School Leavers and Smoking Report

Yingfei Zha

27/10/2020

Report of School Leaver's Data

The data come from book *the Junior School Project* including over 1000 students and 3236 records that are collected over three school years. The response variable is the number of questions that students gets wrong on mathematics tests during the school years, from which we explore its relationship with an individual's gender, social class, grade, school and class.

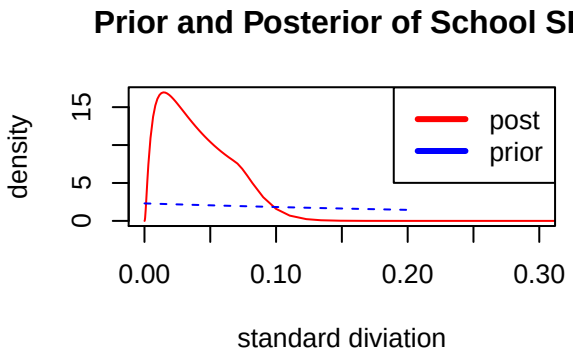
A mixed effects Poisson model is chosen to fit the data. We also use the integrated nested Laplace approximation (INLA) to approximate Bayesian inference. Since we are using Bayesian inference, we assume that students in the 75th percentile has 1.5 times more problems that they get wrong as students in the 25th percentile. So for the random effects, the priors for the standard deviation of school level effect σ_{school}, class level effect σ_{class} and individual level effect $\sigma_{student}$ are Exponential priors with 50% quantile of $\frac{log(1.5)}{\Phi^{-1}(0.75)-\Phi^{-1}(0.25)} \approx 0.301$. It means that the probability of $\sigma > 0.301$ is 0.5 for each random effect. The plots of priors and posteriors for each random effect are shown below. As no information was provided for the priors of the fixed effects gender, social class and grade, we set the priors of them to the default prior, which is β follows a Normal prior with mean equals to 0 and standard deviation equals to 1000.

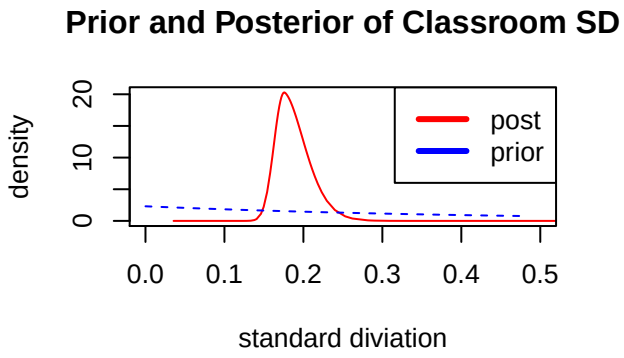
The statistical results are shown in **Table 2**. The number of questions that male students got wrong are 0.998 times less than female students given the other predictor variables in the model are held constant; but since the 95% credible intervals include 1, gender has no statistically significant influences on student performance on math tests. On the other hand, all social classes show significant influences except for social class II and III non-manual. The lower social class students are more likely to get more number of questions wrong. Students from currently unemployment social class have 1.483 times more number of incorrect problems than students from social class I; they have the worst performance on math tests comparing to students from other social classes. Moreover, students in their third Junior school year (grade 2) have 0.656 times less number of incorrect problems comparing to students in their first Junior school year (grade 0). This result is statistically significant, which suggests that grade can be one of the influences on student performance on math tests. In addition, we consider which of the variations among schools, classrooms and students have the highest influences on student performance. There is a little bit of variation amongst schools (= 1.049) and classrooms (= 1.194). Especially for schools, 1 standard deviation increasing only results in 1.049 times more number of incorrect problems and the result is nearly non-significant. On the other hand, there is a lot of variations among student (= 1.579). 1 standard deviation of individual level effect increasing results in 1.579 times more number of incorrect problems.

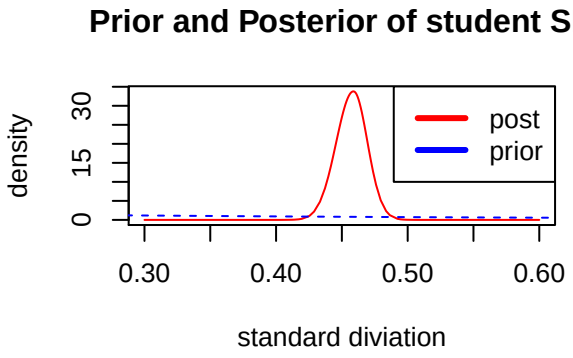
To summary, after analyzing the data from book *the Junior School Project*, gender is not a significant factor that influences student performance on math tests. In addition, students from currently unemployment social class are more likely to have worse performance and students in the third school year are less likely to. There are not really such things as good schools or bad schools, but there is a little bit of variations among classrooms and a lot of variations among students. Therefore, identifying individual weak students and give them extra attention should be a more effective approach to improve student performance on math tests.

