Count Data On Cancer Death In Ohio: A Bayesian Analysis

Walaa Hamdi

Abstract: This paper considers statistical modeling of count data on cancer death in Ohio State. We obtained count data on male and female from a website of the Centers for Disease Control and Prevention and used Bayesian analyses to find suitable models which help us to do inferences and predictions for next year. To assist us in selecting appropriate models, we use criteria such as the DIC. In this paper, we analyze the data to spatial/ longitudinal so we can capture possible correlations. Using our analyses, we make predictions of the numbers of people who will die with cancer in a future year in Ohio State.

Index Terms: Modeling, Cancer death, predictive.

1 INTRODUCTION

Cancer increases when uncontrolled growth and spread of cells in the body. Roughly, 1 of 2 Americans will be diagnosed with cancer at some point in their lives. Some people are screened for cancer with what seems like an ever-increasing frequency. Cancer is one of the most researched and most publicized illnesses. Cancer becomes one of the common diseases in the world and it can result in death. The National Cancer Institute (NCI) estimates that around 13.7 million Americans with a history of cancer were alive on 2012 [6]. The American Cancer society (ACS) estimates that around 25 thousands cancer deaths every year in Ohio State [1]. This project considers statistical modeling of count data on cancer deaths in Ohio State.

2 AIM OF THE PROJECT

In this project will discuss longitudinal data sets on counts to predict the counts in future year for male and female in Ohio State. We will simplify the aforementioned prediction problem by inputting the sizes of death population counts for the future year in which we seek to predict the cancer death counts. The goals will be considered through the Bayesian paradigm.

3 Description of the data set

There is count data collected from Centers for Disease Control and Prevention [4]. I collected cancer death counts and death population counts for male and female in Ohio States from 1999 to 2011. The population counts for male and female are themselves time series. In this project, we aggregated the cancer death count data over the variable male and female. Table 1 provides the data on the aggregated counts $Y_{i,j}$.

TABLE 1

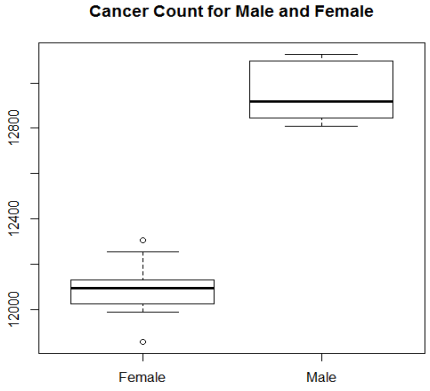
The Count Data for the Cancer Death by Gender and Year from 1999-2011

Year	Male	Female
1999	12928	12305
2000	12857	12131
2001	12808	11996
2002	12920	12253
2003	12826	12250
2004	12846	12094
2005	12844	11858
2006	12854	12121
2007	13098	12132
2008	12960	12038
2009	13126	12023
2010	13096	11987

In general, the cancer death for male is increasing but for female is decreasing.

FIGURE 1

Boxplot of Cancer Death in Ohio State



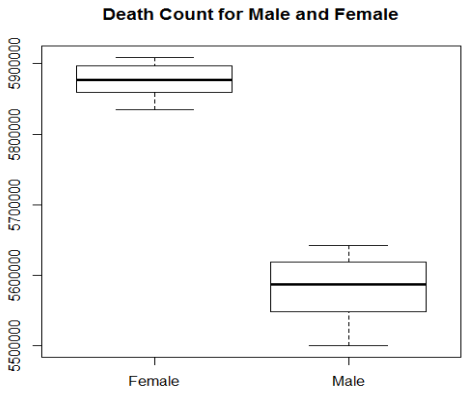
The boxplot showed that cancer death for male is higher than cancer death for female. Table 2 provides the population death on the aggregated counts $Y_{i,j}$. The population death increased for both gender, it is time series.

TABLE 2

The Count Data for Population Death by Gender and Year from 1999-2011 (Multiples of 1000)

Year	Male	Female
1999	5500.437	5835.017
2000	5518.118	5845.425
2001	5533.713	5853.691
2002	5548.504	5859.385
2003	5566.253	5868.535
2004	5577.840	5874.411
2005	5586.617	5876.703
2006	5598.146	5883.067
2007	5609.503	5890.965
2008	5618.797	5896.594
2009	5626.878	5902.018
2010	5636.961	5908.474
2011	5641.673	5908.099

FIGURE 2
Boxplot of Population Death in Ohio State



In general, the population death for male and female are increasing every year. The boxplot showed that population death for female is higher than cancer death for male in Ohio . Now, we used the Bayesian methods to model and analyze the data contained in Table 1 and Table 2.

