Homework 2, Mixed effects models

Huifeng Wu

Bayesian Analysis on Math Performance

Introduction

Parents value more and more on education for their children nowadays, and student performance has been a key indicator in ranking criteria of schools. Given the school leaver's dataset from the University of Bristol, this report delivers a Bayesian analysis to the school board. We are using a generalized mixed effect model to explore the most important influences on student performance on math tests, which include but not limited to Social Class, grade the student is in, and differences between schools. Moreover, we provide a data-driven solution on how to better improve overall student performance by either the levels of schools, classrooms or individual students.

Method

With some exploratory analysis beforehand, we assumed that the response $40 - Math$ follows the Poisson distribution. A generalized mixed effect model of log link is proposed below to interpret the data, and that models the mean number of questions a student gets wrong in a math test, say λ. More specifically, the model is composed of random effects and fixed effects separately. Y_{ijk} denotes the number of mistaken questions in a math test answered by individual k in class j of school i. $X_{ijk}\beta$ indicates the fixed effects associated with the aforementioned covariates, whereas U_i, V_{ij}, Z_{ijk} are i.i.d random variables representing the random effects of unique schools, classes and students, respectively.

$$\begin{aligned} Y_{ijk} &\sim \text{Poisson}(\lambda_{ijk}) \\ \log(\lambda_{ijk}) &= X_{ijk}\beta + U_i + V_{ij} + Z_{ijk} \\ U_i &\sim \text{i.i.d Normal}(0, \sigma_u^2) \\ V_{ij} &\sim \text{i.i.d Normal}(0, \sigma_v^2) \\ Z_{ijk} &\sim \text{i.i.d Normal}(0, \sigma_z^2) \end{aligned}$$

A Bayesian inference allowed us to integrate prior knowledge with the hyperparameters σ_u, σ_v and σ_z, which are the standard deviations of different types of randomness in an observation. We believe it is typical to see a student could perform 20% better or worse than another student. We subjectively fixed all σs to follow the Exponential(0.8) distribution using penalized complexity priors for precision in R-INLA, with parameters log(1.2) and 0.5. Equivalently, P(σ>log(1.2))=0.5.

Results

Table 1 delivers the results of standard deviations (SD) for schools, classes and students. In terms of relative rates, the student-level variation (1.578) is the largest, followed by class (1.200) and school (1.044) random effects in a descending order. That being said, math test scores are most inconsistent among classmates, rather than schoolmates or ones across different schools. It is most crucial to help weak students to improve their math scores. To further interpret, we could expect the relative rate of mistaken questions being 1.578 between students within the same class in a school, when keeping other covariates constant.

For the fixed effects shown in Table 2, the intercept term, particular Social Classes (III, IV, V, long-term unemployed, not currently employed, and father-absent) and grade 3 are significant covariates to math performance in our model. The intercept portraits the baseline group here, a grade 1 female student in Social Class I; such a student is expected to get 10.344 math questions wrong in a test in average. When keeping other variables unchanged, it is evident that a student from low Social Class tend to make more mistakes compared to the reference group, whereas one in Social Class II and III non-manual are indifferent to the reference group. Taking account for other features, the number of wrong answers for a Social Class V student is predicted to be 1.488 times of the baseline group, but the 95% credible interval include 1 for Social Classes II and III non-manual. Also, students in grade 3 show better math performance that they are expected to make 0.656 times fewer mistakes than grade 1, while holding other variables constant.

