# Lab Assignment 3
# PB HLTH 250C: Advanced Epidemiologic Methods

**Katherine Rose Wolf**

**March 10, 2020**

# Questions

## Question One

**Using the R code provided, complete Table 1 using the posterior samples of the odds ratios.** ***(20 points)***

Table 1: Posterior median and 95% credible intervals for odds ratios from logistic regression model of overweight status on smoking, controlling for age, sex, and education level.

| Variable | Vague prior | Informative Prior 1[a] | Informative Prior 2[b] |
|---|---|---|---|
| Current smoker (versus not) | 0.4907 (0.4294, 0.5580) | 0.4879 (0.4299, 0.5552) | 0.6231 (0.5531, 0.7028) |
| Age (per year increase) | 1.1764 (1.1028, 1.2583) | 1.1784 (1.1007, 1.2571) | 1.2028 (1.1257, 1.2822) |
| Male sex (versus female) | 2.1733 (1.9103, 2.4778) | 2.1812 (1.9086, 2.4832) | 2.0469 (1.7944, 2.3466) |
| High school education (versus < high school education) | 0.6474 (0.5566, 0.7474) | 0.6470 (0.5517, 0.7564) | 0.6462 (0.5545, 0.7544) |
| Some college (versus < high school education) | 0.5339 (0.4442, 0.6434) | 0.5348 (0.4431, 0.6438) | 0.5404 (0.4473, 0.6459) |
| College plus (versus < high school education) | 0.5459 (0.4398, 0.6721) | 0.5439 (0.4411, 0.6744) | 0.5568 (0.4501, 0.6825) |

[a] Prior mean for OR of current smoking = 2, prior variance = 1000.

[b] Prior mean for OR of current smoking = 2, prior variance = 0.02. (I believe that the prior variance originally listed on the assignment, 0.08, was an error, and that it arose from a given prior interval for Informative Prior 2 of (1, 3), possibly from a prior version of this assignment, instead of (1.5, 2.67). Proof: $(\log(3) - \log(1))/(2 * 1.96) = 0.079$ whereas $(\log(2.67) - \log(1.5))/(2 * 1.96) = 0.022$.

## Question Two

**Using the parameterization for Informative Prior 1 (IP1), calculate the prior 95% interval for the smoking OR. *Hint: Calculate the interval on the scale of the log-OR ($\beta$) and transform the limits.* In *one or two sentences* describe how this compares to the prior interval for Informative Prior 2 (IP2) stated in the instructions above. *(10 points)***

*Calculations*

Let $\beta_s$ denote the normal prior for the log odds ratio comparing the odds (risk) of overweight (body mass index > 25) in a smoker to that in a nonsmoker. The parameterization for $\beta_s$ given for IP1 states that $\beta_s$ is normally distributed with hyperparameters mean $\mu_s$ and variance $\sigma_s^2$, i.e., $\beta_s \sim N(\mu_s, \sigma_s^2)$, such that the odds ratio $e^{\beta_s}$, or the natural exponentiation of the mean of the log-OR, is 2, i.e., $e^{E[\beta_s]} = e^{\mu_s} = 2$, and $\beta_s$ has a variance of 1000, i.e., $\sigma_s^2 = 1000$.

To get the mean of the log-OR, then, we take the natural logarithm of the natural exponentiation of the mean of the log-OR, i.e., $\mu_s = \log e^{\mu_s} = \log(2)$.

To get the standard deviation of $\beta_s$, $\sigma_s$, we take the square root of the variance $\sigma_s^2$, i.e., $\sigma_s = \sqrt{\sigma_s^2} = \sqrt{1000}$.

Then we can calculate the prior 95% interval for the log-OR by taking 1.96 standard deviations above and below its mean:

- Lower bound on prior 95% interval for $\beta_s$: $\mu_s - 1.96\sigma_s = \log(2) - 1.96\sqrt{1000} = -61.287$
- Upper bound on prior 95% interval for $\beta_s$: $\mu_s + 1.96\sigma_s = \log(2) + 1.96\sqrt{1000} = 62.674$

To get the prior 95% interval for the OR, $e^{\beta_s}$, then, we exponentiate the prior 95% interval for $\beta_s$:

- Lower bound on prior 95% interval for $e^{\beta_s}$: $e^{\mu_s - 1.96\sigma_s} = e^{\log(2) - 1.96\sqrt{1000}} = 2.4165 \times 10^{-27}$
- Upper bound on prior 95% interval for $e^{\beta_s}$: $e^{\mu_s + 1.96\sigma_s} = e^{\log(2) + 1.96\sqrt{1000}} = 1.6553 \times 10^{27}$

Thus the prior 95% interval for IP1 is $(2.4165 \times 10^{-27}, 1.6553 \times 10^{27})$.

