Jiahao Cheng Assignment 2 November 2, 2020

1. Student Leaver's Report: What are the most important influences on student performance on math tests?

Summary

We analyzed the math scores of over 1000 students from three Junior school. We explored that students are most significant on their math scores. Thus, it would be better to identify individual weak students and give them extra attention. In addition, it seems like higher grade students and students who had higher social class perform better. Moreover, the gender is not significant in this investigation.

Introduction

We used data collected from the Junior School Project by Mortimore's group which is about the math score over 1000 students measured over three school years to analyze what are the most significant influences on student performance on their math tests. They investigated each student's gender, school, class, ID, social class, school year, and scores of mathematics tests. We wanted to figure out how do we give extra help to students who did poorly on the test, individually, by class, or by school.

Method

First, we draw a histogram about the number of questions the students gets wrong. In order to do this

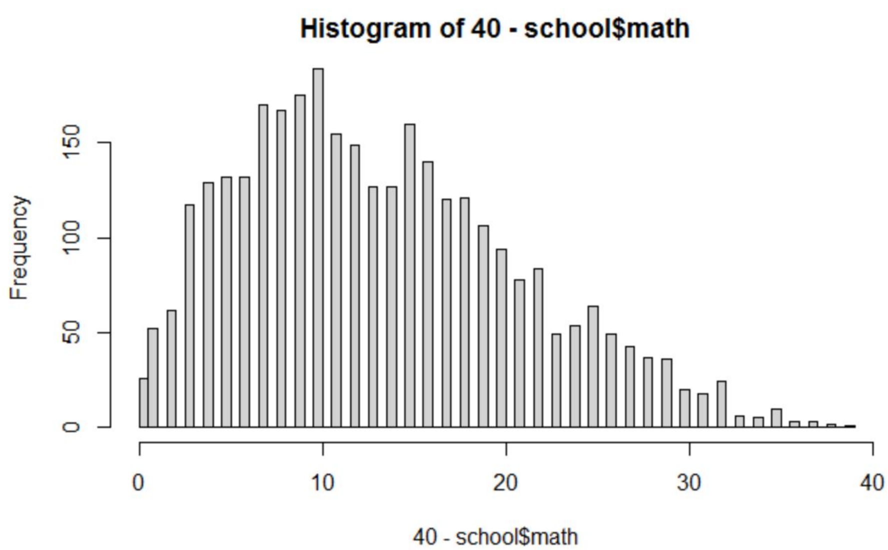

Figure 1: The histogram of the number of questions the students gets wrong.

analysis, we built a generalised linear mixed model using the given dataset which considered the school, the class, and the student itself as the random effect. In addition, we applied a Bayesian analysis for β. We used the same parameters for the Fiji data from slide. Since we have no information about general performances among the school or the class, we used the default setting for the parameters of random effects. Since the number of questions the students gets wrong look like a Poisson distribution, our model would be

$$
\begin{aligned}
Y_{ijkl} &\sim Poisson(\lambda_{ijkl}) \\
log(\lambda_{ijkl}) &= \beta_0 + X_{ijk}\boldsymbol{\beta_1} + U_i + V_{ij} + W_{ijk} \\
U_i &\sim \mathcal{N}(0, \tau_U \Sigma) \\
V_{ij} &\sim \mathcal{N}(0, \tau_V \Sigma) \\
W_{ijk} &\sim \mathcal{N}(0, \tau_W \Sigma)
\end{aligned}
$$

with priors

$$
\begin{aligned}
\beta_0 &\sim \mathcal{N}(0, 10^2) \\
\boldsymbol{\beta_1} &\sim \mathcal{N}(0, 0.2^2) \\
\log(\tau_U) &\sim \log \Gamma(1, 0.00005) \\
\log(\tau_V) &\sim \log \Gamma(1, 0.00005) \\
\log(\tau_W) &\sim \log \Gamma(1, 0.00005)
\end{aligned}
$$

- Y_ijkl indicates the l^{th} math score for k^{th} student from the j^{th} class in the i^{th} school;
- λ_{ijkl} is the proportion of the l^{th} math score for the kth student from the j^{th} class in the i^{th} school;
- X_{ijk} represents a collection of personal information for the k^{th} student from the j^{th} class in the i^{th} school including the the gender, the social class of their father, and the year in the Junior school;
- β_0 is the intercept of the model, which revels the performance of the base line;
- $\boldsymbol{\beta_1}$ are coefficients for all covariates X_{ijk}, which revels the influences of each factor;
- U_i represents the random effect of the i^{th} school.
- V_{ij} represents the random effect of the j^{th} class from i^{th} school;
- W_{ijk} represents the random effect of the k^{th} student from the j^{th} class in the i^{th} school.