Table 1: Estimated Rate Ratio and Exponentiated SD for School Leaver's Data

	0.5quant	0.025quant	0.975quant
(Intercept)	10.344	8.568	12.480
genderm	0.998	0.942	1.059
socialClassII	0.983	0.807	1.198
socialClassIIIn	1.197	0.973	1.472
socialClassIIIm	1.358	1.127	1.636
socialClassIV	1.307	1.064	1.604
socialClassV	1.488	1.205	1.838
socialClasslongUnemp	1.412	1.134	1.758
socialClasscurrUnemp	1.484	1.111	1.982
socialClassabsent	1.402	1.154	1.704
grade1	0.998	0.976	1.019
grade2	0.656	0.639	0.673
SD for school	1.036	1.005	1.105
SD for classUnique	1.203	1.167	1.271
SD for studentUnique	1.580	1.542	1.615







Smoking

Summary

In this study, we investigate the smoking behaviors among youth using a sample from 2019 American National Youth Tobacco Survey. As the result of the primary analysis with regard to states variation and schools variation in the odds of students smoking cigarettes, we find that variation among states is less than variation among schools. This indicates that implementing tobacco control programs in schools with higher rate of students smoking could be a more effective intervention approach. In addition to this, we also find that rural-urban differences are approximately 1.5 times greater than states differences in the rate of smoking. As the probability of smoking in rural areas is 2 times higher than the probability of smoking in urban areas, smoking in rural areas should be concerned.

The results of the secondary analysis with regard to the age effect on smoking among European, African, and Hispanic Americans are different by sex and by rurality. For males living in rural areas, the probability of smoking for European and Hispanic Americans increases over time while African Americans have a fluctuated probability of smoking. These three groups have the largest difference at age 17 but similar probability at age 18. On the other hand, females living in rural areas have the largest difference at age 18. The probability of smoking for female African Americans does not go as high as European and Hispanic Americans; instead, it goes down at age 18. For males living in urban areas, the probability for all three groups steadily increases and their differences are small. On the other hand, females European and Hispanic Americans living in urban areas have increasing probability from age 11 to 18, while African Americans have decreasing probability starting from age 15. Three groups have big differences in smoking at age 17 and 18.

Introduction

In this study, data from a nationally representative sample of over 22000 American school children (2019 American National Youth Tobacco Survey) are analyzed with regard to the use of cigarettes. The primary question is to, first, investigate whether there are significant variations in the rate of students smoking cigarettes among the US states and schools. In addition to this, we want to know if variation among schools is less than variation among states. Second, we are interested in if rural-urban differences in the rate of smoking are much greater than differences between states while keeping students' other demographic characteristics the same. The secondary question is to study the differences between European, African, and Hispanic Americans in the effect of age on smoking. For all above analysis, individual's sex and age are also taken into account as important confounders.

Methods

Model

To investigate the above hypotheses, the mixed effects logistics regression model is chosen as a conventional method to model data with dichotomous response variables. We also use the integrated nested Laplace approximation (INLA) to approximate Bayesian inference. The relationship between the response, fixed effects and random effects is given by:

$$Y \sim Bernoulli(\theta)$$

$$logit(\theta) = \beta_0 + \beta_1 X_{sex} * X_{age} * X_{rurality} * X_{race} + U_{state} + U_{school}$$

$$U_{state} \sim i.i.d.Normal(0, \sigma^2_{state})$$

$$U_{school} \sim i.i.d.Normal(0, \sigma^2_{school})$$

where the response variable Y is 1 when the student smokes and 0 when the student does not smoke. θ represents the probability that a student smokes. For the fixed effect, it is the interaction of sex, age, race and

rurality. It is believed that the effect of age on outcome Y can vary according to other predictor variables. For the random effects, U_{state} is the state-level random effect, it measures the variation between states in the rate of students smoking cigarettes. U_{school} is the school-level random effect, it measures the variation among schools in the rate of students smoking cigarettes.
The data are cleaned by excluding age 9, 10 and 19 from the analysis because the number of observations of these groups are very little. Moreover, the observations of Pacific Americans are also removed due to the same reason. Age 14 is set to the reference level instead of age 11.

Priors

- According to the information provided by some collaborating scientists, the rate of smoking between states could be tripled. So the prior for the standard deviation of state effect σ_{state} is an Exponential prior with 50% quantile of $\frac{log(3)}{\Phi^{-1}(0.90)-\Phi^{-1}(0.10)} \approx 0.429$. It means that the probability of $\sigma_{state} > 0.429$ is 0.5.
- As the collaborating scientists also suggested that the the 'worst' schools are expected to have at most 50% greater rate of smoking than the 'healthiest' schools, the prior for the standard deviation of school effect σ_{school} is an Exponential prior with 50% quantile of $\frac{log(1.5)}{\Phi^{-1}(0.90)-\Phi^{-1}(0.10)} \approx 0.158$. It means that the probability of $\sigma_{school} > 0.158$ is 0.5.
- Since no information was provided with regard to fixed effects priors, we set the priors of fixed effects to the default prior, which is β follows a Normal prior with mean equals to 0 and standard deviation equals to 1000.

