Analyzing When U.S Grade School Students First Begin Smoking Cigarettes

Sri-Amirthan Theivendran

Introduction

Grade school students smoking cigarettes is a big concern for parents. As a result, based on the 2014 American National Youth Tobacco Survey, we aim to answer whether geographic variation (among states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. We also aim to answer whether two non-smoking childen have the same probability of trying cigarettes within the next month regardless of their age controlling for fixed effects such as sex, whether they are an urban/rural student and their ethnicity as well as random effects such as state and school. The first of these questions is of great importance to youth smoking prevention programs as it determines whether money ought to be spent at the state or school level.

The data were obtained based on a school-based, self-administered, pencil-and-paper questionnaire administered to U.S. middle school and high school students in 2014. The questionnaire asked various questions about students beliefs and use practices pertaining to cigars, hookahs, chewing tobacco, and other tobacco based products.

Methodology

To answer the research questions above, we fit a weibull generalized linear mixed model to the time that a student first begins smoking with fixed effects of sex, ethnicity and whether they are a rural/urban student and random effects of state and school and conduct bayesian inference on this model. The time that a student first begins smoking is a censored random variable that is sometimes right censored. In addition, we choose to left censor observations at eight years of age. Models that included interactions yielded results that were similar to the model just mentioned. Hence on the basis of simplicity and interpretability we choose to exclude them to answer the research questions.

The model mentioned above can be written as follows. Let Y be the censored random variable of the time a student first begins smoking. Then the mth observation of the k th school from the j th state is such that $Y_{j,k,m} \mid U_j V_{j,k} \sim \text{Weibull}(\lambda_{j,k,m}, \alpha)$ where $U_j \sim N(0, \sigma^2_{\text{state}})$ is the random effect for the state and $V_{j,k} \sim N(0, \sigma^2_{\text{school}})$ is the school level random effect and $-\log \lambda_{j,k,m} = \beta' X_{j,k,m} + U_j + V_{j,k}$ where $X_{j,k,m}$ is a vector consisting of the covariates indicator for rural, indicator for female, indicator for black, hispanic, asian, native and finally pacific.

We must specify priors on $\sigma_{\text{school}}, \sigma_{\text{state}}, \alpha$ and β. We choose our priors as follows. For α we are given that $P(\alpha < 1) < 0.1$ and $P(\alpha > 7) = 0.1$. Hence we choose the prior on α so that $\log \alpha \sim N(\mu = 1, \sigma = 0.75)$ as the resulting 10 percent and 90 percent quantiles for α are roughly equal to 1 and 7.

Next we consider a prior for σ_{state}. Since it is not expected to the see the 'worst' states having five or 10 times the rate of the healthiest states, it follows that the complement of the event $\exp(\mu + 1.6\sigma_{\text{state}})/\exp(\mu - 1.6\sigma_{\text{state}}) \leq 5$ has probability 0.1 i.e. we choose σ_{state} to be exponentially distributed with $P(\sigma > 0.5) = 0.1$.

Next we consider a prior for σ_{school}. Based on similar considerations as in the previous case, we desire that the complement of the event $\exp(\mu + 1.6\sigma_{\text{state}})/\exp(\mu - 1.6\sigma_{\text{state}}) \leq 1.5$ has probability 0.1. So we choose σ_{school} to be exponentially distributed with $P(\sigma_{\text{school}} > 0.1) = 0.1$. Plots of the prior densities are given below.

