Stat 534 Homework 10

Kenny Flagg

April 17, 2017

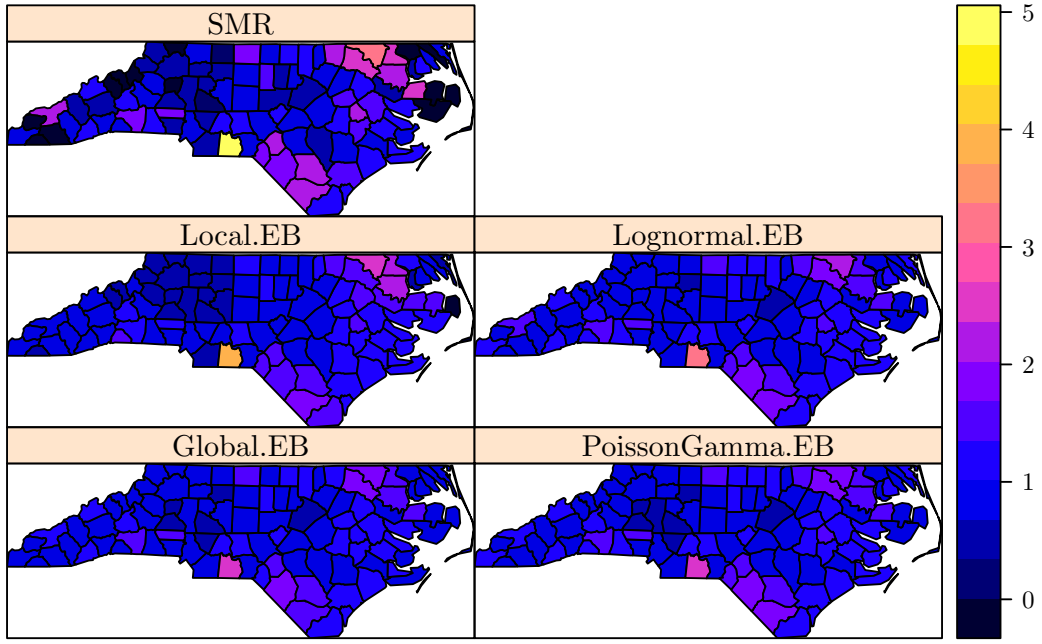
1. *Empirical Bayes – NC SIDS data: This is a well-known data set on the incidence of SIDS (Sudden Infant Death Syndrome) deaths in North Carolina. The data are available in the* ***nc.sids*** *data set in the* ***spdep*** *package. We will use emprical Bayes' methods to produce smoothed disease maps of SIDS incidence in the 1974-78 period. I have been letting you do most of the coding yourself this semester but getting the plots out for a problem like this is too much so you can use the R code below. Briefly discuss the results. In addition to the maps you can look at the raw SMR values and the smoothed estimates and summarize them in some meaningful way. I am particularly interested in counties in which the data dominate and counties in which the empirical Bayes' priors dominate. I will send an R script file with the code included also. I do not know what all those commands mean myself but I know they work.*

```
library(maptools)
library(spdep)
library(rgdal)
nc.sids <- readOGR(system.file('etc/shapes/sids.shp', package = 'spdep'), verbose = FALSE)
proj4string(nc.sids) <- CRS('+proj=longlat +ellps=clrk66')
# The IDs are just numbers...
#row.names(nc.sids) <- sapply(slot(nc.sids, 'polygons'), function(x) slot(x, 'ID'))
row.names(nc.sids) <- as.character(nc.sids$NAME) # County names.

rn <- nc.sids$FIPSNO
r <- sum(nc.sids$SID74) / sum(nc.sids$BIR74)
Expected <- nc.sids$BIR74 * r
SMR <- nc.sids$SID74 / Expected

require(DCluster)
EB.pg <- empbaysmooth(nc.sids$SID74, Expected)$smthrr
EB.LogN <- exp(lognormalEB(nc.sids$SID74, Expected)$smthrr)
EB.Global <- EBest(nc.sids$SID74, Expected)$estmm
ncCR85_nb <- read.gal(system.file('etc/weights/ncCR85.gal',
                                  package = 'spdep' )[1], region.id = rn)
EB.Local <- EBlocal(nc.sids$SID74, Expected, ncCR85_nb)$est

nc <- cbind(nc.sids, SMR, EB.pg, EB.LogN, EB.Global, EB.Local)
names(nc) <- c(names(nc.sids), 'SMR', 'PoissonGamma.EB', 'Lognormal.EB', 'Global.EB', 'Local.EB')
row.names(nc) <- as.character(nc$NAME)
spplot(nc, c('Global.EB', 'PoissonGamma.EB', 'Local.EB', 'Lognormal.EB', 'SMR'))
```




```
# Select 15 counties to show.
counties <- sort(c('Anson', 'Dare', sample(as.character(nc$NAME), 13)))
allSMRs <- data.frame(SMR = nc$SMR, `Local EB` = nc$Local.EB,
                      `Lognormal EB` = nc$Lognormal.EB,
                      `Poisson-Gamma EB` = nc$PoissonGamma.EB,
                      `Global EB` = nc$Global.EB,
                      row.names = nc$NAME, check.names = FALSE)
xtable(allSMRs[counties,], digits = 4)
```

