Project 2: Birth Year

2025-04-17

The next table summarizes, based on a survey, the number of children that had at least one dose of the Varicella vaccine. It gives the number of vaccinated children (Vaccinated) among the number of children in the survey (Sample Size). The information is provided for 3 regions of the US, and split according to birth cohort (2011-2020).

# Question 1

Derive analytically the posterior of the vaccination coverage per birth year and region. Use a conjugate prior that (1) reflects no knowledge on the vaccination coverage, and (2) reflects that vaccination coverage is typically around 90% or higher. Give posterior summary measures of the vaccination coverage per birth year and region. Is the choice of the prior impacting your results?

## Answer:

The experiment can be model using a binomial likelihood:

Where:

* is the number of vaccinated children (observed data),
* is the sample size (total number of children in the survey),
* is the vaccination rate (probability of a child being vaccinated), which is the parameter we want to estimate.

The conjugate prior of a binomial likelihood is a beta distribution. To reflect no knowledge on the vaccination coverage we can set and parameters of the prior distribution equal to 1. This gives us a uniform prior on , meaning that all values of vaccination rate are equally likely for .

Indeed, given the beta distribution:

for and equal to 1, boils down to 1.

The posterior distribution can be found analytically using the formula:

This is a Beta distribution with updated parameters:

We now substitute and with the corresponding value per birth year and region and calculate the posterior summary measures as follows:

**posterior mean**

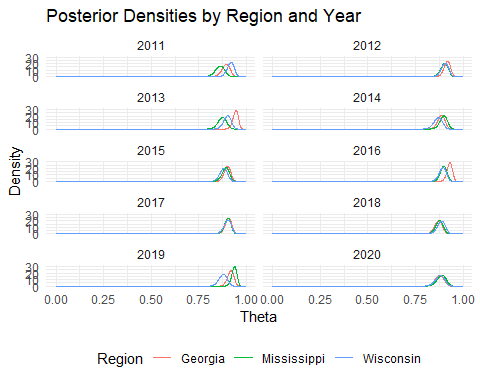
**posterior variance**

**posterior mode** (if and )

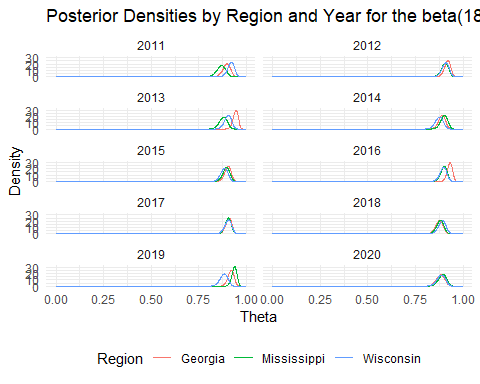
We report the results for the first case (uninformative prior assumption) in table 2. We also plot the posterior densities for each year and each region in figure 1.

We now repeat the same procedure but we assume a prior density that reflects a vaccination rate of 90% or more as most likely. We set and so that the mean and mode of the prior are around 0.9 or more (mean=0.9 and mode=0.944). Results are reported in table 3 and figure 2.

In this second case, since the prior and the likelihood tend to convey similar information, we observe a smaller posterior variance and also a tendency for higher values of posterior mean and median.



Posterior densities by region and year assuming an uninformative prior.



Posterior densities by region and year assuming an expected vaccination rate of 90% or higher.

# Question 2

Investigate whether there is a change in the vaccination coverage over the birth years 2011-2019 using a logistic regression model:

with

where:

* is the location,
* is birth cohort,
* is the vaccination coverage.

Assume non-informative priors for the parameters to be estimated. Write and explain the code in BUGS language

## Answer

The code written in BUGS language is provided below. First we specify the model structure. For each region and year cohort we ask to calculate the binomial likelihood by using a loop function (see chunk below).

for (i in 1:N\_region) { # Loop over regions  
for (j in 1:N\_year) { # Loop over years (cohorts)  
Y[i, j] ~ dbin(pi[i, j], N[i, j]) # Likelihood for region i and year j

Then we specify the logistic function. As it can be seen below we did not index the beta coefficients. This way only one intercept and one beta coefficient for the effect of year of birth will be calculated for all regions as requested in the question.

logit(pi[i, j]) <- beta0 + beta1 \* BirthYear[j] # same beta0, beta1 for all regions

Then a non informative prior is specified. Since we are working with the logit of the vaccination rate, we cannot use a beta distribution as in question one, since this would be bounded between 0 and 1. The support for the logit of the vaccination rate is indeed . Therefore we use a normal distribution centered around zero but with very high variance. In BUGS language this means low precision, hence the code below.

