Bayesian model in JAGS

Tutorial 3 for transition

Zhe Zheng (Gigi)[1]

2023-01-16

[1] zhe.zheng@yale.edu

Outline

- Section 1: Background knowledge of estimating respiratory virus attributable disease burden (hospitalizations/mortality etc) with statistical models.
- Section 2: Background knowledge of Hierarchical Bayesian regression
- Section 3: Hierarchical Bayesian regression to estimate RSV attributable hospitalizations in older adults by age and risk groups and its R code.

Note: In this tutorial, we will focus on using time-series data. Therefore, we will not introduce other methods to estimate disease burden such as the disease pyramid. For other methods, please check out Ginny's class *quantitative method in infectious disease epidemiology* and Dan's class *Public Health Surveillance*.

Section 1: Background knowledge of estimating respiratory virus attributable disease burden with statistical models.

There are 5 main statistical methods to estimate the hospitalization/mortality attributable to respiratory virus infection. In Dan's class, Public Health Surveillance, he teaches all four methods in details. Here, we will only give a brief introduction to each of the method and provide the link to the initial publications.

1. **Serfling regression**
2. **Periseason differences**
3. **Poisson regression with log link**
4. **Negative binomial regression with identity link**
5. **Box-Jenkins transfer function (ARIMA model)**

Brief introduction to serfling regression and periseason differences

1.**Serfling regression**

This type of model identifies the epidemic season and establish a epidemic threshold (seasonal baseline) using historical data. After predicting the expected diseases baseline, observed diseases above the baseline (epidemic threshold) during the epidemic season will be attributable to the viral infection. The basic Serfling regression is given by:
$Y_t = \mu + bt + \sum \alpha_i cos\theta + \sum \beta_i sin\theta$ where Y_t is the expected hospitalizations at time t. μ is the baseline. bt captures time trend. $\sum \alpha_i \cos\theta + \sum \beta_i \sin\theta$ are for the seasonal variations. Serfling regression is a linear model initially using ordinary least square to fit. Click to see an R code example

2.**Periseason differences**

This method calculate the hospitalizations attributable to influenza based on the differences in the hospitalization rates when respiratory virus was circulating and the hospitalization rates when there was no respiratory virus circulation in the community.

Brief introduction to poisson regression with log link

3.**Poisson regression with log link**
Poisson regression is used to model count variables. Therefore, it can be used to model the number of hospitalizations over time. The logarithm of the expected value of the count variable is modeled as a linear combination of the predictor variables. For example, in Thompson's paper[2], the poisson regression with log link is given by:

$$Y = \alpha \exp^{(\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 \sin(2\pi/52)+\beta_4 \cos(2\pi/52)+\beta_5 Flu_A(H1N1)+\beta_6 Flu_A(H3N2)+\beta_7 Flu_B+\beta_8 RSV)}$$

In this example, Y is the observed number of deaths/hospitalization in a particular week, α is the population size as an offset term,β_0 is the baseline hospitalization incidence, $\beta_1 t + \beta_2 t^2$ represents the long-term trend of hospitalizations, $\sin(2\pi/52) + \beta_4 \cos(2\pi/52)$ captures the annual seasonal variations of hospitalizations, the rest are the pathogens that can contribute to the increase of the observed hospitalizations.

---

[2]Thompson WW, Shay DK, Weintraub E, et al. Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States. JAMA. 2003;289(2):179–186. doi:10.1001/jama.289.2.179

Brief introduction to negative binomial regression with identity link

4.**Negative binomial regression with identity link**

Negative binomial regression is also used to model count variables. It is a generalization of the poisson regression, in which the variance is assumed to be equal to the mean. The new approach use an identity link instead of an log link because log link corresponds to a multiplicative relationship among covariates (including baseline hospitalizations, seasonality, and several respiratory viruses) while identity link assumes additive effects. Apparently, additive effects are more realistic than multiplicative effects. The negative binomial regression with identity link is given by:

$$Y \sim NB(p, r), p = \frac{r}{r + \lambda}$$

$$\lambda = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \sin(2\pi/52) + \beta_4 \cos(2\pi/52) + ... + \beta_8 RSV$$

where Y is the number of observed failures (diseases), p is the probability of success (survival/not infected), r is the number of successes (survival/not infected) and λ is the expected value of failures (diseases). All other parameters remain the same as what we introduced in Poisson regression.