4 Methodology

Let $Y_{i,j}$ denote the total number of cancer death in gender j for Year i. We assumed that the sampling model for $Y_{ij}|\lambda_{ij}, D_{ij}$ is Poisson, with $\lambda_{ij} * D_{ij}$, where λ_{ij} is rate per thousand for Year i and gender j, and D_{ij} is the number of people who are died in Year i and gender j, expressed as a multiple of 1000. In this paper, we assumed that the population death count D_{ij} is fixed. We proposed three different models with increasing complexity and hierarchy to try to capture the dynamics and correlations in the data.

4.1 MODEL I

Sampling model:

$Y_{ij}|\lambda_{ij} \sim poisson(\lambda_{ij} * D_{ij})$ are independent, where $i = 1,2,\dots,13$ $j = 1,2$ Let $W_j = (Y_{1j}, Y_{2j}, \dots, Y_{nj})$ where $j = 1,2$. Assume that W_1, W_2 are independent. That is, we assume spatial independence of the discussed counts. Additionally, no temporal dependence among Y_{ij} is assumed.

Priors:

$\mu_j|\mu_o, \tau_1 \sim Normal(\mu_o, \tau_1)$
$\mu_o|c_1, c_2 \sim Normal(c_1, c_2)$ where mean c_1=0.0, and precision $c_2 = .01$
$\tau_1|c_3, c_4 \sim Gamma(c_3, c_4)$ where shape c_3=.001, and rate $c_4 = .001$
$Z_{i,j} = \mu_j$, where $Z_{i,j} = \log(\lambda_{ij})$

4.2 MODEL II

The sampling model:

$Y_{ij}|\lambda_{ij} \sim poisson(\lambda_{ij} * D_{ij})$ where $i = 1,2,\dots,13$ $j = 1,2$
Let $W_j = (Y_{1j}, Y_{2j}, \dots, Y_{nj})$ where $j = 1,2$. Assume that W_1, W_2 are independent.

Priors:

Here, the priors of Model I are expanded to include a first-order auto-regression of the true incidence rate of hepatitis, on the log-scale, using the region-specific auto-regressive parameters $\alpha_j|a, b \sim Uniform(a, b)$, $j = 1,2$; assume $a = -1$, and $b = 1$. α_j is stationary.

$\mu_j|\mu_o, \tau_1 \sim Normal(\mu_o, \tau_1)$
$\mu_o|c_1, c_2 \sim Normal(c_1, c_2)$ where mean c_1=0.0, and precision $c_2 = .01$
$\tau_1|c_3, c_4 \sim Gamma(c_3, c_4)$ where shape c_3=.001, and rate $c_4 = .001$
$Z_{i,j} = \mu_j + \alpha_j * Z_{i-1,j}$, where $Z_{i,j} = \log(\lambda_{ij})$

4.3 MODEL III

Sampling model:

$Y_{ij}|\lambda_{ij} \sim poisson(\lambda_{ij} * D_{ij})$ where $i = 1,2,\dots,13$ $j = 1,2$
Let $W_j = (Y_{1j}, Y_{2j}, \dots, Y_{nj})$ where $j = 1,2$. Assume that W_1, W_2 are independent.

Priors:

Here, the priors of Model II are expanded to include an error term, $\varepsilon_{ij}|\tau_2 \sim Normal(0, \tau_2)$
$\tau_2|c_5, c_6 \sim Gamma(c_5, c_6)$ shape c_5=.001, rate $c_6 = .001$, to explain the true mean number of Cancer death in Ohio, on the log-scale, which is specific to the year and gender of the observation.
$\mu_j|\mu_o, \tau_1 \sim Normal(\mu_o, \tau_1)$
$\mu_o|c_1, c_2 \sim Normal(c_1, c_2)$ where mean c_1=0.0, and precision $c_2 = .01$
$\tau_1|c_3, c_4 \sim Gamma(c_3, c_4)$ where shape c_3=.001, and rate $c_4 = .001$
$\alpha_j|a, b \sim Uniform(a, b)$ where $a = -1$, and $b = 1$
$Z_{i,j} = \mu_j + \alpha_j * Z_{i-1,j} + \varepsilon_{ij}$, where $Z_{i,j} = \log(\lambda_{ij})$

