Small Area Estimation with R

Unit 6: Non-linear models

V. Gómez-Rubio

Department of Mathematics
Universidad de Castilla-La Mancha, Spain

useR! 2016
27th June 2016, Stanford University

Non-linear models

- So far, we have only considered models with a Normal reponse
- Generalized Linear Models have been used to tackle Small Area Estimation problems
- Problems may arise when combining individual and aggregate models
- Specific methods will be required to combine information efficiently (see, for example, Jackson et al., 2007)
- We will show examples using a Bayesian approach

Disease mapping

- Health authorities collect mortality (and morbidity) data on a regular basis
- The aim of disease mapping is to estimate the relative risk of a certain disease
- In addition, to the observed number of cases, and expected number is computed on the population and, possibly, some known risk factors
- Spatial random effects are often considerd because risks are assumed to vary smoothly
- Temporal effects can be included if the data cover several periods of time

Example: SIDS in North Carolina

- Cressie and Chan (1989) have studied the mortality in children by Sudden Infant Death Syndrome (O_i) in 1974-78 and 1979-84
- The administrative aggregation of the data is county level
- In addition, the number of births (N_i) and the proportion of non-white births (NW_i) are available and can be used to compute the expected number of cases
- These data are available in package `spdep`: `data(nc.sids)`
- Overall incidence rate is

$$r = \frac{\sum_i O_i}{\sum_i N_i}$$

- Expected number of cases are computed used indirect standardisation:

$$E_i = N_i \cdot r$$

Besag, York and Mollié (1991)

BYM propose a model that includes spatial and non-spatial random effects to account for different types of unmeasured variables:

$$O_i \sim Po(\mu_i)$$

$$log(\mu_i) = \log(E_i) + \alpha + \beta NW_i + u_i + v_i$$

$$u_i \sim N(0, \sigma_u^2)$$

$$v_i | v_{-i} \sim N(\sum_{j \sim i} v_j / n_j, \sigma_v^2 / n_j)$$

WinBUGS model

```
model
{

  for(i in 1:N)
  {
     observed[i] ~ dpois(mu[i])
     log(theta[i]) <-  alpha + beta*nonwhite[i] + u[i] + v[i]
     mu[i] <- expected[i]*theta[i]

     u[i] ~ dnorm(0, precu)
  }

  v[1:N] ~ car.normal(adj[], weights[], num[], precv)

  alpha ~ dflat()
  beta ~ dnorm(0,1.0E-5)
  precu ~ dgamma(0.001, 0.001)
  precv ~ dgamma(0.1, 0.1)

  sigmau<-1/precu
  sigmav<-1/precv
}
```

Example: SIDS in North Carolina

The following code is used to set the data and run the model to fit

```
> library(maptools)
> library(spdep)
> library(rgdal)
> #Read data from shapefile
> nc <- readShapePoly(system.file("etc/shapes/sids.shp",
+   package = "spdep")[1],
+   ID = "FIPSNO")
> rn <- sapply(slot(nc, "polygons"), function(x) slot(x, "ID"))
> #Adjacency matrix from Cressie and Chan (1989)
> ncCC89nb <- read.gal(system.file("etc/weights/ncCC89.gal",
+   package = "spdep")[1], region.id = rn)
> #Transform adj. matrix into the format required by WB
> nc.nb <- nb2WB(ncCC89nb)
```

Example: SIDS in North Carolina

```
> #Prepare data set
> nc$Observed <- nc$SID74
> nc$Population <- nc$BIR74#Population at risk; number of births
> r <- sum(nc$Observed) / sum(nc$Population)
> nc$Expected <- nc$Population * r
> N <- length(nc$Observed)
> #Computed Standardised Mortality Ratio
> nc$SMR <- nc$Observed / nc$Expected
> #Proportion of non-white births
> nc$nwprop <- nc$NWBIR74 / nc$BIR74
> #Prepare data and initial values for WinBUGS
> d <- list(N = N, observed = nc$Observed, expected = nc$Expected,
+   nonwhite = nc$nwprop,#log(nwprop/(1-nwprop)),
+   adj = nc.nb$adj,  weights = nc.nb$weights, num = nc.nb$num)
> inits <- list(u = rep(0, N), v = rep(0, N), alpha = 0, beta = 0,
+   precu = .001, precv = .001)
```

