An Empirical Bayesian Model in Disease Mapping of TB in Regency Bandung

Elin Marhamah¹, S Padmadisastra² and J Suprijadi³

¹-³) Department of Statistics, Universitas Padjajaran, Indonesia.

Corresponding authors: elin19001@mail.unpad.ac.id

Abstract. In this paper we proposed an Empirical Bayesian model in estimating the relative risk of people getting a Tuberculosis (TB) diseases in a small area. Which is one of the top rank diseases causing death in Indonesia. It is considered that the proposed model will alleviate some of the draw back that exist in the Standardized Mortality Ratio (SMR) method in estimating relative risk. It is shown that the proposed EB method produce an estimate of the risk which is a weighing average of SMR and EB estimate. To give a complete insight into the paper, we give an illustration concerning the case of Tuberculosis (TB), which is still considered as a killer disease in Indonesia.

1. Disease Mapping

Disease mapping is very useful for monitoring disease spread over a certain region, such as Subdistrict or small area in general. Usually over its small area, the number of people get sick was plotted. Therefore a responsible health institution, such as Departemen Kesehatan, can take a safety measure before an outbreak occur. Other useful thing that accrue was an ethology hypotheses that can explain the occurrence of the disease. An example is in Snow (1856) work. He produced a map consists of the addresses of people suffered from cholera, in an outbreak in London. From the pattern of the affected living address, he concluded that the source of the cholera outbreak was the drinking water pump polluted with ecoli bacteriy. Snow’s success story motivates others to apply it for diverse form of disease; such as cancer, Gardner, et. al (1982), and Ishibashi (1999).

Let us suppose that there are $N$ small area comprise a region of interest. And from the $i$th area there were $O_i$ cases of sick person and we assume that its expected number of cases is $E_i$. Traditionally, these two numbers were used as an estimate of relative risk ($\theta_i$) in disease mapping. The estimator is defined as follows, see Lawson (1995),

$$\hat{\theta}_i = \frac{O_i}{E_i} \quad (1)$$

the estimator of $\theta_i$, known as standardized mortality ratio (SMR).

Over the map one plot either the relative risk estimate of getting a certain disease, which was estimated from data, using (1), or its statistical significance. Thus one can adjust their decision based on either of these two information. It turn out that SMR is the maximum Likelihood Estimator (MLE) under an assumption that $O_i$ follow a Poisson distribution with parameter $\lambda_i = E_i\theta_i$. 

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
\[ P(O_i = o) = \frac{\lambda^o}{o!} e^{\lambda_i} \] (2)

But unfortunately, the estimator SMR in (1) has some drawbacks, among other is that it will be small when expected case \( E_i \) is large and large when \( E_i \) small, and zero whenever \( O_i = 0 \). Therefore this situation makes interpretation difficult. This unfortunate situation is one of the undesired properties of a saturated estimation, in which the number of parameter equal to the number of data. In this case we have one parameter \( \theta_i \) and one data \( O_i \) for each \( i \).

Several authors have proposed a solution in the form of a Bayesian approach, in which the parameter \( \theta_i \) is considered to be a random variable. Clayton and Kaldor (1987) proposed an empirical Bayes method of estimation (EB). Bayesian model is build under the following paradigm that states posterior probability distribution must proportionally be equal to the product of likelihood times a prior distribution of the parameter(s) of interest, in disease mapping the parameter of interest is the relative risk \( \theta \).

Prior represent beliefs of the author to the possible values of \( \theta \) initially. This belief then later be revised by data collected. This process produce a posterior distribution, and based on this posterior we can use it. These two parameters can be obtained through the marginal distribution of \( O \).

2. EB Estimate of Relative Risk \( \theta \)

Prior has an important role in the estimation process through posterior distribution. There are some methods of picking the prior form. Gelman (2002) described three ways of choosing the prior; namely subjective, objective, and uninformative. Other proposed a conjugate prior, see Bolstad (2004). This a kind of objective prior. This form of prior will produce a posterior having similar form with the chosen prior.

we will use a conjugate family for the prior. It is well known that a conjugate prior for Poisson is Gamma\((a,b)\) distribution. Thus our model is as follows

\[ O_i \sim \text{Poisson}(E_i \theta_i) \]
\[ \theta_i \sim \text{Gamma}(a, b) \] (3)

Where \( \text{Ga}(a, b) \) is Gamma distribution with shape and scale parameters \( a, b \) respectively, its density is

\[ f(\theta_i) = \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-\theta_i b} \] (4)

