STA442 – Assignment 1

Jeff Xu
1003342545

Question I

1-Introduction:

This report uses data from http://www.bristol.ac.uk that records chemistry test scores from 2,200 schools in the United Kingdom for 31,000 students. This analytic will examine the important factors that causes variations in the GCSE scores, with a hypothesis that students who are born near the end of a calendar year, are a few months older than those born at the start at the calendar year, and should have a slight advantage as being older and mature, thus should achieve a higher grade.

2-Model:

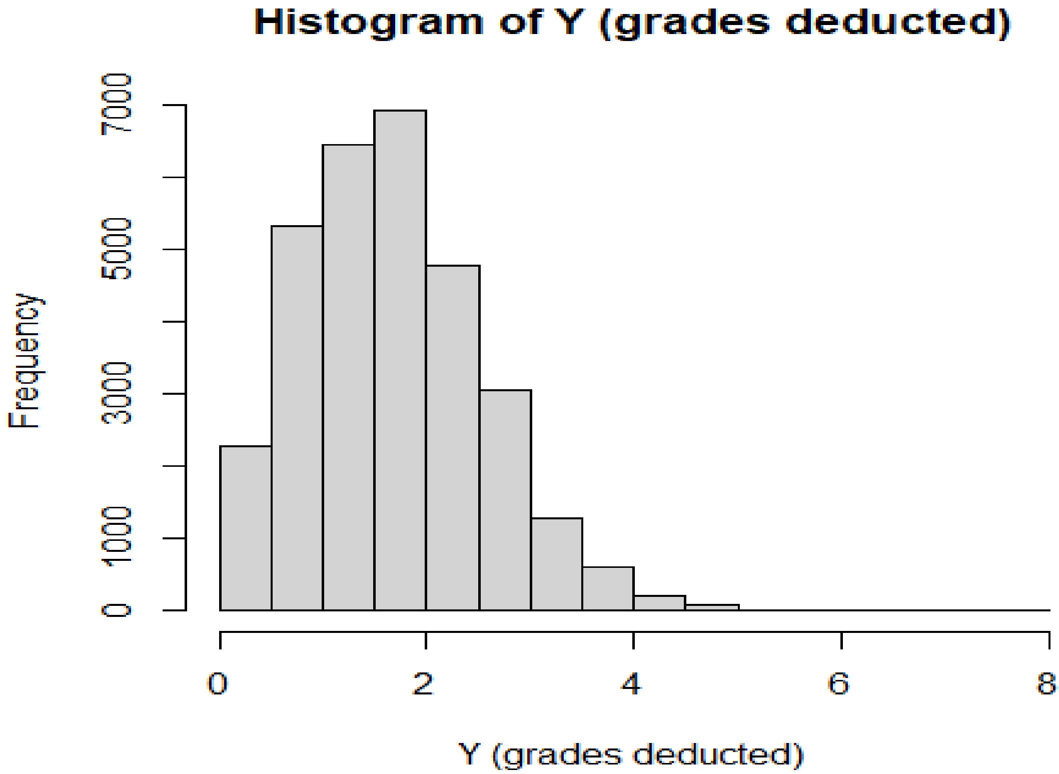

Figure 1, Histogram of Y (grades deducted)

On this histogram of Y, which is presenting the chance of marks deducted. It is clearly positively skewed, that lies nearly in unison with a Gamma distribution, which suggests that the outcome likely follows a gamma distribution. The explanatories such as age and sex will be treated as fixed effect as the ages will be within somewhere ranged from >= 18 but < 19, and only the Male and Female sex is considered for the gender in the data, and region and school will be treated as random effects.

$$Y_i = Gamma({}^{\mu_i}/_{v}, v),$$

P(V > 0.20) = 0.5, with 0.20 being the medium for V

$$e^{V_i - V_j} = 1.2$$

$$\log(\mu_i) = X_i\beta + U_i + V_{ij},$$

$$\mu_i = E(Y_i)$$

$$\frac{male}{female} = \frac{E(Y)}{E(\acute{Y})} = \frac{e^{\beta_0 + \beta_{1\,1} + \cdots}}{e^{\beta_0 + \beta_{2\,1} + \cdots}} = e^{\beta_1 - \beta_2}$$

With X_i as the fixed effects such as age and sex, and U_i as the random effect region, and $V_{ij}$for school in a region.

