Yao Yao
Professor Patrick Brown
STA442
8, October, 2021

Question 1

1.

In this model we take the grade point under the maximum as the response variable. From Figure1, we could observe it is a right skewed histogram. Since the response variable is a continuous variable and it has a right skewed histogram.Thus we could consider this distribution as a gamma distribution, it has a shape parameter and a scale parameter. This means that this model is a logistic regression. And since our observations are students in the schools in each region, the observations may have relation to each other, so we could assume that these observations are actually dependent. So I considered this model as a generalized linear mixed model (GLMM). Then I established the model as follows.

$$log(\mu_i) = X_{ijk}\beta + U_i + V_{ij}$$

where μ_i is the expected grade score below the maximum of student k in school j in region i.
Thus, $log(\mu_i)$ represents the expected probability after the log transformation.
y_i would follow a gamma distribution with shape parameter is 10^{-4} and rate parameter is 10^{-4}.
X_{ijk} has indicator variables for age and sex in school j and region i.
β represents the corresponding parameters of X_{ijk}.
U_i represents the random effect of region i on the tendency of the grade score below the maximum, where it follows Normal distribution(0, σ_u^2).
V_i represents the random effect of school j in region i of getting the grade score below the maximum, where it follows Normal distribution(0, σ_v^2)

This model can also be written as

$$log(\mu_i) = X_{age}\beta_1 + X_{sex}\beta_2 + U_i + V_{ij}$$

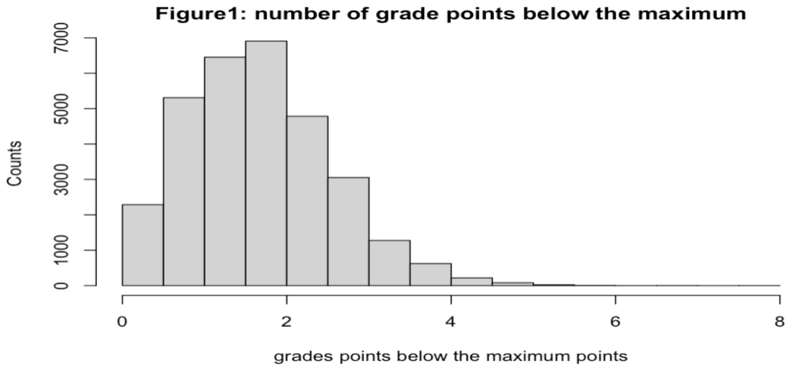


Figure 1

2.

Since changes in grades of 20% are more likely, thus

The expected value of y_i would be $E(y_i) = e^{x\beta+U_i}$, where U_i represents the random effect.

The expected value of y_j would be $E(y_j) = e^{x\beta+U_j}$, where U_j represents the random effect.

Since changes in grades of 20% are more likely, thus the ratio of $E(y_i)$ and $E(y_j)$ is around 1.2, which is

$$\frac{E(y_i)}{E(y_j)} = \frac{e^{x\beta+U_i}}{e^{x\beta+U_j}} = e^{U_i-U_j} = 1.2$$

Since the expected value are obtained by log transformation, thus the actual difference would be $U_i - U_j = ln(0.2) = 0.183$ (round to 3 decimals)

Therefore, the median would be 0.183 and we define the prior of our random effect to be $P(\sigma > 0.183) = 0.5$.

Since the rate parameter equals ln(2)/median, we could obtain the rate parameter is around 3.788.

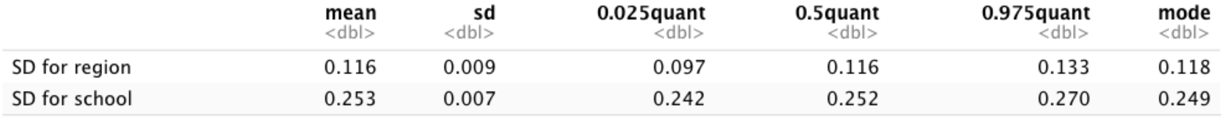
Thus in the new model(with prior)

U_i represents the random effect of region i on the tendency of the grade score below the maximum, where it follows Normal distribution(0, σ_u^2).σ_u^2 follows the exponential distribution(0,3.788)

V_i represents the random effect of school j in region i of getting the grade score below the maximum, where it follows Normal distribution(0, σ_v^2).σ_v^2 follows the exponential distribution(0,3.788)

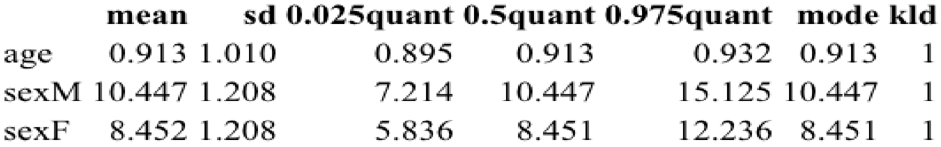
3.