Results

Table 2: Estimated Odds Ratio and Exponentiated SD of State and School for Smoking Data

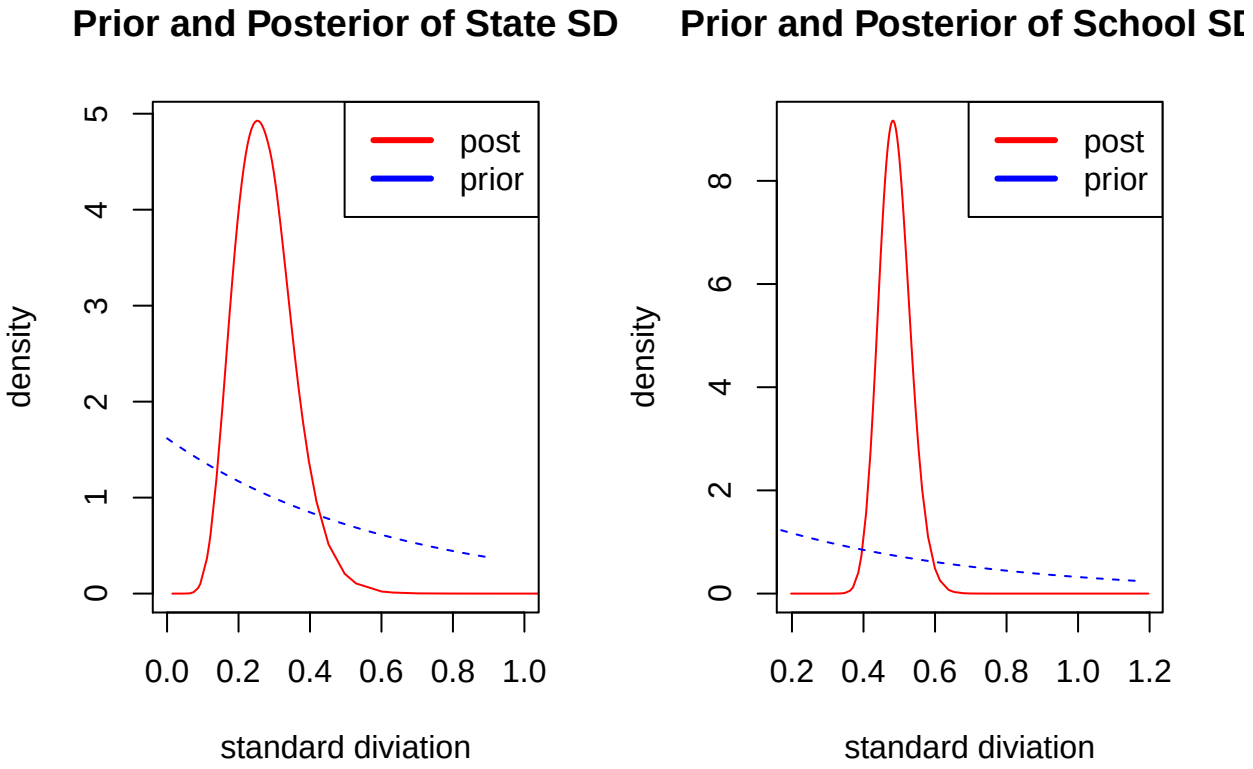
	0.5quant	0.025quant	0.975quant
Baseline odds	0.145	0.098	0.209
Female	0.663	0.387	1.113
African American	1.027	0.548	1.863
Hispanic American	1.637	1.031	2.613
Asian American	0.937	0.314	2.303
Native American	4.862	1.528	15.289
Rural	2.012	1.263	3.231
exp(SD) for state	1.306	1.148	1.571
exp(SD) for school	1.626	1.503	1.787

Primary study

The statistical results to analyze the primary questions are shown in **Table 2**.

The prior distribution and posterior distribution of σ_{state} and σ_{school} are shown in the graphs below. Different states have odds of students smoking cigarettes that are different by $exp(\sigma_{state}) = 1.311$. To be more specific, 1 standard deviation more in state random effect means 1.311 times the odds of students smoking cigarettes comparing to a typical student (i.e $U_{state} = 0$). Similarly, variation among schools in the rate of students smoking cigarettes is different by $exp(\sigma_{school}) = 1.625$. It means that 1 standard deviation more in school random effect means 1.625 times the odds of students smoking cigarettes comparing to a typical student (i.e $U_{school} = 0$). The 95% Credible intervals of both state and school effects indicate significant results. Thus, to answer the first hypothesis, variation among states is less than variation among schools, which indicates that implementing tobacco control programs in schools with higher rate of students smoking cigarettes could

be a more effective intervention approach.