*Comparisons*

The prior 95% interval drawn from IP1 is $(2.4165 \times 10^{-27}, 1.6553 \times 10^{27})$, corresponding to a variance of 1000, whereas IP2 assumes a prior 95% interval of $(1.5, 2.67)$, which, under the assumption of a normal distribution, yields a variance of only $(\log(2.67) - \log(1.5))/(2 * 1.96) = 0.0216$. The drastic difference in variances determines a drastic difference in the size of the prior intervals for which we are 95% confident that we have captured the true parameter: the wide, flat probability distribution of IP1 assigns almost equal proabability of the true parameter value appearing among a wide range of values, whereas IP2's narrow, tall probability distribution assigns much higher probabilities of the true value appearing the closer one gets to its mean.

## Question Three

**What seems to be more influential on the smoking effect, Informative Prior 1 or Informative Prior 2? In *one sentence*, briefly explain what you think is happening? *(5 points)***

*Statement*

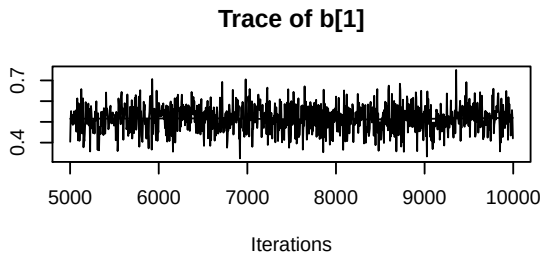
The model run using IP2 gives an estimate of the posterior median closer to 2 but a slightly larger confidence interval around it, 0.6231 (0.5531, 0.7028), than the model run using IP1, 0.4879 (0.4299, 0.5552), showing that IP2 is more influential on the smoking effect than IP1.

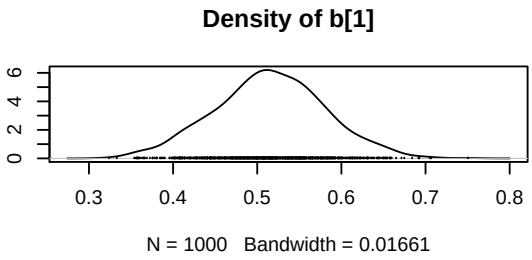
*Explanation*

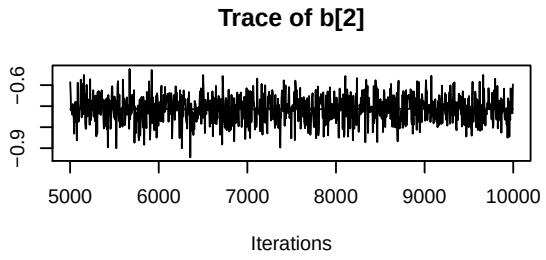
IP2 is more influential because its distribution has a high peak at and assigns high probabilities to values close to its mean (95% of the probability is between 1.5 and 2.67!), which pulls the posterior median toward it, whereas IP1's much more widely distributed probability density assigns almost equal probability to the existence of the true mean anywhere in a big range from $2.4165 \times 10^{-27}$ to $1.6553 \times 10^{27}$, allowing the new data more latitude to assign the peak(s) and thus the posterior median.

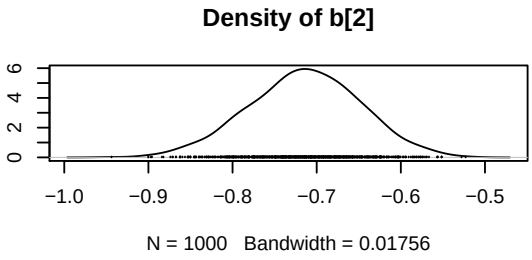
## Question Four

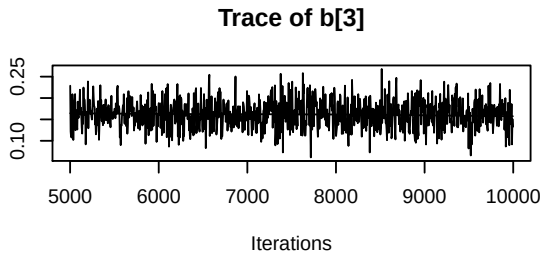
**Using the trace plots, density plots and autocorrelation plots (focus on 1st chain) from the diagnostics for the first model ("Vague prior"), briefly describe any evidence of convergence (or lack of convergence) that you see. Attach these plots (2 pages for trace/density plots; 1 page for autocorrelation plots).** ***(10 points)***