According to Table 3, we could observe the mean on Age is 0.913. Since this summary table is regarding the exponential grade points below the maximum and each student's age, this indicates that if a student's age is one year older than another students with other factors such as their sex fixed, the grade points below the maximum of the student with greater age would be 0.913 times grade points below the maximum of another student. In other words, if the age of a student is one-year greater than another student with the same gender, the losing score would decrease 8.7%. Also from this summary table we could also observe that the mean of Male is 10.447 while the mean of Female is 8.452. Since Sex is a categorical variable, this demonstrates that the lossing score of a female student is 0.809 (8.452/10.447) times the losing score of a male student if their age is the same. This shows that the score that female students lost is 20% lower than male students when they are the same age. From table 1, we could find the standard deviation for region is 0.009 which is greater than school, 0.007. This shows that the difference between regions-level is greater than schools-level. Also, this statement would also be proved by observing the interval of 0.25quant and 0.75quant, since the interval for region is 0.097 to 0.133 and for school is 0.242 to 0.270, and there is no overlapping on both intervals. Overall, if the age of a student is one-year older than another student, then the expected losing point would be 8.7% lower as long as they have the same gender. And if two students are the same age, then the female student would lose 20% lower of expected grade points than the male student. Furthermore, the difference of grade points below the maximum between regions-level is greater than schools-level.



	mean	sd	0.025quant	0.5quant	0.975quant	mode
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
SD for region	0.116	0.009	0.097	0.116	0.133	0.118
SD for school	0.253	0.007	0.242	0.252	0.270	0.249

Table 1



	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
age	0.913	1.010	0.895	0.913	0.932	0.913	1
sexM	10.447	1.208	7.214	10.447	15.125	10.447	1
sexF	8.452	1.208	5.836	8.451	12.236	8.451	1

Table 3

Question 2

Analyzation of Tobacco Usage Among Youth in America

Introduction

Among various classic American TV series, tobacco chewing is a very common phenomenon. In reality, people who chew tobacco are also seen everywhere. In order to know whether people who chewing tobacco is consistent with the accurate reflection in TV series and whether tobacco chewing among high school students is highly regional, I analyzed data from the 2014 American National Youth Tobacco Survey, which is available at http://pbrown.ca/teaching/ appliedstats/data. All the data were collected through surveys from colleges in America, and it contains 22,007 youth data from 207 schools in 34 states. To analyze these two questions, I established two hypotheses. The first hypothesis is that state-level differences in chewing tobacco usage amongst high school students are much larger than differences between school-level, to be more specific, I believe tobacco chewing is common in some states but rare in others. The second hypothesis is that people who chew tobacco are more likely to be white males living in the rural areas, just as described in American TV series. And other ethnic groups and those who live in urban areas have few tobacco chewers. To test these two hypotheses, I will analyze the association between probability of chewing tobacco and some confounders such as age, sex, location, race, state and school status, and predict the distribution of races and locations of tobacco chewers to draw conclusions.

Method

First, I removed the data that were younger than 12 and older than 18, because our main target was the youth group. And then I removed all the missing data to make sure of the accuracy of our model.

By plotting the bar charts of each variable (e.g. Figure1: age, Figure2: sex, Figure3: location, and Figure4: race) and the probability of chewing tobacco, we could find that the associations of these four variables and the probability chewing tobacco might exists, since the differences was quite obvious. Therefore, we chose to use these four variables as our fixed effects and the state and school status as random effects. Based on that, we started to use the package INLA to build the model.

