Project 2: Birth Year

Sofiya Antonyuk
Edoardo Pennesi
Dorothy Chepkoech
Andreea Badache
Steve Omollo

2025-04-15

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).

Table 1: Vaccination Data

Geography	Birth.Year	Vaccinated	Sample.Size
Georgia	2011	196	219
Georgia	2012	248	270
Georgia	2013	261	276
Georgia	2014	252	284
Georgia	2015	276	306
Georgia	2016	311	334
Georgia	2017	265	292
Georgia	2018	246	282
Georgia	2019	251	273
Georgia	2020	165	188
Wisconsin	2011	207	225
Wisconsin	2012	205	226
Wisconsin	2013	212	235
Wisconsin	2014	195	224
Wisconsin	2015	231	262
Wisconsin	2016	246	275
Wisconsin	2017	215	238
Wisconsin	2018	214	241
Wisconsin	2019	197	224
Wisconsin	2020	156	177
Mississippi	2011	171	198
Mississippi	2012	208	230
Mississippi	2013	190	217
Mississippi	2014	215	239
Mississippi	2015	243	272
Mississippi	2016	276	307
Mississippi	2017	290	321
Mississippi	2018	242	276
Mississippi	2019	304	324
Mississippi	2020	161	181

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:

$$P(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Where:

- x is the number of vaccinated children (observed data),
- n is the sample size (total number of children in the survey),
- θ 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 α and β parameters of the *beta* prior distribution equal to 1. This gives us a uniform prior on θ, meaning that all values of vaccination rate are equally likely for $\theta \in [0,1]$.

Indeed, given the beta distribution:

$$f(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1} 1 - \theta^{\beta-1}}{B(\alpha, \beta)}$$

for α and β equal to 1, $f(\theta; \alpha, \beta)$ boils down to 1.

The posterior distribution can be found analytically using the formula:

$$P(\theta|x, n) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

This is a Beta distribution with updated parameters:

$$P(\theta|x, n) \sim \text{Beta}(x+\alpha, n-x+\beta)$$

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

posterior mean

$$\text{Posterior Mean} = \frac{\alpha_{\text{posterior}}}{\alpha_{\text{posterior}} + \beta_{\text{posterior}}}$$

posterior variance

$$\text{Posterior Variance} = \frac{\alpha_{\text{posterior}} \cdot \beta_{\text{posterior}}}{(\alpha_{\text{posterior}} + \beta_{\text{posterior}})^2(\alpha_{\text{posterior}} + \beta_{\text{posterior}} + 1)}$$

posterior mode (if $\alpha_{\text{posterior}} > 1$ and $\beta_{\text{posterior}} > 1$)

$$\text{Posterior Mode} = \frac{\alpha_{\text{posterior}} - 1}{\alpha_{\text{posterior}} + \beta_{\text{posterior}} - 2}$$

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 $\alpha = 18$ and $\beta = 2$ 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.

Table 2: Posterior summary measurs for the beta(1,1) case

Geography	Birth.Year	posterior_mean	posterior_variance	posterior_mode
Georgia	2011	0.8914027	0.0004361	0.8949772
Georgia	2012	0.9154412	0.0002835	0.9185185
Georgia	2013	0.9424460	0.0001944	0.9456522
Georgia	2014	0.8846154	0.0003556	0.8873239
Georgia	2015	0.8993506	0.0002929	0.9019608
Georgia	2016	0.9285714	0.0001968	0.9311377
Georgia	2017	0.9047619	0.0002921	0.9075342
Georgia	2018	0.8697183	0.0003976	0.8723404
Georgia	2019	0.9163636	0.0002777	0.9194139
Georgia	2020	0.8736842	0.0005778	0.8776596
Wisconsin	2011	0.9162996	0.0003364	0.9200000
Wisconsin	2012	0.9035088	0.0003807	0.9070796
Wisconsin	2013	0.8987342	0.0003824	0.9021277
Wisconsin	2014	0.8672566	0.0005071	0.8705357
Wisconsin	2015	0.8787879	0.0004020	0.8816794
Wisconsin	2016	0.8916968	0.0003474	0.8945455
Wisconsin	2017	0.9000000	0.0003734	0.9033613
Wisconsin	2018	0.8847737	0.0004178	0.8879668
Wisconsin	2019	0.8761062	0.0004782	0.8794643
Wisconsin	2020	0.8770950	0.0005989	0.8813559
Mississippi	2011	0.8600000	0.0005990	0.8636364
Mississippi	2012	0.9008621	0.0003833	0.9043478
Mississippi	2013	0.8721461	0.0005069	0.8755760
Mississippi	2014	0.8962656	0.0003842	0.8995816
Mississippi	2015	0.8905109	0.0003545	0.8933824
Mississippi	2016	0.8964401	0.0002995	0.8990228
Mississippi	2017	0.9009288	0.0002755	0.9034268
Mississippi	2018	0.8741007	0.0003944	0.8768116
Mississippi	2019	0.9355828	0.0001843	0.9382716
Mississippi	2020	0.8852459	0.0005521	0.8895028