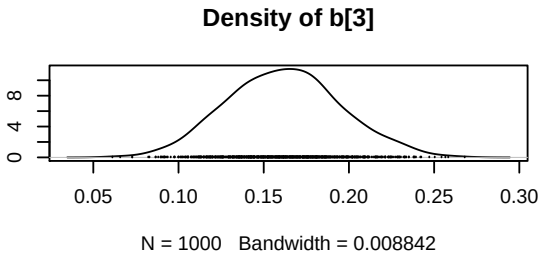


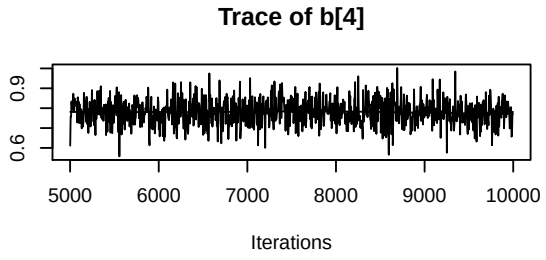



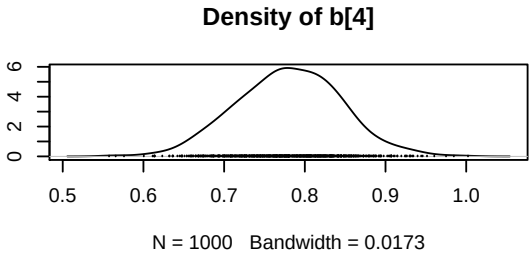



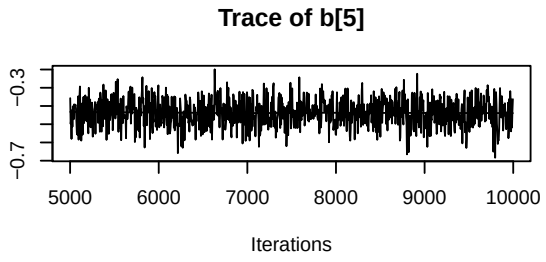



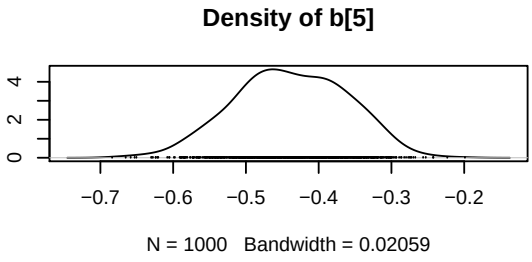



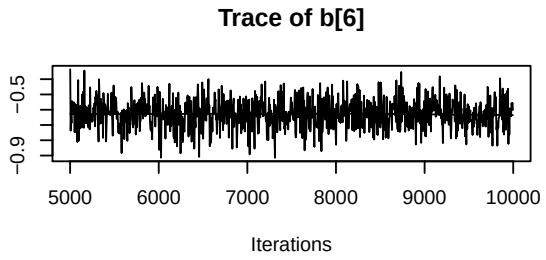



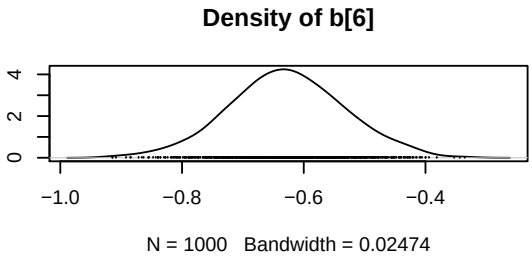



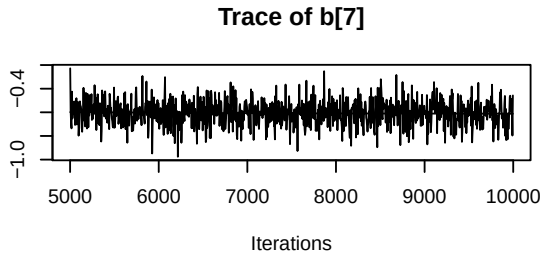



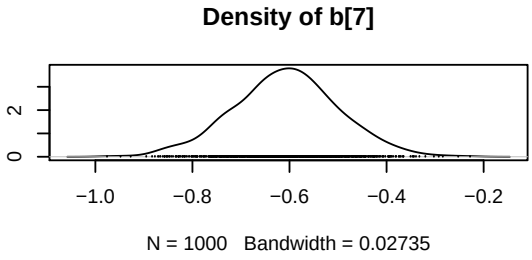

Trace of b[5]
−0.3
−0.7
5000
6000
7000
8000
9000
10000
Iterations

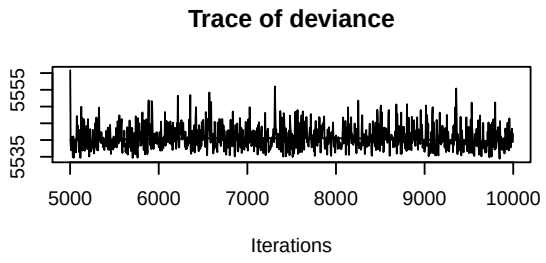

Density of b[5]
0
2
4
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
N = 1000 Bandwidth = 0.02059