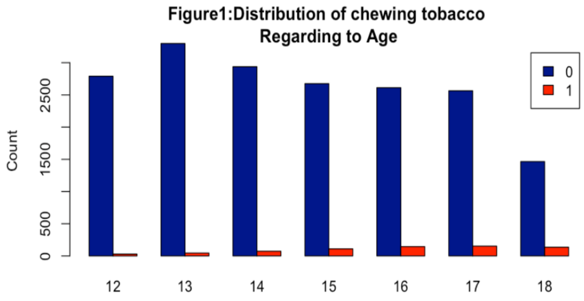


Figure 1

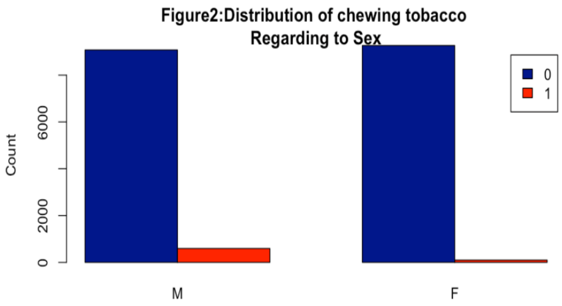


Figure 2

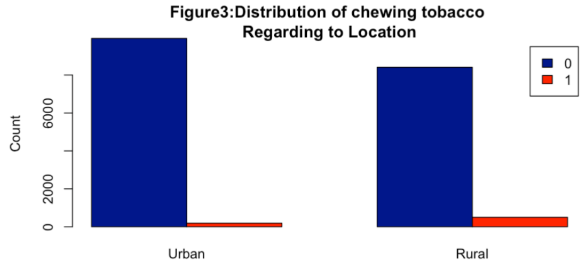


Figure 3

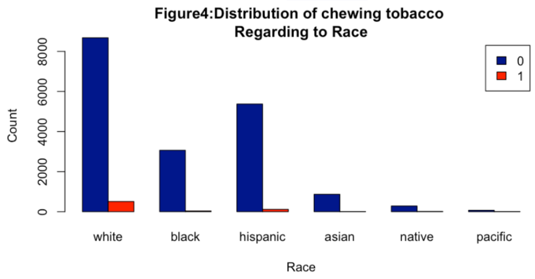


Figure 4

In this model we take whether or not chewing tobacco as the response variable, then the response variable is either 0, not chewing tobacco; or 1, chewing tobacco. Thus there is a binomial distribution between our response variable, which means that this model is a logistic regression. Because our observations are youth in high schools in each district, thus each school or each district's youth may be related to each other, so we could assume that these observations are actually dependent. So I considered this model as a generalized linear mixed model (GLMM). Then I established the model as follows.

$$log(\frac{\mu_{ijk}}{1-\mu_{ijk}}) = \mu + X_{ijk}\beta + U_i + V_j$$

where μ_{ijk} is the probability that student k in school j in state i. $\frac{\mu_{ijk}}{1-\mu_{ijk}}$ is the odds of chewing tobacco, so $log(\frac{\mu_{ijk}}{1-\mu_{ijk}})$ represents the odds after the log transformation.

μ represents when all of the explanatory variables are equals to 0, the intercept would be μ

X_{ijk} has indicator variables for age, sex, living in rural or urban, and race for responding teenager k in school j and state i.

β represents the corresponding parameters of X_{ijk} .

U_i represents random effect of state i on the tendency of chewing tobacco,where it follows Normal distribution(0, σ_u^2)

V_i represents random effect of school j of chewing tobacco,,where it follows Normal distribution(0, σ_v^2)

This model can also be written as

$$log(\frac{\mu_{ijk}}{1-\mu_{ijk}}) = \beta_0 + \beta_1 X_{ageFac} + \beta_2 X_{Sex} + \beta_3 X_{RuralUrban} + \beta_4 X_{Race} + U_i + V_j$$

After building the model, we would analyze it by summarizing the model and plotting the distribution of races and locations of tobacco chewers.

Result

From our summary table (table 1), we could observe that at 0.5quant, which is the median, the state's standard deviation is smaller than the school's standard deviation (0.009<0.803). This means that the variance between states is larger than the variance of schools. So this also shows that the difference between states-level is larger than schools-level.