# Non-informative priors for intercept and slope (shared across regions)  
 beta0 ~ dnorm(0, 0.0001)  
 beta1 ~ dnorm(0, 0.0001)

The rest of the code specifies the matrix to be used as data input and finally the model run commands. We used a burn-in of 500 (meaning that the first 500 samples are discarded), thinning equal to 2 (meaning only every other sample are retained), and then three chains are run. Since BUGS is a declarative language, we have to explicitly tell what the model structure is, then the software will automatically choose the Markov Chain Monte Carlo algorithm (by default Gibbs sampling).

# Run the model  
fit <- jags(  
 data = bugs\_data,  
 parameters.to.save = params,  
 model.file = "logistic\_model.bug",  
 n.chains = 3,  
 n.iter = 5000,  
 n.burnin = 500,  
 n.thin = 2  
)

Overall, the full code is:

# Model structure assuming one intercept and slop for all regions  
model\_structure <- "   
model {  
 for (i in 1:N\_region) { # Loop over regions  
 for (j in 1:N\_year) { # Loop over years (cohorts)  
 Y[i, j] ~ dbin(pi[i, j], N[i, j]) # Likelihood for region i and year j  
 logit(pi[i, j]) <- beta0 + beta1 \* BirthYear[j] # same beta0, beta1 for all regions  
 }  
 }  
  
 # Non-informative priors for intercept and slope (shared across regions)  
 beta0 ~ dnorm(0, 0.0001)  
 beta1 ~ dnorm(0, 0.0001)  
}  
"  
# Save the model structure in a text file  
writeLines(model\_structure, "logistic\_model\_Q.2.bug")  
  
# Prepare the matrix for Y (vaccinated) and N (sample size)   
# by region and year of birth  
vacc\_data <- read.csv("vaccination\_data.csv")  
  
Y <- matrix(vacc\_data$Vaccinated,  
 nrow=length(unique(vacc\_data$Geography)), byrow=TRUE)  
  
N <- matrix(vacc\_data$Sample.Size,  
 nrow=length(unique(vacc\_data$Geography)), byrow=TRUE)  
  
  
# Define the rows as regions and the columns as years  
row.names(Y) <- unique(vacc\_data$Geography)  
colnames(Y) <- min(vacc\_data$Birth.Year):max(vacc\_data$Birth.Year)  
  
row.names(N) <- unique(vacc\_data$Geography)  
colnames(N) <- min(vacc\_data$Birth.Year):max(vacc\_data$Birth.Year)  
  
bugs\_data <- list(  
 Y = Y,  
 N = N,  
 BirthYear = vacc\_data$Birth.Year,  
 N\_region = length(unique(vacc\_data$Geography)),  
 N\_year = length(unique(vacc\_data$Birth.Year))  
)  
  
# Parameters to monitor  
params <- c("beta0", "beta1")  
  
# Run the model  
regression\_Q2 <- jags(  
 data = bugs\_data,  
 parameters.to.save = params,  
 model.file = "logistic\_model\_Q.2.bug",  
 n.chains = 3,  
 n.iter = 5000,  
 n.burnin = 500,  
 n.thin = 2  
)

