442hw1

"Yukun Gao"

2021/10/5

Q1.1

We study the dataset from http://www.bristol.ac.uk/cmm/learning/mmsoftware/data-rev.html#chem97 which contains chemistry test scores for UK students. We are interested in how important the differences of variation of scores are between schools, regions, sex and ages of student. By the first look at the data, the distribution of our response looks somewhat right skewed(Figure 1). It might be appropriate to use Gamma distribution while there should be no intercept. The sex and age of individuals should be treated as fixed effects since each individual is independent from another, whereas schools and regions should be treated as random effects since there might be some correlations between schools and regions. Thus, we should use GLMM model with Gamma distribution:

$$log(E(Y_{ijk})) = X_{ijk}\beta + U_i + V_{ij}$$
$$U_i \sim N(0, \sigma_1^2)$$
$$V_{ij} \sim N(0, \sigma_2^2)$$
$$\beta \sim N(0, 3^2 I)(assumed)$$
$$\sigma_1 \sim exp(\lambda_1)$$
$$\sigma_2 \sim exp(\lambda_2)$$
$$Y_{ijk} \sim Gamma(\frac{\mu_{ijk}}{v}, v)$$
$$Prob(v > 0.18) = 0.5$$

The reason $\sigma_1 \sim exp(\lambda_1)$ and $\sigma_2 \sim exp(\lambda_2)$ is that U_i and V_{ij} follow i.i.d Normal distribution. $E(Y_{ijk}) = \mu_{ijk}$ which represents the average grade deduction of kth student from $school_j$ in $region_i$. X_{ijk} represents the covariates of age and sex respectively, β represents the corresponding parameters of these covariates. U_i is the random effects from $region_i$, V_{ij} is the random effects from $school_j$ in $region_i$.

Q1.2

U_i and V_{ij} are random effects and they follow Normal distribution, thus σ_1 and σ_2 follows exponential distribution. To give a prior distribution for σ_2, suppose $E(Y_{ik}) = e^{X\beta+U_k+V_i}$ and $E(Y_{jk}) = e^{X\beta+U_k+V_j}$. The only difference between these two equation is the variation of school whereas each individual's characteristics and regions are fixed. Then $\frac{E(Y_{ik})}{E(Y_{jk})} = e^{V_i-V_j}$. According to the assumption that changes of 20% of grades are more likely, $e^{V_i-V_j} \approx 1.2$ and $V_i - V_j \approx 0.18$ which means the variation(σ_2) should be approximately 0.18. Since σ_2 follows exponential distribution, treat 0.18 as the median of the distribution and get $\frac{ln2}{\lambda} = 0.18 \Leftrightarrow \lambda_2 \approx 3.851$. By the similar strategy, fix each individual's characteristics and schools, $\frac{E(Y_{ak})}{E(Y_{bk})} = \frac{e^{X\beta+U_a+V_k}}{e^{X\beta+U_b+V_k}} = e^{U_a-U_b} = e^{1.2}$ and thus $\lambda_1 \approx 3.851$. As a result, the prior distributions for σ_1 and σ_2 are given by: $\sigma_1 \sim exp(3.851)$ and $\sigma_2 \sim exp(3.851)$. Additionally, our response follow Gamma distribution and we need to determine the prior for response. However, we don't know much about the prior from previous experience. Thus, we use the PC prior with 0.18 as median instead of log gamma prior.

Q1.3

To summarize, the standard deviation(SD) for region is 0.124 whereas the SD for school is 0.252. This means that the score variations between schools are larger than between regions. Thus, the differences between schools are more obvious than between regions. The 95% confidence interval(CI) of SD for school is (0.241,0.265) whereas the CI of SD for region (0.115,0.133). The CI of SD for school is larger than region while there is no overlap between them, which mean the SD are different between schools and regions. The conclusion is that the differences between schools are more important than between regions in affecting the variation of scores. Additionally, fixed other factors unchanged, as student's age increase by 1 year, the average score deduction becomes 91.4% of the score deduction have the age never increased. This means that student's score will increase as their age increases. Keep other factors constant, the deduction in male students' score is about 1.236($\frac{10.274}{8.311}$) times of female students' score deduction, which means if female students' score deduction is 1, male students' score deduction will be 1.236. This can be inferred from the exponential parameter estimates given in the following table:

Table 1: Summary of SD of Random Effects

	mean	sd	0.025quant	0.5quant	0.975quant	mode
SD for region	0.124	0.005	0.115	0.124	0.133	0.125
SD for school	0.252	0.006	0.241	0.251	0.265	0.250

Table 2: Summary of Fixed Effects(natural scale)

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
age	0.914	1.010	0.896	0.914	0.933	0.914	1
sexM	10.274	1.208	7.093	10.274	14.877	10.274	1
sexF	8.311	1.208	5.738	8.311	12.036	8.311	1

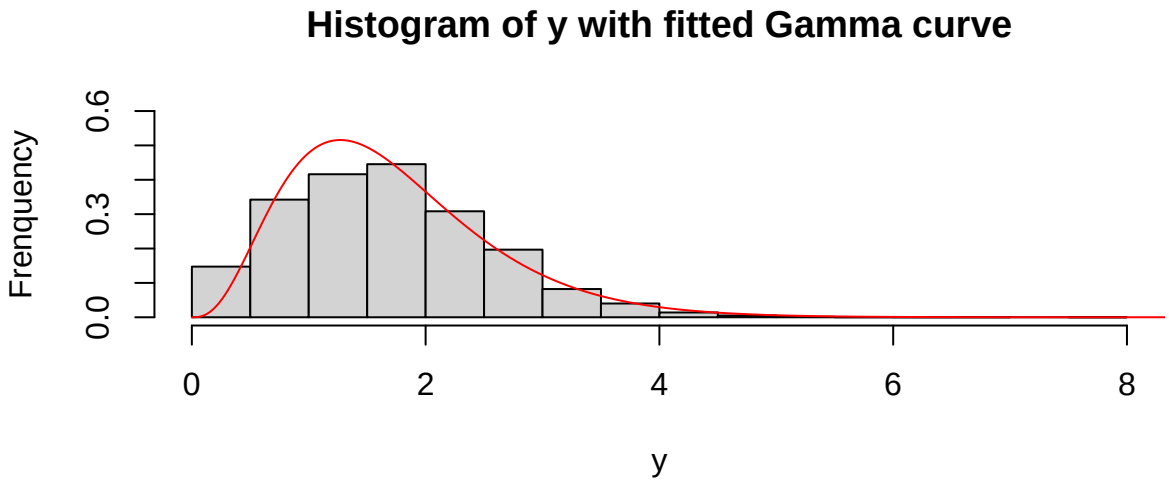

Figure 1: Histogram of y with fitted Gamma curve

2. Report for the Analysis of Smoking data

Summary

We did research on the data set of 2014 American National Youth Tobacco Survey to investigate potential factors that are associated to the trend of chewing tobacco among high school students. A GLMM model is conducted and as a result, the increase in age lead to higher probabilities of chewing tobacco for males except black male who live in urban areas, whereas the increase in age does not associate with the increase in probabilities of chewing tobacco for female. In general, white males living in rural area(Cowboys) has higher probabilities of chewing tobaccos than white females living in rural area. Surprisingly, the probability of chewing tobacco for black female of age 15 who live in rural area is close to 1, more investigation should be done to check whether it is a valid "outlier". On average, tobacco chewing is mostly done by Cowboys, and Cowboys who above 12 years old has higher probabilities of chewing tobacco than any other males of different races no matter they live in urban or rural areas. Additionally, tobacco chewing is more common for whites who live in rural area than urban area, a possible reason is that urban schools have stricter regulation of tobacco products. We also find that differences between schools within a state in chewing tobacco usage are much larger than differences between states among high school students based on both estimation of median and 95% confidence interval. This means more regulation might be put on school level rather than state level, since the variation of differences in chewing tobacco is larger at school level. Last but not the least, students who living in urban areas are less likely to smoke for all ages across all races in general.