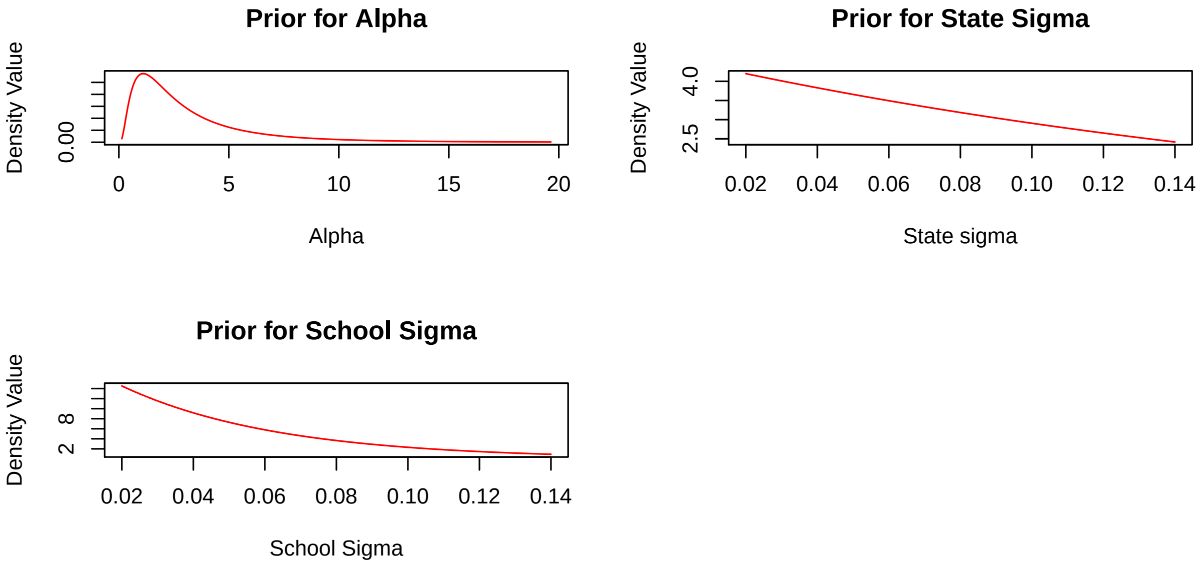
We choose the default INLA priors for β since we are not given sufficient information to meaningfully modify the prior. Also, with the priors chosen, the information that the age at which children first try smoking is different for males and females and earlier in rural areas is confirmed. Hence the conclusions of these parameters are not too sensitive to the prior chosen in this case.

Finally we fit an additional model to the one given above except with all iteractions among sex, ethnicity and rural/urban in order to investigate the difference between white urban males and white rural males and their smoking habits.

```
## Loading required package: Matrix
```

```
## Loading required package: sp
```

```
## This is INLA_18.07.12 built 2019-02-20 16:29:56 UTC.
## See www.r-inla.org/contact-us for how to get help.
```



Results

The coefficients of the model (without any interactions) are given below. Immediately we see that school level variation is much larger than state level variation. Indeed σ_{school} is at least 1.5 times as large as σ_{state}. In fact, the state level random effect is of the same order as the variation in first smoking times between males and females.

Moreover, based on the posterior confidence interval of α we conclude that hazard increases quadratically with age. In particular, as a child ages it becomes more likely that they will begin to smoke.

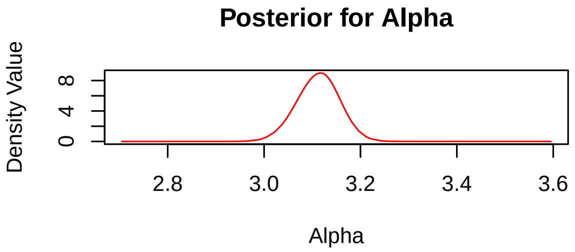
We can interpret the rest of the coefficients as follows. Females take roughly $1.05 = \exp(0.047)$ times as long to begin smoking in comparison to males controlling for confounding variables and random effects. Similarly people in a rural setting take $\exp(-0.107) = 0.90$ times as long to begin smoking in comparison to urban settings controlling for confounding variables. Lastly, among non-whites, asians are least likely to begin smoking. They wil take $1.2 = \exp(0.18462)$ times as long to begin smoking in comparison to whites controling for confounding variables and fixed effects.