	SMR	Local EB	Lognormal EB	Poisson-Gamma EB	Global EB
Alexander	0.0000	0.4940	0.8512	0.6531	0.6634
Anson	4.7264	4.0257	3.1034	2.5935	2.3937
Burke	0.6923	0.8435	0.8623	0.8289	0.8227
Dare	0.0000	0.0000	0.9148	0.8498	0.8345
Davidson	0.7184	0.6257	0.8416	0.8132	0.8093
Duplin	0.7969	1.2066	0.9356	0.9167	0.9013
McDowell	1.2711	1.0188	1.1524	1.1562	1.1153
Montgomery	1.1797	1.1091	1.1002	1.0997	1.0582
Orange	0.6254	1.0120	0.8435	0.7998	0.7954
Rutherford	1.9841	1.2626	1.6139	1.5924	1.5239
Stanly	1.0499	1.0123	1.0519	1.0516	1.0236
Surry	0.7759	0.5372	0.9105	0.8884	0.8771
Tyrrell	0.0000	1.5608	0.9683	0.9456	0.9138
Watauga	0.3739	0.5966	0.8667	0.7964	0.7903
Yadkin	0.3898	0.4960	0.8739	0.8089	0.8013

```
meansds <- sapply(allSMRs, function(x){return(c(Min = min(x), Mean = mean(x),
                                                Max = max(x), SD = sd(x)))})
xtable(meansds, digits = 4)
```

	SMR	Local EB	Lognormal EB	Poisson-Gamma EB	Global EB
Min	0.0000	0.0000	0.5844	0.5158	0.5229
Mean	1.0119	1.0257	1.0924	1.0535	1.0233
Max	4.7264	4.0257	3.1034	2.5935	2.3937
SD	0.7783	0.5177	0.3534	0.3333	0.3030

All of the smoothing methods clearly get the job done, as all four methods result in adjecent counties generally having smoothed values that are more similar than their raw values. The Global method does the most smoothing (having the lowest standard deviation and the smallest range of smoothed values), but the Poisson-Gamma results are similar. The Local method does the least smoothing (biggest SD and range). Unsurprisingly, all the smoothing methods result in smoothed values with similar means as they all pull the observed SMRs toward the observed mean.

Anson county is an outlier in raw SMR, and it dominates in the posterior for each method. For counties with low observed SMRs, the priors tend to dominate and pull the values toward the mean. The Lognormal method allows large SMRs to remain large, but still pulls low SMRs upward. The Local method behaves uniquely, actually smoothing some values *away* from the mean when neighboring counties had more extreme observed SMRs (e.g. Davidson and Surrey counties). Local smoothing actually assigns one county (Dare) a smoothed value of 0 because it and its neighbors all have observed SMRs of zero.