Brief introduction to ARIMA model

5.**Autoregressive integrated moving average (ARIMA) model**
ARIMA model can account for the autorcorrelation of the time-series and also the time-delayed associations. For infectious diseases, the number of the infected individuals in the next time point is correlated with the number of the infected individuals currently. This phenomenon is described as autocorrelated. The influenza infection reported this week may contribute to hospitalizations next week, which means the time-delayed associations. For example, in Gilca's paper[3], ARIMA model is given by:

$$Y_t = \alpha_0 + \alpha_1 Temp_t + \alpha_2 Holiday_t + \beta_1 Flu_t + \sum_{i=0}^{2} \omega_i RSV_{t-i}$$

Y_t is the observed number of hospitalizations in a particular week t, α_0 is the baseline hospitalization incidence,$\alpha_1 Temp_t + \alpha_2 Holiday_t$ are the hospitalizations attributable to seasonal changes and holiday gathering, $\beta_1 Flu_t$ is the flu attributable hospitalizations, and $\sum_{i=0}^{2} \omega_i RSV_{t-i}$ is the RSV attributable hospitalizations. **This example suggested that not only RSV infections this week but also RSV infections reported 1 week and 2 weeks ago associates with hospitalizations reported this week.**

---

[3] Rodica Gilca, Gaston De Serres, Danuta Skowronski, Guy Boivin, David L. Buckeridge, The Need for Validation of Statistical Methods for Estimating Respiratory Virus–Attributable Hospitalization, American Journal of Epidemiology, Volume 170, Issue 7, 1 October 2009, Pages 925–936, https://doi.org/10.1093/aje/kwp195

Initial publications of the method

1. **Serfling regression**

- Methods for Current Statistical Analysis of Excess Pneumonia-influenza Deaths
- The impact of influenza epidemics on mortality: introducing a severity index
- Impact of Influenza Vaccination on Seasonal Mortality in the US Elderly Population

2. **Periseason differences**

- The effect of influenza on hospitalizations, outpatient visits, and courses of antibiotics in children
- Respiratory illness associated with influenza and respiratory syncytial virus infection
- Influenza and the Rates of Hospitalization for Respiratory Disease among Infants and Young Children
- Impact of influenza and respiratory syncytial virus on mortality in England and Wales from January 1975 to December 1990

Initial publications of the method (continued)

3. **Poisson regression with log link**

- Mortality Associated With Influenza and Respiratory Syncytial Virus in the United States
- The Association of Respiratory Syncytial Virus Infection and Influenza with Emergency Admissions for Respiratory Disease in London: An Analysis of Routine Surveillance Data

4. **Negative binomial regression with identity link**

- Modelling the unidentified mortality burden from thirteen infectious pathogenic microorganisms in infants
- Hospitalization Attributable to Influenza and Other Viral Respiratory Illnesses in Canadian Children

5. **Autoregressive integrated moving average (ARIMA) model**

- Community influenza outbreaks and emergency department ambulance diversion
- Time-Series Analysis of the Relation between Influenza Virus and Hospital Admissions of the Elderly in Ontario, Canada, for Pneumonia, Chronic Lung Disease, and Congestive Heart Failure

Other useful references

- The Need for Validation of Statistical Methods for Estimating Respiratory Virus–Attributable Hospitalization

Section 2: Background knowledge of using Hierarchical Bayesian regression to estimate RSV attributable hospitalizations

The benefits of using a Hierarchical Bayesian structure is that it can shrink the uncertainty of inference while allow between-group variability.