3-Prior:

$$V \sim Gamma(10^{-4}, 10^{-4})$$

$$U_i = E(y_i) = e^{x\beta + U_i},\ U_j = E(y_J) = e^{x\beta + U_j}, V_i = E(y_i) = e^{x\beta + V_i},\ V_j = E(y_J) = e^{x\beta + V_j}$$

$$\frac{\mu_i}{\mu_j} = \frac{E(y_i)}{E(y_J)} = e^{U_i - U_j} = 1.2$$

$$\sigma = U_i - U_j \approx (0.1 \sim 0.2)$$

which we will use σ = 0.18, since although rare, a 50% increase or decrease of grades can still maybe happen, so we will keep an upper middle ground of change between 0.1~0.2 as a 18%.

And so: $P(\sigma > 0.18) \approx 0.55$

4-Results:

	MEAN	***SD***	***0.025 quantile***	***0.5 quantile***	***0.975 quantile***	***MODE***	***KLD***
AGE	0.914	1.010	0.896	0.914	0.933	0.914	1
SEXM	10.267	1.207	7.093	10.267	14.859	10.267	1
SEXF	8.306	1.207	5.738	8.306	12.021	8.306	1

Figure 2, table of summary for exp(β) of age, sexM/F

$$\frac{E(male)}{E(female)} = \frac{10.268}{8.306} = 1.236$$

The table above shows the summary for the influence on the grades from the fixed effects. From the table, we can see that age does have an effect on the score, such that as age increase by 1, the chance of losing score reduces to 0.914 as before , which goes in agreement with our hypothesis. And the difference between gender has a notable ratio showing in equation above, with male students are 1.236 times more likely to lose marks than female students.

	SD	*0.025 quantile*	*0.975 quantile*
SD for region	0.007	0.109	0.138
SD for school	0.006	0.242	0.266

Figure 3, table of summary for σ of region and school

As shown in above table, the difference between the standard deviation for σ for school and region is very minor and can potentially be ignored, and from the 0.025 and 0.975 quantile, it can be seen that the standard deviation of school tends to make a bigger difference than the residing region in terms of one's performance on the GCSE.

5-Conclusion:

From the analysis, it is possible for us to conclude that, the hypothesis was correct, that the age of the students matters and does give a slight upper hand on their performance on the GCSE, as all of the participants are around 18 years old only with months difference thus the age difference is never full 1 year, and multiplier should not have taken full effect and would not be as much of an unfair advantage. And the male students are about 1.236 times more likely to lose marks than female students. Lastly, the school that the student attends has a larger impact on their GCSE performance than the region that they reside in.

Question II

1-Introduction

This report features the analytics using R based on the data collected by the center of disease control from the 2014 American National Youth Tobacco Survey, which is available on http://pbrown.ca/teaching/appliedstats/data. Through the general stereotypical image regarding to tobacco culture in the United States through America TV, we form hypothesis that tobacco chewing amongst middle to high school students is strongly regional, state-level variation is much higher than variation within a state. And furthermore, we wish to examine if the stereotypical image delivered through American TV is an accurate reflection of real life, such that tobacco chewing is most common amongst Cowboys, which are described as white raced, male gendered individuals who resides in rural states.

2-Method

For the purpose of the focus of our analysis, our response variable is the smoking status, its outcome can either be yes or no, we can represent the outcome using indicators of 1 for yes and 0 for no.

