3}$$
$$\beta_i \sim \text{Normal}(0, 2.5) \quad (4) \tag{4}$$
$$\gamma_i \sim \text{Normal}(0, 2.5) \quad (5) \tag{5}$$
$$\tag{6}$$

$$y_i \sim \text{NegBinomial}(\mu_i, \phi) \quad (1) \tag{7}$$
$$\mu_i = \exp(\alpha + \beta_i + \gamma_i) \quad (2) \tag{8}$$
$$\alpha \sim \text{Normal}(0, 2.5) \quad (3) \tag{9}$$
$$\beta_i \sim \text{Normal}(0, 2.5) \quad (4) \tag{10}$$
$$\gamma_i \sim \text{Normal}(0, 2.5) \quad (5) \tag{11}$$
$$\phi \sim \text{Exponential}(1) \quad (6) \tag{12}$$

3.2 Model justification

Applying the information from Table 2, it is difficult to generalize whether there would be a positive or negative correlation between the number of deaths and its cause. This is because every year there is a different number of deaths from the same cause. For example, from Figure 1, we can expect that stroke would have negative correlation coefficient because the annual number of deaths has decreased significantly over the 9 years period. On the other hand, we may predict a positive correlation coefficient for the causes like heart disease and lung cancer as they have been showing an increase trend in the number of deaths since 2000 to 2019. It may be difficult to exactly predict how self harm, liver cancer, and stomach cancer without the actual calculations being done due to the ambiguity the data shows.

4 Results

4.1 Overview of models

Our results are summarized in Table 4 and Table 5. We are primarily interested how each cause of death correlates to the annual number of deaths in South Korea. The two models provides us with th estimates for the intercept and coefficient for the causes such as liver cancer, self harm, stomach cancer, stroke, and trachea bronchus lung cancer. The intercept represents the expected log count of deaths when the cause variable is zero. However, with our data set, the cause variable cannot be zero, which makes the interpretation of this quite unclear. The coefficients for the different causes imply the log rate ratio for each cause compared to the reference category. In other words, the coefficient shows the rate at which each cause effects the log count of deaths. Num.Obs represents the number of observations made in the model, Log.lik is the log-likelihood measuring the fitness of the model, where the lower, the better fit it is. ELPD and ELPD s.e. explains the log predictive density and its standard error. LOOIC is an acronym for leave-one-out information criterion which is a measure of model fit. Lower values are better, and the standard error shows the uncertainty of the estimate. WAIC stands for watanabe-akaike information criterion which is another measure of good fit; lower

Table 4: Poisson model of most prevalent cause of deaths in South Korea 2000, 2010, 2015, 2019

	Poisson
Intercept	10.105
Liver Cancer	−0.736
Self Harm	−0.621
Stomach Cancer	−0.826
Stroke	0.286
Trachea, Bronchus, Lung Cancer	−0.349
Num.Obs.	24
Log.Lik.	−8157.665
ELPD	−8835.6
ELPD s.e.	2653.7
LOOIC	17 671.3
LOOIC s.e.	5307.3
WAIC	22 572.8
RMSE	3830.21

values are better fit. RMSE is the room mean squared error measuring the model's predictive performance where the lower values mean more accurate predicts.

4.2 Poisson model results

Table 4 represents the summary of the poisson model. It has an intercept of 10.105 which we can interpret as there will be 10 deaths annually from other factors. Stomach and liver cancer display a strong negative correlation coefficient each -0.826, and -0.736 respectively. Self harm also has a quite strong negative correlation coefficient -0.621. This implies that the model predicts that there will be a decrease in the log count of deaths for these causes, in short, there will be a decrease in the annual number of deaths for causes with a negative correlation coefficient when all else is held the same. Trachea, bronchus, lung cancer has a relatively weaker negative coefficient compared to the other causes mentioned above but it is not a negligible number -0.349. Stroke is interesting because from Figure 1, we expected the coefficient to be strongly negative but it turns out to be slightly positive 0.286. For an unknown reason, the model is predicting that there will be an increase in the annual number of deaths from stroke in the near future. The poisson model has a very low log likelihood of -8157.665 implying the model may be a good fit. However, when we look at the values for LOOIC, WAIC, and RMSE, the values are extremely high each 17671.3, 22572.8, and 3830.21 which makes this model results quite unreliable.