Figure 2 shows the prior distribution for β_0, and some of β_1.
For the data used for the model, we remain the intercept, which means that the base line is the female students who are in the first year in the Junior school with the highest social class.

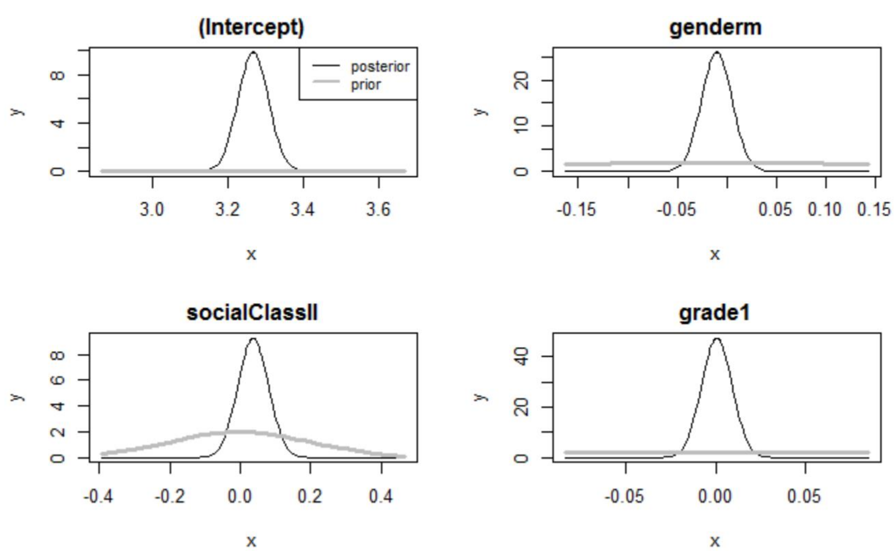

Figure 2: The priors of the intercept, the genderm, the socialClass II, and the grade1.

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	26.223	0.040	3.187	3.267	3.346	3.267	0
genderm	0.990	0.015	-0.040	-0.010	0.020	-0.010	0
socialClassII	1.038	0.043	-0.048	0.037	0.122	0.037	0
socialClassIIIn	0.972	0.046	-0.119	-0.029	0.062	-0.029	0
socialClassIIIm	0.918	0.040	-0.165	-0.086	-0.007	-0.086	0
socialClassIV	0.932	0.046	-0.160	-0.070	0.019	-0.070	0
socialClassV	0.856	0.048	-0.249	-0.156	-0.062	-0.156	0
socialClasslongUnemp	0.888	0.050	-0.216	-0.119	-0.022	-0.119	0
socialClasscurrUnemp	0.864	0.068	-0.279	-0.146	-0.012	-0.146	0
socialClassabsent	0.908	0.043	-0.181	-0.097	-0.013	-0.097	0
grade1	1.001	0.008	-0.016	0.001	0.017	0.001	0
grade2	1.198	0.009	0.164	0.181	0.198	0.181	0

Figure 3: The table of odds of the school leaver's model with confidence intervals.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Precision for school	0.007	0.007	0.028	0.009	0.004	0.017
Precision for classUnique	0.082	0.152	0.110	0.084	0.063	0.087
Precision for studentUnique	0.224	0.913	0.237	0.224	0.211	0.224

Figure 4: The table of precision of the school leaver's model for their random effects with confidence intervals.

Result

Figure 3 shows the exponential coefficients of our model with its confidence intervals, which is convenient to do comparison by computing the odd ratio. We explored that higher grade students are more like to have better performance on math tests. For the second year students, their math scores are about 1.038 ($\frac{1.001+26.223}{26.223}$) times higher than those for the first year students in Junior school, and for the third year student, it is about 1.046 ($\frac{1.198+26.223}{26.223}$) times higher regardless of other factors. In addition, we can see that as the social class gets lower, the mean becomes smaller, which shows that the math scores of students whose father's social class is more likely to be unemployment are lower than those for students whose father has jobs.
After we took into account other factors, we found that only grade2, and some types of the social class are statistically significant because the confidence interval does not include 0. In addition, the confidence intervals of the gender, grade1, social class level II, III, and IV, include 0, which gives that we can ignore those factors.
According to Figure 4, we can see that students play a significant role in our investigation, which reveals that students are most significant comparing with their class and school since the studentUnique has the highest mean and its confidence interval does not overlap that of the other two random effects. The lower bound of its confidence interval is greater than the upper bound of that for the school and class random effects. Thus, we can certain that students themselves are the most important influences on their math scores.