For the second hypothesis, we investigate rural-urban differences and state differences in the rate of smoking. The odds of students smoking cigarettes in rural areas are 2.007 times the odds of smoking in urban areas. Its 95% credible intervals suggest that the differences in the probability of smoking between rural and urban areas are significant. Differences between states are also significant. The odds of smoking are 1.311 times more if 1 standard deviation increases in state random effect comparing to a student with 0 state random effect. Therefore, rural-urban differences are approximately 1.5 times greater than states differences in the rate of smoking.

Secondary Study

To investigate the differences between European, African, and Hispanic Americans in the effect of age on smoking, we look into the effect of age for four groups separately. The four groups are males and females living in rural areas, and males and females living urban areas.

- As **Figure 1** shown below are the estimated probabilities of males smoking in rural areas from age 11 to 18 by racial groups with 95% credible intervals. In general, male European, African, and Hispanic Americans all have increasing probability in smoking as age increases; except for Hispanic Americans at age 18, their probability of smoking drops slightly comparing to those at age 17. Moreover, Hispanic Americans have the highest probability of smoking at age 13, 15, 16 and 17 among these three racial groups. European Americans have the highest probability of smoking at age 11, 14, 16 and 18. So leaving African Americans have the least probability of smoking from age 11 to 18 in general. Unlike Hispanic Americans have the highest smoking probability at age 17, both European and African Americans have the highest smoking probability at age 18. However, the probability for three groups are all closely around 50% at age 18. To summary, the probability of smoking goes up and down for African Americans but there are two big rise at age 15 and 18. The probability of smoking increases steadily for European Americans, the largest rise is at age 16. For Hispanic Americans, the probability also increases over time with a big rise at age 15 and a slight drop at age 18. All three groups have an

increasing probability at age 15. It might suggest that after entering high schools, students are more likely to smoke cigarettes.

- As **Figure 2** shown below are the estimated probabilities of females smoking in rural areas from age 11 to 18 by racial groups with 95% credible intervals. For female European Americans, the probability of smoking goes up constantly and reaches the highest at age 18. However, the probability of smoking fluctuates for female African Americans. There are drops at age 13 and 18. The highest probability is at 17. For Hispanic Americans, the probability also increases over time with a largest rise at age 15. In addition, female Hispanic Americans in rural areas have the highest probability of smoking among these three groups most of the time. Female African Americans in rural areas are less likely to smoke cigarettes. The difference of smoking probability between three groups are larger as age increases.

- As **Figure 3** shown below are the estimated probabilities of males smoking in urban areas from age 11 to 18 by racial groups with 95% credible intervals. For all three groups, the probability of smoking increase steadily. For male European Americans in urban areas, the probability of smoking increases approximately from 1% to 40% at age 17 but drop to 35% at age 18. The probability of smoking for African Americans increases approximately from 1% to 36%. For Hispanic Americans, the probability also increases approximately from 5% to 38% over time. Overall, Hispanic Americans have the highest probability of smoking among these three groups except at age 17. At age 17, European Americans have the highest probability of smoking among three groups from age 11 to 18. However, the smoking probability for all three groups are similar over time.

- As **Figure 4** shown below are the estimated probabilities of females smoking in urban areas from age 11 to 18 by racial groups with 95% credible intervals. The effect of age on smoking for female African Americans in urban areas is very different from European and Hispanic Americans. For European Americans in urban areas, the probability of smoking increases over time with a big rise at age 15 and another one at age 17. The probability of smoking for Hispanic Americans also increases over time. Both European and Hispanic Americans reach their highest probability of smoking around 40% at age 18. However, the probability of smoking for African Americans increases from age 11 to 15 but decreases from age 16 to 18. So African Americans have their highest probability of smoking at age 15 which is around 25%. Overall, the difference of smoking probability between three groups are larger as age increases.