Therefore, the response can be represented by a Binomial (Bernoulli) distribution. And we would also like to compare a few criteria regarding to our hypothesis, such that we ask if the white race are more leaning towards developing behavior of smoking tobacco compared to other races? Then how about the male gendered compared to the female gendered? Etc., in such scenario, a mixed model tends to provide better information. Thus, a generalized linear mixed models (GLMMs) will be best suitable for this analytic, the model can be described as:

$$Y_{ijk} = Ber(\ \mu_{ijk}\ ),$$

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \ X_{age}\ \beta_1 + X_{sex}\ \beta_2 + X_{RuralUrban}\ \beta_3 + X_{race}\ \beta_4 + \ U_{i\ states} + \ V_{ij\ school}$$

$$U(state)\ _i{\sim}N(0,\sigma^2), \frac{1}{\sigma_1^2} = Gamma(\ 1.5\times10^{-5}\ ),$$

$$V(school)\ _{ij}{\sim}N(0,\sigma^2), \frac{1}{\sigma_2^2} = Gamma(\ 1.5\times10^{-5}\ ),$$

Which that:

$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)$ with which $\frac{\mu_{ijk}}{1-\mu_{ijk}}$ is the odds of developing/developed a habit of smoking tobacco,

And X_{sex} represents the gender of the participant in the study, and changes depending on context, and is the fixed effect that we choose for the model.

The random effect that we choose are $U_{i\ states}$ which presents the i_{th} state and the $V_{ij\ school}$ presents j_{th} school in the i_{th} state.

3-Results

Through the R codes (attached in appendix), it generates the following information:

	0.5 quantile	***0.975 quantile***	***0.025 quantile***
Precision for state	0.009	0.004	0.031
Precision for school	0.830	0.707	0.995

Figure 4, table of precision for the random effects

First off, from above table, school has a high precision, thus having lower variance, which implies that the set of data is more concentrated towards the mean, it is lesser dispersed. This goes the opposite for state which has lower precision and higher variance, and are more dispersed, such that there is a greater difference between states, as the data in the set are spread out, far from the mean and from each other. This implies, there will not be as much difference on chances of developing a smoking habit on individuals going to different schools but will for individuals living in different states it will have a huge influence, such that an individual will have a very high chance on developing a smoking habit in a state, but very low chance on developing one if resides in another state, which suggest that the practice of chewing tobacco is more state regional.