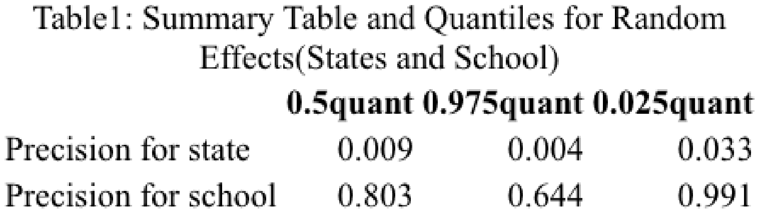


Table1: Summary Table and Quantiles for Random Effects(States and School)

	0.5quant	**0.975quant**	**0.025quant**
Precision for state	0.009	0.004	0.033
Precision for school	0.803	0.644	0.991

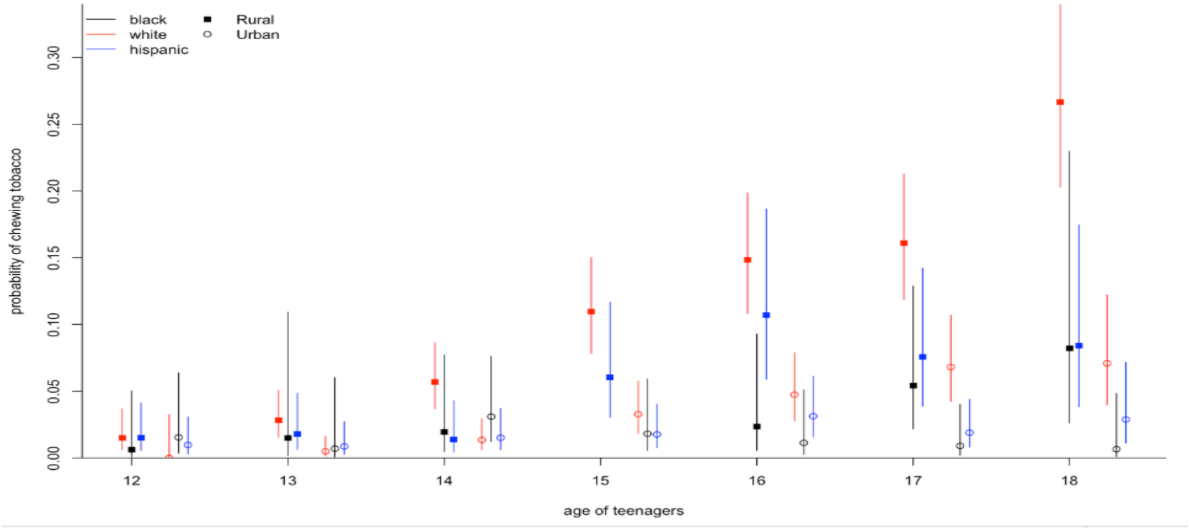
Table 1

At the same time, we observe the values of 0.025quant and 0.975quant. We could find that the states' standard deviation is in the interval of 0.004 to 0.033, while the schools' standard deviation is in the interval of 0.644 to 0.991. Compared to these two intervals, there is no overlapping. This shows that there is a significant difference between the states and schools' confidence interval, which demonstrates that the variance between the two is indeed different.

So from both sides we could prove that state-level differences in chewing tobacco usage amongst high school students are much larger than differences between school-level. Perhaps because many schools are in the same state, and the people in these schools are influenced by the culture of the state.Hence, the differences between these schools are not obvious. But generally speaking, chewing tobacco is strongly regional and tobacco chewing is common in some states but rare in others. Therefore our first hypothesis is true.