Appendix

We used R to built the models, which can be accessed in Appendix 1.

2. Smoking Report: How do we target smoking people and how age affects smoking among different groups?

Summary

We analyzed the smoking situation amongst American students by a mixed logistic regression models using the data from 2014 American National Youth Tobacco Survey. We explored that variation amongst schools is much greater than variation between states. Thus, tobacco control programs should target particular schools to deal with the smoking problem. In addition, rural-urban differences are much greater than differences between states. Moreover, the effect of age plays a significant role on smoking mostly for the Black Americans, followed by Hispanic-Americans, and finally whites from the comparison of these three ethnic groups, that as students getting older, they are more likely to smoke.

Introduction

We used the data collected from the 2014 American National Youth Tobacco Survey to analyze the distribution of smoking people among American youth who are older than 10 in terms of their ethnicity, region, sex, and age. This survey focused on the use of cigars, hookahs, and chewing tobacco among American school children. We wanted to figure out geographic variation and variation amongst schools in the rate of students smoking cigarettes as well as the differences between rural and urban area. In addition, we compared the differences of in the effect of age on smoking for different ethnic groups.

Method

In order to do these analysis, we built two models using the Smoke dataset which eliminating people who are younger than 10. Since the response is a yes or no question, we used a logistic regression model to explore the factors influencing the chance that an individual smokes. The first model used sex and the interaction of the age and the race as covariates, which is

$$
\begin{aligned}
Y_{ijk} &\sim \text{Bernoulli}(\mu_{ijk}) \\
\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) &= X_{ij}\beta + A_i + B_{ij} \\
A_i &\sim \mathcal{N}(0, \tau_A \Sigma) \\
B_{ij} &\sim \mathcal{N}(0, \tau_B \Sigma)
\end{aligned}
$$

- Y_{ijk} indicates whether the k^{th} student who is greater than 10 years old use chewing tobacco, snuff or dip at least once in the past 30 days from the j^{th} school in the i^{th} state;
- μ_{ijk} is the proportion of k^{th} students who is greater than 10 years old using chewing tobacco, snuff or dip at least once in the past 30 days from the j^{th} school in the i^{th} state;
- X_{ij} represents the sex, the region (rural or urban), and the interaction between the age and the race for the students from j^{th} school in the i^{th} state;
- β are coefficients for all covariates X_i, which revels the influences of each factor;
- A_i represents the random effect of the i^{th} state;
- B_{ij} represents the random effect of the j^{th} school from i^{th} state.

We wanted to figure out the prior for the state and the school. For the variability in the rate of smoking between states, since some states having double or triple the rate of smoking update. Thus, we can get $\frac{\exp(U_{\text{worst}})}{\exp(U_{\text{best}})} = 2$ or 3. I chose 3 to do computation here. Since we are suggested to consider the 90^{th} quantile of smoking to the worst and the 10^{th} quantile to the best, we can get $3 = \frac{\exp(1.28\sigma)}{\exp(-1.28\sigma)} \approx \exp(2.5\sigma)$. Thus, $\sigma = \frac{\log(3)}{2.5}$, which means that the prior would be $pc.param(\frac{\log(3)}{2.5}, 0.5)$. Similarly, we

can get the prior for the school would be $pc.param(\frac{1.5}{2.5}, 0.15)$ since I took the median for the differences of 10% to 20% in somking rates that is typical.
Thus, the priors would be

$$\tau_A \sim PC.param(\frac{\log(3)}{2.5}, 0.5) \text{ where } Pr(\sigma > \frac{\log(3)}{2.5}) = 0.5$$
$$\tau_B \sim PC.param(\frac{\log(1.5)}{2.5}, 0.15) \text{ where } Pr(\sigma > \frac{\log(1.5)}{2.5}) = 0.15$$