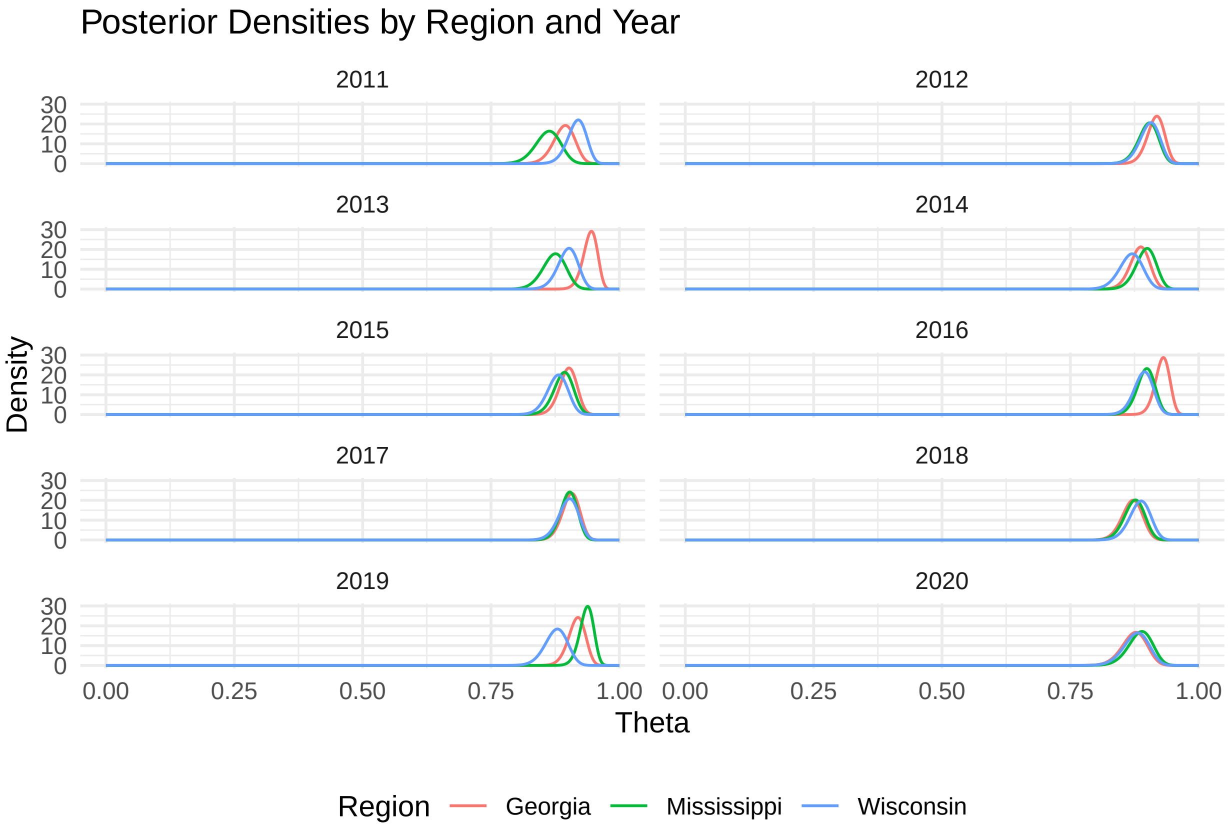


Figure 1: Posterior densities by region and year assuming an uninformative prior.