Introduction

The analyzing process was based on the R version dataset from the 2014 American National Youth Tobacco Survey available on http://pbrown.ca/teaching/appliedstats/data. In this analysis, we investigate the association between the trend of chewing tobacco among high school students and potential factors such as "age", "sex", "RuralUrban(living in rural or urban area)", "race" as well as "state" and "school". To be more specific, we want to find out whether the differences of chewing tobacco between states are much larger than the differences between schools in a state for high school students. Additionally, whether tobacco chewing is mostly done by Cowboys(Male, white and live in rural areas) as reflected in American TV, and whether tobacco chewing is common for whites who live in rural area.

Methods

Generalized linear mixed models(GLMM) are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, GLMM is an extension of generalized linear models to include both fixed and random effects. In our scenario, "age", "sex", "RuralUrban", "race" should be treated as fixed effects. On the other hand, "state" and "school" should be treated as random effects since there might be some correlations for students within a same states or within a same schools. Additionally, the response variable "y", has value of 0 or 1, which is binary. Thus, it is appropriate to consider binomial GLMM model. The model is given by:

$$Y_{ijk} \sim Bernoulli(\mu_{ijk}), log(\frac{\mu_{ijk}}{1-\mu_{ijk}})) = \mu + X_{ijk}\beta + U_i + V_{ij}$$

$$V_{ij} \sim N(0, \sigma_2^2), U_i \sim N(0, \sigma_1^2)$$

$$\frac{1}{\sigma_1^2} \sim Gamma(1, 5 * 10^{-5}), \frac{1}{\sigma_2^2} \sim Gamma(1, 5 * 10^{-5})(default)$$

X_{ijk} represents the covariates of "age", "sex", "RuralUrban" and "race", respectively. U_i and V_{ij} represent the random effects from state and school on the tendency of chewing tobacco. β represents the corresponding parameters of these covariates. μ represents the mean of reference. $log(\frac{\mu_{ijk}}{1-\mu_{ijk}}))$ is the log odds of chewing tobacco. μ_{ijk} is the probability of chewing tobacco for student k from $school_j$ in $state_i$.

Results

From Table 3, the standard deviation(SD) for school is 0.754 whereas the SD for state is 0.01. This means that the variation of difference of chewing tobacco among school within a state are larger than between states. In addition, the CI of SD for school is larger than region while there is no overlap between school's CI and state's CI, this means that the SD for school and states is indeed different. Thus, the conclusion is that the differences of chewing tobacco between schools within a state are much larger than the differences between states for high school students. This contradicts to hypothesis 1. According to Figure 2-4, white male and female are represented by red line. The square represents rural and circle represents urban. We find out that on average, white males living in rural area(Cowboys) has higher probabilities of chewing tobaccos than white females living in rural area. However, there is one outlier for black female who live in rural areas of age 15– the probability of chewing tobacco is about 0.99, which is fairly high. Cowboys above 12 years old has higher probabilities of chewing tobacco than any other males of different races who live in either urban or rural areas. Thus, American TV does reflect the reality that on average, tobacco chewing is mostly done by Cowboys. White females who live in rural area also has higher probabilities of chewing tobacco compared to white females who live in urban area. Thus, tobacco chewing is common for whites who live in rural area. This coincide with hypothesis 2. Another amazing result is that for most of the males, the probabilities of chewing tobacco increase as age increase, except for black male who live in urban area. However, similar trend cannot be found on females. Lastly, students who living in urban areas are less likely to smoke for all ages across all races in general.

Table 3: Summary of SD of random effects

	q0.5	q0.025	q0.975
SD for state	0.010	0.006	0.025
SD for school	0.754	0.624	0.831