However, according to the Figure 5, we found that the prior does not fit the posterior well. Based on

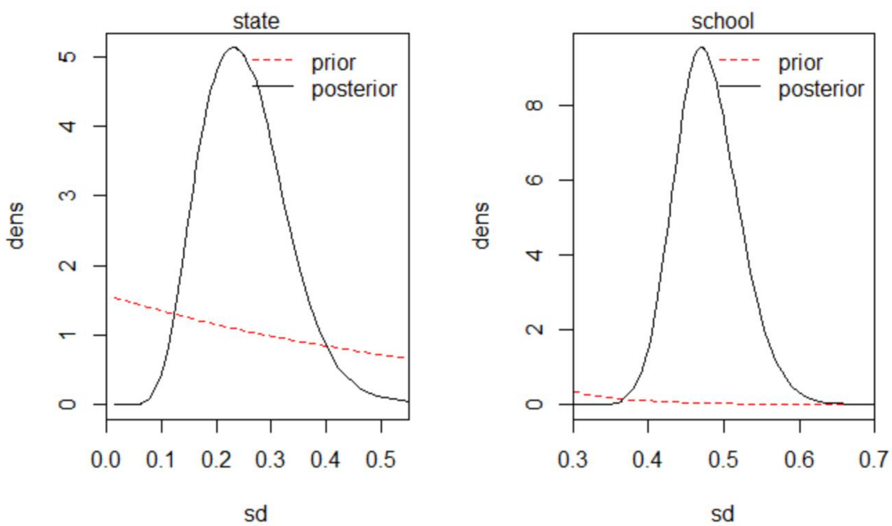

Figure 5: The priors of the state and school for the first smoke model.

the posteriors, we adjusted priors for both random effects. In addition, we wanted to build the model that interacts more confounders. Thus, for the second model, We used sex and the interaction of the age, the region, and the race as covariates, which becomes

$$Y_{ijk} \sim \text{Bernoulli}(\mu_{ijk})$$
$$\log(\frac{\mu_{ijk}}{1-\mu_{ijk}}) = X_{ij}\beta + A_i + B_{ij}$$
$$A_i \sim \mathcal{N}(0, \tau_A \Sigma)$$
$$B_{ij} \sim \mathcal{N}(0, \tau_B \Sigma)$$

with priors

$$\tau_A \sim PC.param(0.4, 0.1) \text{ where } Pr(\sigma > 0.4) = 0.1$$
$$\tau_B \sim PC.param(0.55, 0.1) \text{ where } Pr(\sigma > 0.55) = 0.1$$

- Y_{ijk} indicates whether the k^{th} student who is greater than 10 years old use chewing tobacco, snuff or dip at least once in the past 30 days from the j^{th} school in the i^{th} state;
- μ_{ijk} is the proportion of k^{th} students who is greater than 10 years old using chewing tobacco, snuff or dip at least once in the past 30 days from the j^{th} school in the i^{th} state;
- X_{ij} represents the sex, and the interaction between the age, the region (rural or urban), and the race for the students from j^{th} school in the i^{th} state;
- β are coefficients for all covariates X_i, which revels the influences of each factor;
- A_i represents the random effect of the i^{th} state;
- B_{ij} represents the random effect of the j^{th} school from i^{th} state.

We chose the second model to do analysis since the first model is nested within the second model. Thus, we can get more information from the second one. The prior and the posterior distribution for the state and the school are shown in Figure 6.