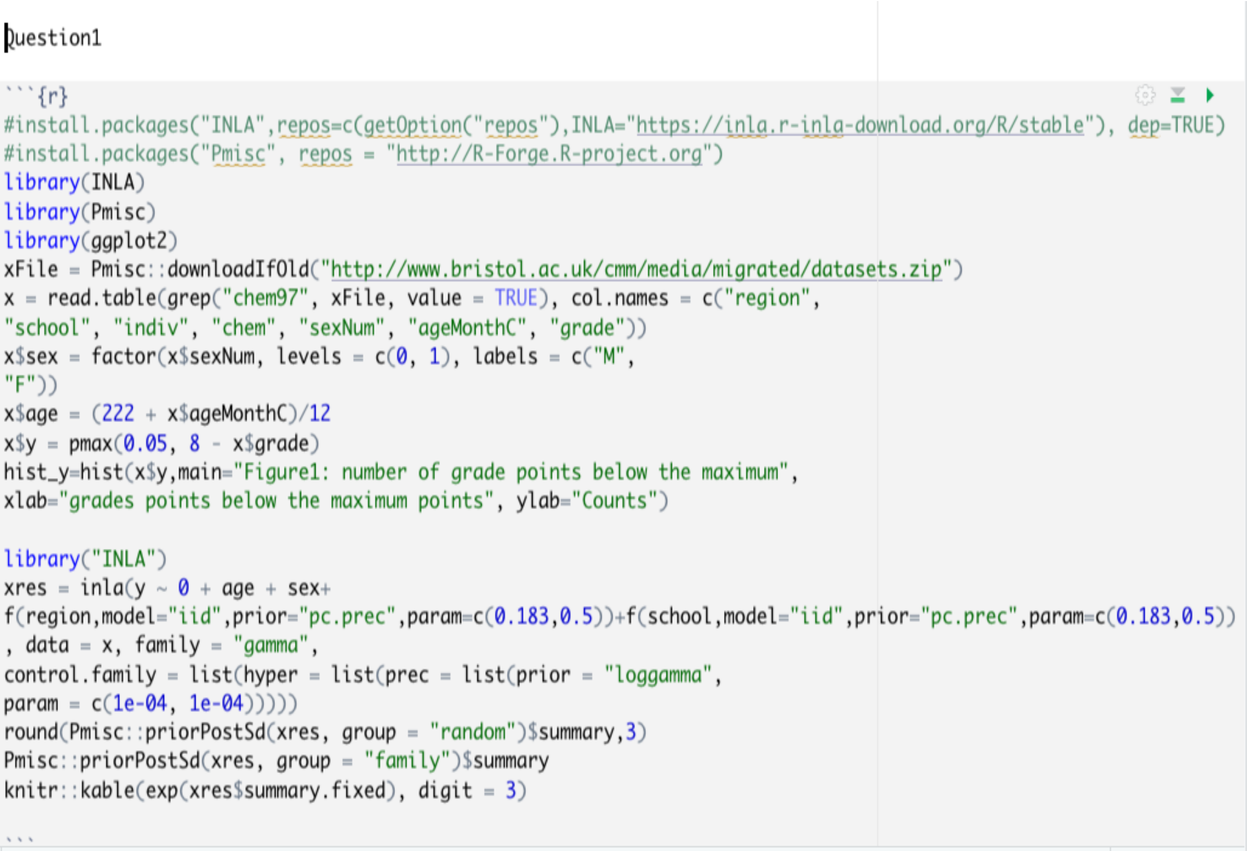
Considering figure5, which represents the probability distribution of races and locations of young tobacco chewers. From that, we could observe that at each age, whites people who live in rural areas always have the highest probability of chewing tobacco. This means that most of the tobacco chewers are white males living in the suburbs, just like the reflections in American TV series. Therefore, the second hypothesis is also held.

