Untitled

Nan Jiang

2020/11/3

Q1

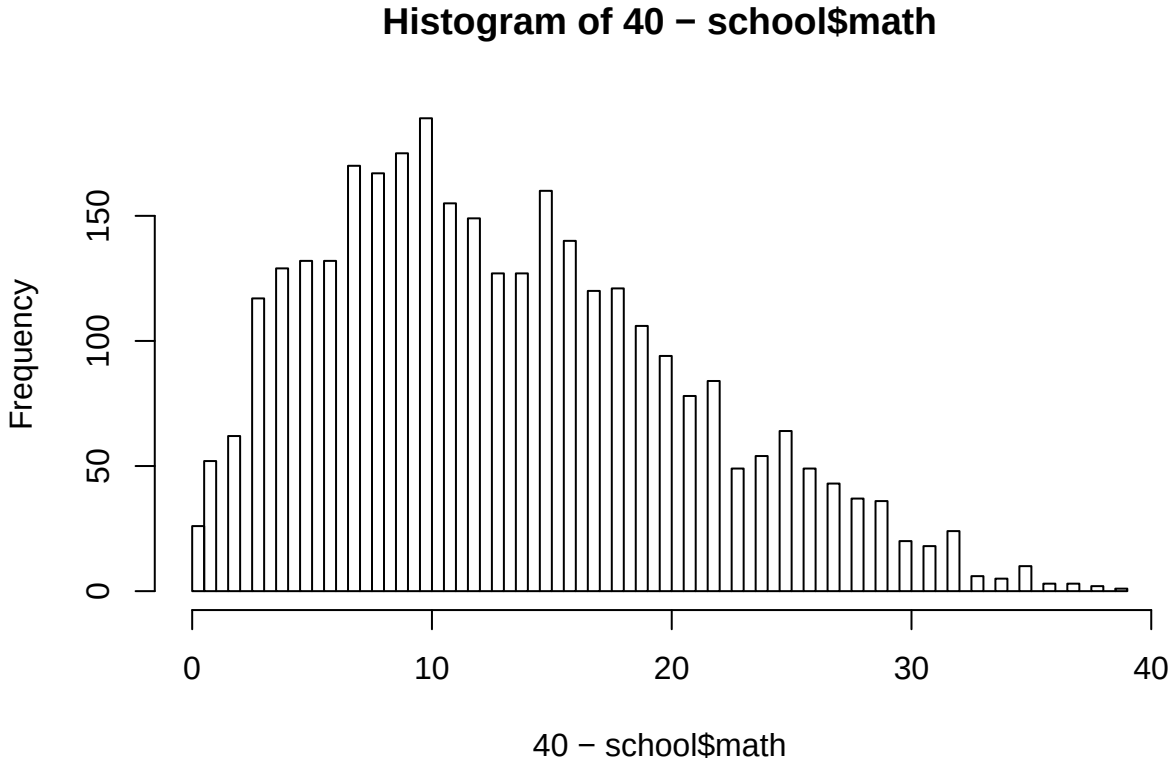
```
## Loading required package: Matrix

## Loading required package: sp

## Loading required package: parallel

## Loading required package: foreach

## This is INLA_20.03.17 built 2020-03-17 08:02:30 UTC.
## See www.r-inla.org/contact-us for how to get help.
## To enable PARDISO sparse library; see inla.pardiso()
```



	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	3.267	0.041	3.187	3.267	3.346	3.267	0
genderm	-0.010	0.015	-0.040	-0.010	0.020	-0.010	0
socialClassII	0.036	0.043	-0.048	0.036	0.121	0.036	0
socialClassIIIn	-0.029	0.046	-0.119	-0.029	0.061	-0.029	0
socialClassIIIm	-0.087	0.040	-0.166	-0.087	-0.008	-0.087	0
socialClassIV	-0.071	0.046	-0.160	-0.071	0.018	-0.071	0
socialClassV	-0.156	0.048	-0.249	-0.156	-0.062	-0.156	0
socialClasslongUnemp	-0.119	0.049	-0.216	-0.119	-0.022	-0.119	0
socialClasscurrUnemp	-0.147	0.068	-0.280	-0.147	-0.013	-0.147	0
socialClassabsent	-0.097	0.043	-0.181	-0.097	-0.014	-0.098	0
grade1	0.001	0.008	-0.016	0.001	0.017	0.001	0
grade2	0.181	0.009	0.164	0.181	0.198	0.181	0

```
##                              mean  0.025quant 0.975quant
## SD for studentUnique 0.22293155 0.210347291 0.23658788
## SD for classUnique   0.08578335 0.065283011 0.11658010
## SD for school        0.01982816 0.003008282 0.04771628
```

introduction

We analyzed the School Data Set by using an R version of the data set available at http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip. This is the data set containing the math scores for different students in different schools. We have two main analysis purposes. We want to find out he most important variable that influences the students' math score. And we want to find the solution to improve the students' test score.