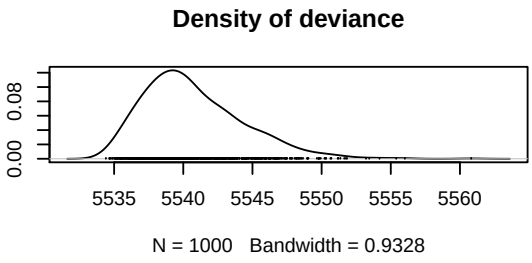

Trace of b[6]
−0.5
−0.9
5000
6000
7000
8000
9000
10000
Iterations


Density of b[6]
0
2
4
−1.0
−0.8
−0.6
−0.4
N = 1000 Bandwidth = 0.02474

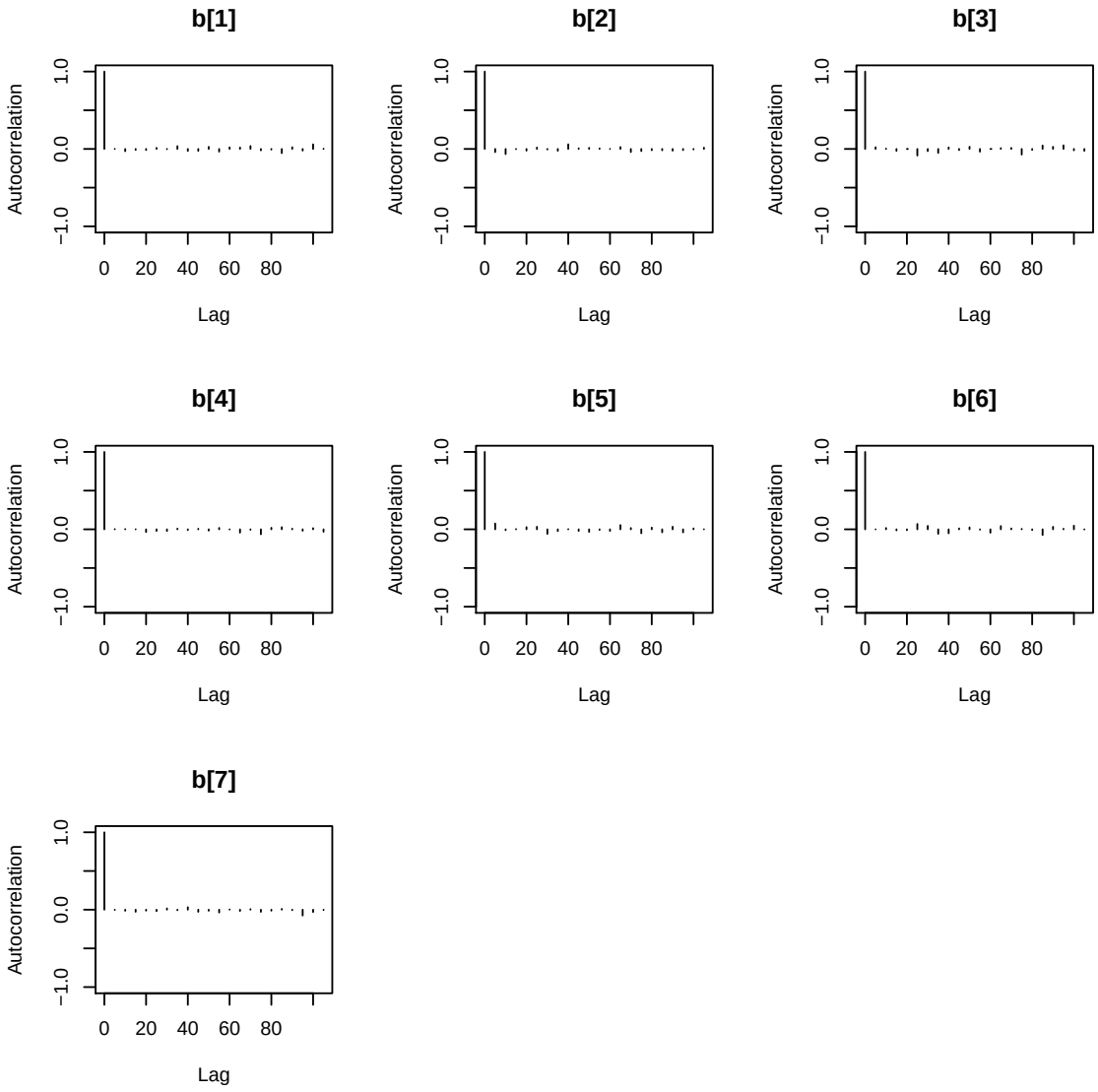

Trace of b[7]
−0.4
−1.0
5000
6000
7000
8000
9000
10000
Iterations


Density of b[7]
0
2
−1.0
−0.8
−0.6
−0.4
−0.2
N = 1000 Bandwidth = 0.02735


Trace of deviance
5555
5535
5000
6000
7000
8000
9000
10000
Iterations


Density of deviance
0.08
0.00
5535
5540
5545
5550
5555
5560
N = 1000 Bandwidth = 0.9328

## Chain 1



## Question Five

**From the results of the Geweke test, is there evidence for lack of convergence? Justify your answer.** ***(5 points)***

The Geweke test tests whether the Markov Chain is constant between early and later parts of the sequence of numbers by comparing subsamples of the random samples for each parameter, here comparing the first 10% of the chain to the last half. It outputs the z-statistic and indicates non-convergence is indicated if the test-statistic is > 1.96 in absolute value.[1]

The Geweke test results here are:

```
## [[1]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##    b[1]    b[2]    b[3]    b[4]    b[5]    b[6]    b[7]
## -1.0951  0.1554  0.8022  0.5155 -0.1102  1.1358  1.7095
##
##
## [[2]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##    b[1]    b[2]    b[3]    b[4]    b[5]    b[6]    b[7]
## -1.2630  0.9628  1.0084  0.1967  2.0785  1.7122  1.0630
##
##
## [[3]]
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##    b[1]    b[2]    b[3]    b[4]    b[5]    b[6]    b[7]
##  0.6734  0.1043  0.3256  0.2065 -0.7789 -1.1426 -1.2885
```

[1] Lecture 5, slide 46.