Figure5:Probability Distribution of Responding Teenages in Race and Location



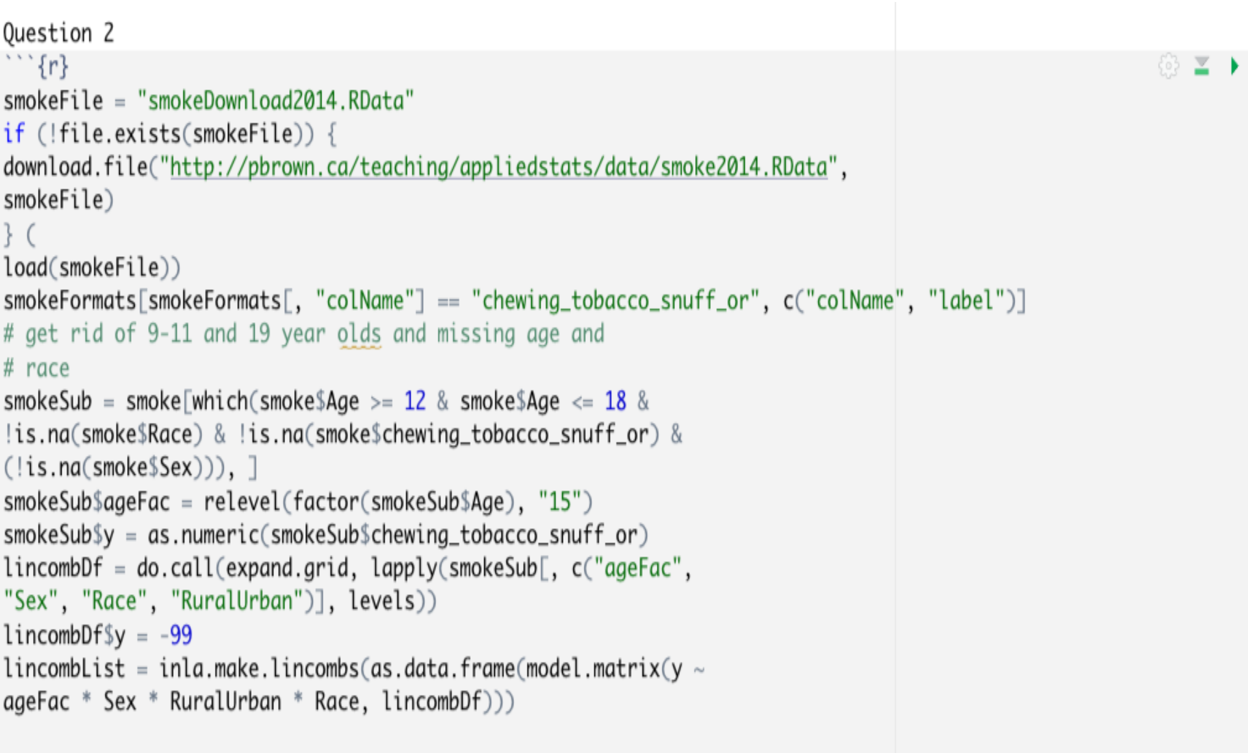
In conclusion, we find that both hypotheses are true. This shows that the stereotype which is shown in TV series is indeed very similar to the reality, that most of the tobacco chewers are white males who live in the rural areas, just like the cowboys, and chewing tobacco is a strong regional action. This is probably because TV series have a strong influence on youth. Since this behavior is managed in a different way in each state, this may result in significant differences. Thus chewing tobacco is very popular in some states and rare in others.

Appendix



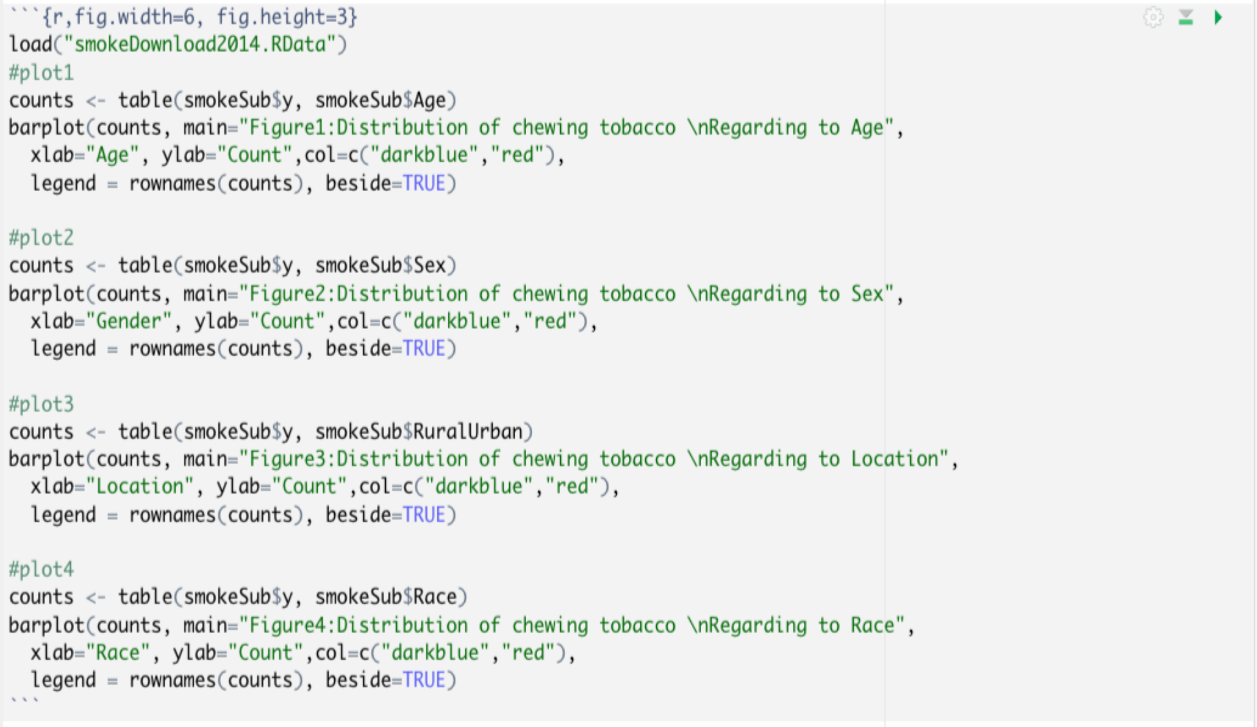
Question1

~~~
```{r}
#install.packages("INLA",repos=c(getOption("repos"),INLA="https://inla.r-inla-download.org/R/stable"), dep=TRUE)
#install.packages("Pmisc", repos = "http://R-Forge.R-project.org")
library(INLA)
library(Pmisc)
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)
hist_y=hist(x$y,main="Figure1: number of grade points below the maximum",
xlab="grades points below the maximum points", ylab="Counts")