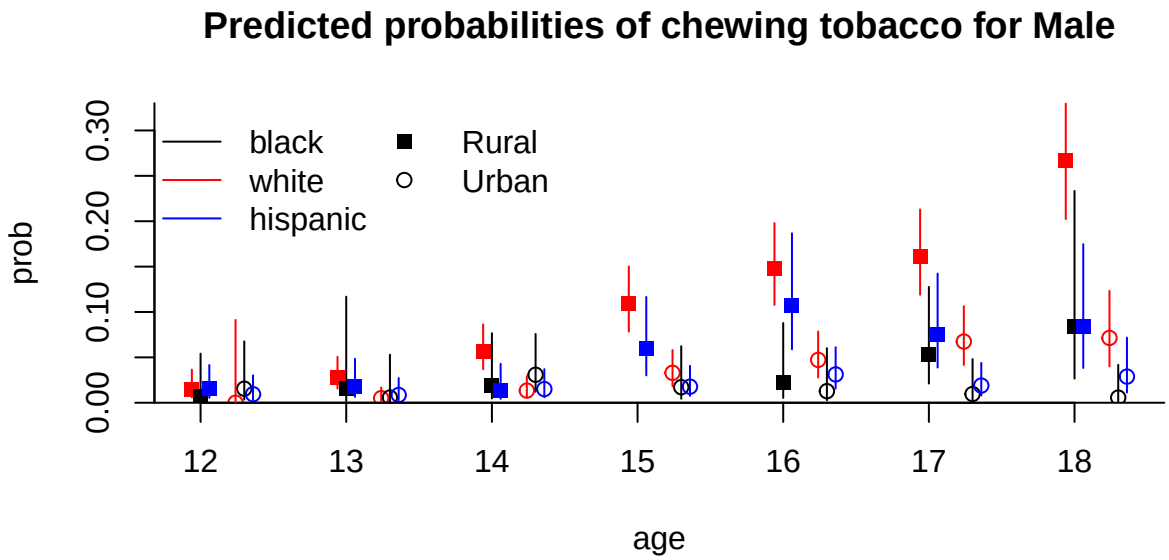

Figure 2: Predicted probabilities of chewing tobacco for Male

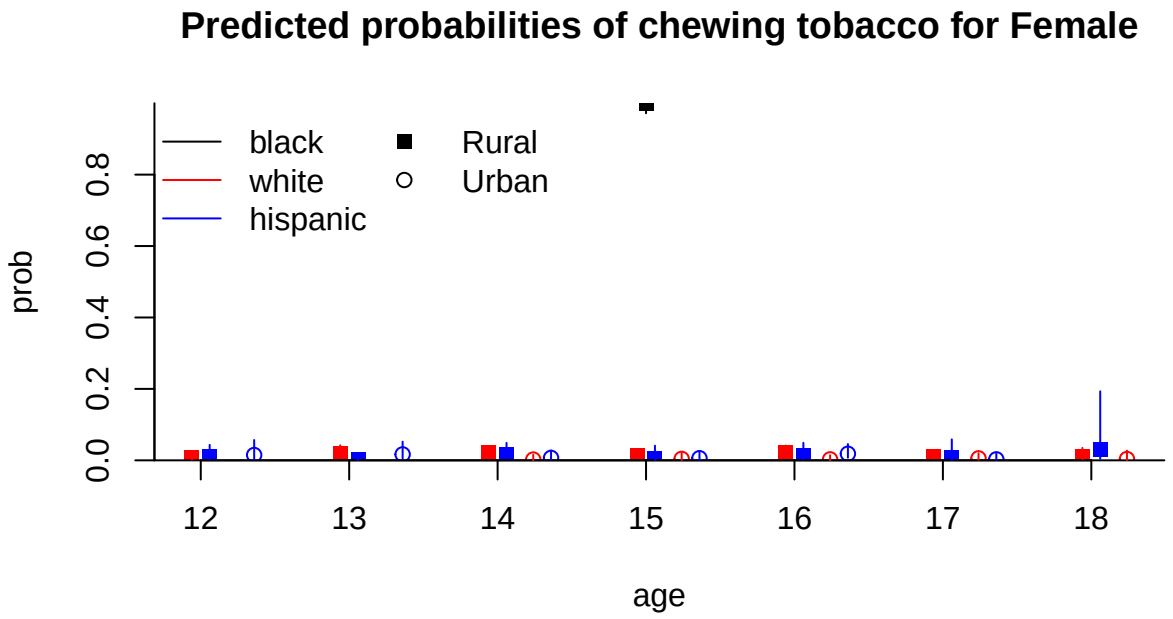

Figure 3: Predicted probabilities of chewing tobacco for Female

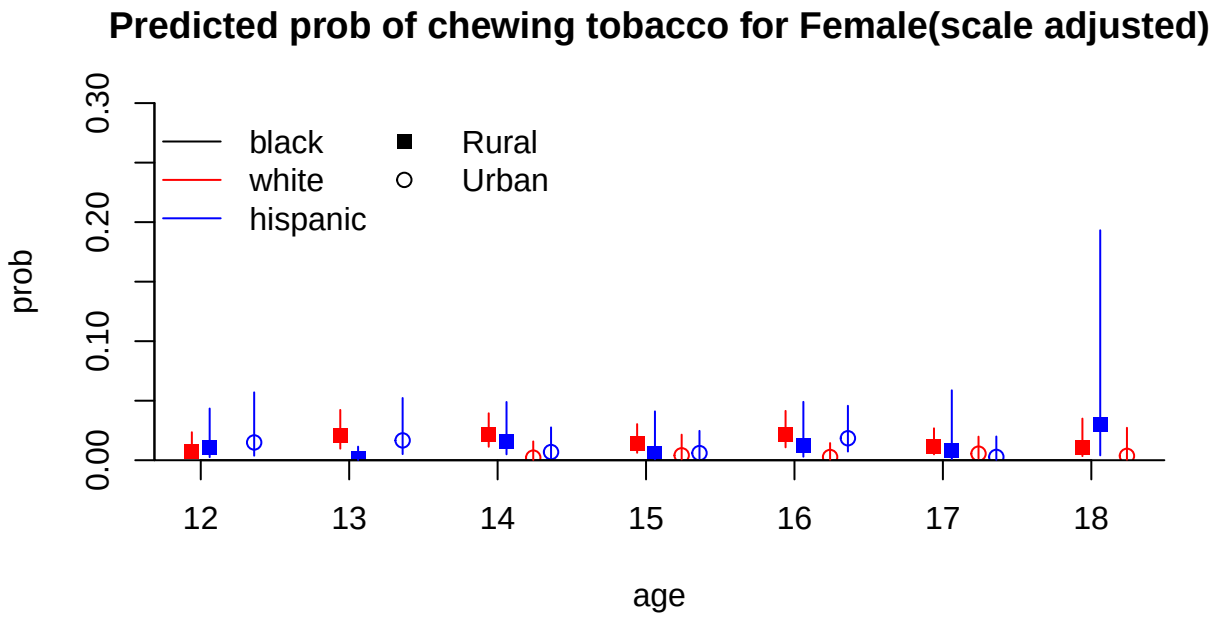

Figure 4: Predicted prob of chewing tobacco for Female(scale adjusted)

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages("INLA")
#install.packages("R.utils")
#install.packages("INLA",repos=c(getOption("repos"),INLA="https://inla.r-inla-download.org/R/stable"),
#install.packages("ggpmisc")
#install.packages("Pmisc", repos = "http://R-Forge.R-project.org", type = "source")
library("INLA")
#devtools::install_github("julianfaraway/brinla")
library("brinla")

#install.packages("R.utils")
#install.packages("ggpmisc")
#install.packages("Pmisc", repos = "http://R-Forge.R-project.org", type = "source")
library(ggpmisc)
library(R.utils)
library(ggplot2)
xFile = Pmisc::downloadIfOld("http://www.bristol.ac.uk/cmm/media/migrated/datasets.zip")
x = read.table(grep("chem97", xFile, value = TRUE),
               col.names = c("region","school", "indiv",
                             "chem", "sexNum", "ageMonthC", "grade"))
x$sex = factor(x$sexNum, levels = c(0, 1), labels = c("M","F"))
x$age = (222 + x$ageMonthC)/12
x$y = pmax(0.05, 8 - x$grade)

xres = inla(y ~ 0 + age + sex+
              f(region,model="iid",prior='pc.prec',param=c(0.18,0.5))+
              f(school,model="iid",prior='pc.prec',param=c(0.18,0.5)),
              control.fixed = list( mean = 0, prec = 1/(3^2) ),
              data = x, family = "gamma", control.family = list
              (hyper = list(prec = list(prior = "pc.prec", param = c(0.18, 0.5)))))




