STA442 Project 2

Aichen Liu

30/10/2020

School leaver's data

Introduction

Mathematics is one of the disciplines that most teenagers struggle to do well in. This report explores the effect of factors such as the social class of the student, gender of student, grade the student is in, school or class the student is in, on student performance on Math test. And also identifies to most efficient ways to improve student's performance from the following three ways: 1: Providing funds to poorly performing schools. 2: Gives extra training to teachers from poorly performing class. 3: Pay more attention to weak students.

The data used is called School leaver's data and it is obtained from the University of Bristol. Figure 1 gives a glimpse of the data. There are some key variables in the dataset: gender of the students(gender), social class(socialClass), math test score(math), grade the student is in(grade), class the student is in (classUnique), individual student ID(studentUnique).

Method

The data contains 3236 observations and 12 variables. Since we are interested in the performance of the math test, I have created a new variable called wrong, which indicates the number of questions students get wrong in the math test. Set wrong to be the dependent variable and plot its distribution below, we can see that it follows the Poisson distribution. To further analyze the research question, Poisson regression is fitted. Variables gender, socialClass and grade are set to be independent variables, and since multiple observation comes from one student, random effects by state, class and individual student are added to the regression. Here is the mathematics equation of the model.

Table 1: Figure 1

socialClass	ravensTest	student	english	math	year	classUnique	studentUnique	grade
absent	23	1	72	23	0	1 1	1 1 1	0
absent	23	1	80	24	1	1 1	1 1 1	1
absent	23	1	39	23	2	1 1	1 1 1	2
II	15	2	7	14	0	1 1	1 1 2	0
II	15	2	17	11	1	1 1	1 1 2	1
II	22	3	88	36	0	1 1	1 1 3	0

$$Y_{ijkl} \mid U_i, V_{ij}, W_{ijk} \sim Poisson(\lambda_{ijkl})$$
$$log(\lambda_{ijkl}) = X_{ijk}\beta + U_i + V_{ij} + W_{ijk}$$

where

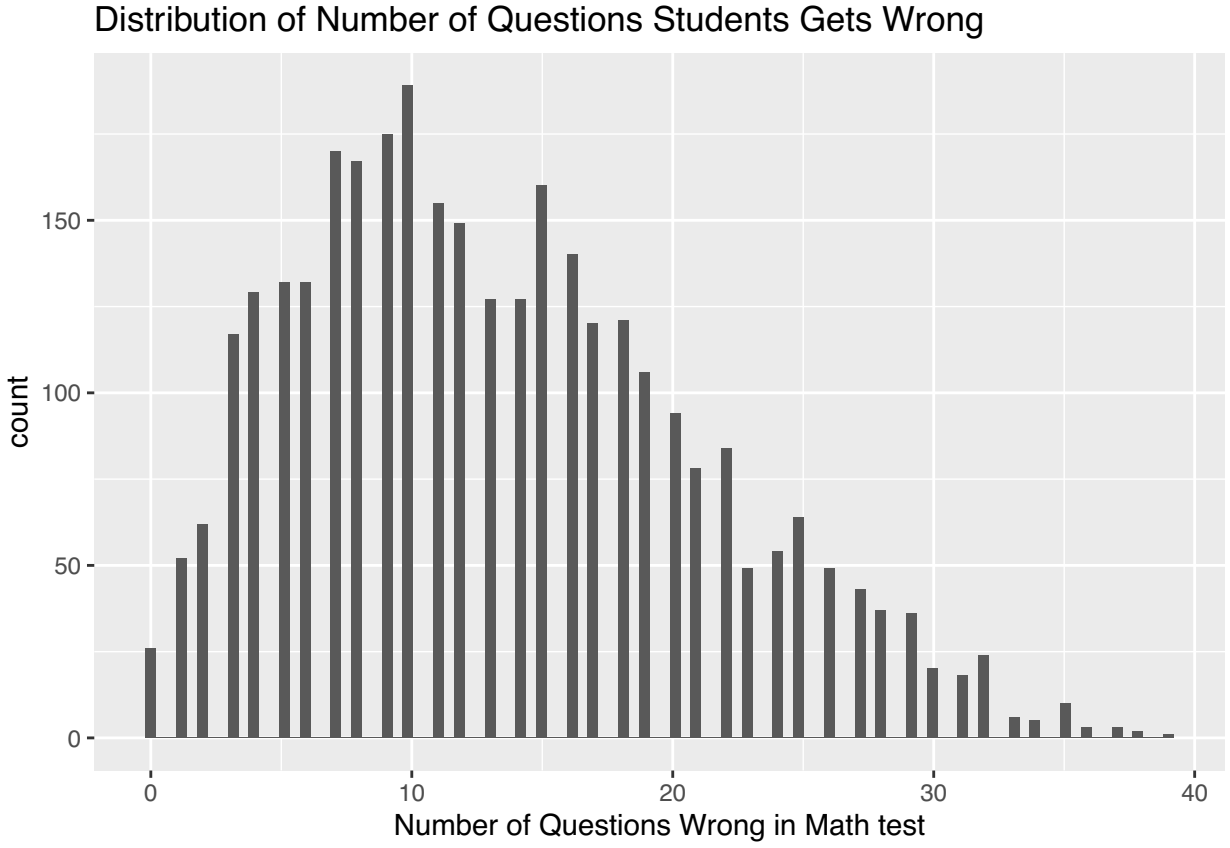
$$U_i \sim MVN(0, \tau_1\Sigma)$$
$$V_{ij} \sim MVN(0, \tau_2\Sigma)$$
$$W_{ijk} \sim MVN(0, \tau_3\Sigma)$$

