Bayesian Prediction for Cancer Rates

February 28, 2019

Bayesian Methods for Kidney Cancer Mortality Rates

- This example is example 2.7 in *Bayesian Data Analysis*, by Andrew Gelman et al.
- Goal: Estimate yearly kidney cancer death rates per county.

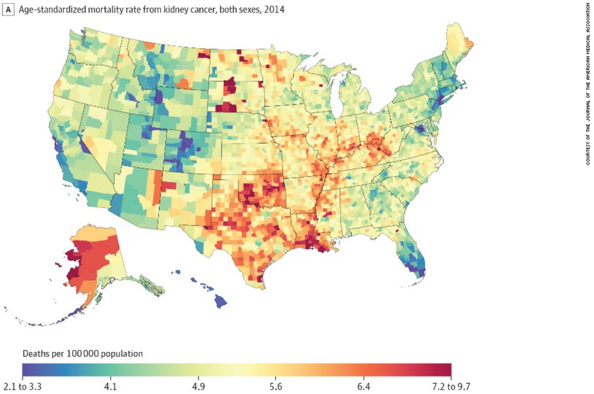

Figure 1: Age-adjusted motrality rates for kidney cancer in 2014. Highest counties are in Kentucky and the south.

Setting up the models

- Let x_j be the number of deaths in county j due to kidney cancer.
- Let n_j be the population of county j
- Let θ_j be the underlying "true" kidney cancer mortality rate for the county.
- Since the units of x_j are counts, use a Poisson distribution as our model, $\mathrm{p}(x_j|\theta_j) \sim \mathrm{Poisson}(10 n_j \theta_j)$.
- For mathematical convenience choose a Conjugate Prior $p(\theta_j) \sim \mathrm{Gamma}(\alpha, \beta)$, where, α and β are parameters TBD.

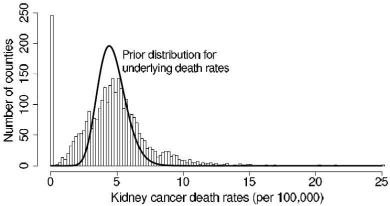

Figure 2: We will estimate the prior from the empirical data in a later slide

Finding the posterior

- Since the prior has been chosen as conjugate, we know the posterior will also be a Gamma distribution.

Finding the posterior

- Since the prior has been chosen as conjugate, we know the posterior will also be a Gamma distribution.
- $p(\theta_j|x_j) = \frac{p(x_j|\theta_j)p(\theta_j)}{p(x_j)}$

$$\propto p(x_j|\theta_j)p(\theta_j) = \frac{(10n_j\theta_j)^{x_j}e^{-10n_j\theta_j}}{x_j!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_j^{\alpha-1}e^{-\beta\theta_j}$$

$$\propto \theta_j^{x_j}e^{-10n_j\theta_j}\theta_j^{\alpha-1}e^{-\beta\theta_j}$$

$$\propto \theta_j^{x_j+\alpha-1}e^{-(10n_j+\beta)\theta_j}$$

$$\propto \mathrm{Gamma}(\alpha + x_j, \beta + 10n_j)$$

Finding the posterior

- Since the prior has been chosen as conjugate, we know the posterior will also be a Gamma distribution.
- $p(\theta_j|x_j) = \frac{p(x_j|\theta_j)p(\theta_j)}{p(x_j)}$

$$\propto p(x_j|\theta_j)p(\theta_j) = \frac{(10n_j\theta_j)^{x_j} e^{-10n_j\theta_j}}{x_j!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j}$$

$$\propto \theta_j^{x_j} e^{-10n_j\theta_j} \theta_j^{\alpha-1} e^{-\beta\theta_j}$$

$$\propto \theta_j^{x_j+\alpha-1} e^{-(10n_j+\beta)\theta_j}$$

$$\propto \text{Gamma}(\alpha + x_j, \beta + 10n_j)$$

- From this, we can see that α is an "average" mortality count and β is an "average" (scaled) county population in the prior.

