---
title: "STA442 A2"
author: "Manyi Luo - 1003799419"
date: "10/28/2020"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
install.packages("INLA", repos=c(getOption("repos"), INLA="https://inla.r-inla-download.org/R/stable"), dep=TRUE)
library(INLA)
```

#1. School leaver's data

```{r, echo = FALSE}
sUrl = "http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip"
dir.create(file.path("~/Downloads", "data"), showWarnings = FALSE)
(Pmisc::downloadIfOld(sUrl, file.path("~/Downloads", "data")))

school = read.fwf("~/Downloads/data/JSP.DAT", widths = c(2, 1, 1, 1, 2, 4, 2, 2, 1), col.names = c("school", "class", "gender",
"socialClass", "ravensTest", "student", "english", "math", "year"))

school$socialClass = factor(school$socialClass, labels = c("I", "II", "IIIn", "IIIm", "IV", "V", "longUnemp", "currUnemp",
"absent"))
school$gender = factor(school$gender, labels = c("f", "m"))
school$classUnique = paste(school$school, school$class)
school$studentUnique = paste(school$school, school$class,
school$student)
school$grade = factor(school$year)
```

Statistical Report - More detailed codes provided in Appendix

Introduction

This report is constructed for a school board about which factors affect the performance on students' math tests, through Bayesian analysis approach. It looks into the most important influences on student performance on math tests between social Class, grade (school year), and differences between schools etc. It also provide discussions and insights about possible solutions toward improving math score: either it's more effective in identifying poorly performing schools and providing them extra government funding, or giving extra trainings for teachers from poorly performing classes, or even paying extra attention to identify individual weak students.

Statistical Method

```{r, echo = FALSE}
school$wrong_ques <- 40 - school$math

school_mod <- inla(wrong_ques ~ gender + socialClass + grade + f(school) + f(classUnique) + f(studentUnique), data = school,
family = "Poisson")

#summary(model1)
```

Since the number of questions students gets wrong associated with math scores follow the Poisson distribution, which is a continuous and countable distribution, a statistical model corresponding to the data set can be expressed as below:

$$
Y_i\sim Poisson (\lambda_i)
$$

$$
log(\lambda_i) = X_{i}\beta + A_j + B_{jk} + C_i
$$

Here, Y_i represents the number of questions that are answered incorrectly (each question corresponds to one point) for the ith student, which has a distribution showned in the graph below

```{r, echo = FALSE}
hist(40 - school$math, breaks = 100)
```

λ_i represents the average number of questions that are answered incorrectly (each question corresponds to one point) for the ith student; A_j represents the random effect of different schools (j); B_{jk} represents the random effect of different classes (k) under school; C_i represents the random effect of different individual student (i).

Last, X_i is a covariate matrix, which represents non-random effect catagorical variables include different gender (male = 1, female = 0), different social classes (I = 1, II = 2, III nonmanual = 3, III manual = 4; IV = 5; V = 6; Long term unemployed = 7; Not currently employed = 8, Father absent = 9), and difference in junior school year (One = 0, Two = 1, Three = 2).

Since no additional information is given, the default prior is used for Bayesian analysis.

Results

From obtaining regression output and getting the odds ratio, it's possible to interpret multiple factors' effect on math scores through analyzing coefficients for both non-random effect (gender, social class, junior school year) and random effect (school, class, individual). Results and interpretations will start from non-random effect.

Here, given that gender, social class, junior school year are all considered as catagorical, statistically significant levels are considered as having p-value that is smaller than 0.05 and credible interval for odds ratio that does not include 1 listed in the table below. It's always important to note that the odds ratio uses 1 as a standard of comparison, given 95% of baysian credible interval.

For gender, female is considered as the reference level. We can observe that the mean ratio here represents the comparison between the average number of questions that are answered incorrectly in male and female. Since the mean ratio for gender equals 0.999, which is smaller than 1, this indicates that male are 0.999 times less likely to answer math questions incorrectly on average, indicating stronger math ability. However, since the credible interval includes 1, this ratio is insignificant, and we can't conclude on the effect of gender toward the average number of math questions that are answered incorrectly.

Similarly, for social class, classI is considered as the reference level. We can observe that the mean ratio here represents the comparison between the average number of questions that are answered incorrectly in other classes and classI, which has the highest social status. Through observing credible intervals, we can see that the ratio for ClassII and ClassIIIn are considered as insignificant, since their credible intervals include 1, meaning that we can't conclude on the effect of ClassII and ClassIIIn toward the average number of math questions that are answered incorrectly. However, starting from lower classes, ClassIIIm, ClassIV, ClassV, their credible intervals do not include 1, meaning that the ratio for ClassIIIm, ClassIV and ClassV are considered as significant, and there is an effect of lower classes toward the average number of math questions that are answered incorrectly. Taking ClassIIIm as an example of interpretation, its mean ratio equals 1.359, which is greater than 1, indicating that ClassIIIm are 1.359 times more likely to answer math questions incorrectly on average. An increase in mean ratio can also be observed, as it increases to 1.497 for ClassV, showing even higher chance of incorrect answers.

Last but not least, for grade, grade0 is considered as the reference level. We can observe that the mean ratio here represents the comparison between the average number of questions that are answered incorrectly in higher grades and the lowest grade, which has only one junior school year. Through observing credible intervals, we can see that the ratio for grade1 (two junior school years) are considered as insignificant, since its credible interval includes 1, meaning that we can't conclude on the effect of having one more year of schooling compared to the baseline toward the average number of math questions that are answered incorrectly. However, for grade2 (three junior school years), its credible interval includes 1, meaning that the effect of having two more years of schooling compared to the baseline is significant toward the average number of math questions that are answered incorrectly. Taking grade2 as an example of interpretation, its mean ratio equals 0.656, which is smaller than 1, indicating that grade2 are 0.656 times less likely to answer math questions incorrectly on average. An decrease in mean ratio can also be observed, and this makes sense because children's brain are constantly developing, so higher grades may associate with higher cognitive abilities, leading to less incorrect math answers.