# R code

```
knitr::opts_chunk$set(echo = FALSE,
                      warning = FALSE,
                      message = FALSE,
                      results = FALSE)


library(knitr)
library(R2jags)
library(coda)
library(foreign)

load("frmgham_recoded_three.Rdata")


# Extract data elements from data frame
bmi <- frmgham_recoded$bmi
overweight <- as.integer(bmi >= 25)
cursmoke <- frmgham_recoded$cursmoke
age.c <- as.numeric(scale(frmgham_recoded$age))
male <- as.integer(frmgham_recoded$sex == 1)

# Create education indicators (a shortcut using the model.matrix command)
X.educ <- model.matrix(~-1 + factor(educ),
                       data=frmgham_recoded)
educ1 <- X.educ[,1]
educ2 <- X.educ[,2]
educ3 <- X.educ[,3]
educ4 <- X.educ[,4]


# JAGS code for the posterior distribution:
overweight.model <- function() {
    for (i in 1:N) {
        logit(pi[i]) <-
            b[1] +
            b[2]*cursmoke[i] +
            b[3]*age.c[i] +
            b[4]*male[i] +
            b[5]*educ2[i] +
            b[6]*educ3[i] +
            b[7]*educ4[i];
        overweight[i] ~ dbin(pi[i], 1);
    }
```

```
# PRIORS ON BETAS
for (j in 1:Nx){
    b[j] ~ dnorm(mu[j],
                 tau[j]); # Independent normal priors
    OR[j] <- exp(b[j]); # Calculate the odds ratios
    }
}


# constants to be passed in
N <- length(overweight); # number of observations to loop over
Nx <- 7; # number of parameters (w/ intercept)
n.iter <- 10000; # number of iterations to run (total)

# Parameters on the priors:
mu <- rep(0,Nx); # Prior mean of betas
tau <- rep(.001,Nx); # Prior precisions

# List of data elements to pass in:
overweight.data <- list("N",
                        "Nx",
                        "overweight",
                        "age.c",
                        "male",
                        "cursmoke",
                        "educ2",
                        "educ3",
                        "educ4",
                        "mu",
                        "tau")

# List of parameters to keep track of:
overweight.parameters <- c("b", "OR")

# Function to generate initial values for each chain:
overweight.inits <- function() {list (b = rnorm(Nx, 0 , sd = 0.5))}


set.seed(123)
overweight.sim <- jags(data = overweight.data,
                       model.file = overweight.model,
                       inits = overweight.inits,
                       parameters.to.save = overweight.parameters,
                       n.iter = n.iter)

print(overweight.sim, digits = 4)
```

```r
overweight.mcmc <- as.mcmc(overweight.sim)

# Traceplot and density plots for regression coefficients
# code will save to PDF in current directory.
# Execute "plot" commands only to plot to screen.
pdf("Traceplot_LogisticReg1.pdf") # Write what comes next to PDF file
plot(overweight.mcmc[1][, 1:4]) # For beta1-4

pdf("Traceplot_LogisticReg2.pdf") # Write what comes next to PDF file
plot(overweight.mcmc[1][, 5:8]) # For beta5-7 and deviance

dev.off() # Stop writing to the PDF file

# Autocorrelation plots for the regression coefficients
pdf("ACF_LogisticReg.pdf")
par(omi=c(.25, .25, .25, .25)) # Create an outer margin (room for title)

autocorr.plot(overweight.mcmc[1][, 1:7]) # For chain 1
title("Chain 1", outer=T) # Place title in outer margin of page

autocorr.plot(overweight.mcmc[2][, 1:7]) # For chain 2 (optional)
title("Chain 2", outer=T)

autocorr.plot(overweight.mcmc[3][, 1:7]) # For chain 3 (optional)
title("Chain 3", outer=T)

dev.off()

geweke.diag(overweight.mcmc[,1:7]) # Geweke test


# Informative prior 1 (Change prior mean to log(2) for b[2])
mu[2] <- log(2)

set.seed(123)
overweight.sim.inform1 <- jags(data = overweight.data,
                               model.file = overweight.model,
                               inits = overweight.inits,
                               parameters.to.save = overweight.parameters,
                               n.iter = n.iter)

print(overweight.sim.inform1, digits = 4)


# Informative prior 2 (Change prior precision to 1/0.1225 vor beta[4])
sd.prior <- (log(2.67) - log(1.5))/(2*1.96) # SD for beta2 on log-scale
tau[2] <- 1/sd.prior^2 # Convert to precision (reciprocal of variance)
```

```r
set.seed(123)

overweight.sim.inform2 <- jags(data = overweight.data,
                               model.file = overweight.model,
                               inits = overweight.inits,
                               parameters.to.save = overweight.parameters,
                               n.iter = n.iter)

print(overweight.sim.inform2, digits=4)

            # b[1] +
            # b[2]*cursmoke[i] +
            # b[3]*age.c[i] +
            # b[4]*male[i] +
            # b[5]*educ2[i] +
            # b[6]*educ3[i] +
            # b[7]*educ4[i];

# vague prior
vague_smoke <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[2]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim$BUGSoutput$summary["OR[2]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim$BUGSoutput$summary["OR[2]", "97.5%"],
                      4), nsmall = 4),
         ")")

vague_age <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[3]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim$BUGSoutput$summary["OR[3]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim$BUGSoutput$summary["OR[3]", "97.5%"],
                      4), nsmall = 4),
         ")")