For example, if we would like to estimate the baseline hospitalizations β_{0jk} in age j and socioeconomic group k, we can treat this parameter as (1) independent, meaning that the baseline hospitalizations are different in each group and we will need to estimate 27 parameters $\beta_{0j=1,k=1}$, $\beta_{0j=1,k=2} \ldots \beta_{0j=9,k=3}$, (2) the same, meaning that the baseline hospitalizations are the same in all age and socioeconomic group and we will estimate one parameter β_0, and (3) random draws from the a normal distribution, meaning that baseline hospitalizations are normally distributed in the whole population and baseline hospitalization of each subpopulation jk is an realization of the distribution (a point in the distribution). In this case, we will estimate the mean and variations of the normal distribution and then get the subpopulation estimates.

Background knowledge of Hierarchical Bayesian regression: Benefits

If we model the parameters in Bayesian framework, the third scenario corresponds to a Hierarchical Bayesian model.

Compared with treating the parameters as group-specific, Bayesian Hierarchical structure pools information across groups to help shrink the uncertainty of the single group estimates.

Compared with treating the group specific parameters as homogeneous, Bayesian Hierarchical structure gives the flexibility and possibility of estimating group specific parameters. It can help evaluate the hypothesis of between-group variation.

Visualization of a Hierarchical Bayesian structure:

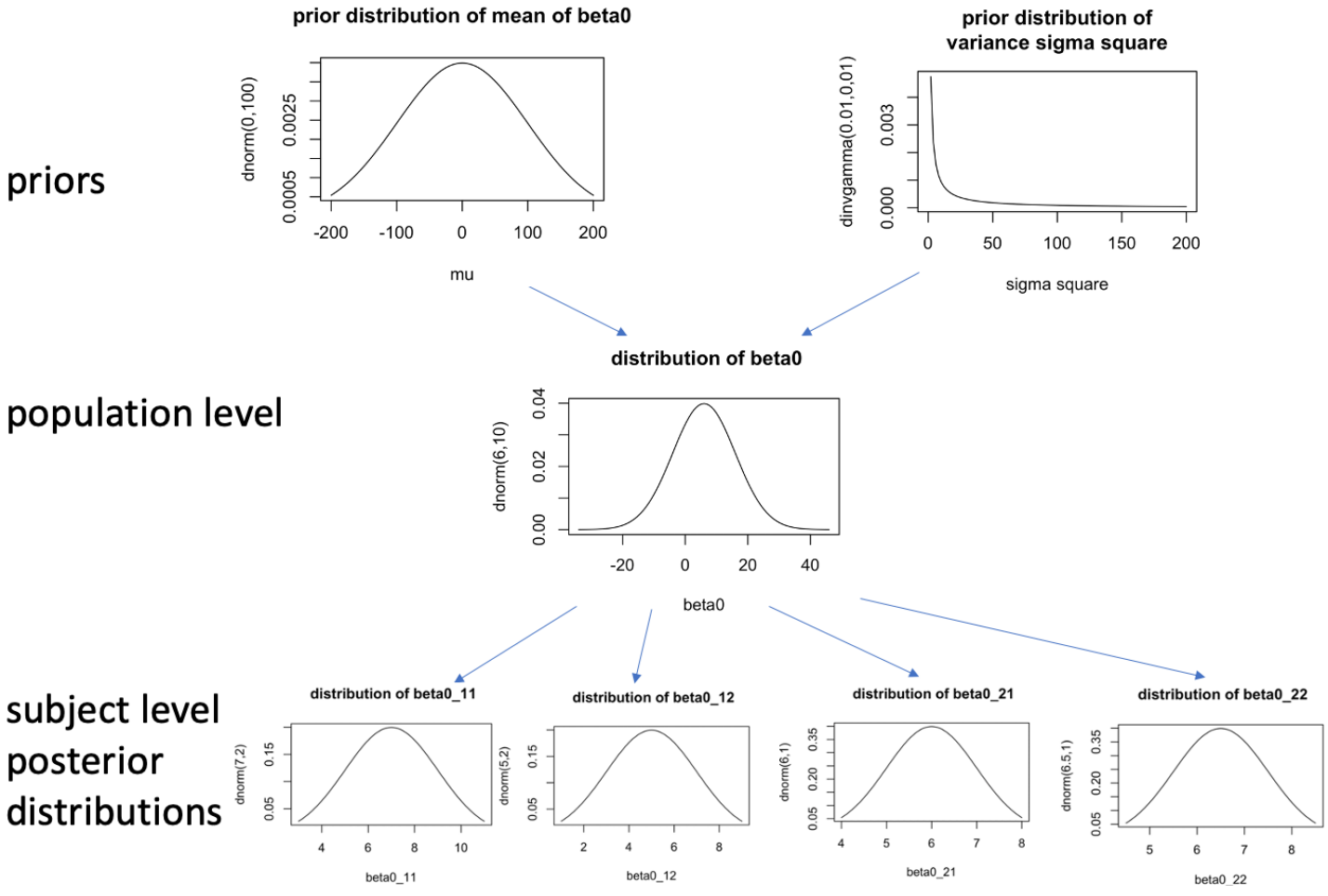
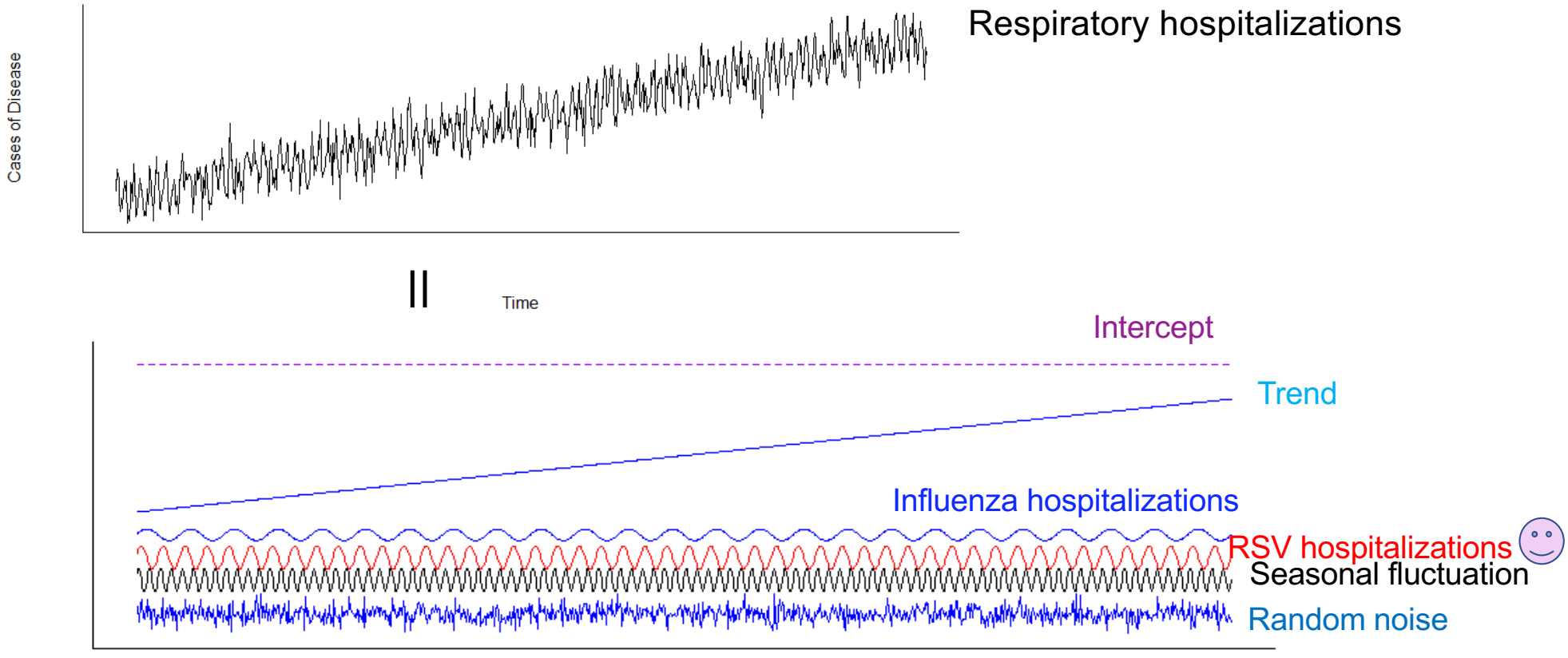