method

We changed the response of gender to male and female. And changed the response of social class from number to different level of descripction from level I to absent.
We used the INLA method to make the Bayesian inference analysis. We treat the gender(M or F), social class(I to absent) and grade(0 to 2) as fixed effect while the school , student unique, class unique and school as the random effect.Y ijkl means the ith math score for the j'th student in the k'th class and the l'th school.

$$h(Y_{ijkl}) \sim X_{ijkl}\beta + U_i + V_{ij} + W_{ijk}$$
$$\beta \sim N(0, 0.2^2), \beta_1 \sim N(0, 10^{-2})$$
$$h(Y) = log(E(Y_{ijkl}))$$
$$U_i \sim MVN(0, \Sigma(\theta_1))$$
$$V_{ij} \sim MVN(0, \Sigma(\theta_2))$$
$$W_{ijk} \sim MVN(0, \Sigma(\theta_3))$$

Result

From the table, we notice that the gender, some social class and grade 1 are not statistically significant since their credible intervals contains 0. The social class II students have a positive impact on their math scores, usually 0.036 improvement on the log ratios.
We can see the credible intervals for standard diviation of individual is the highest, between 0.21 and 0.237, which means the biggest differences is between students. The credible intervals of standard diviation for schools has the lowest range between 0.03 and 0.048. It means the schools have the lowest influences on the math score within the three factors.

Conclusions

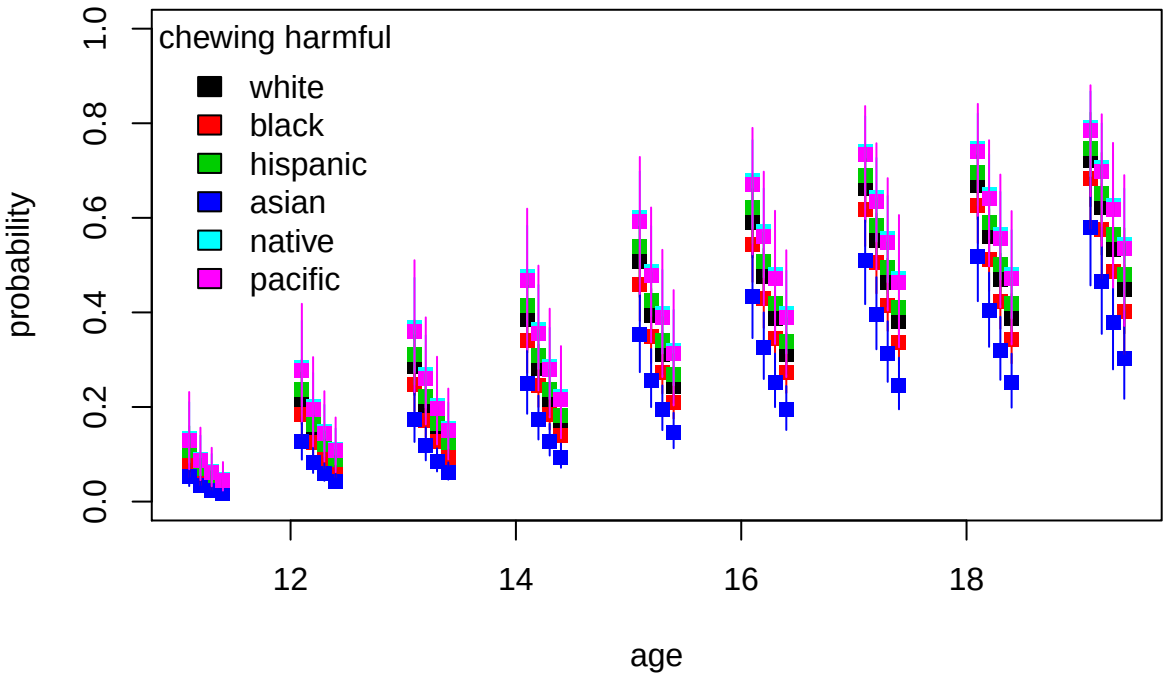
The effect level are individual > class> schools. Such that individual have the highest impact on the math score, and school have the least impact. Teachers are encouraged to find the individual with the poor test mark and gave them support.