#random effect
mytable_1 = Pmisc::priorPostSd(xres, group = "random")$summary
knitr::kable(mytable_1, caption = "Summary of SD of Random Effects",
             align = "ccc", digits = 3)

#fixed effect
knitr::kable(exp(xres$summary.fixed),
             caption = "Summary of Fixed Effects(natural scale)",
             align = "ccc", digits = 3)


x$y = pmax(0.05, 8 - x$grade)
shape = (mean(x$y))^2/var((x$y))
scale = mean(x$y)/shape
hist(x$y, freq=FALSE,xlab = "y", ylab= "Frenquency",
     main="Histogram of y with fitted Gamma curve",ylim=c(0,0.6))
a <- seq(0, 9, by = 0.01)
```

```
lines(a,dgamma(a, scale = scale , shape = shape, log = FALSE),col="red")
smokeFile = "smokeDownload2014.RData"
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
smokeFile)
}
(load(smokeFile))
smokeFormats[smokeFormats[, "colName"] == "chewing_tobacco_snuff_or",
c("colName", "label")]
# get rid of 9-11 and 19 year olds and missing age and
# race
smokeSub = smoke[which(smoke$Age >= 12 & smoke$Age <= 18 &
!is.na(smoke$Race) & !is.na(smoke$chewing_tobacco_snuff_or) &
(!is.na(smoke$Sex))), ]
smokeSub$ageFac = relevel(factor(smokeSub$Age), "15")
smokeSub$y = as.numeric(smokeSub$chewing_tobacco_snuff_or)
lincombDf = do.call(expand.grid, lapply(smokeSub[, c("ageFac",
"Sex", "Race", "RuralUrban")], levels))
lincombDf$y = -99
lincombList = inla.make.lincombs(as.data.frame(model.matrix(y ~
ageFac * Sex * RuralUrban * Race, lincombDf)))
library("INLA", quietly = TRUE)
smokeModel = inla(y ~ ageFac * Sex * RuralUrban * Race +
f(state) + f(school), lincomb = lincombList, data = smokeSub,
family = "binomial")
library(brinla)
knitr::kable( bri.hyperpar.summary(smokeModel)[,c(4,3,5)],align="ccc",
              digits = 3 ,caption="Summary of SD of random effects")#random

smokePred = smokeModel$summary.lincomb.derived[,
paste0(c(0.5, 0.025, 0.975), 'quant')]
smokePred = exp(smokePred)/(1+exp(smokePred))
smokePred$diff = smokePred$'0.975quant' - smokePred$'0.025quant'
lincombDf$Age = as.numeric(as.character(lincombDf$ageFac))
lincombDf$ageShift = lincombDf$Age + 0.06*(as.numeric(lincombDf$Race)-2) +
0.3*(lincombDf$RuralUrban == 'Urban')
Spch = c('Rural' = 15, 'Urban' = 1)
Scol = c(black = 'black', white = 'red', hispanic='blue')
toPlot = (lincombDf$Race %in% names(Scol)) & (smokePred$diff < 0.9) &
lincombDf$Sex == 'M'
lincombDfHere = lincombDf[toPlot,]
smokePredHere = smokePred[toPlot,]
plot(
lincombDfHere$ageShift,
smokePredHere$'0.5quant',
pch = Spch[as.character(lincombDfHere$RuralUrban)],
col = Scol[as.character(lincombDfHere$Race)],
# log='y',
ylim = c(0,0.33),
xlab='age', ylab='prob', #yaxt='n',
yaxs='i', bty='l',main="Predicted probabilities of chewing tobacco for Male")
#forY = 1/c(4,10,25,100,500)
#axis(2, at=forY, mapply(format, forY), las=1)
```