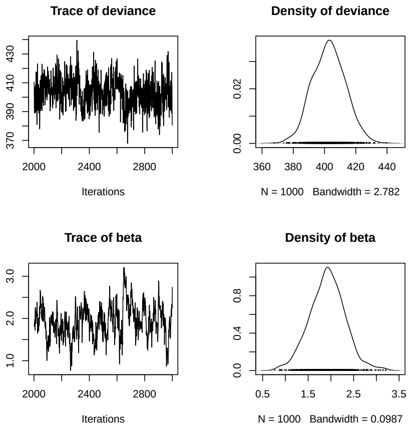
Example: SIDS in North Carolina

```
> library(R2WinBUGS)
> bymmodelfile<-paste(getwd(), "/BYM-model.txt", sep="")
> wdir<-paste(getwd(), "/BYM", sep="")
> if(!file.exists(wdir)){dir.create(wdir)}
> BugsDir <-
+    "/Users/virgil/.wine/dosdevices/c:/Program Files/WinBUGS14"
> MCMCres<- bugs(data=d, inits=list(inits),
+    working.directory=wdir,
+    parameters.to.save=c("theta", "alpha", "beta", "u", "v",
+      "sigmau", "sigmav"),
+    n.chains=1, n.iter=30000, n.burnin=20000, n.thin=10,
+    model.file=bymmodelfile, bugs.directory=BugsDir,
+    WINEPATH="/usr/local/bin/winepath")
> #Load the data obtained by running WinBUGS in Windows
> nc$BYMmean<-MCMCres$mean$theta
> nc$BYMumean<-MCMCres$mean$u
> nc$BYMvmean<-NA
> nc$BYMvmean[nc.nb$num>0]<-MCMCres$mean$v
```

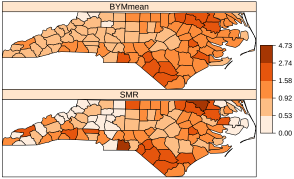
Example: SIDS in North Carolina

Assessing converenge of MCMC methods

```
> library(coda)
> ncoutput <- read.coda("BYM/coda1.txt", "BYM/codaIndex.txt")
> plot(ncoutput[,c("deviance", "beta")])
```



Example: SIDS in North Carolina



```
> library(RColorBrewer)
> brks <- quantile(nc$SMR, seq(0, 1, 1/5))
> #Used method proposed by Nicky Best
> logSMR <- log(nc$SMR[nc$SMR > 0])
> nsteps <- 5
> step <- (max(logSMR) - min(logSMR)) / nsteps
> brks <- exp(min(logSMR) + (0:nsteps) * step)
> brks[1] <- 0
> cols <- brewer.pal(5, "Oranges")
> atcol <- (0:5) * max(nc$SMR)/5
> key.labels <- as.character(c(formatC(brks, format = "f", dig = 2)))
> colorkey <- list(labels = key.labels, at = atcol,  height = .5)
> print(spplot(nc, c("SMR", "BYMmean"), at = brks, col.regions = cols,
+   axes = TRUE, colorkey = colorkey))
```