Figure 1: Hierarchical Bayesian structure

Further readings on Hierarchical Bayesian regression

- Bayesian hierarchical modeling by wikipedia
- Chapter 10 Bayesian Hierarchical Modeling by Drs. Jim Albert and Jingchen Hu
- Hierarchical Bayesian Modeling by Dr. Angie Wolfgang

Section 3: Hierarchical Bayesian regression to estimate RSV attributable hospitalizations in older adults by age and risk groups and its R code.



Courtesy of Dan's lecture note

Hierarchical Bayesian regression structure

The model structure of our hierarchical Bayesian regression is given as:

$$Y_{ijk} \sim \text{Negative Binomial}\left(p_{ijk}, r\right),\ p_{ijk} = \frac{r}{r + \lambda_{ijk}}$$

$$\lambda_{ijk} = \beta_{0jk} + \alpha_{1g(i)} + \alpha_{2m(i)} + \beta_{1jk} RSV_{ik} + \beta_{2g(i)jk} Flu_{ik}$$

Intercept

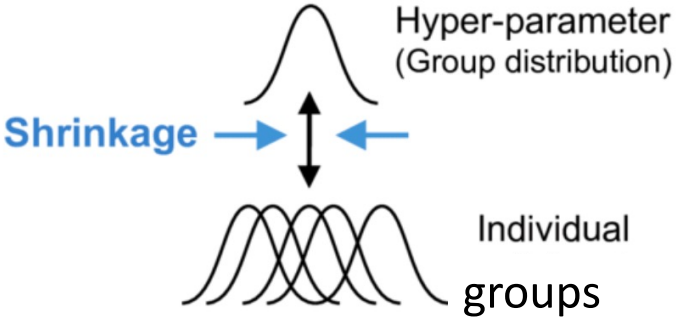
yearly variations

monthly variations

group-specific effect of RSV

RSV hospitalizations in children < 2 years old

group-specific effect of flu (varies by epidemiologic year)



time *i*
age group *j*
SES group *k*

The model structure of our hierarchical Bayesian regression is given as:

$$Y_{ijk} \sim \text{Negative Binomial}\left(p_{ijk}, r\right), \;\; p_{ijk} = \frac{r}{r + \lambda_{ijk}}$$

where Y_{ijk} denotes the number of all-cause respiratory hospitalizations at time *i*, in age group *j*, and SES group *k*; the expected value of Y_{ijk} is λ_{ijk}; and the variance of Y_{ijk} is $\lambda_{ijk}\left(1 + \lambda_{ijk}/r\right)$. The parameter $r > 0$ serves as an overdispersion parameter with $r \to \infty$ indicating that the mean and variance are the same, as in a typical Poisson regression framework. We define the expected value as a function of covariates and random effects such that

$$\lambda_{ijk} = \beta_{0jk} + \alpha_{1g(i)} + \alpha_{2m(i)} + \beta_{1jk} RSV_{ik} + \beta_{2g(i)jk} Flu_{ik} \quad (1)$$

where β_{0jk} is the intercept parameter for age group *j* and SES group *k*; $\alpha_{1g(i)}$ represents an intercept term that varies by epidemiologic year, where *g(i)* is a function that maps time to epidemiologic year (defined from July in the previous year to June in the next year); $\alpha_{2m(i)}$ represents a similar intercept term which varies by month, where *m(i)* is a function that maps time to the corresponding month category; β_{1jk} describes the group-specific effect of RSV, which associates RSV infections to respiratory hospitalizations; $\beta_{2g(i)jk}$ describes the association between influenza infections and respiratory hospitalizations, and varies by epidemiologic year in addition to the groups [28].

