Bayesian Evidence Synthesis:opioid crisis

Hyeongcheol Park* & Paul Gustafson* & Micheal A Irvine*[1]

January 29, 2020

Contents

List of Figures

List of Tables

Abstract

1 Introduction

The opioid crisis is a major issue in North America including Canada. There were 1,490 deaths and 15,598 paramedic- attended overdose events during 2017 alone. [1] (need to know about bib in latex, change statistics to 2018 later) The goal of this project is to apply Bayesian evidence synthesis to understand better the opoid crisis in Vancouver, Canada.

[Need to give a backstory-why isn't this observable? The number of overdoses] [specify a geographic area]

All examples here were performed in Python 3.7 using the library pyMC (reference) and JAGS (reference). Training was performed using No U-Turn Sampling (NUTS) over two chains with 1000 iterations (is it sample size?). Fitting was performed on a GHz Intel Core i5 with 8GM of LPDD3 RAM and typically had wall times under ten minutes. Data processing was carried out using the Pandas and SciPy library [reference]. Data visualization was performed using the libraries Seaborn and Matplotlib [ref]. Code for all examples in this study are provided.

[explain Bayesian statistics, MCMC and related topics]

2 methods

The number of overdoses is our ultimate interest of estimation. Let O_t the number of overdoses in a given month t. Suppose there was a survey conducted to estimate the proportion of ambulance call p_A among the overdoses. p_A is assumed constant across time for simplicity. Let n_A the sample size of the survey and x_A to be the total number who confirmed they did call ambulance. It is assumed that x_A follows a Binomial distribution:

$$x_A \sim Bin(n_A, p_A) \text{ ambulance call-outs model} \tag{1}$$

The total overdoses need to be modeled. The simplest conceptual model is to take an underlying log-rate z_t that is independent and identically distributed across months according to a normal distribution with mean μ and variance σ^2. [1] Denote λ_t^{OD} the rate of overdose at time t. It is assumed that the total overdose O_t follows Poisson distribution where the population of the region of interest is N. [Give a graphical model representation? state that(?)]

$$\left.\begin{aligned} z_t &\sim N(\mu, \sigma^2) \\ \lambda_t^{OD} &= \exp(z_t) \\ O_t &\sim Poi(\lambda_t^{OD} N) \end{aligned}\right\} \text{ overdose model} \tag{2}$$

Estimation of O_t is not straightforward since none of the variables (μ, σ, N) determining O_t is known. Hence O_t should be inferred from using U_t and p_A,

where p_A is the ambulance call out rate and U_t is the number of ambulance-attended overdoses at a time point t. In general, the data of ambulance-attended overdoses U_t can be obtained. It is assumed that U_t follows Binomial distribution:

$$U_t \sim Bin(O_t, p_A) \tag{3}$$

Now O_t can be estimated as p_A can be infered by survey data and the data regarding U_t is given. We suggest a simple model as a start where the model only combines Ambulance Call-outs Model (1) and Overdose Model (2).

The next step is to run some simulations to figure out how different types of inputs lead some changes of output. To do so, the simple model is illustrated below.

2.1 Simulation

The first simulation simplifies the assumptions of variables as much as possible; We assumed $N = 10000, n_A = 1000$. Note that the population size, N, is unknown in reality but it is assumed known for the investigator here for simplicity. The assumptions will change later to see the impact of the likelihood over the posterior distributions of variables of interest; The total population size for a region, N, could vary over time or it can be staritified for a better realization of the real world. We vary the survey size as $n_A = 100$ or $n_A = 10000$.

2.1.1 Likelihood

There exist two data sets; survey data (n_A, x_A), and ambulance attended overdose data (U_t). The two data sets are simulated as follows. The true value of p_A was set $p_A = 0.8$ for the survey data. It is assumed that the data was collected for a year (t=1,2,3, ..., 12) and x_t values were independentally generated from the Binomial distribution (1). In terms of overdose data, It is assumed that the true values of parameters for overdose model were $\mu = \log 0.05$, $\sigma = 1$. The vector of O_t was generated following the overdose model (2). The vector of U_t was generated from the Binomial relation of the two variables (3). The two generated vectors have the same length with the survey data (t=1,2,3, ..., 12). [Delete: Then the 12 sueveys were combined into a single survey since p_A is assumed fixed across time in this contexts.] Note that only U_t and x_t are known as the likelihood and p_A needs to be estimated first so as to estimate O_t which is the ultimate interest of the research.

2.1.2 Prior Distributions

Noninformative prior distributions are presumed as a start for simplicity.

$$p(p_A) \sim Beta(1, 1) \text{ noninformative prior of ambulance model} \tag{4}$$

$$\left.\begin{aligned}\mu &\sim U(-10, 0)\\ \sigma &\sim U(0, 5)\end{aligned}\right\} \text{ noninformative prior of overdose model} \tag{5}$$

This leads the posterior distribution of variables of interest to heavily depend on the likelihood. Later, the noninformative priors will be changed and the impact of the changes over posteriors will be investigated.