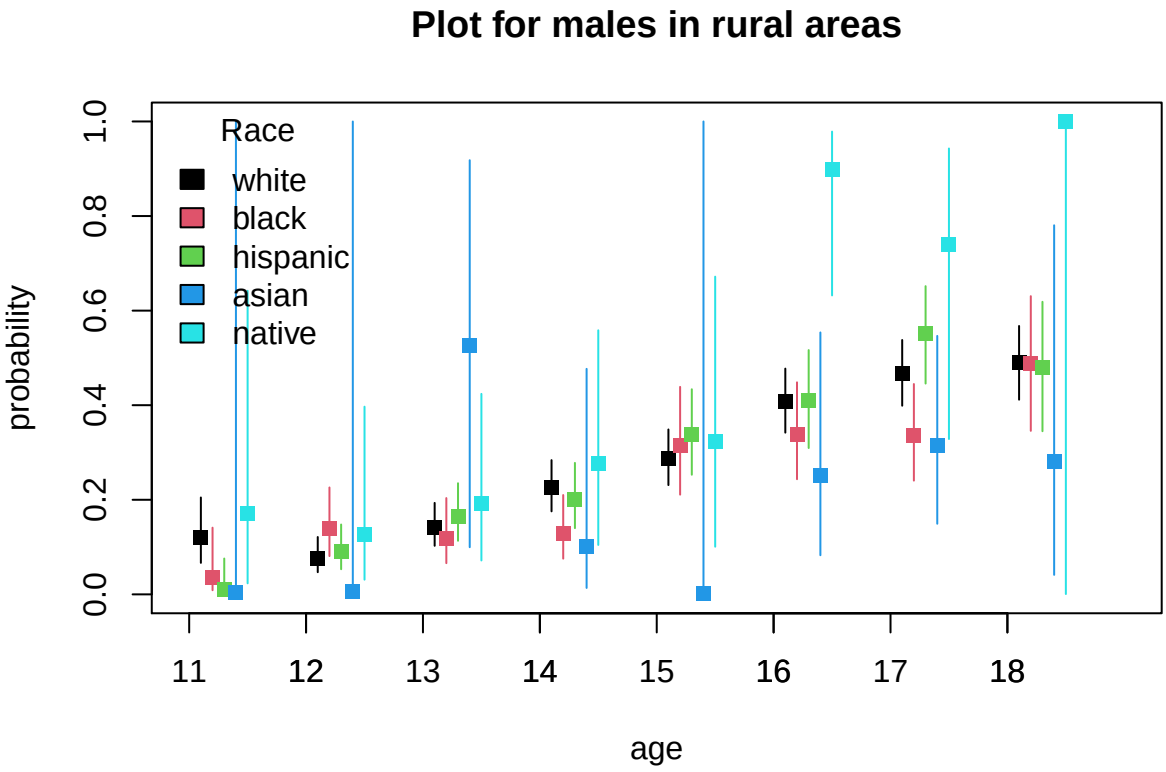


Figure 1: Estimated probabilities of males smoking in rural areas from age 11 to 18 by racial groups with 95% CI

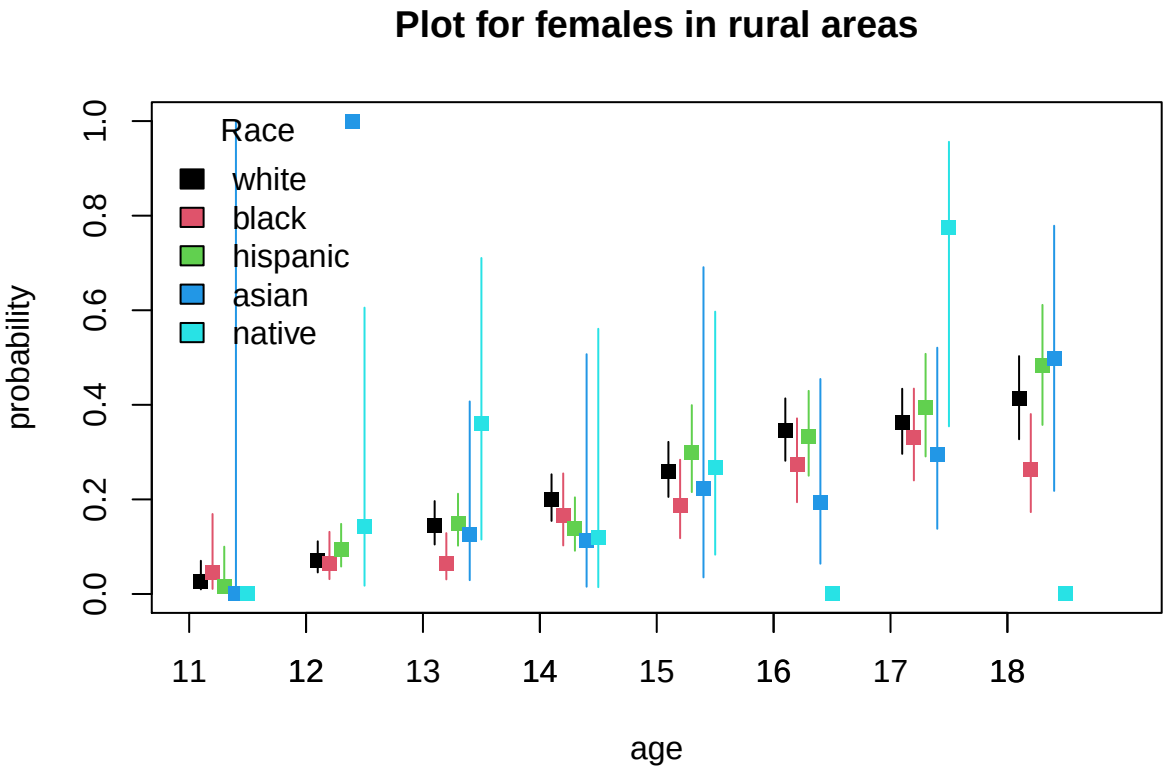


Figure 2: Estimated probabilities of females smoking in rural areas from age 11 to 18 by racial groups with 95% CI