Q2

```
##   Age ever_cigarettes Sex    Race state       school RuralUrban
## 1  18            TRUE   F   white    AL mdr_00013045      Rural
## 2  18            TRUE   M pacific    AL mdr_00013045      Rural
## 3  16            TRUE   M   white    AL mdr_00013045      Rural
```

```
## Warning in writeChar(lines[i], fp, nchars = nchar(lines[i]), eos = NULL):
## problem writing to connection
```

```
##                                         mean  0.025quant  0.975quant
## (Intercept)                       -2.713122756 -3.17800244 -2.24696478
## SexF                              -0.114454276 -0.19090916 -0.03806369
## ageFac12                           0.944243401  0.56238347  1.32578723
## ageFac13                           1.325518483  0.95242987  1.69829627
## ageFac14                           1.774972855  1.40343178  2.14619572
## ageFac15                           2.276664999  1.89547828  2.65742987
## ageFac16                           2.613699747  2.23149793  2.99549985
## ageFac17                           2.918176804  2.53608641  3.29988142
## ageFac18                           2.949239988  2.56067755  3.33739676
## ageFac19                           3.203224097  2.69601294  3.70999170
## Raceblack                         -0.190409023 -0.31875445 -0.06230574
## Racehispanic                       0.128623047  0.02466680  0.23243895
## Raceasian                         -0.631860149 -0.86134434 -0.40258372
## Racenative                         0.379567030  0.05970390  0.69906680
## Racepacific                        0.343086680 -0.20427168  0.88994980
## RuralUrbanRural                    0.465566174  0.07368177  0.85645790
## chewingHarmequal                  -0.433136186 -0.72313528 -0.14336479
## chewingHarmmore                   -0.819433007 -1.07941911 -0.55964362
## chewingHarmdunno                  -0.958143883 -1.21250963 -0.70396857
## RuralUrbanRural:chewingHarmequal -0.028698282 -0.43361605  0.37587816
## RuralUrbanRural:chewingHarmmore  -0.002046562 -0.36816836  0.36376776
## RuralUrbanRural:chewingHarmdunno -0.202514442 -0.56146134  0.15612717
## SD for state                       0.233784848  0.10257088  0.40972396
## SD for school                      0.494563666  0.41441275  0.58858019
```



Introduction

We analyzed the Smoke Data Set by using an R version of the data set available at http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData. This task concerns the 2014 American National Youth Tobacco Survey. This is the data set containing the smoke data from diffent teenagers in differnet area, school and religion. We have two questions.The first is which area should we target to avoid the teenager smoking problem, by state or by school. The second question is the difference between rural and urban and between states, which is bigger.

Method

We cleaned the data by exclude the year 9 and year 10 childrens since their data does not make sense. We changed the response of chewing harm from numerical to description, less, equal, more and dunno.
We used the INLA method to make the Bayesian inference analysis. We treat the Sex,ageFac,Race and the intersection of race and agefac, RuralUrban,chewingHarm and the intersection of RuralUrban and chewing Harm as fixed effect while the state, school as the random effect.Y ijk means the k'th for the k'th person's smoking in the i'th school and the j'th state.

$$Y_i \sim Binomial(N_i, \mu_i)$$

$$logit(\frac{\mu_{ijk}}{1-\mu_{ijk}}) \sim X_{ijk}\beta + U_i + V_{ij}$$

We made two null hypothesis. The first is eographic variation (between states) in the rate of students smoking cigarettes is substantially greater than variation amongst schools. As a result, tobacco control programs should target the states with the most smoking and not concern themselves with nding particular schools where smoking is a problem. The second is Rural-urban differences are much greater than di erences between states.