```
segments(lincombDfHere$ageShift, smokePredHere$'0.025quant',
lincombDfHere$ageShift, smokePredHere$'0.975quant',
col = Scol[as.character(lincombDfHere$Race)])
legend('topleft', bty='n',
ncol = 2,
pch=c(rep(NA, length(Scol)), Spch),
lty = rep(c(1,NA), c(length(Scol), length(Spch))),
col = c(Scol, rep('black', length(Spch))),
legend=c(names(Scol), names(Spch)))
smokePred = smokeModel$summary.lincomb.derived[,
paste0(c(0.5, 0.025, 0.975), 'quant')]
smokePred = exp(smokePred)/(1+exp(smokePred))
smokePred$diff = smokePred$'0.975quant' - smokePred$'0.025quant'
lincombDf$Age = as.numeric(as.character(lincombDf$ageFac))
lincombDf$ageShift = lincombDf$Age + 0.06*(as.numeric(lincombDf$Race)-2) +
0.3*(lincombDf$RuralUrban == 'Urban')
Spch = c('Rural' = 15, 'Urban' = 1)
Scol = c(black = 'black', white = 'red', hispanic='blue')
toPlot = (lincombDf$Race %in% names(Scol)) & (smokePred$diff < 0.9) &
lincombDf$Sex == 'F'
lincombDfHere = lincombDf[toPlot,]
smokePredHere = smokePred[toPlot,]
plot(
lincombDfHere$ageShift,
smokePredHere$'0.5quant',
pch = Spch[as.character(lincombDfHere$RuralUrban)],
col = Scol[as.character(lincombDfHere$Race)],
# log='y',
ylim = c(0,max(smokePredHere)),
xlab='age', ylab='prob', #yaxt='n',
yaxs='i', bty='l',main="Predicted probabilities of chewing tobacco for Female")
#forY = 1/c(4,10,25,100,500)
#axis(2, at=forY, mapply(format, forY), las=1)
segments(lincombDfHere$ageShift, smokePredHere$'0.025quant',
lincombDfHere$ageShift, smokePredHere$'0.975quant',
col = Scol[as.character(lincombDfHere$Race)])
legend('topleft', bty='n',
ncol = 2,
pch=c(rep(NA, length(Scol)), Spch),
lty = rep(c(1,NA), c(length(Scol), length(Spch))),
col = c(Scol, rep('black', length(Spch))),
legend=c(names(Scol), names(Spch)))
smokePred = smokeModel$summary.lincomb.derived[,
paste0(c(0.5, 0.025, 0.975), 'quant')]
smokePred = exp(smokePred)/(1+exp(smokePred))
smokePred$diff = smokePred$'0.975quant' - smokePred$'0.025quant'
lincombDf$Age = as.numeric(as.character(lincombDf$ageFac))
lincombDf$ageShift = lincombDf$Age + 0.06*(as.numeric(lincombDf$Race)-2) +
0.3*(lincombDf$RuralUrban == 'Urban')
Spch = c('Rural' = 15, 'Urban' = 1)
Scol = c(black = 'black', white = 'red', hispanic='blue')
toPlot = (lincombDf$Race %in% names(Scol)) & (smokePred$diff < 0.9) &
lincombDf$Sex == 'F'
```

```
lincombDfHere = lincombDf[toPlot,]
smokePredHere = smokePred[toPlot,]
plot(
lincombDfHere$ageShift,
smokePredHere$'0.5quant',
pch = Spch[as.character(lincombDfHere$RuralUrban)],
col = Scol[as.character(lincombDfHere$Race)],
# log='y',
ylim = c(0,0.3),
xlab='age', ylab='prob', #yaxt='n',
yaxs='i', bty='l',main="Predicted prob of chewing tobacco for Female(scale adjusted)")
#forY = 1/c(4,10,25,100,500)
#axis(2, at=forY, mapply(format, forY), las=1)
segments(lincombDfHere$ageShift, smokePredHere$'0.025quant',
lincombDfHere$ageShift, smokePredHere$'0.975quant',
col = Scol[as.character(lincombDfHere$Race)])
legend('topleft', bty='n',
ncol = 2,
pch=c(rep(NA, length(Scol)), Spch),
lty = rep(c(1,NA), c(length(Scol), length(Spch))),
col = c(Scol, rep('black', length(Spch))),
legend=c(names(Scol), names(Spch)))
```