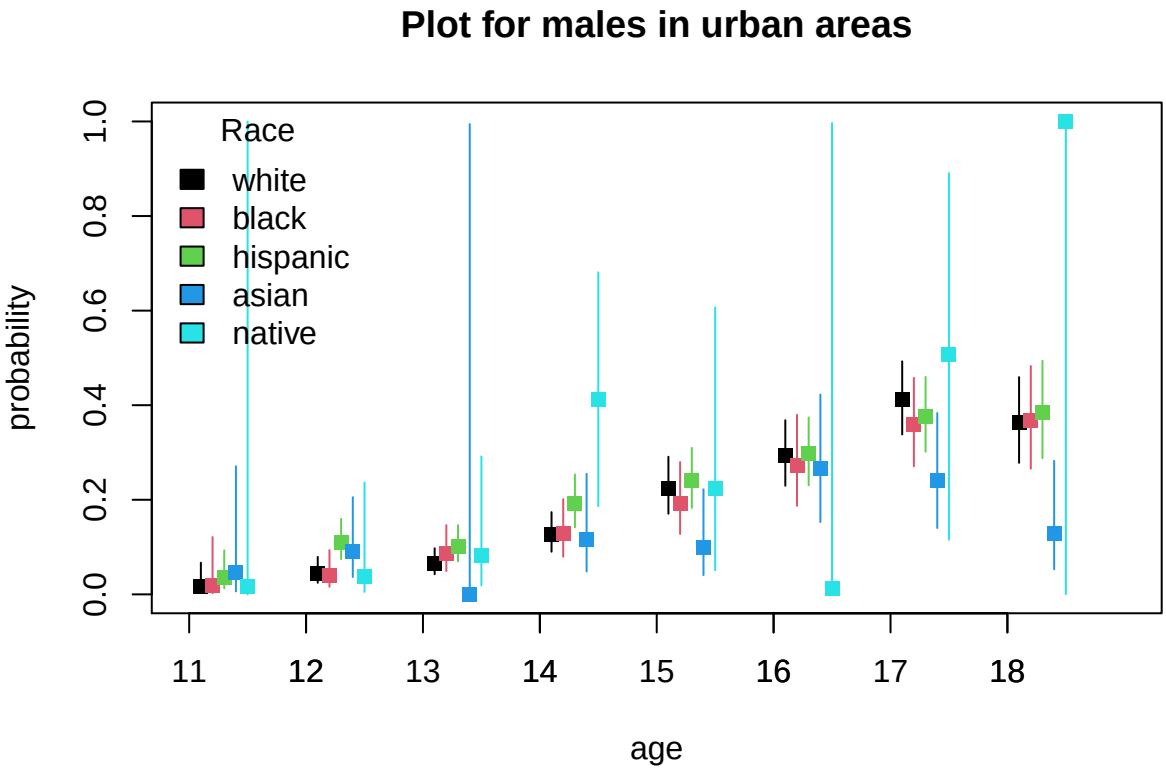


Figure 3: Estimated probabilities of males smoking in urban areas from age 11 to 18 by racial groups with 95% CI