Result

From the table, we notice that the racepacific,RuralUrbanRural:chewingHarmequal are not statistically significant sincetheir credible intervals contains 0. The female tends to have a 2.71 times lower possibility of smoking than male. And it appears that possibility of smoking gets higher as the age of teenger gets older. The asian race people have a 0.63 times lower possibility of smoking than the baseline. The native race people have a 0.37 times higher possibility of smoking than the baseline which is the highest. And the people think chewing makes dunno harm has the lowest possibility of smoking among the population. Compare to the state and the schools, the range of credible intervals for standard deviation of schools is much higher than the range of credible intervals for state.
Since the ruralurban variable are categorical variable, we can easily compare it with the random effect. Such that we can see the schools have a highest effect among the three variables(schools, urbanrural and states), the stae have the lowest impact, and the area is in between. Such that the hypothesis one is false and the hypothesis two is true.
I tried to combine the age and area variables in the model and found the coefficients does not make any sense according to the credit interval. Such that I divided the variables to two different independent variables. We can see the effect of race based on the graph. The graph shows that pacific people tends to have higher probability of smoking, and the asian people have the lowest probability of smoking among all ages.

conclusion

The pacific people have the highest probability of smoking among all ages. And the asian people have the lowest probability.According to the table, the schools have the highest impact on the teenager smoking problem. We suggest to take action to avoid the teenager smoking among the worst performence schools instead of states since the school have a higher impact on the probability of teenager smoking.

HW2

Nan Jiang

2020/11/2

```
knitr::opts_chunk$set(echo = FALSE)
sUrl = "http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip"
#dir.create(file.path("..", "data"), showWarnings = FALSE)
#(Pmisc::downloadIfOld(sUrl, file.path("..", "data")))
school = read.fwf("/Users/nanjiang/Desktop/R/sta442a2/jsp/JSP.DAT", widths = c(2,
  1, 1, 1, 2, 4, 2, 2, 1), col.names = c("school",
  "class", "gender", "socialClass", "ravensTest",
  "student", "english", "math", "year"))
school$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)
#schoolLme = glmmTMB::glmmTMB(math ~ gender + socialClass +
#  grade + (1 | school) + (1 | classUnique) + (1 |
#  studentUnique), data = school)
#summary(schoolLme)
hist(40 - school$math, breaks = 100)
library("INLA")
```

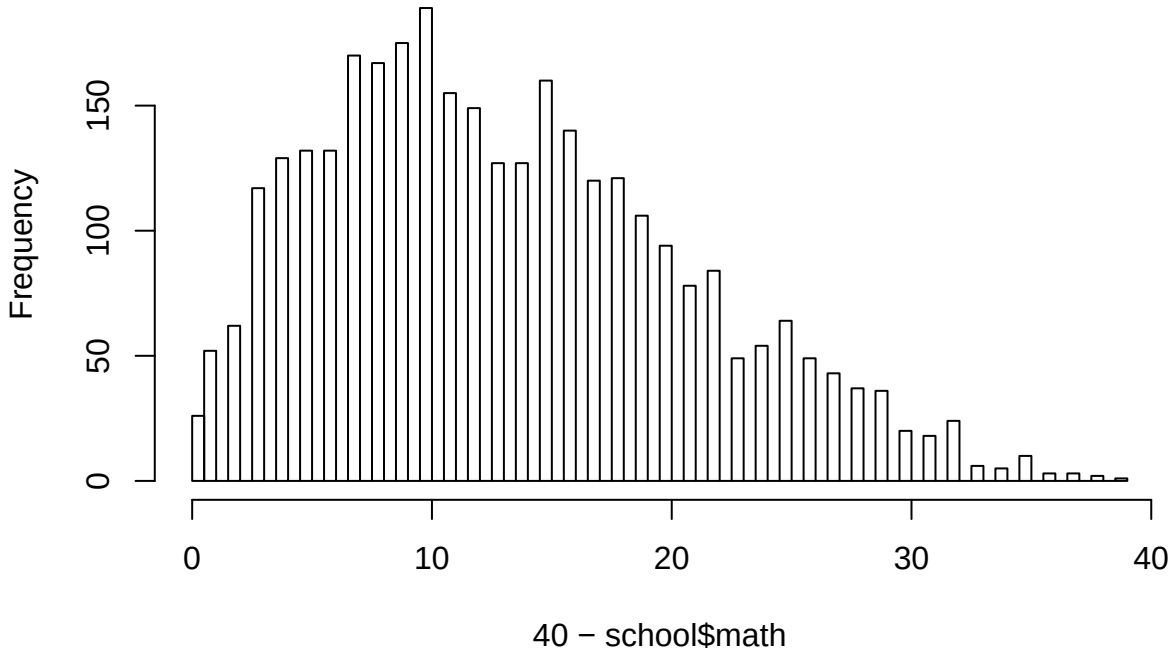
```
## Loading required package: Matrix

## Loading required package: sp

## Loading required package: parallel

## Loading required package: foreach

## This is INLA_20.03.17 built 2020-03-17 08:02:30 UTC.
## See www.r-inla.org/contact-us for how to get help.
## To enable PARDISO sparse library; see inla.pardiso()
```




```
schoolformula = math ~ gender + socialClass +
  grade + (1 | school) + (1 | classUnique) + (1 |
  studentUnique)
school = school[!is.na(school$student)& !is.na(school$class)&!is.na(school$student),]
SchoolINLA = inla(math ~ gender + socialClass +
  grade +f(studentUnique, hyper = list(prec = list(prior = 'pc.prec',param = c(0.1,0.05))))+f(classUniqu

knitr::kable(rbind(SchoolINLA$summary.fixed[,c("mean", "0.025quant", "0.975quant")],Pmisc::priorPostSd(S
```