Lawson, et. al. (1995) studied several methods of estimating relative risk trough simulation, he found that a mixture of gamma model as in (3) produced better results compare to some other model he used. This mixture will produce the following posterior for \( \theta = (\theta_1, \cdots, \theta_N) \)

\[ P(\theta|O) \propto \prod_{i=1}^{m} \frac{e^{-E_i \theta_i}}{\theta_i^q_i} (E_i \theta_i)^{a_i} \times \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-\theta_i b} \]

Simplified becomes

\[ P(\theta|O) \propto \prod_{i=1}^{m} \theta_i^{(a_i + b) - 1} e^{-\theta_i (E_i + b)} \] (5)

Thus the posterior distribution of vector \( \theta \) is a product of gamma \( \text{Ga}(\alpha^*, b^*) \) distribution, as expected, with

\[ \alpha^* = (a_i + a) \text{ dan } b^* = (E_i + b) \] (6)

Hence the EB estimate of \( \theta_i \), say \( \widehat{\theta}_i \), can be set equal to its posterior mean,

\[ \widehat{\theta}_i = E(\theta_i) \]
\[ = \frac{a_i + a}{E_i + b} \] (7)

But in (7) there are two parameter that we do not know, \( a \) and \( b \), therefore we have to find them first before we can use it. These two parameters can be obtained through the marginal distribution of \( O \).
\[ g(o_i) = \int e^{-E_i \theta_i} \frac{\theta_i^{o_i}}{o_i!} \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-\theta_i b} d\theta_i \]
\[ = (o_i + a - 1) \left( \frac{E_i}{E_i + b} \right)^{o_i} \left( \frac{b}{E_i + b} \right)^a \]

Which is a binomial negative distribution. And thus, the MLE for \( a \) and \( b \) is a solution of the following system equation

\[
\frac{\delta l}{\delta a} = m \log(b) - \sum_{i=1}^{m} \log(E_i + b) + \frac{\Psi(o_i + a)}{\Gamma(o_i + a)} - \frac{\Psi(a)}{\Gamma(a)}
\]

\[
\frac{\delta l}{\delta b} = -\sum_{i=1}^{m} o_i \left( \frac{1}{E_i + b} \right) + \frac{ma}{b} - a \sum_{i=1}^{m} \frac{1}{(E_i + b)}
\]

\( \Psi(.) \) is a digamma function; which is a derivative of the logarithm of gamma function, see Abramowitz and Stegun (1972).

\[ \Psi(x) = \frac{d}{dx} \ln(\Gamma(x)) \]

Use Newton-Raphson to solve (8).

3. TB Cases Example

In the following, we give an illustration of the EB estimation for the relative risk in case of Tuberculosis TB disease. The region that we consider is Regency Bandung, located in the south of Bandung city, province West Java, Indonesia, comprises of 31 Subdistrict. The number of people living in each Subdistrict and suffered form TB is given in the following table.

| Subdistrict   | Population Number \( (N_i) \) | Number of Cases \( (O_i) \) | \( E_i = N_i \frac{n}{N} \) | \( \hat{\theta}_i = \frac{o_i}{E_i} \) | \( \frac{E(\theta_i)}{E_i} \) = \( \frac{o_i + a}{E_i + b} \) | \( V(\theta_i) = \frac{o_i + a}{(E_i + b)^2} \) |
|---------------|-------------------------------|-----------------------------|-----------------------------|--------------------------------|---------------------------------|---------------------------------|
| Ciwidey       | 79.974                        | 143                         | 112.65                      | 1.27                          | 1.29                            | 0.01                            |
| Rancabali     | 52.072                        | 58                          | 73.35                       | 0.79                          | 0.87                            | 0.01                            |
| Pasirjambu    | 87.932                        | 151                         | 123.86                      | 1.22                          | 1.25                            | 0.01                            |
| Cimalang      | 81.182                        | 164                         | 114.35                      | 1.43                          | 1.45                            | 0.01                            |
| Pangalengan   | 152.735                       | 184                         | 215.13                      | 0.86                          | 0.88                            | 0.00                            |
| Kertasari     | 71.755                        | 45                          | 101.07                      | 0.45                          | 0.52                            | 0.00                            |
| Paseh         | 112.197                       | 186                         | 158.03                      | 1.18                          | 1.20                            | 0.01                            |
| Ibun          | 84.330                        | 45                          | 118.78                      | 0.38                          | 0.45                            | 0.00                            |
| Cikacunkung   | 134.004                       | 283                         | 188.75                      | 1.50                          | 1.51                            | 0.01                            |
| Cicalengka    | 93.880                        | 177                         | 132.23                      | 1.34                          | 1.36                            | 0.01                            |
| Majalaya      | 122.374                       | 116                         | 172.37                      | 0.67                          | 0.71                            | 0.00                            |
| Solokanjeruk  | 54.035                        | 86                          | 76.11                       | 1.13                          | 1.18                            | 0.01                            |
| Majalaya      | 487.904                       | 313                         | 264.67                      | 1.18                          | 1.20                            | 0.00                            |
| Solokanjeruk  | 168.698                       | 359                         | 237.62                      | 1.51                          | 1.52                            | 0.01                            |
| Ciparay       | 101.750                       | 193                         | 143.32                      | 1.35                          | 1.36                            | 0.01                            |
| Banjanan      | 67.619                        | 211                         | 181.27                      | 1.16                          | 1.18                            | 0.01                            |
| Cangkuang     | 76.761                        | 101                         | 108.12                      | 0.93                          | 0.98                            | 0.01                            |
| Pameungpeuk   | 78.783                        | 132                         | 110.97                      | 1.19                          | 1.22                            | 0.01                            |
| Katapang      | 130.012                       | 320                         | 183.13                      | 1.75                          | 1.75                            | 0.01                            |
The table above also shows some quantities that were needed in estimating the risk. In Column (4) we calculate the expected cases for each Subdistrict using the formula