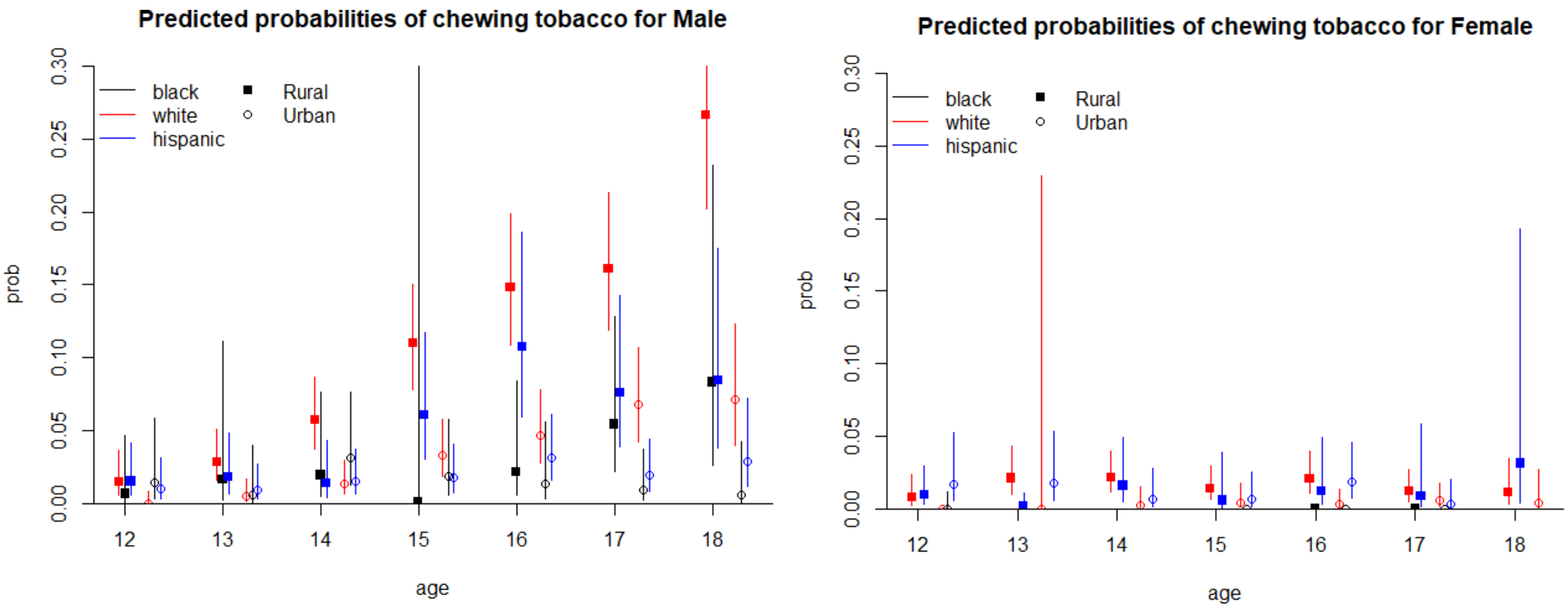

Figure 5, histogram of probability of smoking tobacco in relative to age group, race, and region of habiting for different school and states, for different gender

Through above histogram, we can easily conclude that the chance of an individual developing the habit of smoking tobacco will increase as the age of the individual increases, males have much higher chance than female, and that white Americans, individuals living in a rural region has a visibly more significant chance to develop a smoking habit than the other demographic groups.

In conclusion, the analysis lies in unison with our hypotheses formed, that tobacco chewing is strongly regional, that it is more commonly practiced in a certain state and lesser common in other states, this might be due to the state's policy and laws regarding to smoking being tougher than others, or due to local culture passed on from older generations such that the youth follow suit on their family adult's habit. And the general stereotypical image from the American TV and media, that white, male Americans who live in a rural region of the United States or Cowboys, has a higher chance on developing a habit of smoking tobacco.

Appendix

Please refer to the R codes used below

```
##################################################################
################### STA442 ASSIGNMENT 1 ###########################
##################
### QUESTION 1 ###
##################
library(Pmisc)

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)
library("INLA")
xres = inla(y ~ 0 + age +
       f(region,model="iid",prior='pc.prec',param=c(0.2,0.5))+
       f(school,model="iid",prior='pc.prec',param=c(0.2,0.5))+
       sex,
      control.fixed = list( mean = 0, prec = 1/(3^2) ),
      data = x, family = "gamma",
      #control.family = list(hyper = list(prec = list(prior = "loggamma", param = c(1e-04, 1e-04)))))
      control.family = list(hyper = list(prec = list(prior = "pc.prec", param = c(0.18, 0.5)))))

####
Pmisc::priorPostSd(xres, group = "random")$summary
Pmisc::priorPostSd(xres, group = "family")$summary
####

### plot the data
#hist(x$y, main = "Histogram of Y (grades)", xlab = "Y (grades)")

###summary table for β
knitr::kable(exp(xres$summary.fixed), digits = 3)
###summary table for σ
knitr::kable(round(Pmisc::priorPostSd(xres, group = "random")$summary[,c(2,3,5)],3))

##################################################################
################### STA442 ASSIGNMENT 1 ###########################
##################
### QUESTION 2 ###
##################

library(Pmisc)
library("INLA")
```

```
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")]

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
library("INLA", quietly = TRUE)
lincombList = inla.make.lincombs(as.data.frame(model.matrix(y ~ageFac * Sex * RuralUrban * Race,
lincombDf)))
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)]), digits = 3)
#########################################################################################
###########################

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')

###################################Male
plot####################################################
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)],
  ylim = c(0,0.3), #max(smokePredHere)
  xlab='age', ylab='prob',
  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$Race)])
legend('topleft', bty='n',
      ncol = 2,
      pch=c(rep(NA, length(Scol)), Spch),
      lty = rep(c(1,NA), c(length(Scol), length(Spch))),
      col = c(Scol, rep('black', length(Spch))),
      legend=c(names(Scol), names(Spch)))
##################################Female
plot################################################
toPlot = (lincombDf$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',
  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)))
```