~~~
```{r, echo=FALSE}
#default priors
parTable = rbind(school_mod$summary.fixed[,c("mean", "0.025quant",
"0.975quant")], Pmisc::priorPostSd(school_mod)$summary[,
c("mean", "0.025quant", "0.975quant")])

#parTable

#odds ratio
knitr::kable(exp(parTable), digits = 3, caption = "Odds ratios and confidences intervals of the school model")
```
~~~

For random effect, since we are concerning either it's more effective in identifying poorly performing schools and providing them extra government funding, or giving extra trainings for teachers from poorly performing classes, or even paying extra attention to identify individual weak students, we have to determine which level (school, class, individual) is the most problematic, which is indicated by the largest standard deviation and variance. Standard deviation and variance are used to measure change and dispersion, representing the difference between good performing and bad performing schooles, classes, as well as individuals.

Through observing the table below, variance of school and class are negligible comparing to individual, as individual variance occupies 31.729 / (31.729 + 14.357 + 4.520 + 0.000) = 62.698% of the total variation. Another signifiant effect comes from class, as class variance occupies 4.520 / (31.729 + 14.357 + 4.520 + 0.000) = 8.932% of the total variation, which make sense as well: class is the smallest unit of aggregation for individuals, as problematic students may come from the same class as well. On the other hand, the school variance is nearly zero when keeping three decimal places, meaning that the variation in school is the most insignificant effect to concern. Therefore, it's more effective to pay extra attention to identify individual weak students and provide them direct help with math scores. After focusing on individuals, we may consider to boost the standard of teaching in class level.

~~~
```{r, echo=FALSE}
#summary(schoolLme)

variance <- matrix(c(0.000, 0.002, 4.520, 2.126, 31.729, 5.633, 14.357, 3.789),ncol=2,byrow=TRUE)
colnames(variance) <- c("Variance","Std.Dev.")
rownames(variance) <- c("school","classUnique","studentUnique", "Residual")
variance <- as.table(variance)
variance
```
~~~

#2. Smoking

~~~
```{r}
dataDir = "../data"
smokeFile = file.path(dataDir, "smoke2014.RData")
if (!file.exists(smokeFile)) {download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData", smokeFile)}

forInla = smoke[smoke$Age > 10, c("Age", "ever_cigarettes",
"Sex", "Race", "state", "school", "RuralUrban",
"Harm_belief_of_chewing_to")]
forInla = na.omit(forInla)
forInla$y = as.numeric(forInla$ever_cigarettes)
forInla$ageFac = factor(as.numeric(as.character(forInla$Age)))
forInla$chewingHarm = factor(forInla$Harm_belief_of_chewing_to,
levels = 1:4, labels = c("less", "equal", "more",
"dunno"))
library("INLA")
~~~

~~~~
```

Hypothesis 1:
default prior
```{r}
model2 <- inla(y ~ f(state, model = "iid") + f(school, model = "iid"), data = forInla, family= "binomial")

rbind(model2$summary.fixed[, c("mean", "0.025quant","0.975quant")], Pmisc::priorPostSd(model2)$summary[, c("mean",
"0.025quant", "0.975quant")])
```

```{r}
install.packages("remotes")
remotes::install_github("andrewzm/INLA")
```