5 MODEL SELECTION

In order to compare the relative performances of Model I, Model II and Model III, given data $Y_{i,j}$, for each model, define deviance information criterion (DIC), given below:

$$DIC \equiv \overline{D(\theta)} + P_D = 2\overline{D(\theta)} - D(\bar{\theta})$$

where $\bar{\theta}$ is a vector of the posterior mean of the parameter.
$P_D \equiv \overline{D(\theta)} - D(\bar{\theta})$ where $\bar{\theta}$ is the posterior mean.
$D(\theta) = -2\log[L(\theta)] + C$ where C is a constant, which does not depend on θ.

The DIC demonstrated using posterior samples generated by Markov chain Monte Carlo (MCMC) algorithms. Models with the smaller of the DIC are preferable for Bayesian inferences [3]. Based on the WinBUGS outputs from fitting Models I, II, and III to data, we obtain the DIC for Model I, Model II, and Model III as 567.615, 564.107 and 344.219, respectively. Hence, according to the DIC, the best performing model is Model III. However, many authors have warned against casual use of DIC as a criterion to discriminate among hierarchical Bayesian models [5.[I could propose a Bayesian predictive goodness-of-fit criterion to assess how well the models fit the data on cancer death counts and death population counts to get better model selection [2], but it is hard to use this way.

6 PREDICTIVE

I used the Model III to predict the cancer death counts in 2012 for both male and female in Ohio State. Since I added the population death for 2012 to Predictive for numbers of cancer death for both male and female for year 2012 is 25220 which close to actual number is 25261.

TABLE 3

Predict the cancer death counts in 2012

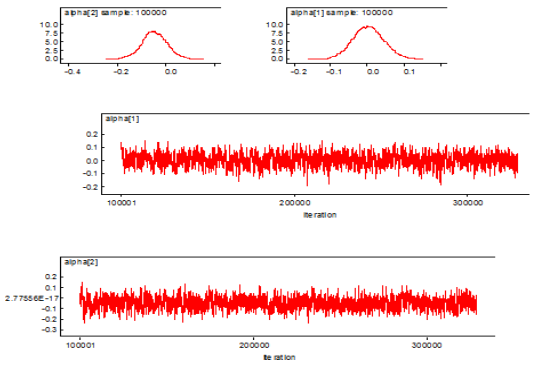
node	mean	Sd	MC error
2012	25220.0	1239.0	59.61

7 POSTERIOR DENSITIES AND THE HISTORIES

As seen in Figure 3 the posterior densities look symmetric and normal. The histories of the chain, indicated by different colors, are overlapping one another and look like "straight caterpillars."

FIGURE 3

Posterior densities and the histories after burn-in



8 FUTURE WORK

In this project, I assumed that the sampling model for $Y_{ij}|\lambda_{ij}, D_{ij}$ is Poisson; future research could assume that the model is following normal distribution. In addition, future research could assume that the population counts D_{ij} are random. Furthermore, future research could model population death counts to get better predictions of cancer death counts by predict D_{ij} to predict Y_{ij}. Also, research could use the joint density $Y_{ij}, \lambda_{ij}, D_{ij}$to modeling. This paper is not just for people who are interested in the cancer death studies; others could use it in count data.

9 REFERENCE

[1] American Cancer society. Retrieved from http://www.cancer.org/

[2] Berkhof J., Mechelen, I. E., and Gelmanm A. (2003).A Bayesian Approach to the Selection and Testing of Mixture Models. Statistica sinica 13, 4232-442.

[3] Carlin, B., and Louis, T. (2009). Bayesian Methods for Data Analysis. New York: Chapman and Hall. Book.

[4] Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov

[5] Millar, R. B. (2009).Comparison of Hierarchical Bayesian Model for Overdispersed Count data Using DIC and Bayes' Factors. Biometrics 65, 962-969.

[6] National Cancer Institute. Retrieved from http://www.nih.gov/about-nih/what-we-do/nih-almanac/national-cancer-institute-nci