Table 5: Negative Binomial model of most prevalent cause of deaths in South Korea 2000, 2010, 2015, 2019

	Negative Binomial
Intercept	10.102
	(0.190)
Liver Cancer	−0.726
	(0.266)
Self Harm	−0.610
	(0.268)
Stomach Cancer	−0.814
	(0.276)
Stroke	0.296
	(0.274)
Trachea, Bronchus, Lung Cancer	−0.341
	(0.272)
Num.Obs.	24
Log.Lik.	−237.440
ELPD	−240.6
ELPD s.e.	2.4
LOOIC	481.2
LOOIC s.e.	4.8
WAIC	481.0
RMSE	3832.16

4.3 Negative binomial model results

Table 5 displays the results of the negative binomial test. Looking into the results, there is one thing different compared to what saw from Table 4. The negative binomial model includes the standard error in parentheses which indicates uncertainty or variability and smaller standard error implies more precision. The intercept is 10.102 with a standard error of 0.190 suggesting the intercept is accurate. Liver cancer, self harm, stomach cancer, and trachea bronchus lung cancer is showing a negative correlation coefficient each of which is -0.726, -0.610, -0.814, and -0.314. Each has a standard error of 0.266, 0.268, 0.276, 0.272. Liver cancer, self harm, and stomach cancer exhibit the strongest negative correlation coefficient in which we can interpret as in the future, there will be a significant decrease in the number of deaths from these causes if all else is equal. Trachea, bronchus, lung cancer show a relatively weaker correlation coefficient which implies that it is likely that there will be a small decrease in the number of deaths in the future. Stroke is the only cause that has a positive correlation coefficient 0.296 with a standard error of 0.274. We can say that there is a high probability that there will be a slight increase in the number of deaths caused by stroke in the future. The log likelihood is -237.44 suggesting

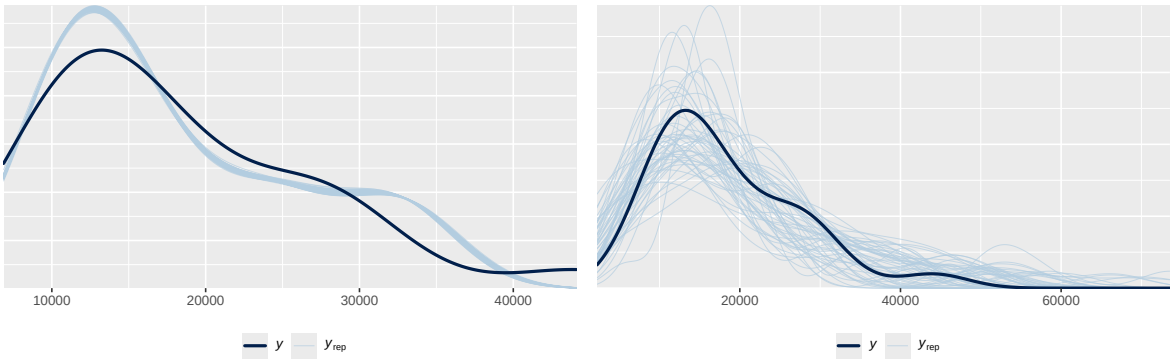
the results of this model is accurate. The LOOIC and WAIC are relatively lower than the ones from Table 4 encourages that the negative binomial model is a better fit than the poisson model. Since the RMSE is quite similar, we can think that the negative binomial model is a better fit than the poisson model which makes the results from the negative binomial more accurate.

4.4 Model results summary

Comparing the values of the two data sets, and the goodness of fit of the models, we can confirm the validity of our model results. Looking at the two values we from each model for each cause, all causes are in very close proximity. Through this we can see that the results from Table 4 and Table 5 are credible. Thus, we can confidently say that liver cancer has a negative correlation coefficient of -0.73, self harm's coefficient is around -0.61, stomach cancer is around -0.82, stroke is around 0.29, lung cancer is around -0.341 all of which have a standard error around 0.27. These results suggest that the model estimates that liver cancer, self harm, stomach cancer, and lung cancer will have a decrease in the log number of deaths while stroke will have an increase in the log number of deaths.

4.5 Other model results

Figure 2 is posterior predictive check and Figure 3 is a leave-one-out cross validation. These will be further discussed in Section B.1.



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 2: Examining how the model fits, and is affected by, the data

```
                                        elpd_diff se_diff
cause_of_death_south_korea_neg_binomial     0.0       0.0
cause_of_death_south_korea_poisson      -8595.0    2652.1
```

Figure 3: Checking the convergence of the MCMC algorithm

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

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/.