Table 3: Posterior summary measurs for the beta(18,2) case

Geography	Birth.Year	posterior_mean	posterior_variance	posterior_mode
Georgia	2011	0.8953975	0.0003903	0.8987342
Georgia	2012	0.9172414	0.0002609	0.9201389
Georgia	2013	0.9425676	0.0001823	0.9455782
Georgia	2014	0.8881579	0.0003257	0.8907285
Georgia	2015	0.9018405	0.0002707	0.9043210
Georgia	2016	0.9293785	0.0001849	0.9318182
Georgia	2017	0.9070513	0.0002694	0.9096774
Georgia	2018	0.8741722	0.0003630	0.8766667
Georgia	2019	0.9180887	0.0002558	0.9209622
Georgia	2020	0.8798077	0.0005060	0.8834951
Wisconsin	2011	0.9183673	0.0003048	0.9218107
Wisconsin	2012	0.9065041	0.0003431	0.9098361
Wisconsin	2013	0.9019608	0.0003454	0.9051383
Wisconsin	2014	0.8729508	0.0004527	0.8760331
Wisconsin	2015	0.8829787	0.0003651	0.8857143
Wisconsin	2016	0.8949153	0.0003177	0.8976109
Wisconsin	2017	0.9031008	0.0003379	0.9062500
Wisconsin	2018	0.8888889	0.0003770	0.8918919
Wisconsin	2019	0.8811475	0.0004275	0.8842975
Wisconsin	2020	0.8832487	0.0005208	0.8871795
Mississippi	2011	0.8669725	0.0005266	0.8703704
Mississippi	2012	0.9040000	0.0003458	0.9072581
Mississippi	2013	0.8776371	0.0004512	0.8808511
Mississippi	2014	0.8996139	0.0003473	0.9027237
Mississippi	2015	0.8938356	0.0003239	0.8965517
Mississippi	2016	0.8990826	0.0002766	0.9015385
Mississippi	2017	0.9032258	0.0002556	0.9056047
Mississippi	2018	0.8783784	0.0003597	0.8809524
Mississippi	2019	0.9360465	0.0001735	0.9385965
Mississippi	2020	0.8905473	0.0004825	0.8944724