2.2 Early Result

The result from the simple case scenario is illustrated below.

2.2.1 Posterior Distribution

Figure 1 is the boxplot of posterior samples of O_t. It is shown that our posterior estimates of O_t are fairly accurate since (1) the boxplots contain actual values of O_t within their interquartile range (IQR) and (2) the ranges of IQR and 95% range seem narrow. Notice that the range of the boxplot from a higher O_t values (t=1, 7, 11) is wider than the other ranges of the boxplots from smaller estimates of O_t (all t values but 1, 7, 11)

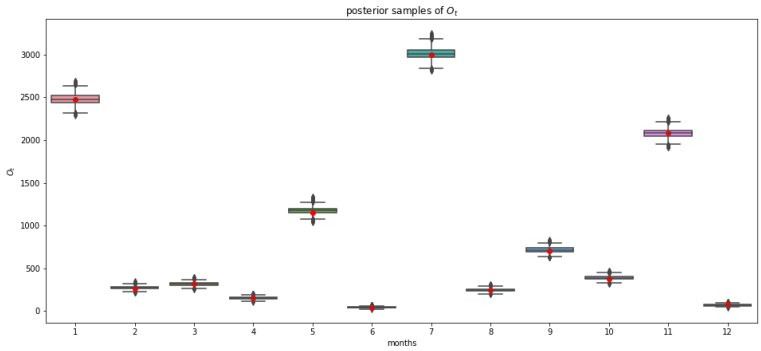
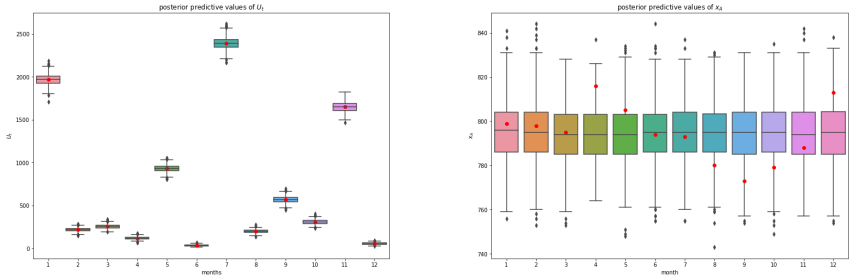


Figure 1: Boxplot of posterior samples of O_t (2000 samples for each month) with actual data points of simulated O_t values. The simulated values are shown as red dots.

2.2.2 Posterior Predictive Check

Figure 2a is the boxplot of posterior predictive samples of U_t. It is shown that the posterior predictive estimates of U_t is failry accurate with the same two reasons regarding the accuracy of the posterior distribution of O_t. It is more obvious here that the range of the boxplots from higher O_t values (t=1, 7, 11) is wider than the other ranges of the boxplots from smaller estimates of O_t (all t values but 1,7, 11). Notice that the relative ranges of figure 2a follow the ones of figure 1; For months where U_t values are higher, the values of O_t are also higher than the average (t=1, 7, 11)



(a) ppc of U_t

(b) ppc of p_A

Figure 2: Boxplots of posterior predictive samples of x_A (1000 samples) with the actual data point of simulated U_t,x_A value. The simulated values are shown as a red dots.

Figure 2b is the boxplots of posterior predictive samples of x_A. It is shown that the posterior predictive estimates of x_A is tolerably accurate since (1) the boxplots contain actual values of O_t within their lines connecting the maximum and the minimum and (2) the ranges of IQR seem modereately narrow.

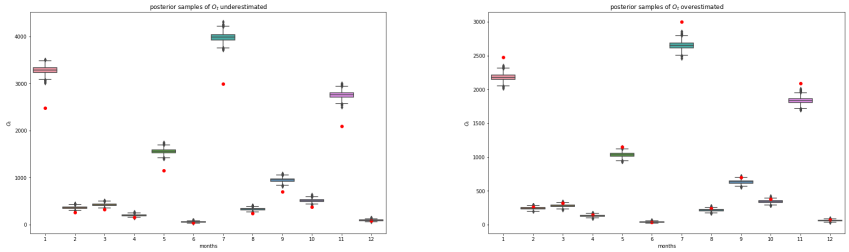
2.3 Early Result: Contamination of p_A

One of the attention of this research project is to investigate how robust the model is from contaminations of the data sets. The first inspection is to check an impact of a contamination of p_A; what would happen if the estimation of p_A is biased? It is assumed that the survey data gives us a wrong estimate of p_A such that it would be underestimated or overestimated. We then want to see how the biased estimation of p_A affects the estimate of O_t, the total overdose.