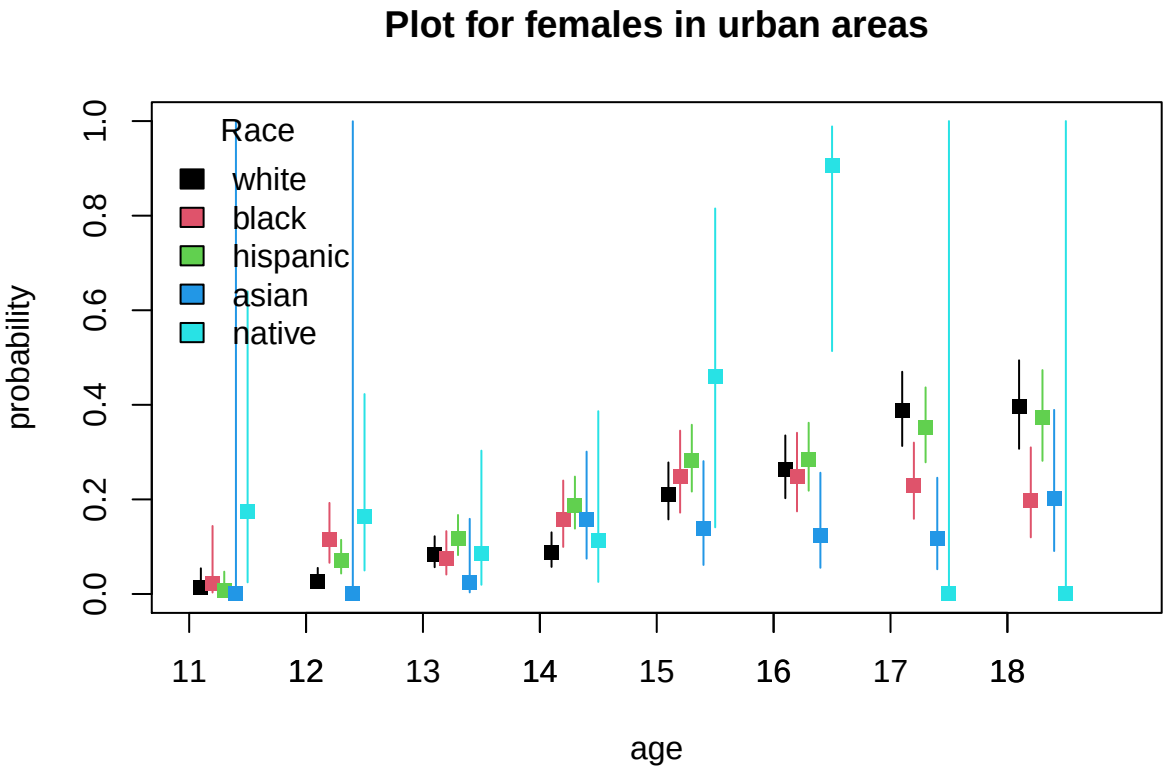


Figure 4: Estimated probabilities of females smoking in urban areas from age 11 to 18 by racial groups with 95% CI