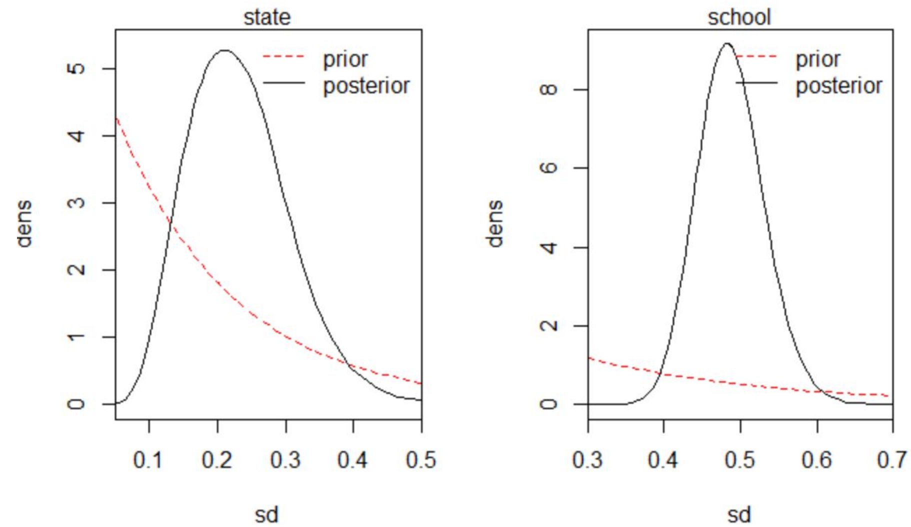

Figure 6: The priors of the state and school for the second smoke model.

Result

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	0.019	0.506	-4.982	-3.989	-2.996	-3.989	0
SexF	0.846	0.039	-0.243	-0.167	-0.091	-0.167	0
ageFac12	2.293	0.546	-0.242	0.830	1.901	0.830	0
ageFac13	4.774	0.517	0.549	1.563	2.577	1.563	0
ageFac14	7.193	0.516	0.960	1.973	2.985	1.973	0
ageFac15	16.169	0.516	1.771	2.783	3.795	2.783	0
ageFac16	22.257	0.515	2.091	3.103	4.113	3.103	0
ageFac17	38.235	0.514	2.635	3.644	4.652	3.644	0
ageFac18	34.881	0.520	2.530	3.552	4.573	3.552	0
ageFac19	144.144	0.967	3.072	4.971	6.868	4.971	0
RuralUrbanRural	4.203	0.576	0.305	1.436	2.566	1.436	0
Raceblack	0.678	1.113	-2.575	-0.389	1.795	-0.389	0
Racehispanic	1.376	0.676	-1.008	0.319	1.645	0.319	0
Racenative	3.736	1.144	-0.929	1.318	3.563	1.318	0

Figure 7: The table of odds for some fixed effects part 1.

In order to to do the comparison between students for different ages among white, Hispanic, and black ethic group, we set the 11-year-old urban male students as the baseline of our model. Thus, according to Figure 7, 8, and 9, which show the exponential coefficients of the second logistic model, we can easily do comparison by computing the odds ratio between different groups. We can get that older students are more likely to smoke in general since the odds as well as its confidence intervals are strictly increasing except age 18 as the year goes up. For the black students, the effect of age is more significant

ageFac12:Raceblack	3.456	1.159	-1.035	1.240	3.514	1.240	0
ageFac13:Raceblack	1.559	1.141	-1.797	0.444	2.683	0.444	0
ageFac14:Raceblack	1.984	1.135	-1.543	0.685	2.911	0.685	0
ageFac15:Raceblack	1.439	1.131	-1.857	0.364	2.583	0.364	0
ageFac16:Raceblack	1.364	1.130	-1.908	0.310	2.527	0.310	0
ageFac17:Raceblack	0.867	1.127	-2.356	-0.143	2.068	-0.143	0
ageFac18:Raceblack	0.975	1.136	-2.255	-0.025	2.202	-0.025	0
ageFac19:Raceblack	0.295	1.466	-4.099	-1.222	1.654	-1.222	0
ageFac12:Racehispanic	1.905	0.726	-0.781	0.644	2.068	0.644	0
ageFac13:Racehispanic	1.124	0.700	-1.257	0.116	1.489	0.116	0
ageFac14:Racehispanic	1.395	0.696	-1.033	0.333	1.698	0.333	0
ageFac15:Racehispanic	0.903	0.692	-1.461	-0.102	1.256	-0.102	0
ageFac16:Racehispanic	0.770	0.691	-1.619	-0.261	1.095	-0.261	0
ageFac17:Racehispanic	0.626	0.690	-1.823	-0.468	0.886	-0.468	0
ageFac18:Racehispanic	0.729	0.701	-1.692	-0.316	1.059	-0.316	0
ageFac19:Racehispanic	0.415	1.138	-3.113	-0.879	1.353	-0.879	0

Figure 8: The table of odds for some fixed effects part 2.