library("INLA")
xres = inla(y ~ 0 + age + sex+
f(region,model="iid",prior="pc.prec",param=c(0.183,0.5))+f(school,model="iid",prior="pc.prec",param=c(0.183,0.5))
, data = x, family = "gamma",
control.family = list(hyper = list(prec = list(prior = "loggamma",
param = c(1e-04, 1e-04)))))
round(Pmisc::priorPostSd(xres, group = "random")$summary,3)
Pmisc::priorPostSd(xres, group = "family")$summary
knitr::kable(exp(xres$summary.fixed), digit = 3)

```
~~~

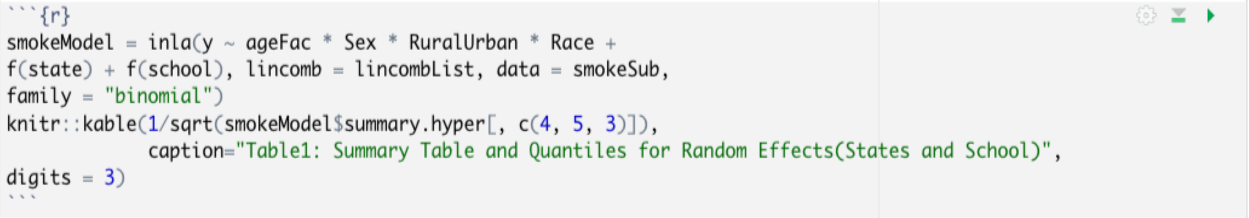


Question 2

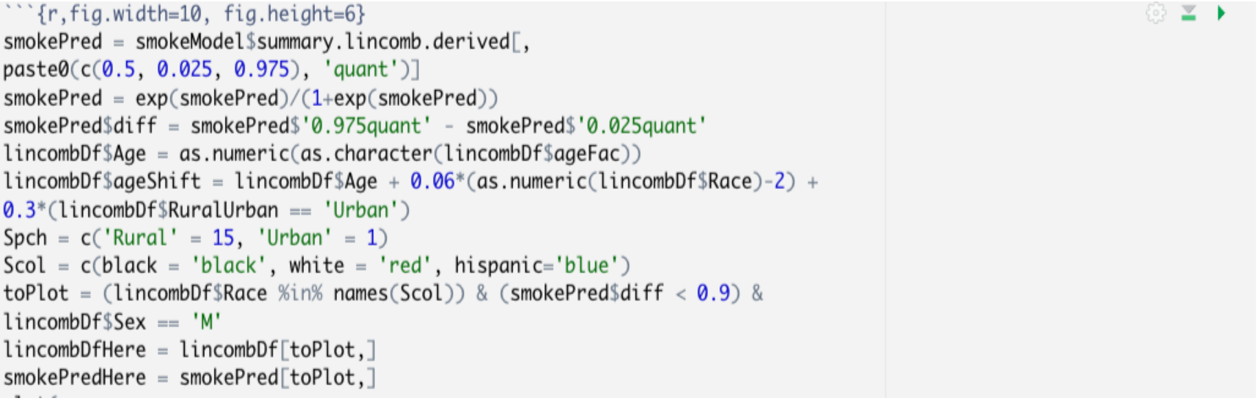
~~~
```{r}
smokeFile = "smokeDownload2014.RData"
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
smokeFile)
} (
load(smokeFile))
smokeFormats[smokeFormats[, "colName"] == "chewing_tobacco_snuff_or", c("colName", "label")]
# get rid of 9-11 and 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)))
~~~



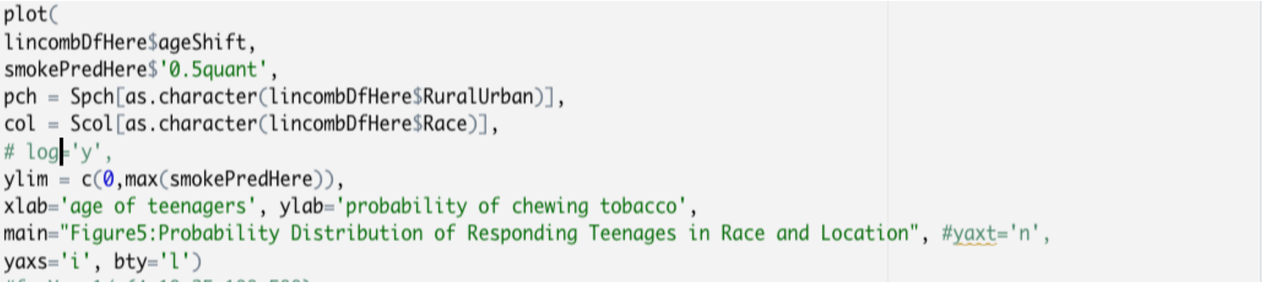
~~~
```{r,fig.width=6, fig.height=3}
load("smokeDownload2014.RData")
#plot1
counts <- table(smokeSub$y, smokeSub$Age)
barplot(counts, main="Figure1:Distribution of chewing tobacco \nRegarding to Age",
  xlab="Age", ylab="Count",col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)

#plot2
counts <- table(smokeSub$y, smokeSub$Sex)
barplot(counts, main="Figure2:Distribution of chewing tobacco \nRegarding to Sex",