Estimation of unemployment

- Some surveys record the unemployment status of the population in the sample
- Other socio-economic covariates can be recorded
- A logictic regression can be used to model the probability of being unemployed in the area
- In order to produce small area estimates, the model must be fit using area level covariates
- The Office for National Statistics (UK) used the following model in some reports:

$$y_{ij} \sim Binom(p_i, N_i);\ j = 1, \dots, n_i$$

$$\mathrm{logit}(p_i) = \alpha + \beta X_i;\ i = 1, \dots, K$$

Unemployment in Sweden

- We have simulated a data set mimicking a survey on the population
- The target variable is the employment status: 1-employed, 0-unemployed
- The covariates, which are based on the area level, are
    - Age (average age)
    - Sex (proportion of males)
    - Higher education status (proportion of persons with higher education)
- Unit level covariates are not used in this model because combining them with area level covariates is not straighforward
- If the mode is fit with unit level covariates and the small area estimates are computed by 'plugging-in' the area level covariates a bias is introduced

WinBUGS model

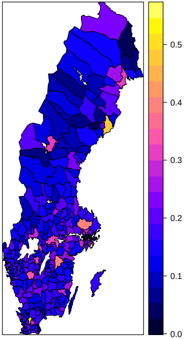
```
model
{
for(i in 1:totssize)
{
emp[i]~dbern(p[i])

logit(pp[i])<-alpha+bage*AGE[area[i]]+bsex*SEX[area[i]]+beduc*EDUC[area[i]]+u[area[i]]+v[area[i]]

p[i]<-max(0.000000001, min(pp[i], 0.999999999))
}

for(i in 1:N)
{
u[i] ~ dnorm(0, precu)

logit(rateemp[i])<-alpha+bage*AGE[i]+bsex*SEX[i]+beduc*EDUC[i]+u[i]+v[i]
rateunemp[i]<- 1-rateemp[i]
}

v[1:N] ~ car.normal(adj[], weights[], num[], precv)

precu ~ dgamma (a0,b0)
precv ~ dgamma (a1,b1)

alpha ~ dflat()
bage ~ dflat()
bsex ~ dflat()
beduc ~ dflat()

sigmau<-1/precu
sigmav<-1/precv
}
```

Example: Unemployment in Sweden

```
> dunemp <- source("Unemployment/WBdata/datasp.txt")$value
> dsp <- source("Unemployment/WBdata/spdata.txt")$value
> initsunemp1 <- source("Unemployment/inits/initssp-1.txt")$value
> initsunemp2 <- source("Unemployment/inits/initssp-2.txt")$value
> bymmodelfile <- paste0(getwd(), "/Unemployment/models/modelsp.txt"
> wdir <- paste0(getwd(), "/BYM-Unemp")
> if(!file.exists(wdir)){dir.create(wdir)}
> #BugsDir <-
> #  "/Users/virgiliogomezgislab/.wine/dosdevices/c:/Program Files/
> MCMCresunemp<- bugs(data = c(dunemp, dsp),
+    inits = list(initsunemp1, initsunemp2),
+    working.directory = wdir,
+    parameters.to.save = c("p", "alpha", "bage", "bsex", "beduc"),
+    n.chains = 2, n.iter = 3000, n.burnin = 2000, n.thin = 1,
+    model.file = bymmodelfile,
+    bugs.directory = BugsDir,
+    WINEPATH = "/usr/bin/winepath")
```

```
> load("MCMCresunemp.RData")
```

Example: Unemployment in Sweden



```
> library(maptools)
> Sweden <- readShapePoly(fn = "Sweden_municipality")
> Sweden <- unionSpatialPolygons(Sweden, Sweden$KOD83_91)
> Sweden <- SpatialPolygonsDataFrame(Sweden,
+    data.frame(unemp = 1 - unique(MCMCresunemp$mean$p)),
+    match.ID = FALSE )
> print(spplot(Sweden, "unemp", cuts=20))
```

References

- Besag, J., J. C. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59.
- Cressie, N, Chan NH (1989) Spatial modelling of regional variables. *Journal of the American Statistical Association* **84**: 393-401
- Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal. *Statistical Science* **9**(1), 55–76.
- Jackson, C., Best, N. and Richardson, S. (2007). Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A*, **171(1)**, 159–178.