Table 1: Random effects posterior means and 95% credible intervals in terms of relative rates

	mean	0.025quant	0.975quant
SD for school	1.044	1.008	1.108
SD for classUnique	1.200	1.147	1.265
SD for studentUnique	1.578	1.543	1.619

Table 2: Fixed effects posterior means and 95% credible intervals in terms of relative rates

	mean	0.025quant	0.975quant
(Intercept)	10.344	8.577	12.469
genderm	0.998	0.942	1.058
socialClassII	0.984	0.808	1.197
socialClassIIIn	1.197	0.975	1.471
socialClassIIIm	1.358	1.127	1.635
socialClassIV	1.307	1.065	1.603
socialClassV	1.488	1.207	1.836
socialClasslongUnemp	1.413	1.136	1.757
socialClasscurrUnemp	1.483	1.112	1.978
socialClassabsent	1.402	1.155	1.702
grade1	0.998	0.976	1.019
grade2	0.656	0.639	0.673

Conclusion

Overall, the data suggest that administrators should identify individual weak students and provide them with extra practice, for the best interest. Secondly, it is worthwhile training teachers who are less experienced in teaching, and then consider to offer additional funding to poorly performance schools, for the least priority. Our research shows that Social Class in the UK has significant influences on student math performance, especially negative in math scores for ones from lower Social Class. Moreover, students are predicted to do better gradually in math tests once they are in grade 3.

Appendix

```
## School leaver’s data ##
library(knitr)
library(glmmTMB)
library(INLA)

sUrl ="http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip"

school = read.fwf("./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)

# Below is a Generalized Linear Mixed Model fit to the data
schoolLme =glmmTMB::glmmTMB(
  math~gender+socialClass+grade+(1|school)+(1|classUnique)+(1|studentUnique),data=school)
# summary(schoolLme)
# hist(40-school$math,breaks =100)

# Bayersian
school_inla = inla((40-math) ~ gender+socialClass+grade+
                     f(school, model='iid',
                       hyper=list(prec=list(prior="pc.prec",param=c(log(1.2),0.5))))+
                     f(classUnique, model='iid',
                       hyper=list(prec =list(prior="pc.prec",param=c(log(1.2),0.5))))+
                     f(studentUnique, model='iid',
                       hyper=list(prec =list(prior="pc.prec",param=c(log(1.2),0.5)))),
                   family = "poisson",
                   data=school, control.inla = list(strategy='laplace', fast=FALSE))

table1 <-
  round(exp(Pmisc::priorPostSd(school_inla)$summary[, c('mean','0.025quant','0.975quant')]), 3)
table2 <-
  round(exp(school_inla$summary.fixed[, c('mean','0.025quant','0.975quant')]), 3)

# Table 1
kable(table1,
      booktabs = TRUE,
      caption = "Random effects posterior means and 95%
      credible intervals in terms of relative rates")
# Table 2
kable(table2,
      booktabs = TRUE,
      caption = "Fixed effects posterior means and 95%
      credible intervals in terms of relative rates")

# plot(school_inla$marginals.hyperpar$'Precision for state', type='l')
# plot(school_inla$marginals.hyperpar$'Precision for school', type='l')
```

```
#
# plot(theSd$school$posterior, type='l',
#      xlab='sd for school', ylab='dens', xlim = c(0,1), col='blue')
# lines(theSd$school$prior, col='blue', lty=2)
#
# plot(theSd$state$posterior, type='l',
#      xlab='sd for state', ylab='dens', xlim = c(0,1), col='blue')
# lines(theSd$state$prior, col='blue', lty=2)
```

The Effects of Factor Variations on Youth Smoking

Introduction

The 2014 American National Youth Tobacco Survey highlights a significant proportion of cigarette smoking history amongst students, and that could severely harm their health status at an early age. Taking the advantage of some prior knowledge about the dataset, the report evaluates two research hypotheses that greater variation is found in student smoking rates between states versus amongst schools, as well as rural-urban differences versus differences between states. Additionally, we investigate the effect of age on student smoking for different races, with varying sexes and rurality.

Method

The incidence of ever smoking cigarettes for a student, say Y_i, follows the Bernoulli distribution. Considering observations on the basis of schools or states, repeated measurements occur in the dataset. Thus, a comprehensive generalized mixed effect model of logit link is used below to answer the research questions.