$$E_i = N_i \frac{n}{N}$$

Where \( n \) is the total number of cases in Regency as a whole, whereas \( N_i \) is the number of people in the \( i^{th} \) Subdistrict, and \( N \) is the population number in the Regency. As can be seen from the last row there are \( N = 3,657,701 \) people living in the regency in the year 2018, and \( N_1 = 79,794 \) people in Subdistrict Ciwidey with \( O_1 = 143 \) cases. So its expected number is equal to \( E_1 = 112.64 \) people, not much different from its observed cases.

Comparing observed to its expected cases, the table above shows no general pattern; some of Subdistrict have higher observed number but the other show lower observed than its expected cases. In Subdistrict with higher number of people we see the number of cases is also higher, similarly its expected number. That is why, it may happen that two Subdistrict with the same SMR but have different numerator and denominator. This is also the reason of using Bayesian approach to alleviate it.

Regarding the calculation of risk estimate which is given in column (6), we calculate \( a \) and \( b \) from equation (8) using Newton Raphson method in \( R \) (CRAN), see appendix for its complete code. The initial values that we used for \( a=2 \) and \( b=1 \), in the end the computational processes converge to \( a=11.58 \) and \( b=6.72 \). Using these two values the expected value of risk parameter is calculated using equation (7), and the results for each Subdistrict is placed in column 6. It can be seen that the result is expected, its values is always larger than SMR estimate. See the following figure.

![Figure 1 Estimate of relative risk](image)

4. References

[1] Abramowitz A and Stegun I 1970 *Handbook of Mathematical Functions*

[2] Bolstad W 2004 *An Introduction to Bayesian Statistics*, John Wiley & Son, Inc. New Jersey

[3] BPS 2018 *Regency Bandung Dalam Angka*, Regency Bandung

[4] Clayton D and Kaldor J 1987 *Biometrics* 43, 671-681

[5] Gelman A 2002 *Encyclopedia of Environmetrics* 3 1634-1637

[6] Howe GM 1971 *The mapping of disease in history*: In: Clarke E (ed) Modern Methods in the history of medicine. London: University of London, Athlone Press, Ch. 20.
[7] Ishibashi N, et. Al 1999 Ulcerative colitis and colorectal cancer: a follow-up study in Fukuoka, Japan. International Journal of Epidemiology, Vol 28, Issue 4, pages 609-613.

[8] Lawson A B, Biggeri A B, Boehning D, LesaJre E, Viel J F, Clark A, Schlattmann P, Divino F 2000 Statist. Med. 19, 2217–2241.

[9] Lawson A B, Biggeri A B, Boehning D, LesaJre E, VielJ-F, Bertollini R. (1999). Disease Mapping and Risk Assessment for Public Health. Wiley, NewYork.

[10] Lawson A B and Cressie N A C 2000 Spatial statistical methods in environmental epidemiology. In Handbook of Statistics: Bioenvironmental and Public Health Statistics, vol 18, 357-396, Sen P K, Rao, C R (eds). North Holland,.

[11] Snow J 1856 Cholera and Water Supply in the South Districts of London in 1854 Journal of Public Health and Sanitary Review. 1856 Oct; 2(7):239-PMCID.