How to choose prior?
Use "pc.prec" precision prior for the $log(\tau) = log(\frac{1}{\sigma^2})$

$\text{precision} = \tau = \frac{1}{\sigma^2} = \frac{1}{\SD^2}$

Note:
1. Bayesian: like to use precision, and different precisions for state and school;
2. pc prior for precision penalized complexity with log scale prob distribution (given different range for sd);
3. Binary outcomes：  logistic regression model

```{r}
#another model

model2_1 <- inla(y ~ f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec", pram = c(99, 0.05)))) +
f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec", pram = c(99, 0.05)))),
              data = forInla, family= "binomial")

rbind(model2_1$summary.fixed[, c("mean", "0.025quant","0.975quant")], Pmisc::priorPostSd(model2_1)$summary[, c("mean",
"0.025quant", "0.975quant")])
```

Graphs of prior and posterior densities of model parameters
```{r}

theSd <- Pmisc::priorPost(model2_1)

plot(theSd$'sd for state'$posterior, type ='l', xlab = 'sd', ylab = 'dens',
     xlim=c(0,1), ylim = c(0, 7.5), col= 'black')
lines(theSd$'sd for state'$prior, col= 'black', lty = 2)
lines(theSd$'sd for school'$prior, col= 'red', lty = 1)
lines(theSd$'sd for school'$prior, col= 'red', lty = 2)
legend("topright", col = c("black", "black", "red", "red"), lty= c(2,1,2,1),
legend = c("prior for state's SD",
           "posterior for state's SD",
           "prior for school's SD",
           "posterior for school's SD"), bty="n",
           title = "Priors and Posterios")




#plot(model2$marginals.fixed$state, type = 'l')
#lines(
  #model2$marginals.fixed$state[, 'x'],
  #dnrom(model2$marginals.fixed$state[, 'x'], mean = 0, sd=0.2),
 # col= 'blue', lwd=3
#)

```

The model and prior distribution
```{r}
#To choose the percision for school

toget <- rlnorm(10000,
               meanlog= -2, sdlog= 0.8)
hist(inv.logit(toget))
mean(toget)
quantile(toget, 0.95)
```

Model parameters
```{r}
toPredict = expand.grid(ageFac = levels(forInla$ageFac),
  RuralUrban = levels(forInla$RuralUrban), Race = levels(forInla$Race),
  Sex = levels(forInla$Sex))
forLincombs = do.call(inla.make.lincombs, as.data.frame(model.matrix(~Sex +
  ageFac * RuralUrban * Race, data = toPredict)))

fitS2 = inla(y ~ Sex + ageFac * RuralUrban * Race +
  f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec",
    param = c(log(1.1), 0.05)))), data = forInla, family = "binomial",
  control.inla = list(strategy = "gaussian"), lincomb = forLincombs)
~~~~

~~~~
rbind(fitS2$summary.fixed[, c("mean", "0.025quant",
  "0.975quant")], Pmisc::priorPostSd(fitS2)$summary[,
  c("mean", "0.025quant", "0.975quant")])

# create matrix of predicted probabilities
theCoef = exp(fitS2$summary.lincomb.derived[, c("0.5quant",
  "0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)

# create an x axis, shift age by race
toPredict$Age = as.numeric(as.character(toPredict$ageFac))