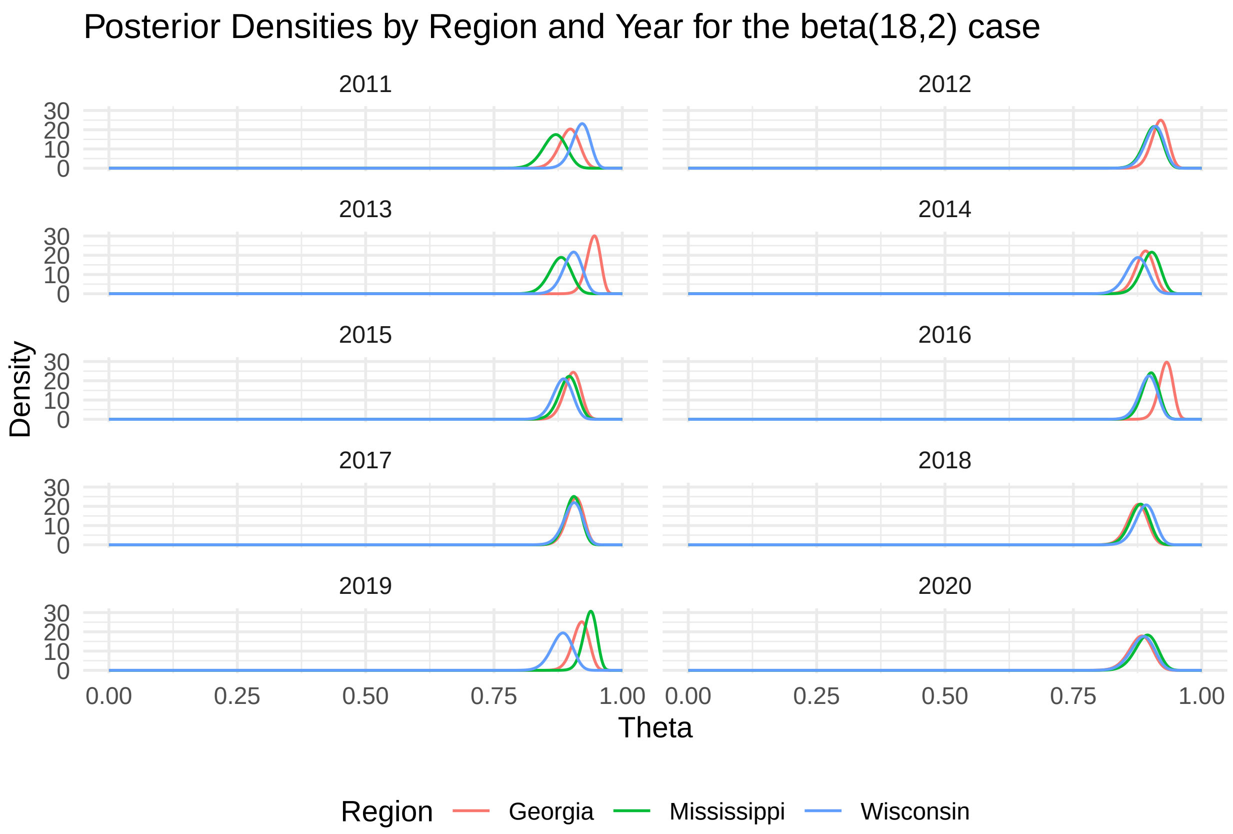


Figure 2: Posterior densities by region and year assuming an expected vaccination rate of 90% or higher.

Question 2

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

$$Y_{ij} \sim Binom(\pi_{ij}, N_{ij})$$

with

$$\text{logit}(\pi_{ij}) = \beta_{0i} + \beta_{1i} \cdot \text{BirthYear}_j$$
$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

where:

- i is the location,
- j is birth cohort,
- π_{ij} 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 $(-\infty, \infty)$. 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)),
```

```r
  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
)
```

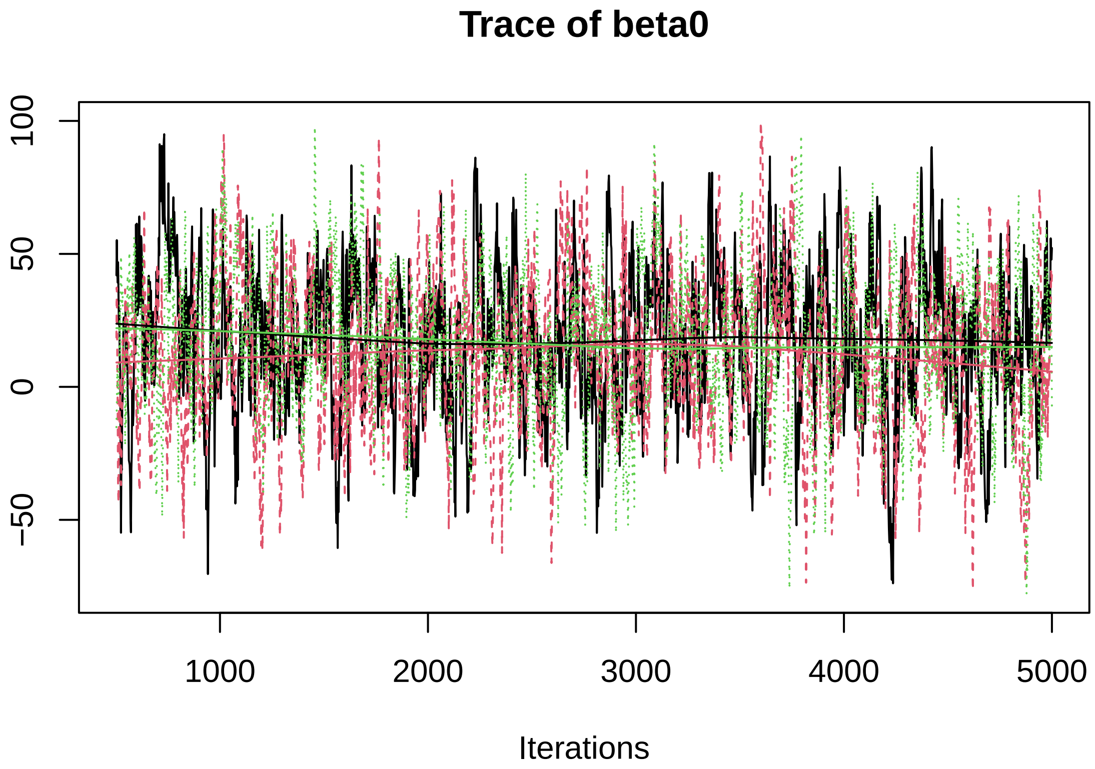
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 30
##    Unobserved stochastic nodes: 2
##    Total graph size: 126
##
## Initializing model
```