vague_sex <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[4]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim$BUGSoutput$summary["OR[4]", "2.5%"],
                      4), nsmall = 4),
```

```
        ", ",
        format(round(overweight.sim$BUGSoutput$summary["OR[4]", "97.5%"],
                     4), nsmall = 4),
        ")")

vague_high <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[5]", "50%"],
                     4), nsmall = 4),
        " (",
        format(round(overweight.sim$BUGSoutput$summary["OR[5]", "2.5%"],
                     4), nsmall = 4),
        ", ",
        format(round(overweight.sim$BUGSoutput$summary["OR[5]", "97.5%"],
                     4), nsmall = 4),
        ")")

vague_some <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[6]", "50%"],
                     4), nsmall = 4),
        " (",
        format(round(overweight.sim$BUGSoutput$summary["OR[6]", "2.5%"],
                     4), nsmall = 4),
        ", ",
        format(round(overweight.sim$BUGSoutput$summary["OR[6]", "97.5%"],
                     4), nsmall = 4),
        ")")

vague_college <-
  paste0(format(round(overweight.sim$BUGSoutput$summary["OR[7]", "50%"],
                     4), nsmall = 4),
        " (",
        format(round(overweight.sim$BUGSoutput$summary["OR[7]", "2.5%"],
                     4), nsmall = 4),
        ", ",
        format(round(overweight.sim$BUGSoutput$summary["OR[7]", "97.5%"],
                     4), nsmall = 4),
        ")")

# informative one
info1_smoke <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[2]", "50%"],
                     4), nsmall = 4),
        " (",
        format(round(overweight.sim.inform1$BUGSoutput$summary["OR[2]", "2.5%"],
                     4), nsmall = 4),
        ", ",
        format(round(overweight.sim.inform1$BUGSoutput$summary["OR[2]", "97.5%"],
```

```
                        4), nsmall = 4),
          ")")

info1_age <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[3]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[3]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[3]", "97.5%"],
                        4), nsmall = 4),
          ")")

info1_sex <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[4]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[4]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[4]", "97.5%"],
                        4), nsmall = 4),
          ")")

info1_high <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[5]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[5]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[5]", "97.5%"],
                        4), nsmall = 4),
          ")")

info1_some <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[6]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[6]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform1$BUGSoutput$summary["OR[6]", "97.5%"],
                        4), nsmall = 4),
          ")")
```

```
info1_college <-
  paste0(format(round(overweight.sim.inform1$BUGSoutput$summary["OR[7]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim.inform1$BUGSoutput$summary["OR[7]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim.inform1$BUGSoutput$summary["OR[7]", "97.5%"],
                      4), nsmall = 4),
         ")")

# informative two
info2_smoke <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[2]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[2]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[2]", "97.5%"],
                      4), nsmall = 4),
         ")")

info2_age <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[3]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[3]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[3]", "97.5%"],
                      4), nsmall = 4),
         ")")

info2_sex <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[4]", "50%"],
                      4), nsmall = 4),
         " (",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[4]", "2.5%"],
                      4), nsmall = 4),
         ", ",
         format(round(overweight.sim.inform2$BUGSoutput$summary["OR[4]", "97.5%"],
                      4), nsmall = 4),
         ")")

info2_high <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[5]", "50%"],
```

```
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[5]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[5]", "97.5%"],
                        4), nsmall = 4),
          ")")

info2_some <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[6]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[6]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[6]", "97.5%"],
                        4), nsmall = 4),
          ")")

info2_college <-
  paste0(format(round(overweight.sim.inform2$BUGSoutput$summary["OR[7]", "50%"],
                        4), nsmall = 4),
          " (",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[7]", "2.5%"],
                        4), nsmall = 4),
          ", ",
          format(round(overweight.sim.inform2$BUGSoutput$summary["OR[7]", "97.5%"],
                        4), nsmall = 4),
          ")")


info1_prior_95_upper <- signif(exp(log(2) + 1.96*sqrt(1000)), 5)

info1_prior_95_lower <- signif(exp(log(2) - 1.96*sqrt(1000)), 5)


knitr::include_graphics("Traceplot_LogisticReg1.pdf")

knitr::include_graphics("Traceplot_LogisticReg2.pdf")

include_graphics("ACF_LogisticReg.pdf")


geweke.diag(overweight.mcmc[,1:7])
```