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(\mu_{ij}) \\
\text{logit}(\mu_{ij}) &= X_{ij}\beta + U_i + V_{ij} \\
U_i &\sim \text{i.i.d Normal}(0, \sigma_u^2) \\
V_{ij} &\sim \text{i.i.d Normal}(0, \sigma_v^2)
\end{aligned}
$$

To investigate the hypotheses of interest, the model considers random effects of schools and states, and fixed effects featuring *sex*, *rurality*, *ethnicity* and *age*. Even though the first three variables are susceptible of confounders, their interaction effects are still taken into account for part of the total fixed effects. Age is transformed to categorical variable here. Analytically, Y_{ij} denotes the predicted probability of smoking by a student of school j in state i. $X_{ij}\beta$ indicates the fixed effects associated with the aforementioned covariates, whereas U_i and V_{ij} represents the random effects of unique states and schools, respectively.

We assumed U_i and V_{ij} are independent and identically distributed random variables that follow the Normal distribution. Based on prior information from collaborating scientists, it is unlikely to observe a 5-fold or 10 fold difference of smoking rates between the 'worst' and 'healthiest' states, although we may see the rate doubled or tripled for comparable students between states. Within a given state, the largest difference of smoking rates should be 50% at most between the 'worst' and 'healthiest' schools, and it is more prevalent to see differences from 10% to 20%. Therefore, we set σ_u to follow the Exponential(2.5) distribution, and σ_v to follow the Exponential(8.5) distribution. Both of the hyperparameters were fine-tuned using penalized complexity priors for precision in R-INLA, with parameters (log(10)/2.5, 0.1) and (log(1.5), 0.1), correspondingly. Mathematically, P(σ_u>log(10)/2.5)=0.1 and P(σ_v>log(1.5))=0.1 for the prior assumptions. Note that the scale factor of 2.5 roughly takes consideration into the odds of 'worst' and 'best' schools.

Last but not least, the proposed model illustrates the effect of age differences on smoking for specific races, such as white, Black, and Hispanic Americans. These effects are different by sex and by rurality, and we visualized that in comparable groups.

Results

Table 3 presents the results of standard deviations (SD) for schools and states, and Table 4 exhibits fixed effects outlining posterior means and their 95% credible intervals. Overall, the intercept term, particular ethnicity (Hispanic, white and Asian) and rurality and age are significant covariates to odds of student smoking. It is worth-noting that all interaction effects are insignificant to our model. The intercept describe the baseline group that a 11 year-old Caucasian male student living in a urban area; his predicted odds of smoking is 0.028.

To address the research hypotheses, we firstly found that the school-level variation (1.645) is larger than states (1.284). Another word, the events of student smoking are rather varied by school environment, instead of geographic variation (between states). Thus, policy makers should execute their tobacco control programs targeting ‘bad’ schools.

Later, we used odds ratios in order to compare rural-urban differences and differences between states. To further interpret, the odds of smoking increase by a factor of 1.714, considering comparable students who live in rural areas to those in cities. However, 1 SD change in the random effect of states merely leads to a multiplicative increase of 1.284 in the odds of smoking. In a natural scale, we say the odds ratio with rurality variation (from urban to rural) arises by 71% approximately, while 1 SD of random effects for states increase around 30%. Thus, these results can only prove the partial correctness of the research statement that rural-urban differences are much greater than states, since their credible intervals intersect with each other. The random effect of states may still exceed the rural-urban variation.