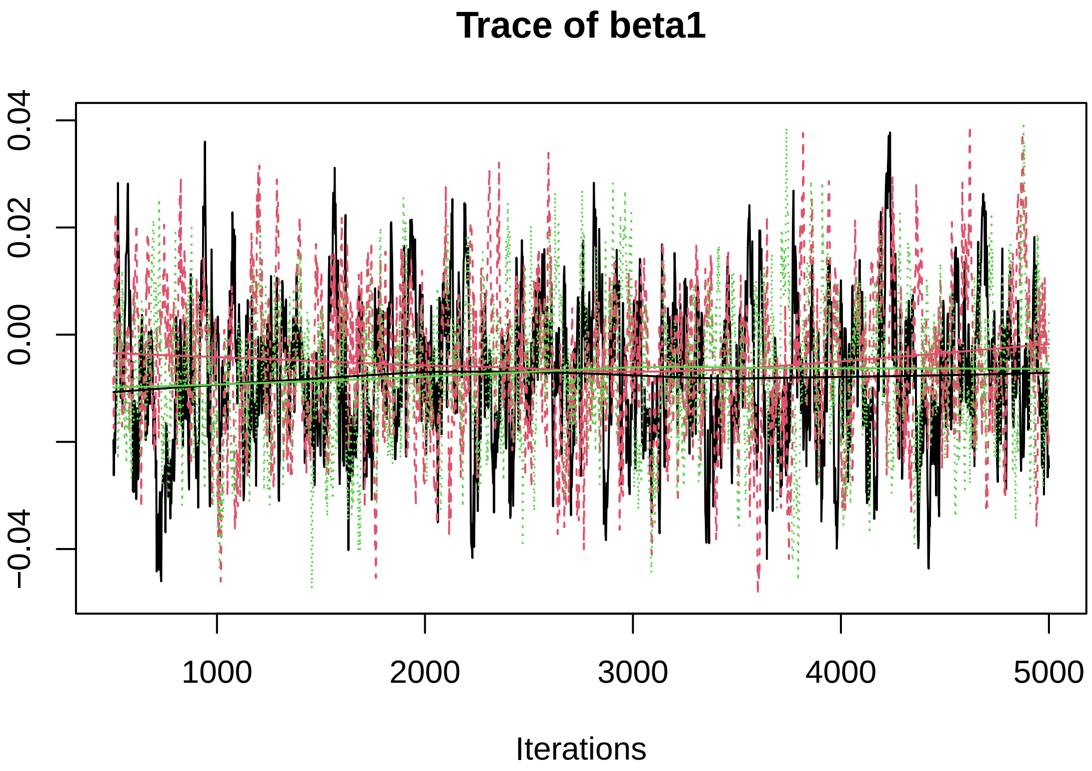
Question 3

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 β_0 and β_1 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.





The plot reports on the y axis the variance between multiple chains and traces its reduction as the iterations on the x axis increase. In essence, a decrease towards a value of 1 is a sign of convergence. In our case, for both β_0 and β_1 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.