Finding the right prior parameters

- In general, we would assign a prior to the parameters α, and β and compute something called a hierarchical model, but that is beyond the scope of this class so we will take a different approach to figure out the prior.

Finding the right prior parameters

- In general, we would assign a prior to the parameters α, and β and compute something called a hierarchical model, but that is beyond the scope of this class so we will take a different approach to figure out the prior.
- The marginal distribution of x_j is $p(x_j) = \frac{p(x_j|\theta_j)p(\theta_j)}{p(\theta_j|x_j)}$, by Bayes Theorem

Finding the right prior parameters

- In general, we would assign a prior to the parameters α, and β and compute something called a hierarchical model, but that is beyond the scope of this class so we will take a different approach to figure out the prior.
- The marginal distribution of x_j is $p(x_j) = \frac{p(x_j|\theta_j)p(\theta_j)}{p(\theta_j|x_j)}$, by Bayes Theorem
- We know all these distributions!

$$p(x_j) = \frac{\text{Poisson}(10n_j\theta_j)\text{Gamma}(\alpha,\beta)}{\text{Gamma}(\alpha + x_j, \beta + 10n_j)}$$

$$= \frac{\Gamma(\alpha + x_j)\beta^\alpha}{\Gamma(\alpha)x_j!(10n_j + \beta)^{\alpha + x_j}}$$

$$= \binom{\alpha + x_j - 1}{x_j}\left(\frac{\beta}{\beta + 10n_j}\right)^\alpha\left(\frac{1}{\beta + 10n_j}\right)^{x_j}$$

$$\sim \text{Neg} - \text{Binomial}(\alpha, \frac{\beta}{10n_j})$$

Finding the right prior parameters cont.

- To find α and β, use the expectation and variance of x_j based on the Negative Binomial Distribution.

Finding the right prior parameters cont.

- To find α and β, use the expectation and variance of x_j based on the Negative Binomial Distribution.
- $\mathbb{E}(x_j) = 10n_j \frac{\alpha}{\beta}$
- $\mathrm{var}(x_j) = 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2}$

Finding the right prior parameters cont.

- To find α and β, use the expectation and variance of x_j based on the Negative Binomial Distribution.
- $\mathbb{E}(x_j) = 10n_j \frac{\alpha}{\beta}$
- $\mathrm{var}(x_j) = 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2}$
- Setting these equal to the sample mean and variance yields $\alpha \approx 20$, $\beta \approx 430,000$.

Finding the right prior parameters cont.

- To find α and β, use the expectation and variance of x_j based on the Negative Binomial Distribution.
- $\mathbb{E}(x_j) = 10n_j\frac{\alpha}{\beta}$
- $\mathrm{var}(x_j) = 10n_j\frac{\alpha}{\beta} + (10n_j)^2\frac{\alpha}{\beta^2}$
- Setting these equal to the sample mean and variance yields $\alpha \approx 20$, $\beta \approx 430,000$.
- Then the estimated of mortality rates for each county are given by

$$\mathbb{E}(\theta_j|x_j) = \frac{20 + x_j}{430000 + 10n_j}$$

$$\mathrm{var}(\theta_j|x_j) = \frac{20 + x_j}{(430000 + 10n_j)^2}$$

- As the county population increases, the mortality rate goes down. As the number of recorded deaths increases, the mortality rate increases.

Comments

- We used a ”poor man’s hierarchical model” to estimate the prior distribution. A better way to find the parameters is to assign them their own prior and apply bayesian inference again.
- All the data are used to estimate the prior parameters, but each county’s mortality rate is estimated individually.
- Figure 1 Clearly shows some spatial correlation. A more involved model could make use of something called a variogram.
- To do this idea model mortality rates as a function of location, $x_j = x(\text{county j location})$.
- The variogram is given by $v(x_i, x_j) = \frac{1}{2}\mathbb{E}[(x(\text{county i location}) - x(\text{county j location}))^2]$
- Then describe the likelihood as non-independent poisson random variables with covariances given by the variogram, pick a prior, and repeat the process.
- Other models might include information about other risk factors determined by doctors.