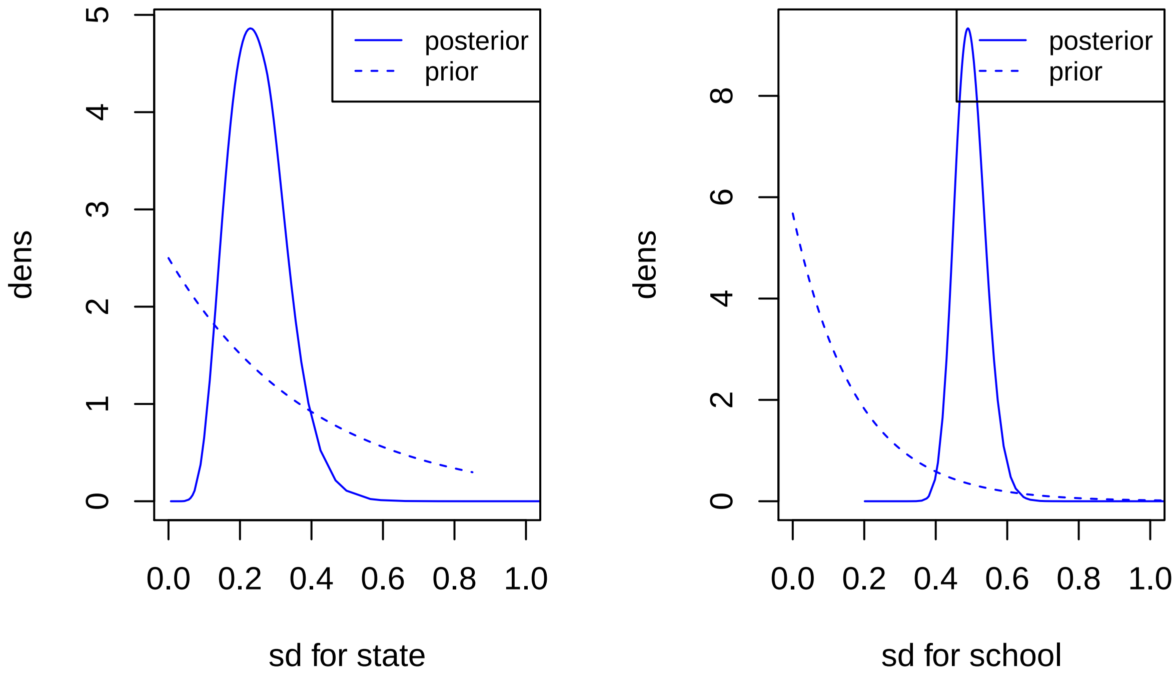
Table 3: Random effects posterior means and 95% credible intervals in terms of odds ratio

	mean	0.025quant	0.975quant
SD for state	1.284	1.122	1.530
SD for school	1.645	1.521	1.804

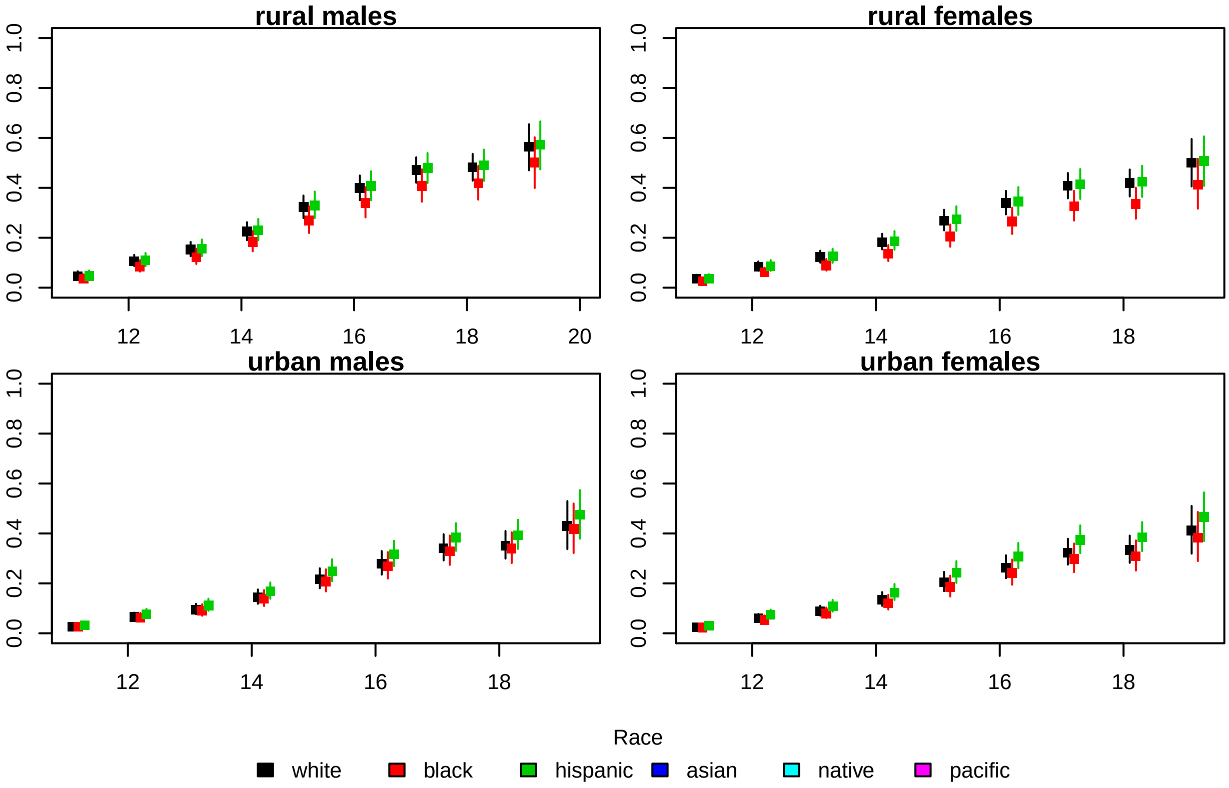
Table 4: Fixed effects posterior means and 95% credible intervals in terms of odds ratio