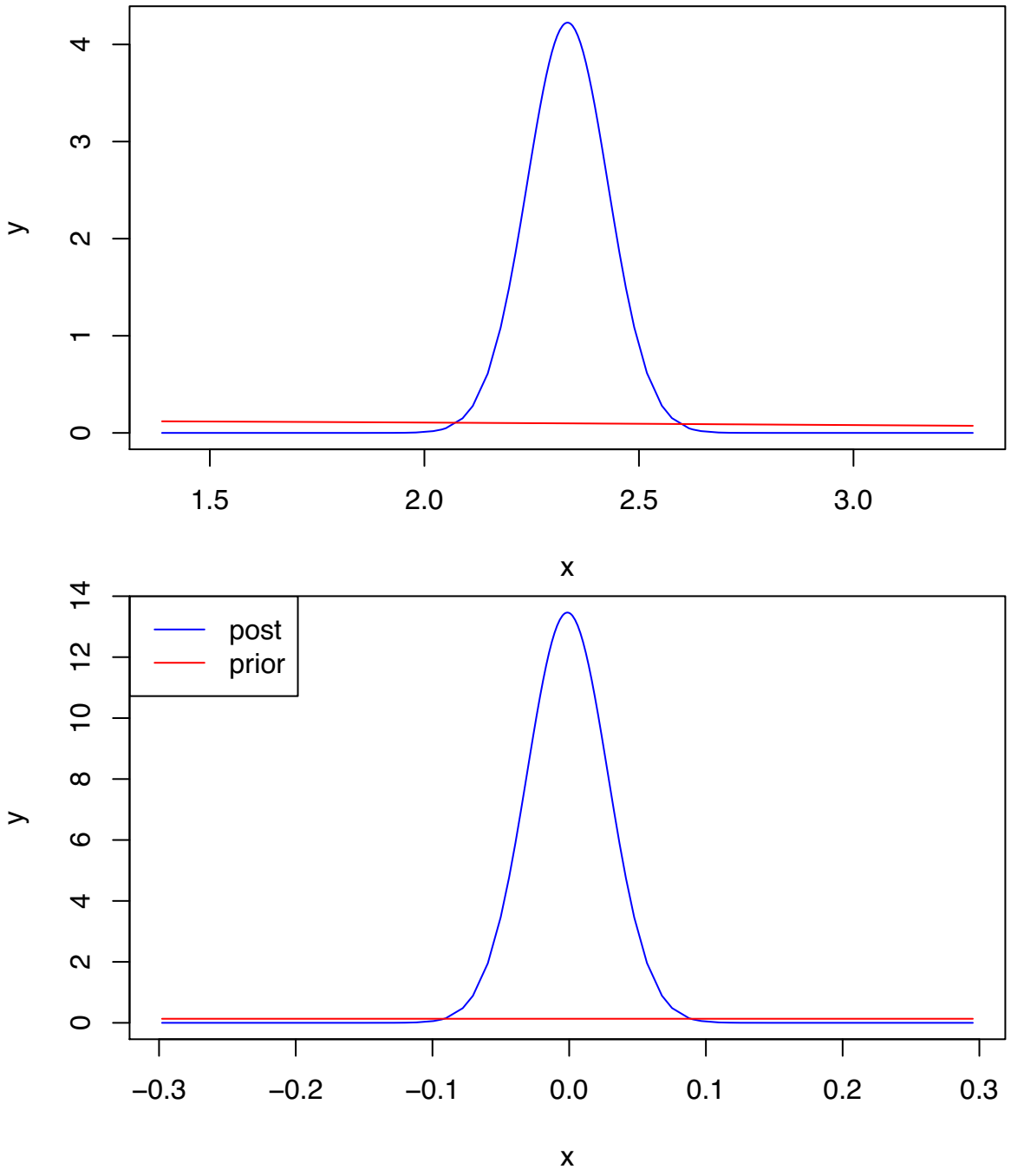
Priors :

$$\beta_0 \sim N(0, 10^2)$$
$$\beta_1 ... \beta_p \sim N(0, 8^2)$$
$$log(\tau_i) \sim logGamma()$$

In this model: Y_{ijkl} is the lth math score of the kth student from the jth class from the ith school. X_{ijk} are variables gender, socialClass and grade. U_i is the school specific random effect. V_i is the class in school specific random effect. W_i is the individual student in class in school specific random effect. βs are the parameters we are interested in.

In the meanwhile, we don't have any prior knowledge for the variables, the prior of the model is set base on my own approximation. The prior for the intercept is set with mean 0 and precision $frac110^2$, the prior for other βs are set with mean 0 and precision $frac18^2$, and the prior for random effect is default. The following two figures are the prior and posterior for intercept and one of the β





Result

The following tables provide a statistical analysis of the model. The first column of Figure 3 is the estimated median of the ratio of the number of questions wrong for that certain group compare to the reference group, and the second and third columns give their credible intervals (CI). The reference group for this model indicates females in grade 1 who have social class-I.

In figure 3 if the CI contains one, then its corresponding subgroup is not significant, and hence doesn't imply any useful information. Therefore, we focus on subgroups which their CI doesn't contain 1 (figure 2). According to figure 3, the number of questions that get wrong for socialClassIIIm is 1.359 times to those for the reference group. Therefore students from socialClassIIIm have lower math scores than students from the

Table 2: Figure 2

	Explanation of groups
socialClassIIIm	people who have a manual occupation
socialClassIV	people who have social class 4
socialClassV	people who have social class 5
socialClasslongUnemp	people who are long-term-unemployed
socialClasscurrUnemp	people who are currently unemployed
socialClassabsent	people whose father is absent
grade 2	people in grade 3

reference group. The same pattern observed for groups socialClassIV, socialClassV, socialClasslongUnemp, socialClasscurrUnemp, socialClassabsent; the questions that get wrong for these groups are 1.307, 1,497, 1.417, 1.483, 1.404 times to the questions get wrong for reference group respectively. In the contrast, for people who are in year3, the number of questions they get wrong is 0.656 times to reference group's, hence students in year 3 perform better than the reference group. To sum up, we don't have enough information to conclude that gender has an effect on math test performance, but we know that students in year 3 do better in math tests than students in year 1. As for social class, we don't have enough information to conclude anything for socialClassII and socialClassIIIn, but all other social class groups have poorer performance than socialclassI, especially for socialClassV has the worst performance.