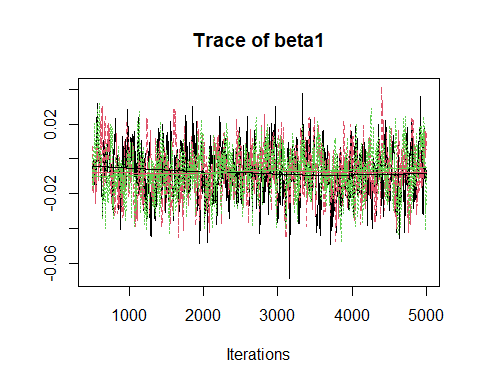
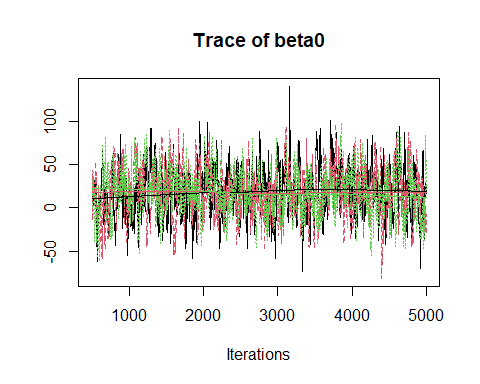
## Compiling model graph  
## Resolving undeclared variables  
## Allocating nodes  
## Graph information:  
## Observed stochastic nodes: 30  
## Unobserved stochastic nodes: 2  
## Total graph size: 126  
##   
## Initializing model

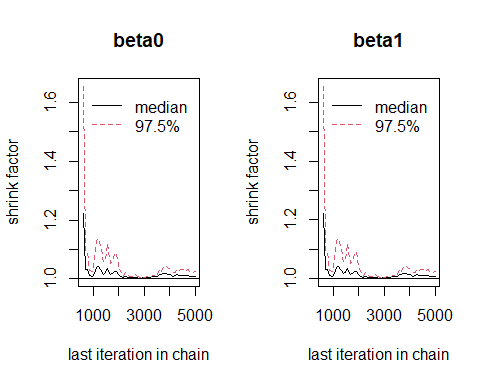
# Question 3

Run the MCMC method and check convergence of the MCMC chains. Give the details on how you checked convergence.

## Answer

Looking at the trace plots below, for both and we can see that the chains jump around the same mean and visit different areas of the parameter space. No drift is apparent. All chains seem to oscillate within a similar range of values. A visual check favours a good convergence of the model. We then present Gelman-Rubin diagnostic plots which compare the variance within each Markov chain to the variance between multiple chains.





Gelman plots for the logistic model in question 2

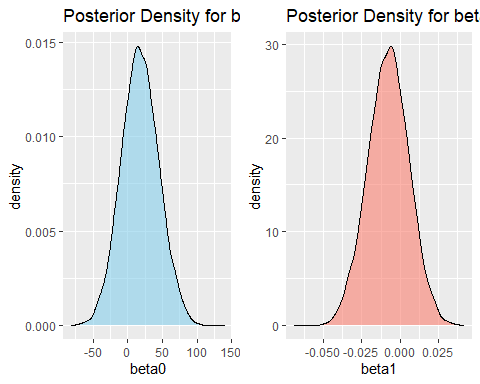
The plot reports on the axis the variance between multiple chains and traces its reduction as the iterations on the axis increase. In essence, a decrease towards a value of 1 is a sign of convergence. In our case, for both and we see a quick drop of the shrink factor (variance within and between the chains for each parameter) before 1000 iterations and the shrink factor stabilizes around 1 thereafter particularly after 4000 iterations when also the 97.5th percentile of the shrink factor seems to be close to 1. This can be interpreted as a good convergence.

# Question 4

Make a plot of the posterior densities and give summary measures of the posterior distributions of the model parameters. Interpret the results.

## Answer

The posterior density for should captures the baseline vaccination coverage when year of birth is zero. Therefore, here it does not have a direct interpretation in terms of vaccination coverage, but it is still necessary for the model. Looking at posterior density we see that is centered very close to 0, and its mass spans both positive and negative values. This means that the effect of year of birth is limited and there is high uncertainty regarding its estimate. Posterior summary measures for and are provided in table 3. If we then apply the inverse logit transformation to the linear predictor we can obtain the estimated vaccination coverage per year by using as and their mean. Data are reported in the table 4 below and indeed reflect an approximate vaccination around 90%.



Posterior density plots for the logistic model in question 2

# Question 5

Give the posterior estimate of the vaccination coverage per birth year. Compare with the analytically results you obtained in Question 1.

## Answer

Posterior estimate of the vaccination coverage per birth year calculated in question 1 assuming a non informative prior are reported below side by side with those estimated in question 4. Since in question 1 we actually had the break-down of the coverage per year and per region we averaged the yearly data over the three regions for an easier comparison. The differences are small and mostly limited to approximately one percentage point.