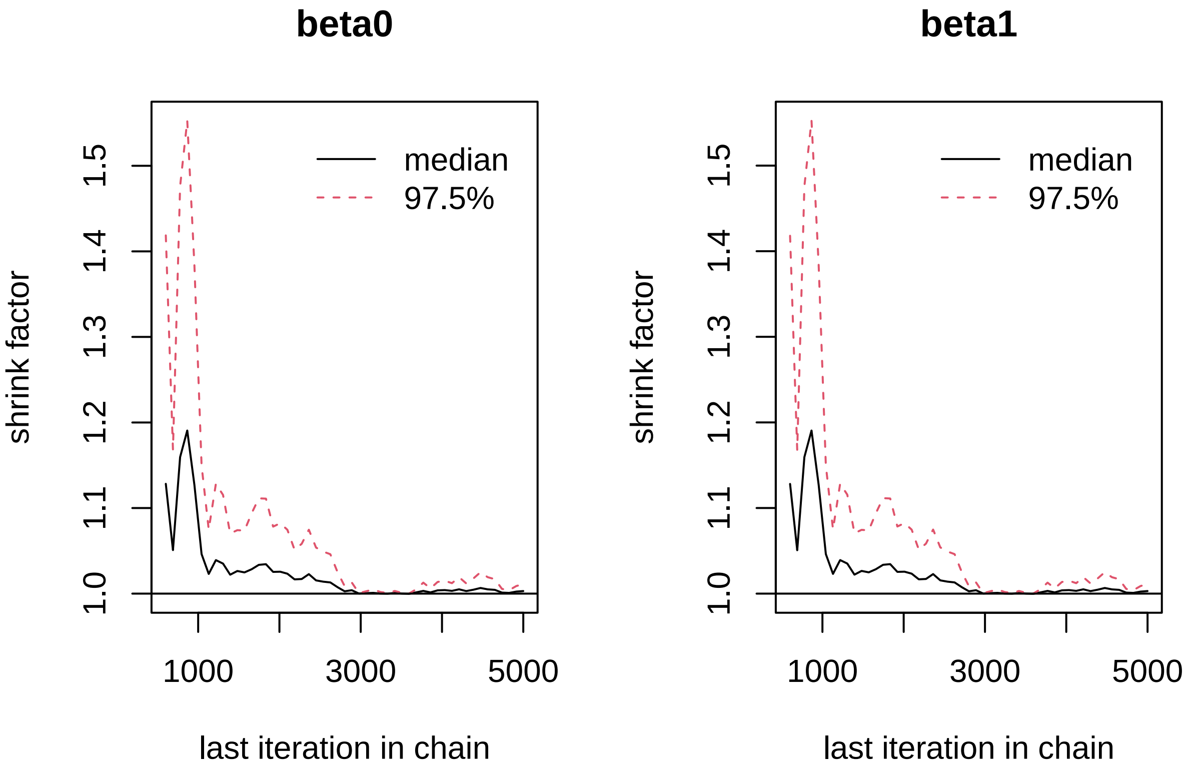


Figure 3: Gelman plots for the logistic model in question 2

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 β_0 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 $\beta_1$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 β_0 and β_1 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 β_0 and β_1 their mean. Data are reported in the table 4 below and indeed reflect an approximate vaccination around 90%.