	mean	0.025quant	0.975quant
(Intercept)	0.028	0.019	0.042
RuralUrbanRural	1.718	1.334	2.207
Raceblack	0.948	0.754	1.191
Racehispanic	1.199	1.001	1.435
Raceasian	0.569	0.402	0.805
Racenative	1.410	0.747	2.657
Racepacific	1.782	0.708	4.480
SexF	0.924	0.780	1.093
ageFac12	2.496	1.729	3.602
ageFac13	3.762	2.629	5.383
ageFac14	6.037	4.225	8.624
ageFac15	9.943	6.890	14.343
ageFac16	13.872	9.600	20.038
ageFac17	18.604	12.876	26.874
ageFac18	19.456	13.380	28.284
ageFac19	27.093	16.645	44.083
RuralUrbanRural:Raceblack	0.817	0.592	1.126
RuralUrbanRural:Racehispanic	0.862	0.661	1.124
RuralUrbanRural:Raceasian	0.802	0.379	1.696
RuralUrbanRural:Racenative	1.271	0.562	2.873
RuralUrbanRural:Racepacific	1.268	0.334	4.801
RuralUrbanRural:SexF	0.837	0.671	1.043
Raceblack:SexF	0.937	0.695	1.264
Racehispanic:SexF	1.040	0.819	1.321
Raceasian:SexF	0.775	0.461	1.303
Racenative:SexF	1.188	0.490	2.881
Racepacific:SexF	0.335	0.056	2.008
RuralUrbanRural:Raceblack:SexF	0.967	0.634	1.475
RuralUrbanRural:Racehispanic:SexF	0.956	0.672	1.359
RuralUrbanRural:Raceasian:SexF	1.881	0.703	5.028
RuralUrbanRural:Racenative:SexF	0.474	0.135	1.668
RuralUrbanRural:Racepacific:SexF	1.706	0.164	17.741

We here use the following figure to elaborate on more details about the prior distributions. Both posterior distributions rather describe the absolute difference of odds between states and between school. Neither of the prior distributions do not look strong to the posterior distributions. Although the medians agree between prior and posterior distributions (median(σ_u^2)=0.583, median($\sigma_u^2|Y = y$)=0.429, and median(σ_v^2))=0.447, median($\sigma_v^2|Y = y$)=0.476), the tail probabilities part away (90th-quantile(σ_u^2)=0.865, 90th-quantile($\sigma_u^2|Y = y$)=4.458, and 90th-quantile(σ_v^2)=1.503, 90th-quantile($\sigma_v^2|Y = y$)=8.107).