In figure 4, it gives the estimated quantiles of standard deviation for school level, class level and individual student level. These standard deviations indicate how much variability there is between each level across all samples. The larger the standard deviation, the higher the variability, and the more efficient that the math test score would improve when that level received help. The estimated median of standard deviation for school, class and individual is 0.008, 0.175, 0.457 respectively where individual student level has the highest standard deviation, which means giving more attention to individual students who have weak performance would improve the overall math score the most. Besides, the 0.025-0.975 quantile range for an individual student is the highest and with no overlap with the school's [0.004, 0.025] and class's [0.134, 0.228]. Therefore, the conclusion is strengthened that giving more attention to individual students who have weak performance is the most efficient way to improve overall math performance.

Summary

This research explored what factors that most influent student's math test scores and provide a managerial solution to improve students' performance. Base on the school leaver's data from the University of Bristol and the model stated above, we get an unexpected result. Research shows that social class and grades the student has a significant effect on math grade; Compare with students from grade 1, grade-3 students, in general, have a better performance on math. On the other hand, students from socialclass-I have the best performance on the math test, people who have social class 4 or social class 5, people who are unemployed and people whose father is absent perform not as good as students from socialclass-I. Whereas students from socialClassV perform the worst. There are three ways to improve the overall performance, they are, 1: Fund poorly performing schools. 2: Give extra training to teachers from poorly performing class. 3: Pay more attention to weak students. Research shows that instead of identifying poorly performing schools or classes, it is more worthwhile to identify and provide more help to weak individuals.

Table 3: Figure 3:Fix Effect for the Model

	0.5quant	0.025quant	0.975quant
Baseline prob	0.912	0.895	0.925
genderm	0.999	0.942	1.059
socialClassII	0.983	0.808	1.197
socialClassIIIn	1.199	0.976	1.472
socialClassIIIm	1.359	1.129	1.636
socialClassIV	1.307	1.066	1.603
socialClassV	1.497	1.214	1.846
socialClasslongUnemp	1.417	1.140	1.762
socialClasscurrUnemp	1.483	1.111	1.978
socialClassabsent	1.404	1.157	1.703
grade1	0.998	0.976	1.019
grade2	0.656	0.639	0.673

Table 4: Figure 4: Mixed Effect for the Model

	0.5quant	0.025quant	0.975quant
SD for school	0.008	0.004	0.025
SD for classUnique	0.175	0.134	0.228
SD for studentUnique	0.457	0.434	0.481

Smoking

Introduction

This research explores the smoking pattern among American youths. Specialists believe there exists some variation amongst the US states and schools, and they also hypothesize that variation of states is greater than variations among schools, thus if tobacco control wants to reduce the population who smokes, they should target the states with the most smoking rather than schools. Also, they believe that Rural-urban differences are more significant than differences between states. Base on prior research, we already knew that age is a significant factor influencing smoking. In this research, we would categorize the effect of age by race, sex, and regions and further explore the smoking pattern. Data from the 2014 American National Youth Tobacco Survey is used. It contains 22007 observations and 160 variables, however, a subset of data with only the key variables are created and shown below in figure 5.

In Figure 5, variable y indicates whether the people smoke and ageFac is the age of the individual.

Method

To test the above hypotheses, set y to be the dependent variable and ageFac, Race, RuralUrban, sex to be the independent variables. To further analyze the hypotheses, a logistic model is fitted, and since we are interested in the variation between states and schools, two random effects are added to the model. The mathematical equation of the model is as follows.

Table 5: Figure 5

Age	ever_cigarettes	Sex	Race	state	school	RuralUrban	y	ageFac
18	TRUE	F	white	AL	mdr_00013045	Rural	1	18
18	TRUE	M	pacific	AL	mdr_00013045	Rural	1	18
16	TRUE	M	white	AL	mdr_00013045	Rural	1	16
17	TRUE	F	white	AL	mdr_00013045	Rural	1	17
17	FALSE	F	white	AL	mdr_00013045	Rural	0	17
15	FALSE	F	white	AL	mdr_00013045	Rural	0	15

$$Y_{ijk} \mid U_i, V_{ij} \sim Bernolli(\lambda_{ijk})$$
$$logit(\lambda_{ijk}) = \mu + X_{ijk}\beta + U_i + V_{ij}$$

where

$$U_i \sim MVN(0, \tau_1\Sigma)$$
$$V_{ij} \sim MVN(0, \tau_2\Sigma)$$

$Priors$:

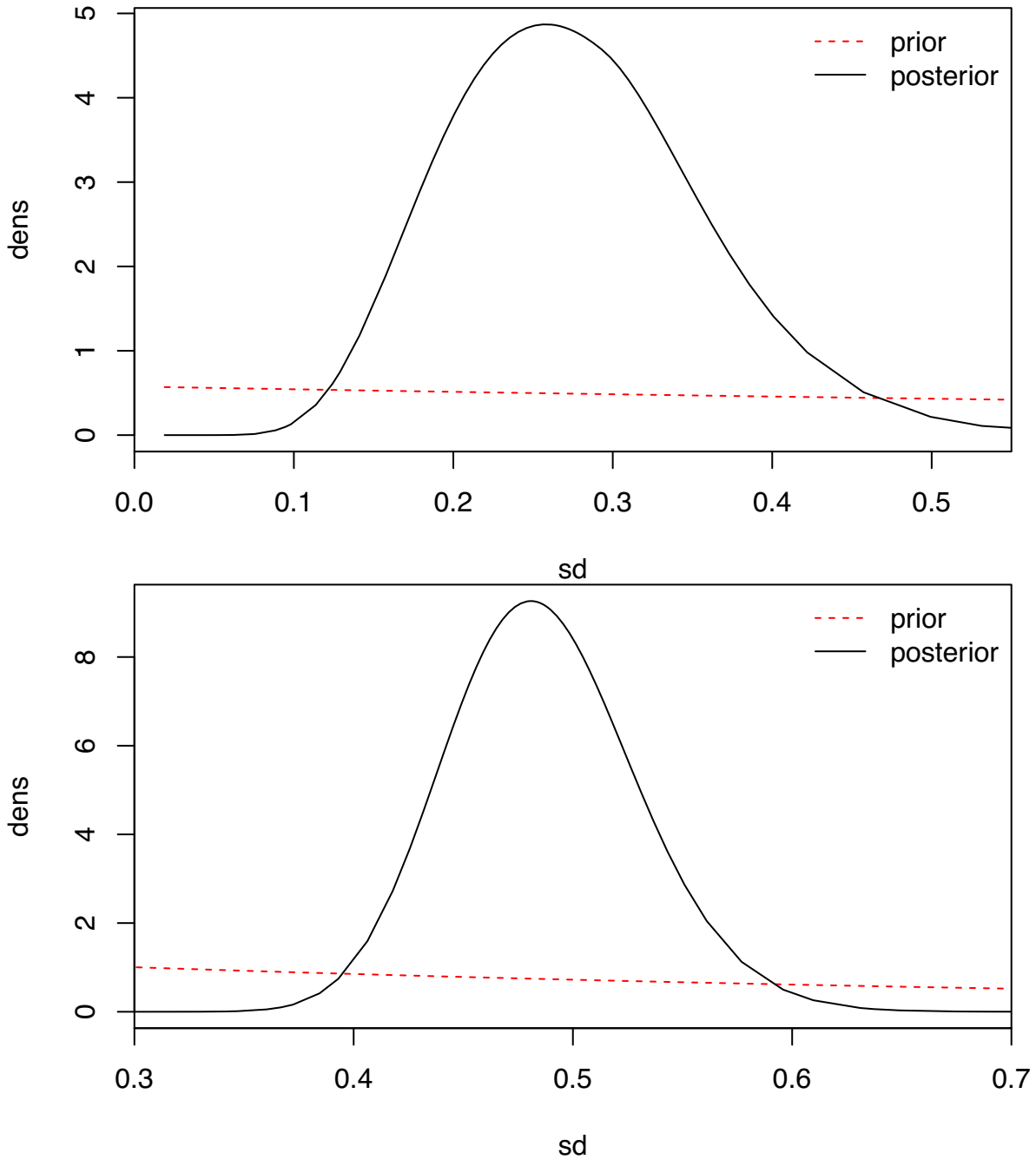
$$P(\sigma_1 > 4) = 0.1$$
$$P(\sigma_2 > 1.4) = 0.1$$

where

$$\sigma = \frac{1}{\sqrt{\tau}} \; and$$
$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}exp(-\lambda\tau^{-1/2}) \; \tau > 0$$

In this model, Y_{ijk} is the kth student from the jth school from the ith state. X_{ijk} are variables ageFac, Race, RuralUrban and sex $logit(p)$ represents the log of odds, it is the log of odds of smoking cigarette. $\lambda_{ijk}$is the estimated probability that the kth individual smokes. U_i is the state specific random effect. V_i is the school in state specific random effect. βs are the parameters we are interested in.

Collaborating scientists provide that the variability of the rate of smoking between states are in general between two or three times when compared with each other, and it is not common to see that the variability is higher than 5 times or 10 times even when comparing the healthiest state to the worst state. Therefore I decided to set the prior for state to be $p(\sigma > 4) = 0.1$ which can be interpreted as that the probability of the rate of smoking excessing 400% when comparing two states is approximately equal to 10%. However, within a given state, the typical variability of the rate of smoking between schools is 1.1 or 1.2 times when comparing with each other. Whereas when comparing the worst school with the healthiest school, the rate is at most 1.5%. Hence, I set the prior for school to be $P(\sigma > 1.4) = 0.1$ which means that the likelihood of the rate of smoking excessing 1.5 times when comparing two schools is approximately 10%. The following figures give the plot for prior and posterior for state and school