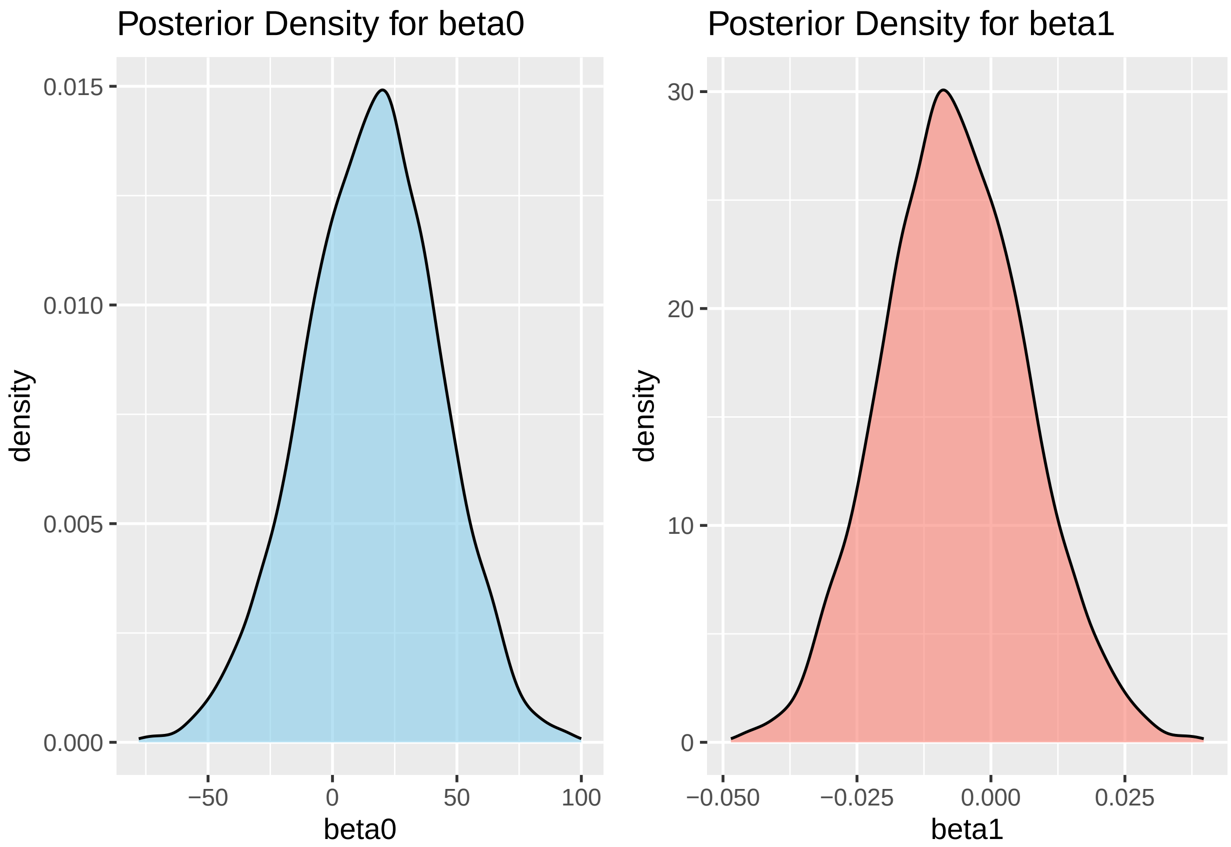


Figure 4: Posterior density plots for the logistic model in question 2

Table 4: Posterior summary measures for baysan regression model in Q.2

Parameter	Mean	Median	SD	Median Abs.Dev	5th Percentile	95th Percentile
beta0	15.5844	16.3513	27.0670	27.2437	-30.1900	59.9502
beta1	-0.0066	-0.0070	0.0134	0.0135	-0.0286	0.0161
deviance	186.3918	185.8367	1.8680	1.4039	184.5727	190.1744

Table 5: Vaccination coverage estimates per year based on model in Q2

Coverage	Year of birth
2011	90.22
2012	90.16
2013	90.11
2014	90.05
2015	89.99
2016	89.93
2017	89.87
2018	89.81
2019	89.74
2020	89.68

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.

Table 6: Vaccination coverage estimates per year based on model in Q1 and Q5

Yea of birth	estimate from Q1	estimate from Q5	difference (Q1-Q5)
2011	88.92	90.22	-1.30
2012	90.66	90.16	0.50
2013	90.44	90.11	0.33
2014	88.27	90.05	-1.78
2015	88.95	89.99	-1.04
2016	90.56	89.93	0.63
2017	90.19	89.87	0.32
2018	87.62	89.81	-2.19
2019	90.94	89.74	1.20
2020	87.87	89.68	-1.81

Question 6

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

$$\text{logit}(\pi_{ij}) = \beta_{0i} + \beta_{1i} \cdot \text{BirthYear}_j$$

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.

```
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 for a given year j and region i
    logit(pi[i,j])<- beta0[i] + beta1[i] * BirthYear[j] # regression function
    }

    # Non-informative priors for intercept and slope
    beta0[i] ~ dnorm(0, 0.001)
    beta1[i] ~ dnorm(0, 0.001)
  }

}
"
```

Question 7

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

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.

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.

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.