If I knew enough about SIDS to expect spatial autocorrelation, I would use Local smoothing to accomodate this. Otherwise, I would use either Lognormal or Global smoothing depending on whether or not I have prior information suggesting that large observed SMRs are real.

2. *Continuing with the NC SIDS data. We assume the Poisson-Gamma model.*

   (a) *Find the posterior distribution assuming a more informative Gamma prior with* $\alpha = \beta = 4$.

   We assume $Y_i|\theta \sim \text{Poisson}(E_i\theta)$, $\theta \sim \text{Gamma}(4, 4)$. We observe $\sum_{i=1}^{n} y_i = 667$ SIDS cases in 1974 across all counties. I define the exposure to be the number of births in thousands, so the response and the exposure are on similar scales. Thus $\sum_{i=1}^{n} E_i = 329.962$.

   Then the posterior distribution is $\theta|\mathbf{y} \sim \text{Gamma}(671, 333.962)$.

   (b) *Find the posterior distribution assuming a non-informative Gamma prior with* $\alpha = \beta = 0.1$.

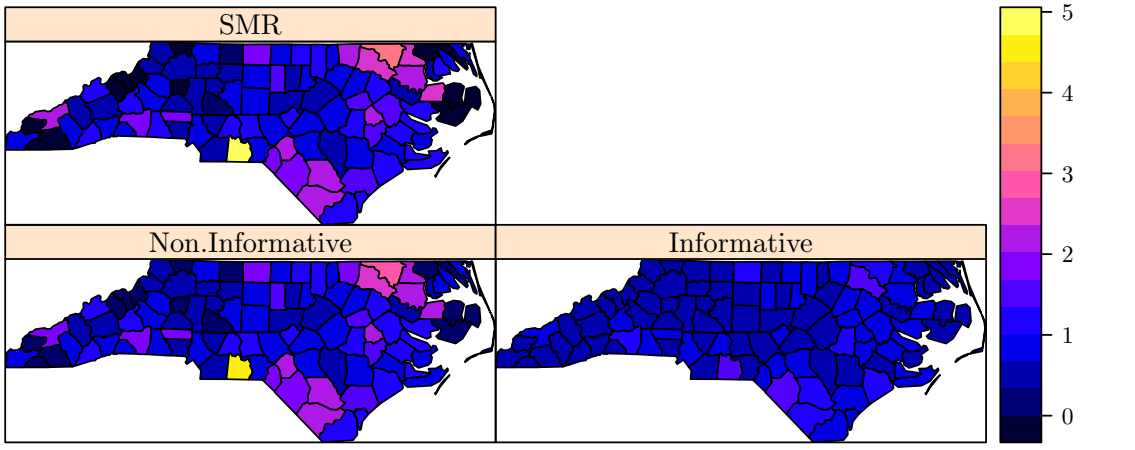
   The posterior distribution is $\theta|\mathbf{y} \sim \text{Gamma}(667.1, 330.062)$.

(c) *Summarize the role of the data and the prior in the posterior distributions by comparing the observed SMR and the posterior means from the 2 Bayesian approaches. A plot would help and you should be able to modify the code above to produce those plots. Just add the vectors of the poterior distribution to* `nc.sids` *using* `cbind` *and then use* `spplot` *to get plots of SMR, and the pure Bayes' posterior means.*

Both prior distributions have mean 1, and the observed SIDS counts have mean 6.67. The informative prior results in a posterior mean of $671/333.962 = 2.009$ and the non-informative prior yields a posterior mean of $667.1/330.062 = 2.021$. The informative prior pulls the posterior mean toward 1, but the data dominate over the non-informative prior resulting in a posterior mean close to the data mean.