We model the coefficients of RSV infections (β_{1jk} parameters) as a multiplicative combination of age and SES effects, such that

$$\beta_{1jk} = \exp\{\omega_{1j} + \gamma_{1k} + \epsilon_{1jk}\}$$

where ω_{1j} represents the age group effects; γ_{1k} represents the SES effects; and $\epsilon_{1jk} \sim \mathrm{N}(0, \sigma_{\epsilon 1}^2)$ accounts for other unexplained variation. A similar structure is applied to the coefficients of influenza infections ($\beta_{2g(i)jk}$), with an additional yearly effect included, such that

$$\beta_{2g(i)jk} = \exp\{\omega_{2j} + \gamma_{2k} + \xi_{2g(i)} + \epsilon_{2g(i)jk}\}$$

where ω_{2j} represents the age group effects; γ_{2k} represents the SES effects; $\xi_{2g(i)}$ represents the epidemiologic year effects potentially due to differences in the severity of the circulating strain; and $\epsilon_{2g(i)jk} \sim \mathrm{N}(0, \sigma_{\epsilon 2}^2)$ accounts for other unexplained variation. We assign weakly informative prior

distributions to the remaining model parameters while ensuring that the remaining intercept parameters (i.e., $\beta_{0jk}, \alpha_{1g(i)}, \alpha_{2m(i)}$) are positive.

Posterior samples were collected using a Markov chain Monte Carlo (MCMC) algorithm. To make posterior inference, three chains of 12,500 MCMC iterations were used following an initial burn-in of 62,500 iterations per chain. The combined set of 37,500 MCMC iterations were then thinned by a factor of 10, resulting in 3,750 less correlated posterior samples from the joint posterior distribution with which to make inference. Convergence was assessed by examining individual parameter trace plots and Gelman-Robin diagnostics [38, 39]. Posterior means and 95% equal-tailed quantile-based credible intervals were calculated using the samples. The model was fitted using the rjags package [40]

The hyperparameters for $\beta_{0jk}, \alpha_{1g(i)}, \alpha_{2m(i)}, \omega_{1j}, \omega_{2j}, \gamma_{1k}, \gamma_{2k}$ are as follows:

$$\beta_{0jk} = \exp\{\mu_{0jk}\}$$

$$\mu_{0jk} \sim \mathrm{N}(\mu_0, \sigma_0^2)$$

$$\alpha_{1g(i)} = \exp\{\mu_{1g(i)}\}$$

$$\mu_{1g(i)} \sim \mathrm{N}(0, \sigma_1^2)$$

$$\alpha_{2m(i)} = \exp\{\mu_{2m(i)}\}$$

$$\mu_{2m(i)} \sim \mathrm{N}(0, \sigma_2^2)$$

$$\omega_{1j} \sim \mathrm{N}\left(0, \sigma_{\omega_1}^2\right)$$

$$\omega_{2j} \sim \mathrm{N}\left(0, \sigma_{\omega_2}^2\right)$$

$$\gamma_{1k} \sim \mathrm{N}\left(0, \sigma_{\gamma_1}^2\right)$$

$$\gamma_{2k} \sim \mathrm{N}\left(0, \sigma_{\gamma_2}^2\right)$$

$$\sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_{\omega_1}^2, \sigma_{\omega_2}^2, \sigma_{\gamma_1}^2, \sigma_{\gamma_2}^2 \sim \text{Inverse Gamma}(0.01,0.01)$$

$$\mu_0 \sim \mathrm{N}(0, 1000)$$

After we fitted the model, the estimated "true" number of hospitalizations attributable to RSV infections for each time point and stratum were estimated by multiplying posterior samples of β_{1jk} by RSV_{ik}. The average annual incidence of RSV was estimated by dividing the sum of $\beta_{1jk} RSV_{ik}$ over nine epidemiologic years by the age- and site-specific population. The recording ratios were