Result

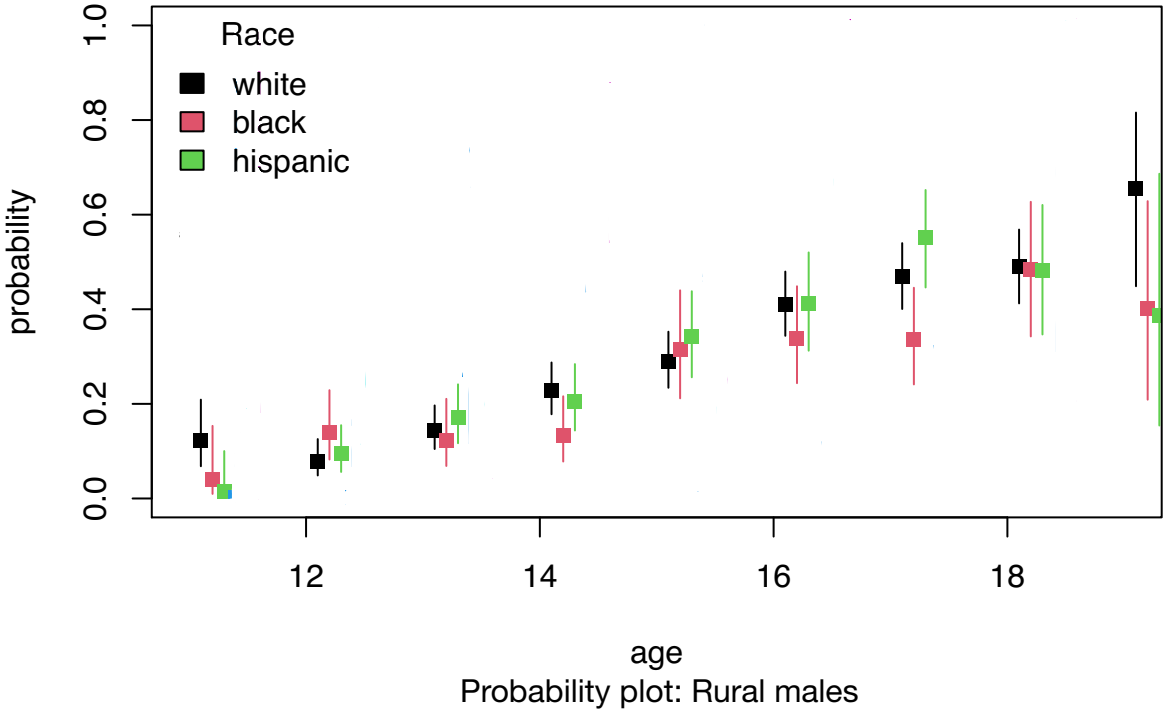
First, let's address the secondary task. Focus on figure 7, we know the reference group for this model is 14-year-old white males who live in urban areas. We observed in figure 7 that the odds of smoking cigarettes for Black males live in rural areas whose age is 11 is 0.799 times comparing with the reference group. Hence 11-year-old Black males live in rural areas are less likely to smoke. The same pattern is observed for 19-year-old Black males who live in rural areas. In contrast, Black males live in rural areas whose age is in between 12-18 has a higher odds of smoking comparing with the reference group. A similar pattern observed for race Hispanic; Hispanic males that live in rural areas whose age is 11, 12 or 19 has a lower odds of smoking cigarettes comparing with the reference group, but those whose age is between 13-18 are more likely to smoke

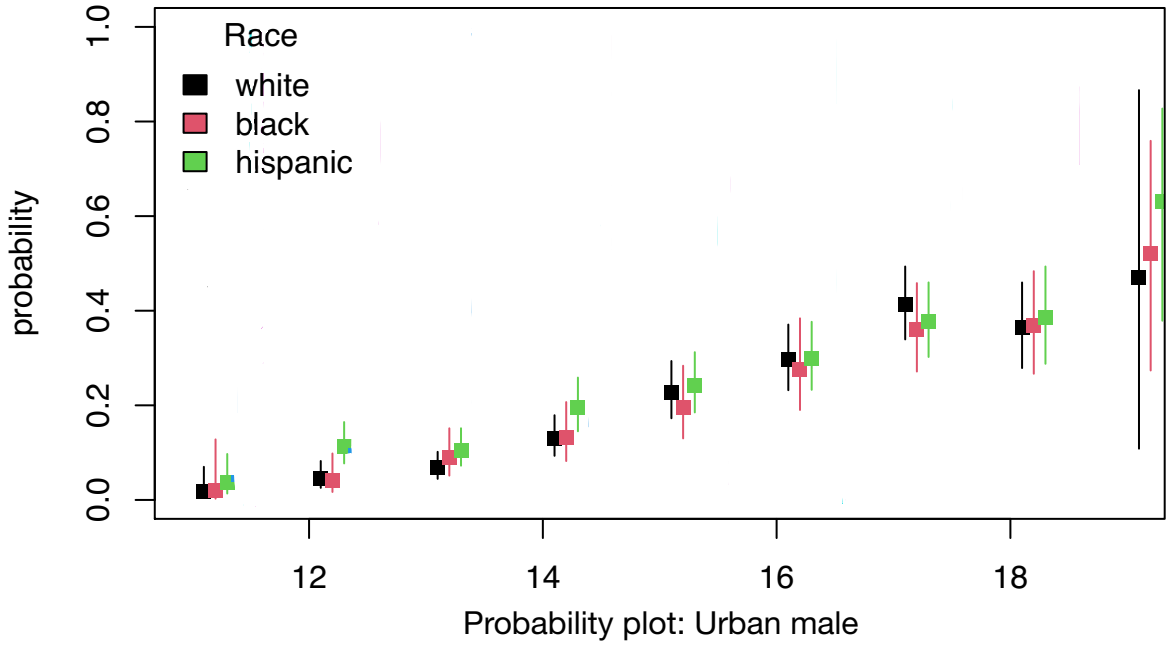
cigarettes. In addition, if we focus the age effect on race and further categorize by gender, we found that except black females whose age is 12 that have a higher odds of smoking cigarettes when comparing to the reference group, all other age groups (11, 13-19) have a much lower odds of smoking cigarettes. Whereas all Hispanic females have lower odds of a smoking cigarette than the reference group.