```

Plots
```{r}

# for male in Rural area
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban ==
  "Rural"
plot(toPredict[toPlot, "Age"], theCoef[toPlot, "0.5quant"],
  xlab = "age", ylab = "probability", ylim = c(0,
    1), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "Age"], theCoef[toPlot, "0.025quant"],
  y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot,
    "Race"])
legend("topleft", fill = 1:nlevels(toPredict$Race),
  legend = levels(toPredict$Race), bty = "n",
  title = "Race")


# for male in Urban Area
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban ==
  "Urban"
plot(toPredict[toPlot, "Age"], theCoef[toPlot, "0.5quant"],
  xlab = "age", ylab = "probability", ylim = c(0,
    1), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "Age"], theCoef[toPlot, "0.025quant"],
  y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot,
    "Race"])
legend("topleft", fill = 1:nlevels(toPredict$Race),
  legend = levels(toPredict$Race), bty = "n",
  title = "Race")
```

```{r}
#for female in rural area
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban ==
  "Rural"
plot(toPredict[toPlot, "Age"], theCoef[toPlot, "0.5quant"],
  xlab = "age", ylab = "probability", ylim = c(0,
    1), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "Age"], theCoef[toPlot, "0.025quant"],
  y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot,
    "Race"])
legend("topleft", fill = 1:nlevels(toPredict$Race),
  legend = levels(toPredict$Race), bty = "n",
  title = "Race")

#for female in Urban area
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban ==
  "Urban"
plot(toPredict[toPlot, "Age"], theCoef[toPlot, "0.5quant"],
  xlab = "age", ylab = "probability", ylim = c(0,
    1), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "Age"], theCoef[toPlot, "0.025quant"],
  y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot,
    "Race"])
legend("topleft", fill = 1:nlevels(toPredict$Race),
  legend = levels(toPredict$Race), bty = "n",
  title = "Race")
```
~~~~

Consulting Report - More detailed codes provided in Appendix

Introduction

To carry about multiple research questions about smoking behaviors among children with different origins and sex, the 2019 American National Youth Tobacco Survey is used to conduct the following analysis, as it includes data of the use of cigars, hookahs, and chewing tobacco amongst American school children. There are two main research hypotheses: The first hypothesis is that given white Americans more likely to live in rural areas and cigar smoking is a rural phenomenon, the chance of smoking of cigars, cigarillos or little cigars is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans. The second hypothesis is that the chance of used electronic cigarettes on at least one occasion is the same among people of different sex when controlling their age, ethnicity, and other demographic characteristics.

Statistical Method

According to the research hypothesis, two variables are used, 1. "ever_cigars_cigarillos_or", representing the binary output from the question: "Have you ever tried smoking cigars, cigarillos, or little cigars, such as Swisher Sweets, Black and Mild, Garcia y Vega, Cheyenne, White Owl, or Dutch Masters, even one or two puffs?"; 2. "ever_ecigarette", representing the binary output from the question: "Have you ever used an e-cigarette, even once or twice?" Therefore, logistic regression is the most appropriate model to test out multiple factors that may influence dichotomous outputs (Yes = 1, No = 0).

In logistic model, the response variable (Y, with $P(Y = 1) = p$) is represented by the logarithmic of smoking odds, either ever tried smoking cigars, cigarillos, little cigars or ever used an e-cigarette. Specifically, odds represents the chance of ever tried smoking over the chance of never tried smoking ($logit[Y = 1] = log(\frac{P(Y = 1)}{P(Y = 0)}) = log(\frac{p}{1-p})$).

Explanatory variables (X_1 to X_4) include: X_1. RuralUrban: a catagorical variable representing whether the school the respondent attended was rural or urban (binary output, Rural = 1); X_2. variable representing Race: also catagorical, including Black (binary, Black = 1), Hispanic (binary, Hispanic = 1), Asian (binary, Asian = 1), Native (binary, Native = 1), Pacific (binary, Pacific = 1), whereas White is considered as the reference level; X_3. Sex: a catagorical variable representing male and female (binary output, Female = 1); X_4. Age: a categorical variable that has been converted to years. Note that the counfounding variable between RuralUrban and Race may appear in the first hypothesis to facilitate our regression, since the hypothesis considers the joint affect between Race and RuralUrban, specifically under Rural.

Therefore, the model for the first hypothesis can be expressed as: $logit[Y = 1] = \alpha + \beta_1 X_1*X_2 + \beta_2 X_3 + \beta_3 X_4$, whereas the model for the second hypothesis can be expressed as: $logit[Y = 1] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

Results - First Research Question

For the first hypothesis, regression output and odds ratio can be obtained from the table below to see how region (RuralUrban), Race, Sex, and Age may influence the ever smoking odds of children. Here, all four independent variables are statistically significant related to the ever smoking odds, as their p-values are smaller than 0.05 and confidence intervals for odds ratio do not include 1, specifically RuralUrbanRural, Raceblack, Raceasian, Racepacific, SexF, Age.

Using RuralUrban as an example of significant categorical variables, the corresponding odds ratio for Rural is compared toward the reference level, Urban. Its odds ratio is 1.602, which is greater than 1, meaning that there exist greater odds when increasing by one unit. While keeping all other variables fixed, this means that comparing children from Rural region to children from Urban region, the odds of ever smoking is 1.602 times more relative to that of children from Urban region, meaning that there is more chance for them to ever tried smoking cigars, cigarillos, or little cigars. Similarly, for Race, Black children odds of ever smoking is 1.436 times more and Pacific children odds of ever smoking is 2.070 times more relative to that of white children; Asian children odds of ever smoking is 0.328 times less relative to that of white children. For sex, female children's odds of ever smoking is 0.685 times less relative to that of male children.

For continuous variable specifically Age, it's odds ratio is 1.454, which is greater than 1. This means that, while keeping all other variables fixed, when age increases by 1 year, the odds of ever smoking will increase by 1.454 times, indicating there is more chance of smoking for greater age.