calculated by dividing the number of recorded ICD-9-CM diagnoses for RSV in each age and SES group by the modeled estimates in the same group. The attributable percent of RSV was calculated by dividing the sum of $\beta_{1jk}RSV_{ik}$ by the sum of the model-predicted number of all-cause respiratory hospitalizations in each age and SES group (λ_{ijk}) over the entire study period. (Note that $E(Y_{ijk}) = \lambda_{ijk}$ and Y_{ijk} in each group are almost identical; $E(Y_{ijk})$ captures 99.98% of the variability in Y_{ijk}.) For each measure, we obtained and summarized samples from the posterior distributions of interest.

R code (JAGS code) for Hierarchical Bayesian regression structure

Please open Jags_model.R for the whole R script.

$k = 1, 2, 3$ represent each socioeconomic groups, $j = 1, ..., 9$ represent each age groups, and $i = 1, ..., 108$ represent each month.

Hierarchical Bayesian model are specified in jags as follows:

```
model {
for(k in 1:n.group){
  for (j in 1:n.age){
      for (i in 1:n.date) {

 y[i,j,k] ~ dnegbin(prob[i,j,k],r)
 # negative binomial
  prob[i,j,k]<- r/(r+lambda[i,j,k])  ## likelihood
```

R code (JAGS code): expected respiratory hospitalizations

```
# expected respiratory hospitalizations lambda
    lambda[i,j,k] <- rd0[j,k]+ # baseline hosp
exp(epi[epi.year[i]]) + # year to year variation
# epi.year[i] is an input indicator for the calendar year
# epi.year[i] = 1,2,...,9
exp(delta[month[i]]) + # monthly variations
# month[i] is an input indicator for the epi month
# month[i] = 1,2,...,12 (1 for July and 12 for June)
rsv[i,j,k]*rd2[j,k] + # RSV attributable hosp
flu[i,j,k]*rd1[epi.year[i],j,k] # flu attributable hosp
      }

# baseline hospitalizations
#(must be greater than or equal to 0)
  rd0[j,k] <- exp(beta0[j,k])
  beta0[j,k]~ dnorm(mu0,tau0)
# hierarchical structure: shared mean and variance
```

The coefficient of flu in each age and SES group

```
 for (p in 1:n.year) {
# epi-year effects for antigen shift
# coefficient of influenza-associated respiratory
# hospitalization varies annually
    rd1[p,j,k] <- exp(beta1[p,j,k])
    # ensure positive coefs
    beta1[p,j,k] ~ dnorm(beta1_mean[p,j,k],tau.flu)
    # shared variance
    # the mean of the coefficient is
# a multiplicative combination of
    beta1_mean[p,j,k] <- gamma_flu[k]+ omega_flu[j]+ xi_flu[p]
    # SES effects gamma_flu[k]
    # age effects omega_flu[j]
    # epi-year (antigen) xi_flu[p]
    }
```

The coefficient of RSV in each age and SES group

```
# this coefficient depends on SES and age
   rd2[j,k] <- exp(beta2[j,k]) # ensure positive
   beta2[j,k] ~ dnorm(beta2_mean[j,k],tau.rsv)
   beta2_mean[j,k]<- gamma[k]+omega[j]
  # SES effects gamma[k]
 # age effects omega[j]
```

hyperparameters

```
# n.year = 1,2,...,9
for (p in 1:n.year) {
epi[p] ~ dnorm(0, tau.epi)
# prior for yearly variation of baseline hosp
xi_flu[p] ~ dnorm(0,tau7)}
# prior for flu annual variation (antigen shift)

  for (m in 1:12){
    delta[m] ~ dnorm(0,disp.m)
  }
# prior for monthly variation of baseline hosp
```

hyperparameters (continued)

```
for(k in 1:n.group){
gamma_flu[k] ~ dnorm(0,tau3)
gamma[k] ~ dnorm(0, tau4)
}
# prior for SES effects of flu and RSV

for(j in 1:n.age){
omega[j] ~ dnorm(0, tau5)
omega_flu[j] ~ dnorm(0,tau6)
}
# prior for age effects of flu and RSV
```