Furthermore, by observing the probability plot, we can see the probability of smoking for white males living in both rural and urban areas increases as the age increases. The same pattern observed for black males living in rural and urban area and hispanic males living in urban area as well. Whereas the probabiliry of smoking for hispanic males living in rural area first increases as age reaches 17 than decreases as age go to 19. On the other aspect, for urban hispanic female, rural black female , urban black females, and rural white females, their probability of smoking is increases as age increases. For urban white females have the same pattern but with a drastic increase of smoking when reach the age 19. Whereas for rural hispanic females, they also have the same pattenr but with a big decreases of smoking when reach the age 19

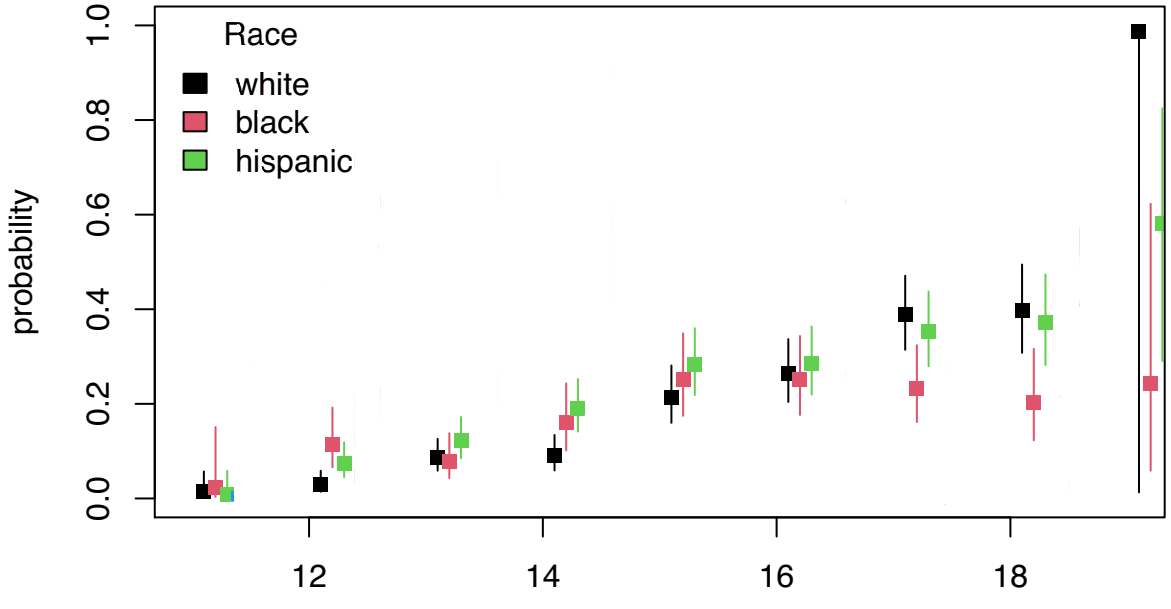
In figure 8, the first column of figure 8 gives the estimated median of standard deviation for state level and school level, and the second and third columns give their credible intervals. As indicated in figure 8, the standard deviation for school level is greater than the standard deviation for state level, which means that the variation among schools is greater than variation among states. If we look at the credible intervals, even though the credible interval for school level [0.406, 0.577] is higher than the credible interval for state [0.141, 0.457], there is a small overlap of the CI. Therefore the variation for school is in general higher than the variation for states, but there exists a small chance that it isn't. We recommend tobacco control to target the schools to reduce the rate of student smoking, but it is not certain.

The first column of figure 7 gives the estimated median for the log ratio of that subgroup compare with the reference group, and the second and third column gives the credible interval. To compare the differences between Rural-urban and states, we extract the odd ratio for rural-urban, 1.977, and the standard deviation for state, 0.272. A one standard deviation change in the random effect result in a 1.977 change in rural-urban and a 0.272 change in state, hence Rural-urban differences is greater than the differences between states. In addition, their credible intervals are not overlapping([1.236, 3.158] and [0.141, 0.457]), therefore hypothesis two is correct, rural-urban differences are indeed greater than differences between states.