	mean	0.025quant	0.975quant
(Intercept)	3.267	3.187	3.346
genderm	-0.010	-0.040	0.020
socialClassII	0.036	-0.048	0.121
socialClassIIIn	-0.029	-0.119	0.061
socialClassIIIm	-0.087	-0.166	-0.008
socialClassIV	-0.071	-0.160	0.018
socialClassV	-0.156	-0.249	-0.062
socialClasslongUnemp	-0.119	-0.216	-0.022
socialClasscurrUnemp	-0.147	-0.280	-0.013
socialClassabsent	-0.097	-0.181	-0.014
grade1	0.001	-0.016	0.017
grade2	0.181	0.164	0.198
SD for studentUnique	0.223	0.210	0.237
SD for classUnique	0.086	0.065	0.117
SD for school	0.020	0.003	0.048

```
##   Age ever_cigarettes Sex    Race state       school RuralUrban
```

```
## 1  18            TRUE   F   white    AL mdr_00013045      Rural
## 2  18            TRUE   M pacific    AL mdr_00013045      Rural
## 3  16            TRUE   M   white    AL mdr_00013045      Rural

## Warning in writeChar(lines[i], fp, nchars = nchar(lines[i]), eos = NULL):
## problem writing to connection

##                                          mean  0.025quant  0.975quant
## (Intercept)                       -2.713023524 -3.17789519 -2.24687801
## SexF                              -0.114453788 -0.19090860 -0.03806327
## ageFac12                           0.944241859  0.56238114  1.32578644
## ageFac13                           1.325519783  0.95243099  1.69829772
## ageFac14                           1.774978499  1.40343793  2.14620089
## ageFac15                           2.276678513  1.89549570  2.65743991
## ageFac16                           2.613710587  2.23151308  2.99550674
## ageFac17                           2.918186760  2.53610066  3.29988741
## ageFac18                           2.949250812  2.56069267  3.33740364
## ageFac19                           3.203230869  2.69602298  3.70999534
## Raceblack                         -0.190407650 -0.31875308 -0.06230445
## Racehispanic                       0.128634778  0.02467929  0.23244976
## Raceasian                         -0.631848522 -0.86133158 -0.40257315
## Racenative                         0.379582687  0.05972081  0.69908125
## Racepacific                        0.343099000 -0.20425791  0.88996071
## RuralUrbanRural                    0.465462707  0.07360661  0.85632234
## chewingHarmequal                  -0.433139070 -0.72313769 -0.14336816
## chewingHarmmore                   -0.819438242 -1.07942381 -0.55964941
## chewingHarmdunno                  -0.958149861 -1.21251505 -0.70397513
## RuralUrbanRural:chewingHarmequal  -0.028695593 -0.43361277  0.37588027
## RuralUrbanRural:chewingHarmmore   -0.002040796 -0.36816196  0.36377290
## RuralUrbanRural:chewingHarmdunno  -0.202507922 -0.56145420  0.15613309
## SD for state                       0.233784848  0.10257088  0.40972396
## SD for school                      0.494563666  0.41441275  0.58858019
```