# Question 6

Secondly, investigate whether the vaccination coverage trends are distinct at the different locations by adding a location-specific intercept and slope:

Use data from the years 2011-2019. Assume non-informative priors for the parameters to be estimated. Write the code in BUGS language. Give a brief summary of the convergence checks you performed. Give the posterior estimates of this model.

## Answer

To evaluate whether trends in vaccination coverage differ across regions, we use a hierarchical logistic regression model. The number of vaccinated children in each region and birth year is assumed to follow a Binomial distribution with region-specific probabilities.

Where:

* is the number of vaccinated children in region and year ,
* is the number of children surveyed in region and year ,
* is the probability of being vaccinated,
* is the region-specific intercept, capturing baseline coverage,
* is the region-specific slope, capturing the change in coverage over time.

To reflect minimal prior knowledge, we use vague, non-informative priors for the intercepts and slopes:

This hierarchical model structure allows each region to have its own baseline coverage and trend while still sharing the same model form.

We implemented this model in **JAGS** using three MCMC chains with 5000 iterations, a burn-in of 500, and thinning of 2. Convergence was assessed using **trace plots** and **Gelman-Rubin diagnostics**, confirming good mixing and for all parameters.

model\_structure <- "  
model {  
 for (i in 1:N\_region) { # number of regions  
 for (j in 1:N\_year) { # number of year cohorts  
 Y[i, j] ~ dbin(pi[i, j], N[i, j]) # likelihood  
 logit(pi[i, j]) <- beta0[i] + beta1[i] \* BirthYear[j] # regression  
 }  
  
 beta0[i] ~ dnorm(0, 0.001)  
 beta1[i] ~ dnorm(0, 0.001)  
 }  
}  
"

## Compiling model graph  
## Resolving undeclared variables  
## Allocating nodes  
## Graph information:  
## Observed stochastic nodes: 27  
## Unobserved stochastic nodes: 6  
## Total graph size: 154  
##   
## Initializing model

# Question 7

What is the probability (a posteriori) that there is an increase in vaccination coverage (per location)?

## Answer

## Interpretation

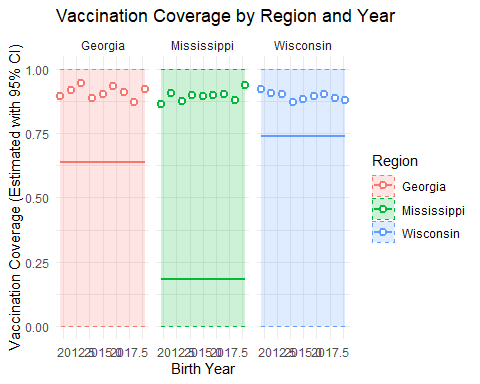
The posterior probabilities represent the likelihood that vaccination coverage is increasing in each region, based on the posterior distribution of the slope . A probability close to 1 suggests strong evidence of a positive trend over time, while values near 0.5 reflect uncertainty or no clear directional change. If a region exhibits a posterior probability above 0.95, it provides strong Bayesian evidence for an increase in vaccination coverage. On the other hand, probabilities near or below 0.5 may indicate stability or even a potential decline.

# Question 8

Make a plot of the estimated vaccination coverage (per location and birth year), including the uncertainty on the estimates. Include also the observed vaccination proportion in the plot.

## Answer

The following plot shows the estimated vaccination coverage by region and birth year, along with 95% credible intervals for the model estimates. The shaded regions represent uncertainty around the predicted coverage. Observed vaccination proportions are overlaid as white points with outlines. If the observed data points fall within the shaded bands, this indicates good model fit. Differences in slope and uncertainty width across regions reflect regional variation in trends and data size.



# Question 9

Investigate whether the observed number of vaccinated children in 2020 is in line with the expectations from earlier years. For this, compare the observed number of vaccinated children in 2020 with the prediction intervals for number of vaccinated children in 2020.

## Answer

# Question 10

Make pairwise comparisons of the vaccination coverage in 2019 by estimating the ratio of the vaccination coverage in 2019 in two locations. Interpret the results.

## Answer

# Question 10

Make pairwise comparisons of the vaccination coverage in 2019 by estimating the ratio of the vaccination coverage in 2019 in two locations. Interpret the results.

## Answer