To make a smoothed spatial map, I use the hierarchical model where $Y_i|\gamma_i \sim \text{Poisson}(E_i\gamma_i)$ and $\gamma_i \sim \text{Gamma}(\alpha, \beta)$. Then the posterior distribution for each county's mean is $\gamma_i|y_i \sim \text{Gamma}(\alpha + y_i, \beta + E_i)$. The result is that the informative prior dominates (doing a lot of smoothing) while the non-informative prior allows the data to dominate (resulting in a map that is barely distinguishable from map of the data). To create maps on SMR scale, I multiplied the posterior means by the exposure and divided by the expected counts.

```
nc.bayes <- cbind(nc,
                  (nc$BIR74 / 1000) * (nc$SID74 + 0.1) / (nc$BIR74 / 1000 + 0.1) / Expected,
                  (nc$BIR74 / 1000) * (nc$SID74 + 4) / (nc$BIR74 / 1000 + 4) / Expected)
names(nc.bayes) <- c(names(nc), 'Non.Informative', 'Informative')
spplot(nc.bayes, c('Non.Informative', 'Informative', 'SMR'))
```



3. *See the notes for the details on the Global Estimator (page 85). Marshall assumed, with E_i being the expected count that*

$$Z_i|\gamma_i \sim \text{Poi}(E_i\gamma_i)$$

but made no assumption about the distribution of the spatially varying relative risks or SMR values γ_i. We denote the estimates of SMR by $r_i = Z_i/E_i$. We make no distributional assumptions about the γ_i values but do assume that the prior means and variances exist. Denote them m_{γ_i} and v_{γ_i}, respectively. Marshall found method-of-moments estimators of the marginal or unconditional mean and variance of the γ_i.

(a) *Find the conditional mean and variance of r_i. That is find $E(r_i|\gamma_i)$ and $Var(r_i|\gamma_i)$.*

$$E(r_i|\gamma_i) = E\left(\frac{Z_i}{E_i}\middle|\gamma_i\right) = \frac{E(Z_i|\gamma_i)}{E_i} = \frac{E_i\gamma_i}{E_i} = \gamma_i$$

$$Var(r_i|\gamma_i) = Var\left(\frac{Z_i}{E_i}\middle|\gamma_i\right) = \frac{Var(Z_i|\gamma_i)}{E_i^2} = \frac{E_i\gamma_i}{E_i^2} = \frac{\gamma_i}{E_i}$$

(b) *Show that the marginal mean of r_i is equal to the prior mean of γ_i.*

$$\begin{aligned} E(r_i) &= E(E(r_i|\gamma_i)) \\ &= \int_0^\infty E(r_i|\gamma_i)f(\gamma_i)d\gamma_i \\ &= \int_0^\infty \gamma_i f(\gamma_i)d\gamma_i \\ &= E(\gamma_i) \end{aligned}$$

(c) *Find the marginal variance of r_i. (Hint: it will be a function of the prior mean and variance of γ_i.)*

$$\begin{aligned} Var(r_i) &= E\left((r_i - E(r_i))^2\right) \\ &= E\left((r_i - E(\gamma_i))^2\right) \\ &= E\left(r_i^2\right) - (E(\gamma_i))^2 \\ &= E\left(E\left(r_i^2|\gamma_i\right)\right) - (E(\gamma_i))^2 \\ &= \int_0^\infty E\left(r_i^2|\gamma_i\right) f(\gamma_i)d\gamma_i - (E(\gamma_i))^2 \\ &= \int_0^\infty \left(Var\left(r_i|\gamma_i\right) + \left(E\left(r_i|\gamma_i\right)\right)^2\right) f(\gamma_i)d\gamma_i - (E(\gamma_i))^2 \\ &= \int_0^\infty \left(\frac{\gamma_i}{E_i} + \gamma_i^2\right) f(\gamma_i)d\gamma_i - (E(\gamma_i))^2 \\ &= E\left(\frac{\gamma_i}{E_i}\right) + E\left(\gamma_i^2\right) - (E(\gamma_i))^2 \\ &= \frac{E(\gamma_i)}{E_i} + Var(\gamma_i) \end{aligned}$$