Besides, we conveyed the differences in the effect of age on smoking for particular races, with respect to different genders and living areas. It is evident that smoking rate goes up in a linear trend with the age of students. Hispanic American students are leading in smoking rate, followed by Caucasian American and then African American students. Also, it is noticeable that the gap between Hispanic Americans and Caucasian American students is very small for those in rural area, and males seem to smoke more frequently than women.



Conclusion

Based on the 2014 American National Youth Tobacco Survey data, the report takes advantage of linear mixed modelling to differentiate effects contributed from randomness and fixed variations. Over the course, we learn that rurality, particular races and age exert significant influences on student smoking. Moreover, geographic variation of states in the smoking rate of students is substantially lower than variation amongst schools, so administrators should target schools for tobacco control. The in-between states variation on smoking is also lower than rural-urban differences. Ultimately, we found the strong correlation between age and the rate of students smoking.

Appendix

```
## Smoking ##
rm(list = ls())
library(INLA)

dataDir ="./data"
smokeFile =file.path(dataDir,"smoke2014.RData")
load(smokeFile)

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

# Some models
# fit1 = inla(y~RuralUrban*Age*Race+Sex+
#              f(state,model ="iid",
#                hyper =list(prec =list(prior ="pc.prec",param =c(99,0.05))))+
#              f(school,model ="iid",
#                hyper =list(prec =list(prior ="pc.prec",param =c(99,0.05)))),
#            data =forInla,
#            family ="binomial")
# rbind(fit1$summary.fixed[,c("mean","0.025quant","0.975quant")],
#       Pmisc::priorPostSd(fit1)$summary[,c("mean","0.025quant","0.975quant")])
#
# fit2 = inla(y~RuralUrban+ageFac+Race+
#              f(state,model ="iid",
#                hyper =list(prec =list(prior ="pc.prec",param =c(99,0.05)))),
#            data =forInla,
#            family ="binomial")
# rbind(fit2$summary.fixed[,c("mean","0.025quant","0.975quant")],
#       Pmisc::priorPostSd(fit2)$summary[,c("mean","0.025quant","0.975quant")])

# Final model
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(~RuralUrban*Race*Sex+ageFac, data=toPredict)))
final_fit= inla(y~RuralUrban*Race*Sex+ageFac+
                  f(state,model ="iid",
                    hyper =list(prec =list(prior="pc.prec",param=c(log(10)/2.5,0.1))))+
                  f(school,model ="iid",
                    hyper =list(prec =list(prior="pc.prec",param=c(log(1.5),0.1)))),
                data =forInla,
                family ="binomial",
                control.inla = list(strategy='gaussian'),
                lincomb = forLincombs)

table3 <-
  round(exp(Pmisc::priorPostSd(final_fit)$summary[, c('mean','0.025quant','0.975quant')]), 3)
table4 <-
  round(exp(final_fit$summary.fixed[, c('mean','0.025quant','0.975quant')]), 3)
```

```r
# Table 3
kable(table3,
      booktabs = TRUE,
      caption = "Random effects posterior means and 95%
      credible intervals in terms of odds ratio")
# Table 4
kable(table4,
      booktabs = TRUE,
      caption = "Fixed effects posterior means and 95%
      credible intervals in terms of odds ratio")

theSd= Pmisc::priorPostSd(final_fit)

# plot(final_fit$marginals.hyperpar$'Precision for state', type='l')
# plot(final_fit$marginals.hyperpar$'Precision for school', type='l')

par(mfrow = c(1,2))