Appendix

```
knitr::opts_chunk$set(echo=FALSE, warning = FALSE, message = FALSE)
sUrl = "http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip"
dir.create(file.path("~/Downloads/", "data"), showWarnings = FALSE)
Pmisc::downloadIfOld(sUrl, file.path("~/Downloads/", "data"))
school = read.fwf("/Users/AnnieZha/Downloads/data/JSP.DAT", widths =
                    c(2, 1, 1, 1, 2, 4, 2, 2, 1), col.names =
                    c("school", "class", "gender", "socialClass", "ravensTest", "student",
                      "english", "math", "year"))
school$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)

hist(school$math, breaks = 100)
hist(40 - school$math, breaks = 100)
library(INLA)
school$mathWrong = 40 - school$math
n=qnorm(0.75)-qnorm(0.25)
mod = inla(mathWrong ~ gender + socialClass + grade
           + f(school, model = "iid", prior='pc.prec', param = c(log(1.5)/n, 0.5))
           + f(classUnique, model = "iid", prior='pc.prec', param = c(log(1.5)/n, 0.5))
           + f(studentUnique, model = "iid", prior='pc.prec', param = c(log(1.5)/n, 0.5)),
           data = school, family = 'poisson', control.inla =
             list(strategy='laplace',fast=FALSE))
knitr::kable(rbind(
  exp(mod$summary.fixed[, c('0.5quant','0.025quant','0.975quant')]),
    exp(Pmisc::priorPostSd(mod)$summary[, c('0.5quant','0.025quant','0.975quant')])
), digits = 3, caption="Estimated Rate Ratio and Exponentiated SD for School Leaver's Data")
theSd1 = Pmisc::priorPost(mod)
par(mfrow=c(2,2))
plot(theSd1$"sd for school"$posterior, xlim=c(0,0.3), xlab='standard diviation', ylab='density',
     main= 'Prior and Posterior of School SD',type='l', col='red')
lines(theSd1$"sd for school"$prior, col='blue', lty=2)
legend("topright", col=c("red","blue"),lty=1, lwd=3, legend=c("post", "prior"))

plot(theSd1$"sd for classUnique"$posterior, xlab='standard diviation', ylab='density',
     main='Prior and Posterior of Classroom SD',type='l', col='red',xlim=c(0,0.5))
lines(theSd1$"sd for classUnique"$prior, col='blue', lty=2)
legend("topright", col=c("red","blue"),lty=1, lwd=3, legend=c("post", "prior"))

plot(theSd1$"sd for studentUnique"$posterior, xlab='standard diviation', ylab='density',
     main='Prior and Posterior of student SD',type='l', col='red')
lines(theSd1$"sd for studentUnique"$prior, col='blue', lty=2)
legend("topright", col=c("red","blue"),lty=1, lwd=3, legend=c("post", "prior"))
#######Smoking########
load("~/Downloads/smoke2014.RData")
smoke[1:3,c('Age','ever_cigarettes','Sex','Race',
        'state','school', 'RuralUrban')]
```

```r
forInla = smoke[,c('Age','ever_cigarettes','Sex','Race',
        'state','school', 'RuralUrban')]

forInla$y = as.numeric(forInla$ever_cigarettes)

#Data cleaning
forInla <- forInla[ !(forInla$Age == 9), ] #Take out age 9 data
forInla <- forInla[ !(forInla$Age == 10), ]#Take out age 10 data
forInla <- forInla[ !(forInla$Age == 19), ]#Take out age 19 data
forInla <- forInla[ !(forInla$Race == 'pacific' ), ]#Take out pacific data

forInla = na.omit(forInla)
forInla$ageFac = relevel(factor(forInla$Age), '14') #new baseline = 14 years old
forInla$Race = factor(forInla$Race)
x=qnorm(0.90)-qnorm(0.10)
library(INLA)
toPredict = expand.grid(Sex = levels(forInla$Sex), ageFac = levels(forInla$ageFac),
                        RuralUrban = levels(forInla$RuralUrban),
                        Race = levels(forInla$Race))
forLincombs = do.call(inla.make.lincombs,
                      as.data.frame(model.matrix(~Sex * ageFac * RuralUrban * Race,
                                                 data = toPredict)))
smokeMod = inla(y ~ Sex * ageFac * RuralUrban * Race
                + f(state, model='iid', hyper=list(prec=list(prior='pc.prec',
                                                             param=c(log(3)/x, 0.5))))
                + f(school, model = "iid", hyper=list(prec=list(prior='pc.prec',
                                                                param=c(log(1.5)/x, 0.5)))),
                data=forInla, family='binomial',lincomb = forLincombs)
smokeTable=rbind(
    "Baseline odds" = 1/exp(-smokeMod$summary.fixed[1,c('0.5quant','0.025quant','0.975quant')]),
    exp(smokeMod$summary.fixed[, c('0.5quant','0.025quant','0.975quant')])
  )
knitr::kable(smokeTable,digits = 3)

k = rbind("Baseline odds" =
            1/exp(-smokeMod$summary.fixed[1,c('0.5quant','0.025quant','0.975quant')]),
          exp(smokeMod$summary.fixed['SexF',c('0.5quant','0.025quant','0.975quant')]),
          exp(smokeMod$summary.fixed[c('Raceblack','Racehispanic','Raceasian','Racenative'),
                                     c('0.5quant','0.025quant','0.975quant')]),
          exp(smokeMod$summary.fixed['RuralUrbanRural',
                                     c('0.5quant','0.025quant','0.975quant')]),
          exp(Pmisc::priorPostSd(smokeMod)$summary[, c('0.5quant','0.025quant','0.975quant')]))
row.names(k)=c("Baseline odds", 'Female','African American','Hispanic American','Asian American',
               'Native American','Rural', 'exp(SD) for state', 'exp(SD) for school')
knitr::kable(k, digits = 3, caption="Estimated Odds Ratio and Exponentiated SD
             of State and School for Smoking Data")
theSd = Pmisc::priorPost(smokeMod)
par(mfrow=c(1,2))
plot(theSd$"sd for state"$posterior, xlim=c(0,1), xlab='standard diviation', ylab='density',
     main= 'Prior and Posterior of State SD',type='l', col='red')
lines(theSd$"sd for state"$prior, col='blue', lty=2)
legend("topright", col=c("red","blue"),lty=1, lwd=3, legend=c("post", "prior"))
```

```
plot(theSd$"sd for school"$posterior, xlab='standard diviation', ylab='density',
     main='Prior and Posterior of School SD',type='l', col='red')
lines(theSd$"sd for school"$prior, col='blue', lty=2)
legend("topright", col=c("red","blue"),lty=1, lwd=3, legend=c("post", "prior"))
theCoef = exp(smokeMod$summary.lincomb.derived[, c("0.5quant", "0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$Age = as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX = as.numeric(toPredict$Race)/10
toPredict$x = toPredict$Age + toPredict$shiftX
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban == "Rural"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"], xlab = "age", ylab = "probability",
     ylim = c(0,1), xlim = c(11,19), pch = 15, col = toPredict[toPlot, "Race"],
     main="Plot for males in rural areas")
segments(toPredict[toPlot, "x"], 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")
axis(side = 1, at=11:18)
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban == "Rural"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"], xlab = "age", ylab = "probability",
     ylim = c(0,1), xlim = c(11,19), pch = 15, col = toPredict[toPlot, "Race"],
     main="Plot for females in rural areas")
segments(toPredict[toPlot, "x"], 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")
axis(side = 1, at=11:18)
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban == "Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"], xlab = "age", ylab = "probability",
     ylim = c(0,1), xlim = c(11,19), pch = 15, col = toPredict[toPlot, "Race"],
     main="Plot for males in urban areas")
segments(toPredict[toPlot, "x"], 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")
axis(side = 1, at=11:18)
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban == "Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"], xlab = "age", ylab = "probability",
     ylim = c(0,1), xlim = c(11,19), pch = 15, col = toPredict[toPlot, "Race"],
     main="Plot for females in urban areas")
segments(toPredict[toPlot, "x"], 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")
axis(side = 1, at=11:18)
```