ageFac12:Racenative	0.701	1.285	-2.877	-0.355	2.165	-0.355	0
ageFac13:Racenative	0.346	1.275	-3.565	-1.062	1.439	-1.062	0
ageFac14:Racenative	0.801	1.232	-2.640	-0.221	2.195	-0.221	0
ageFac15:Racenative	0.450	1.287	-3.325	-0.798	1.727	-0.798	0
ageFac16:Racenative	0.382	1.322	-3.557	-0.962	1.631	-0.962	0
ageFac17:Racenative	0.165	1.437	-4.622	-1.802	1.016	-1.802	0
ageFac18:Racenative	44.445	10.101	-16.037	3.794	23.609	3.794	0
ageFac19:Racenative	1.000	31.623	-62.086	-0.001	62.034	0.000	0

Figure 9: The table of odds for some fixed effects part 4.

	mean	**sd**	**0.025quant**	**0.5quant**	**0.975quant**	**mode**
SD for state	0.231	0.075	0.107	0.225	0.395	0.210
SD for school	0.487	0.044	0.407	0.486	0.579	0.482

Figure 10: The table of odds for the random effects.

on smoking because the odds ratio are over 1 until age 16. Similarly, the odds ratio of the Hispanic students are over 1 until age 14. Focusing on the white Americans, the effect on age does not play a significant role while there are too many smoking people among the group of age 18 comparing with other ethic groups.
According to Figure 10, we can see that the standard error of the schools within a state, 0.487, is greater than that of the state, 0.231. Since the overall state confidence interval is lower than that of school within a state, we can conclude that there is reasonable certainty that the differences between schools within a state in chewing tobacco usage amongst American Youth are larger than that for the state-level differences. In addition, since the rural-urban differences are about $\frac{4.203}{0.019} = 221$ times, and both confidence intervals do not include 0, we can conclude that rural-urban differences are much greater than the state differences.

Appendix

We used R to built the models, which can be accessed in Appendix 2.

Appendix

Appendix 1 R code for Question 1

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

schoolLme = glmmTMB::glmmTMB(math ~ gender + socialClass +
                                grade + (1 | school) + (1 | classUnique) +
                                (1 | studentUnique), data = school)
summary(schoolLme)
hist(40 - school$math, breaks = 100)
schoolInla = inla(math ~ gender + socialClass +
                    grade + f(school, model='iid') + f(classUnique, model='iid') +
                    f(studentUnique, model='iid'), data = school,
                  family='poisson',
                  control.fixed=list(mean=0, mean.intercept=0, prec=0.2^(-2),
                                     prec.intercept=10^(-2)),
                  control.compute=list(return.marginals=TRUE))
schoolTable <- schoolInla$summary.fixed
schoolTable[, 1] <- exp(schoolTable[, 1])
knitr::kable(schoolTable, digits=3)

schoolTable2 <- 1/sqrt(schoolInla$summary.hyper)
# schoolTable2[, 1] <- exp(schoolTable2[, 1])
knitr::kable(schoolTable2, digits=3)
# draw prior + posterior
par(mfrow=c(2, 2))
names <- c(names(schoolInla$marginals.fixed)[1: 3],
           names(schoolInla$marginals.fixed)[11])
plot(schoolInla$marginals.fixed$`(Intercept)`, type='l', main=names[1])
lines(schoolInla$marginals.fixed$`(Intercept)`[, 'x'],
      dnorm(schoolInla$marginals.fixed$`(Intercept)`[, 'x'], mean=0, sd=10),
      col='grey', lwd=3)
legend("topright", legend=c("posterior", "prior"),
       col=c("black", "grey"), lty=1:1, lwd=1:3, cex=0.8)
plot(schoolInla$marginals.fixed$genderm, type='l', main=names[2])
lines(schoolInla$marginals.fixed$genderm[, 'x'],
      dnorm(schoolInla$marginals.fixed$genderm[, 'x'], mean=0, sd=0.2),
      col='grey', lwd=3)
plot(schoolInla$marginals.fixed$socialClassII, type='l', main=names[3])
lines(schoolInla$marginals.fixed$socialClassII[, 'x'],
      dnorm(schoolInla$marginals.fixed$socialClassII[, 'x'], mean=0, sd=0.2),
      col='grey', lwd=3)
plot(schoolInla$marginals.fixed$grade1, type='l', main=names[4])
lines(schoolInla$marginals.fixed$grade1[, 'x'],
      dnorm(schoolInla$marginals.fixed$grade1[, 'x'], mean=0, sd=0.2),
      col='grey', lwd=3)

par(mfrow=c(2, 2))
schoolPrior = Pmisc::priorPostSd(schoolInla)
do.call(matplot, schoolPrior$school$matplot)
do.call(legend, schoolPrior$legend)
mtext("school", side=3)

do.call(matplot, schoolPrior$classUnique$matplot)
do.call(legend, schoolPrior$legend)
mtext("classUnique", side=3)
```