plot(theSd$school$posterior, type='l',
     xlab='sd for school', ylab='dens', xlim = c(0,1), col='blue')
lines(theSd$school$prior, col='blue', lty=2)
legend("topright", legend = c("posterior", "prior"),
       lty = c(1,2), col ="blue", cex = 0.8)

plot(theSd$state$posterior, type='l',
     xlab='sd for state', ylab='dens', xlim = c(0,1), col='blue')
lines(theSd$state$prior, col='blue', lty=2)
legend("topright", legend = c("posterior", "prior"),
       lty = c(1,2), col ="blue", cex = 0.8)
```

```r
# median(theSd$state$prior)
# median(theSd$state$posterior)
# quantile(theSd$state$prior, 0.9)
# quantile(theSd$state$posterior, 0.9)
#
# median(theSd$school$prior)
# median(theSd$school$posterior)
# quantile(theSd$school$prior, 0.9)
# quantile(theSd$school$posterior, 0.9)

toPredict = expand.grid(ageFac=levels(forInla$ageFac),
                        RuralUrban=levels(forInla$RuralUrban),
                        Race=levels(forInla$Race),
                        Sex=levels(forInla$Sex))
index <- (toPredict$Race=="white" | toPredict$Race=="black" | toPredict$Race=="hispanic")

# create matrix of predicted probabilitiesthe
theCoef=exp(final_fit$summary.lincomb.derived[,c("0.5quant","0.025quant","0.975quant")])
theCoef=theCoef/(1+theCoef)
theCoef <- theCoef[index, ]
toPredict <- toPredict[index, ]

# create an x axis, shift age by categories groupto
toPredict$Age=as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX=as.numeric(toPredict$Race)/10
```

```r
toPredict$x=toPredict$Age+toPredict$shiftX

m <- matrix(c(1,2,3,4,5,5),nrow = 3,ncol = 2,byrow = TRUE)
layout(mat = m,heights = c(0.6,0.6,0.3))

par(mar = c(2,2,1,1))
# plot rural males
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(11,20),
     pch =15,col=toPredict[toPlot,"Race"],
     main="rural males")
segments(toPredict[toPlot,"x"],
         theCoef[toPlot,"0.025quant"],
         y1 =theCoef[toPlot,"0.975quant"],col =toPredict[toPlot,"Race"])

# plot rural females
toPlot=toPredict$Sex=="F" & toPredict$RuralUrban=="Rural"
plot(toPredict[toPlot,"x"],theCoef[toPlot,"0.5quant"],
     xlab ="age",ylab ="probability",
     ylim =c(0,1),
     pch =15,col=toPredict[toPlot,"Race"],
     main="rural females")
segments(toPredict[toPlot,"x"],
         theCoef[toPlot,"0.025quant"],
         y1 =theCoef[toPlot,"0.975quant"],
         col =toPredict[toPlot,"Race"])

# plot urban males
toPlot=toPredict$Sex=="M" & toPredict$RuralUrban=="Urban"
plot(toPredict[toPlot,"x"],theCoef[toPlot,"0.5quant"],
     xlab ="age",ylab ="probability",
     ylim =c(0,1),
     pch =15,col=toPredict[toPlot,"Race"],
     main="urban males")
segments(toPredict[toPlot,"x"],
         theCoef[toPlot,"0.025quant"],
         y1 =theCoef[toPlot,"0.975quant"],
         col =toPredict[toPlot,"Race"])

# plot urban females
toPlot=toPredict$Sex=="F" & toPredict$RuralUrban=="Urban"
plot(toPredict[toPlot,"x"],theCoef[toPlot,"0.5quant"],
     xlab ="age",ylab ="probability",
     ylim =c(0,1),
     pch =15,col=toPredict[toPlot,"Race"],
     main="urban females")
segments(toPredict[toPlot,"x"],
         theCoef[toPlot,"0.025quant"],
         y1 =theCoef[toPlot,"0.975quant"],
         col =toPredict[toPlot,"Race"])

plot(1, type = "n", axes=FALSE, xlab="", ylab="")
plot_colors <- c("blue","black", "green", "orange", "pink")
legend("top",inset = 0,fill =1:nlevels(toPredict$Race),
       legend =levels(toPredict$Race),bty ="n",title ="Race",
       cex=1, horiz = TRUE)
```