Probability plot: Urban male



Probability plot: Urban females

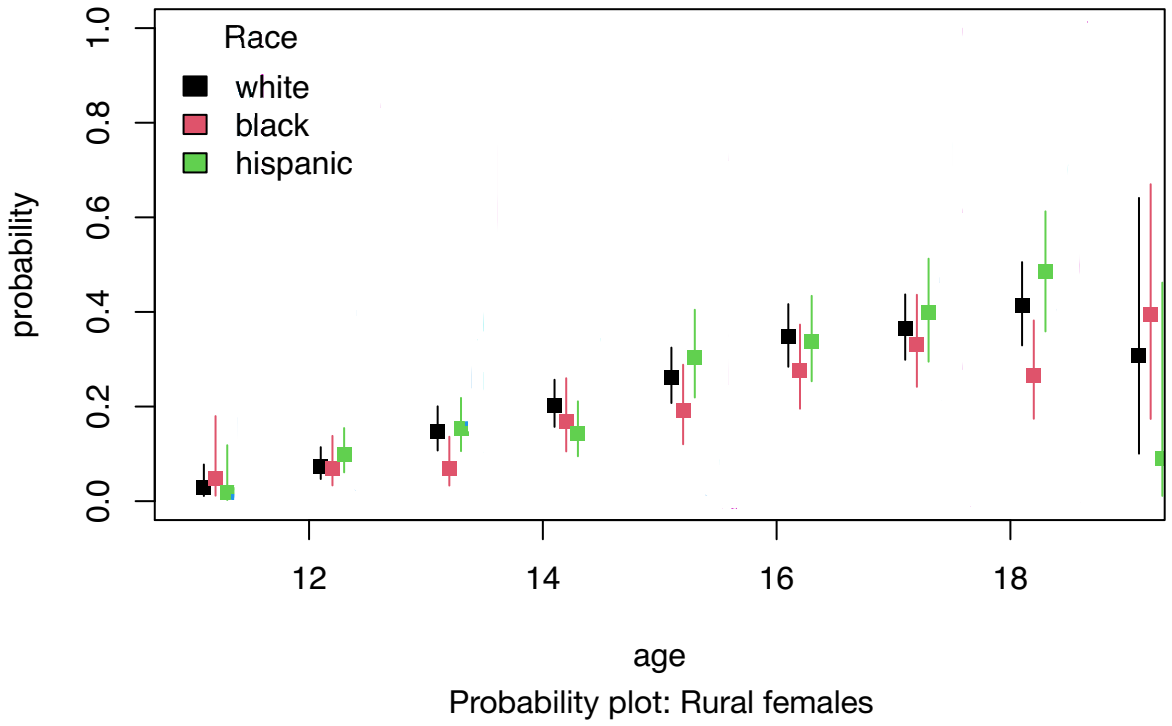


Figure 8

	0.5quant	0.025quant	0.975quant
SD for state	0.272	0.141	0.457
SD for school	0.484	0.406	0.577

Figure 7

	0.5quant	0.025quant	0.975quant
RuralUrbanRural	1.977	1.236	3.158
ageFac11:Raceblack	1.061	0.087	12.872
ageFac12:Raceblack	0.873	0.255	2.983
ageFac13:Raceblack	1.326	0.536	3.277
ageFac15:Raceblack	0.809	0.361	1.812
ageFac16:Raceblack	0.889	0.398	1.982
ageFac17:Raceblack	0.778	0.367	1.647
ageFac18:Raceblack	0.997	0.439	2.262
ageFac19:Raceblack	1.202	0.119	12.140
ageFac11:Racehispanic	1.285	0.216	7.627
ageFac12:Racehispanic	1.630	0.708	3.748
ageFac13:Racehispanic	1.000	0.496	2.017
ageFac15:Racehispanic	0.671	0.364	1.237
ageFac16:Racehispanic	0.625	0.342	1.140
ageFac17:Racehispanic	0.528	0.293	0.953
ageFac18:Racehispanic	0.674	0.336	1.349
ageFac19:Racehispanic	1.185	0.123	11.443
ageFac11:Raceblack:RuralUrbanRural	0.552	0.027	11.191
ageFac12:Raceblack:RuralUrbanRural	4.224	0.884	20.160
ageFac13:Raceblack:RuralUrbanRural	1.212	0.332	4.421
ageFac15:Raceblack:RuralUrbanRural	2.693	0.844	8.577
ageFac16:Raceblack:RuralUrbanRural	1.605	0.522	4.929
ageFac17:Raceblack:RuralUrbanRural	1.420	0.479	4.209
ageFac18:Raceblack:RuralUrbanRural	1.890	0.570	6.258
ageFac19:Raceblack:RuralUrbanRural	0.567	0.038	8.345
ageFac11:Racehispanic:RuralUrbanRural	0.098	0.006	1.559
ageFac12:Racehispanic:RuralUrbanRural	0.856	0.258	2.835
ageFac13:Racehispanic:RuralUrbanRural	1.390	0.517	3.729
ageFac15:Racehispanic:RuralUrbanRural	2.169	0.890	5.281
ageFac16:Racehispanic:RuralUrbanRural	1.852	0.757	4.526
ageFac17:Racehispanic:RuralUrbanRural	3.008	1.248	7.241
ageFac18:Racehispanic:RuralUrbanRural	1.647	0.587	4.621
ageFac19:Racehispanic:RuralUrbanRural	0.322	0.021	5.015
ageFac11:Raceblack:SexF	0.799	0.023	27.170
ageFac12:Raceblack:SexF	2.572	0.492	13.418
ageFac13:Raceblack:SexF	0.350	0.094	1.297
ageFac15:Raceblack:SexF	0.791	0.253	2.466
ageFac16:Raceblack:SexF	0.542	0.177	1.655
ageFac17:Raceblack:SexF	0.317	0.107	0.938
ageFac18:Raceblack:SexF	0.201	0.060	0.675
ageFac19:Raceblack:SexF	0.002	0.000	15.469
ageFac11:Racehispanic:SexF	0.190	0.009	3.933
ageFac12:Racehispanic:SexF	0.686	0.187	2.510
ageFac13:Racehispanic:SexF	0.616	0.225	1.684
ageFac15:Racehispanic:SexF	0.908	0.371	2.221
ageFac16:Racehispanic:SexF	0.746	0.310	1.796
ageFac17:Racehispanic:SexF	0.682	0.287	1.618
ageFac18:Racehispanic:SexF	0.561	0.209	1.508
ageFac19:Racehispanic:SexF	0.006	0.000	50.825

```
#install.packages("INLA", repos=c(getOption("repos"), INLA="https://inla.r-inla-download.org/R/stable"), dep=TRUE)
library("INLA")
#install.packages("glmmTMB", type="source")
library(glmmTMB)
library(dplyr)
library(ggplot2)
library(cowplot)
library(kableExtra)
#read data
school=read.fwf("file:///Users/aichenliu/Desktop/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$year)
school=school %>%
  mutate(wrong = 40-math)
formula = wrong ~ gender + socialClass + grade + f(school,model = "iid") + f(classUnique,model = "iid") +
f(studentUnique,model = "iid")
fResP = inla(formula,
 data=school, family='poisson',
 control.fixed = list(
 mean = 0, mean.intercept=0,
 prec = 8^(-2), prec.intercept = 10^(-2)
 ),
 control.compute = list(return.marginals=TRUE)
 )
#knitr::kable(rbind(fResP$summary.fixed[,c("mean","0.025quant","0.975quant")],Pmisc::priorPostSd(fResP)$summary[,c(1,3,5)
]),digits = 3)
table_fix = rbind('Baseline prob' = 1/(1+exp(-fResP$summary.fixed[1,c(4,3,5)])),exp(fResP$summary.fixed[-1,c(4,3,5)]))
table_r = Pmisc::priorPostSd(fResP)$summary[,c(4,3,5)]


exp(fResP$summary.fixed[,c("0.5quant","0.025quant","0.975quant")])
ggplot(school, aes(x= wrong))+geom_histogram(bins=100)+ xlab("Number of Questions Wrong in Math test")
+ggtitle("Distribution of Number of Questions Students Gets Wrong")
```
```
```{r,, echo=FALSE,message=FALSE}
#names(fResP$marginals.fixed)