```

do.call(matplot, schoolPrior$studentUnique$matplot)
do.call(legend, schoolPrior$legend)
mtext("studentUnique", side=3)
```

Appendix 2 R code for Question 2

```
dataDir = "../data"
smokeFile = file.path(dataDir, "smoke2014.RData")
if (!file.exists(smokeFile)) {
  download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
                smokeFile)
}

load(smokeFile)
smoke[1:3, c("Age", "ever_cigarettes", "Sex", "Race",
             "state", "school", "RuralUrban")]

forInla = smoke[smoke$Age > 10, c("Age", "ever_cigarettes",
                                  "Sex", "Race", "state", "school", "RuralUrban",
                                  "Harm_belief_of_chewing_to")]
forInla = na.omit(forInla)
forInla$y = as.numeric(forInla$ever_cigarettes)
forInla$ageFac = factor(as.numeric(as.character(forInla$Age)))
# forInla$ageFac = relevel(factor(forInla$Age), '14')
forInla$chewingHarm = factor(forInla$Harm_belief_of_chewing_to,
levels = 1:4, labels = c("less", "equal", "more", "dunno"))


toPredict = expand.grid(ageFac = levels(forInla$ageFac),
RuralUrban = levels(forInla$RuralUrban), Race = levels(forInla$Race),
Sex = levels(forInla$Sex))
forLincombs1 = do.call(inla.make.lincombs,
                       as.data.frame(model.matrix(~Sex + RuralUrban + ageFac * Race,
                                                  data = toPredict)))
smokeModel1 = inla(y ~ Sex + RuralUrban + ageFac * Race +
               f(state, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(log(3)/2.5, 0.5))))
               + f(school, model = "iid", hyper =
                     list(prec = list(prior = "pc.prec", param = c(log(1.5)/2.5, 0.15))
                         )),
             data = forInla, family = "binomial", control.inla =
               list(strategy = "gaussian"), lincomb = forLincombs1)

par(mfrow=c(1, 2))
smokePrior1 = Pmisc::priorPostSd(smokeModel1)
do.call(matplot, smokePrior1$state$matplot)
do.call(legend, smokePrior1$legend)
mtext("state", side=3)

do.call(matplot, smokePrior1$school$matplot)
do.call(legend, smokePrior1$legend)
mtext("school", side=3)

forLincombs2 = do.call(inla.make.lincombs,
                       as.data.frame(model.matrix(~Sex + ageFac *
                                                    RuralUrban * Race,
                                                  data = toPredict)))
smokeModel2 = inla(y ~ Sex + ageFac * RuralUrban * Race +
               f(state, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(0.4, 0.1))))
               + f(school, model = "iid", hyper =
                     list(prec = list(prior = "pc.prec", param = c(0.55, 0.1)))),
             data = forInla, family = "binomial", control.inla =
               list(strategy = "gaussian"), lincomb = forLincombs2)

```

```
par(mfrow=c(1, 2))
smokePrior2 = Pmisc::priorPostSd(smokeModel2)
do.call(matplot, smokePrior2$state$matplot)
do.call(legend, smokePrior2$legend)
mtext("state", side=3)

do.call(matplot, smokePrior2$school$matplot)
do.call(legend, smokePrior2$legend)
mtext("school", side=3)

rbind(smokeModel2$summary.fixed[, c("mean", "0.025quant", "0.975quant")],
      Pmisc::priorPostSd(smokeModel2)$summary[, c("mean", "0.025quant",
                                                  "0.975quant")])
# create matrix of predicted probabilities
theCoef2 = exp(smokeModel2$summary.lincomb.derived[, c("0.5quant",
                                                "0.025quant", "0.975quant")])
theCoef2 = theCoef2/(1 + theCoef2)

smokeTable <- smokeModel2$summary.fixed
smokeTable[, 1] <- exp(smokeTable[, 1])
knitr::kable(smokeTable[c(1:13, 15, 25:40, 49:56),], digits=3)
knitr::kable(Pmisc::priorPostSd(smokeModel2)$summary, digits=3)
```