Both of underestimation and overestimation were conducted for the analysis. In terms of underestimation, the simulated survey data (n_A, x_A) was generated

with $p_A = 0.6$ while the true value of p_A is 0.8, and all the other assumptions hold the same. That is, x_A is generated from $x_A \sim Bin(n_A, 0.6)$, while U_t is generated from $U_t \sim Bin(O_t, 0.8)$ for every t. For overestimation, the simulated survey data was generated with $p_A = 0.9$ while the true value of p_A is 0.8, and all the other assumptions hold the same.

2.3.1 Posterior Distribution



(a) ppc of O_t: underestimated p_A

(b) ppc of O_t: overestimated p_A

Figure 3: Boxplot of posterior samples of O_t (1000 samples) where survey data is contaminated. The actual data point of simulated O_t values are shown as red dots.

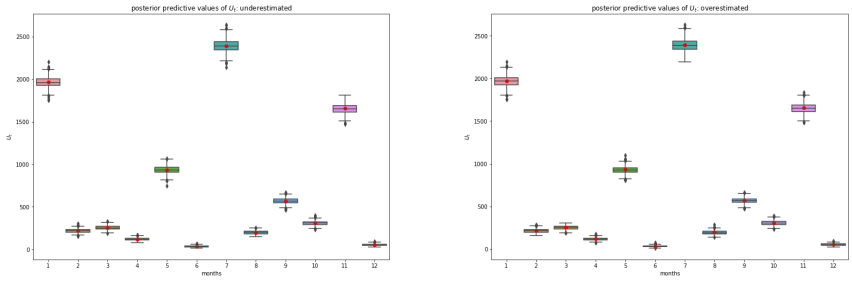
Figure 3 is the boxplots of posterior samples where p_A is contaminated from the survey data. It is seen that the underestimation of p_A (3a) leads to an overestimation of O_t as the boxplots are above the red dots. This is justifiable considering the given data sets (likelihoods) and the relationship between the two models (3); O_t is generated by multiplying U_t and the inverse of p_A where p_A is underestimated; Hence $\frac{1}{p_A}$ is overestimated which leads overestimation of O_t.

$$O_t = U_t \frac{1}{p_A} \tag{6}$$

Figure 3b shows the opposite case. Overestimation of p_A leads underestimation of the inverse of p_A which causes underestimation of O_t. From both figures it is seen that the bias increases as the estimated values and the actual values get large.

2.3.2 Posterior Predictive Check

Figure 4 are the boxplots of posterior predictive samples where p_A is underestimated and overestimated respectively from the survey data. It is seen that the none of the contaminations of p_A leads an effect U_t. [Explain why is this]



(a) ppc of U_t: underestimated p_A

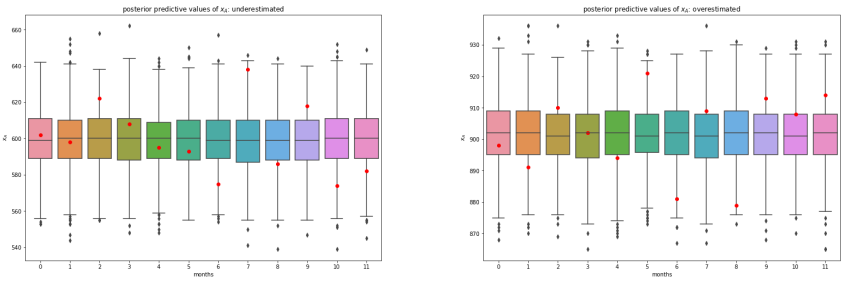
(b) ppc of U_t: overestimated p_A

Figure 4: Boxplot of posterior samples of U_t (1000 samples) where survey data is contaminated. The actual data point of simulated U_t values are shown as red dots.

Figure 5 are the boxplots of posterior predictive samples of x_A where p_A is underestimated and overestimated respectively from the survey data. It is seen that none of the contaminations of p_A leads an effect of contamination on x_A; the actual simulated points (red dots) are close to the medians from the two boxplots. However, notice that the range of the estimated values are different between the two boxplots; the median from Figure 5a is around 600 whereas the median from Figure 5b is around 900. [I think this should go to the previous of the previous paragraph: This is because overdose model (2) does not affect the resulf from the ambulance model (1); Only the ambulance model has an effect on the overdose model.]

References

[1] Buxton J Balshaw R Otterstatter M Macdougall L et al. Irvine MA, Kuo M. Modelling the combined impact of interventions in averting deaths during a synthetic-opioid overdose epidemic. *Addiction*, 2019.



(a) ppc of x_A: underestimated p_A

(b) ppc of x_A: overestimated p_A

Figure 5: Boxplot of posterior samples of x_A (1000 samples) where survey data is contaminated. The actual data point of simulated x_A values are shown as red dots.