for(D in c('(Intercept)', 'genderm')) {
plot(fResP$marginals.fixed[[D]],type='l',col='blue')
xseq = fResP$marginals.fixed[[D]][,'x']
lines(xseq, dnorm(xseq, sd=3), type='l', col='red')
}
legend('topleft',lty=1,col=c('blue','red'),legend=c('post','prior'))
table11 <- matrix(c("people who have a manual occupation",
                    "people who have social class 4",
                    "people who have social class 5",
                    "people who are long-term-unemployed",
                    "people who are currently unemployed",
                    "people whose father is absent",
                    "people in grade 3"),
                  ncol = 1,byrow = TRUE)
row.names(table11) <-
c("socialClassIIIm","socialClassIV","socialClassV","socialClasslongUnemp","socialClasscurrUnemp","socialClassabsent","grad
e 2")
colnames(table11) <- c("Explanation of groups")
table11<-as.table(table11)
```
```

```
knitr::kable(head(school[,c(4,5,6,7,8,9,10,11,12)]),caption = "Figure 1", align = "ccccccccccc")
knitr::kable(table11,caption = "Figure 2" )
knitr::kable(table_fix,caption = "Figure 3:Fix Effect for the Model", digits=3)
knitr::kable(table_r,caption = "Figure 4: Mixed Effect for the Model",digits=3)
#read data
smokeFile = 'smoke2014.RData'
if(!file.exists(smokeFile)){
        download.file(
                        "http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
                        smokeFile)
}
load(smokeFile)
#smoke[1:3, c("Age", "ever_cigarettes", "Sex", "Race",
#"state", "school", "RuralUrban")]

forInla = smoke[smoke$Age>10,c('Age','ever_cigarettes','Sex','Race',
    'state','school', 'RuralUrban')]
forInla = na.omit(forInla)
forInla$y = as.numeric(forInla$ever_cigarettes)
forInla$ageFac = relevel(factor(forInla$Age), '14')

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

fitS1 = inla(y ~  ageFac*Race*RuralUrban*Sex+f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec",
param = c(4, 0.1))))+f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",
param = c(1.4, 0.1)))),data = forInla, family = "binomial",
control.inla = list(strategy = 'Gaussian'), lincomb = forLincombs)

table_fix2 = rbind('Baseline prob' = 1/(1+exp(-fitS1$summary.fixed[1,c(4,3,5)])),exp(fitS1$summary.fixed[-1,c(4,3,5)]))
sub_table_fix2 =
table_random2 = Pmisc::priorPostSd(fitS1)$summary[,c(4,3,5)]
knitr::kable(head(forInla),caption = "Figure 5", align = "ccccccccc")


theSD = Pmisc::priorPostSd(fitS1)
do.call(matplot,theSD$state$matplot)
do.call(legend,theSD$legend)

do.call(matplot,theSD$school$matplot)
do.call(legend,theSD$legend)
knitr::kable(table_fix2[c(15,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,1
24,125,126,127,128,129,130,131,132,133,134,135,136,137,138,139),],caption = "Figure 7", align = "ccccccccc",digits = 3)
knitr::kable(table_random2,caption = "Figure 8", align = "ccccccccc",digits = 3)
#knitr::kable(table_fix2,caption = "Figure 7", align = "ccccccccc",digits = 3)

# create matrix of predicted probabilities
theCoef = exp(fitS1$summary.lincomb.derived[, c("0.5quant",
"0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$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", xlim = c(11,19),ylim = c(0,
1), pch = 15, col = toPredict[toPlot, "Race"])
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")
# create matrix of predicted probabilities
theCoef = exp(fitS1$summary.lincomb.derived[, c("0.5quant",
"0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$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 ==
"Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
xlab = "age", ylab = "probability", xlim = c(11,19),ylim = c(0,
1), pch = 15, col = toPredict[toPlot, "Race"])
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")
# create matrix of predicted probabilities
theCoef = exp(fitS1$summary.lincomb.derived[, c("0.5quant",
"0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$Age = as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX = as.numeric(toPredict$Race)/10
toPredict$x = toPredict$Age + toPredict$shiftX
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban ==
"Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
xlab = "age", ylab = "probability", xlim = c(11,19),ylim = c(0,
1), pch = 15, col = toPredict[toPlot, "Race"])
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")
# create matrix of predicted probabilities
theCoef = exp(fitS1$summary.lincomb.derived[, c("0.5quant",
"0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by chewing harm group
toPredict$Age = as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX = as.numeric(toPredict$Race)/10
toPredict$x = toPredict$Age + toPredict$shiftX
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban ==
"Rural"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
xlab = "age", ylab = "probability", xlim = c(11,19),ylim = c(0,
1), pch = 15, col = toPredict[toPlot, "Race"])
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")
```