Priors

```
 r ~ dunif(0,250)
#r > 0 serves as an overdispersion parameter
  mu0 ~ dnorm(0,0.0001) # 0.0001 = 1/variance
# In jags, dnorm specify precision

# the conjugate priors for variance of normal distribution
# is an inverse Gamma distribution
# In jags, we put priors on precision
  tau0 ~ dgamma(0.01, 0.01)
  tau1 ~ dgamma(0.01, 0.01)
  tau2 ~ dgamma(0.01, 0.01)
  tau3 ~ dgamma(0.01, 0.01)
  tau4 ~ dgamma(0.01, 0.01)
  tau5 ~ dgamma(0.01, 0.01)
  tau6 ~ dgamma(0.01, 0.01)
  tau7 ~ dgamma(0.01, 0.01)
  tau8 ~ dgamma(0.01, 0.01)
  tau.epi ~ dgamma(0.01, 0.01)
  tau.flu ~ dgamma(0.01, 0.01)
  tau.rsv ~ dgamma(0.01, 0.01)
  disp.m ~ dgamma(0.01, 0.01)
```

Clarification for input data

Note: we coded as rsv[i,j,k] and flu[i,j,k] for the purpose of a clear dimension for multiplication. This is a 1 to 1 multiplication, not a matrix multiplication. rsv[i,j,k] are the same for j=1,. . . ,9, representing ICD-9 recorded RSV hospitalizations in children under 2 in each SES group in each month. flu[i,j,k] are the same for j=1,. . . ,9, representing the total ICD-9 recorded flu hospitalizations in the entire age spectrum in each SES group in each month.

In the equation, we simplify them to RSV_{ik} and Flu_{ik}.

Specify input data

```
epi.year <- as.factor(rep(1:9, each=12))
# indicator for calendar year
month <- rep(1:12,9)
# indicator for epi month (1 for July and 12 for June)

dataset <- list('y' = y_income_whole,
# respiratory hospitalizations in each age and
# SES group in each month
"rsv"=rsv_income_whole,
# ICD-9 recorded RSV hospitalizations
# in children under 2 in each SES group in each month
"flu"=flu_income_whole,
# the total ICD-9 recorded flu hospitalizations in
# the entire age spectrum in each SES group in each month
'epi.year'=epi.year,"month"=month,
n.age=9,n.date=108,n.year=9,"n.group"=3)
# number of age/SES groups; number of months and years
```

Posterior sampling

```
jags_post <- jags.model(
  textConnection(model_string),
  data = dataset,n.chains = 3)
# specify the number of chains

update(jags_post, n.iter=82500) # burn-in period

rsv_resp <- coda.samples(jags_post,
                         variable.names=
  c("rd2","lambda","rd1", "rd0","epi","delta"),
                       thin = 10,n.iter = 12500)
# posterior samples
```

Estimate the RSV-attributable hospitalizations in each age and SES group

```
## Remember to check convergence first
## use: gelman.diag() see 3_JAGS_Bayesian_model
## Calculate RSV attributable respiratory
## hospitalization incidence and percent
  post1 <- as.data.frame(as.matrix(rsv_resp[[1]]))
  post2 <- as.data.frame(as.matrix(rsv_resp[[2]]))
  post3 <- as.data.frame(as.matrix(rsv_resp[[3]]))
  post <- bind_rows(post1,post2,post3)
## A total of 3750 posterior samples
  lambda <- post[, grep("lambda", colnames(post), fixed=T)]
  rd2.resp <- post[, grep("rd2[", colnames(post), fixed=T)]

rsv_count <- array(data = NA,dim = c(3750,108,9,3))
for (i in 1:108) {
  for (j in 1:9) {
    for (k in 1:3) {
      rsv_count[,i,j,k] <- rsv_income[i,j,k]*rd2.resp[,j+9*(k-1)
## posterior samples of RSV-attributable hospitalizations
## in each SES and age group in each month
```