  xlab="Gender", ylab="Count",col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)

#plot3
counts <- table(smokeSub$y, smokeSub$RuralUrban)
barplot(counts, main="Figure3:Distribution of chewing tobacco \nRegarding to Location",
  xlab="Location", ylab="Count",col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)

#plot4
counts <- table(smokeSub$y, smokeSub$Race)
barplot(counts, main="Figure4:Distribution of chewing tobacco \nRegarding to Race",
  xlab="Race", ylab="Count",col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)
```
~~~



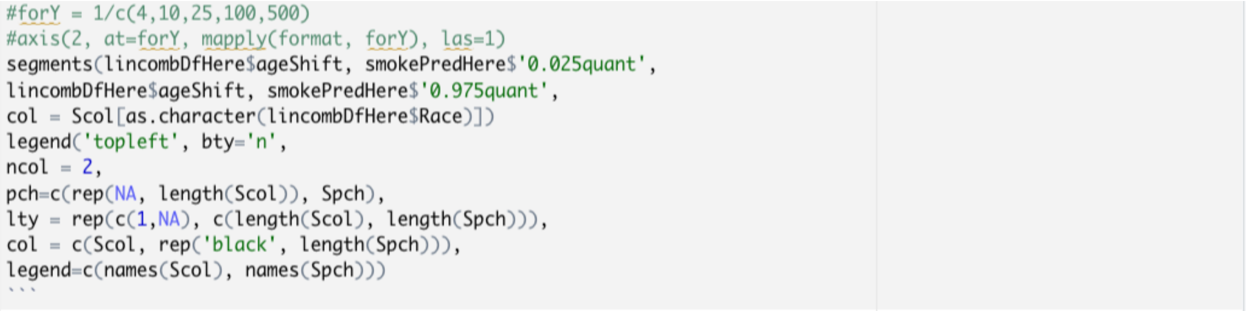
~~~
```{r}
smokeModel = inla(y ~ ageFac * Sex * RuralUrban * Race +
f(state) + f(school), lincomb = lincombList, data = smokeSub,
family = "binomial")
knitr::kable(1/sqrt(smokeModel$summary.hyper[, c(4, 5, 3)]),
             caption="Table1: Summary Table and Quantiles for Random Effects(States and School)",
digits = 3)
```
~~~



~~~
```{r,fig.width=10, fig.height=6}
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,max(smokePredHere)),
xlab='age of teenagers', ylab='probability of chewing tobacco',
main="Figure5:Probability Distribution of Responding Teenages in Race and Location", #yaxt='n',
yaxs='i', bty='l')
```



~~~
#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)))
```
~~~