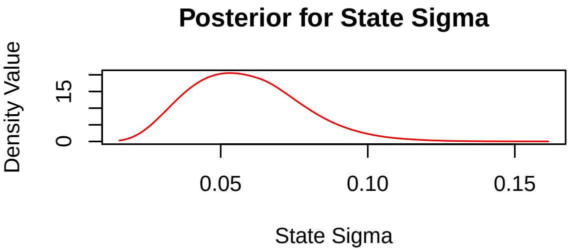
Table 1: Parameter Estimates of Weibull GLMM for First Smoking Time with Covariates Rural/Urban, Sex, Race as well as School and State Level Random Effects

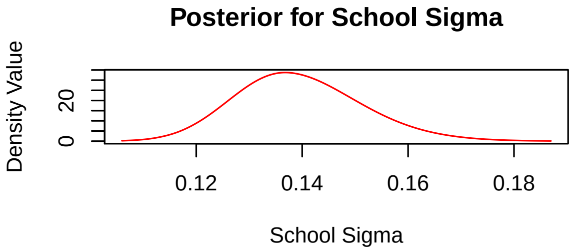
	mean	0.025quant	0.975quant
(Intercept)	-0.595	-0.647	-0.542
RuralUrbanRural	0.107	0.051	0.163
SexF	-0.047	-0.066	-0.028

	mean	0.025quant	0.975quant
Raceblack	-0.054	-0.086	-0.022
Racehispanic	0.036	0.010	0.062
Raceasian	-0.184	-0.249	-0.121
Racenative	0.090	0.012	0.164
Racepacific	0.121	-0.017	0.245
alpha for weibullsurv	3.112	3.022	3.196
sd for school	0.140	0.118	0.165
sd for state	0.059	0.027	0.100

The posterior densities of α, σ_{state} and σ_{school} are given below.







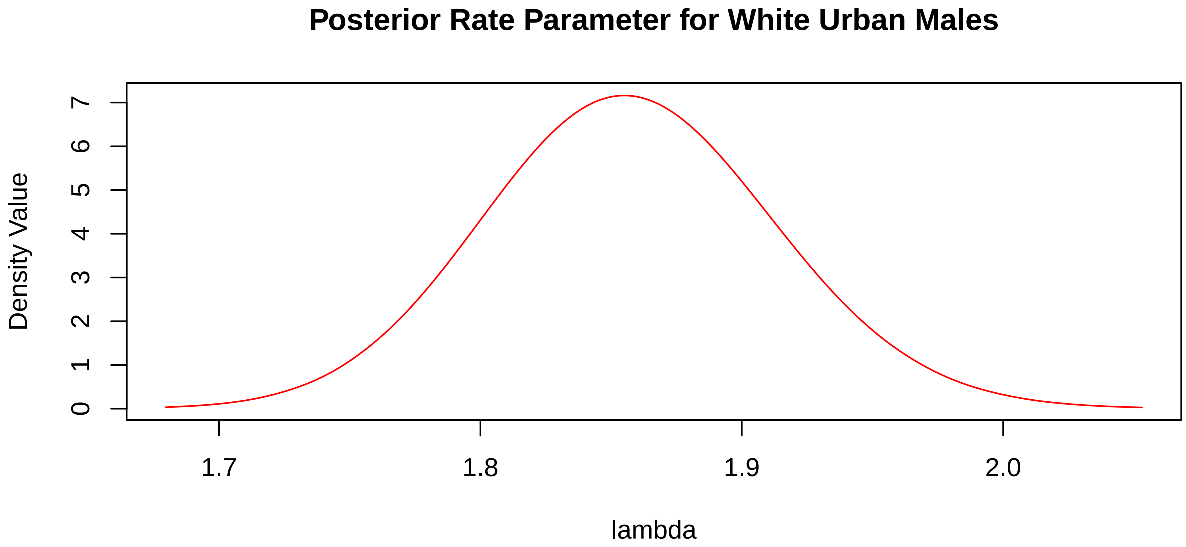
We fit an additional model (not related to the research question) in order to investigate the differences between white urban males and white rural males and their smoking uptake habits. As mentioned above this model is identical to the one above except that it includes all interactions among the variables. The coefficients of the model with interactions is given below.

From the table below we see that $\exp(-\text{intercept})$ represents the rate λ at which white urban males begin smoking (which is proporitional to their expected time to begin smoking) controlling for random effects. Similarly, $\exp(-(\text{intercept} + \text{Rural}))$ represents the rate at which white rural males begin smoking. Hence we see that white urban males take $\exp(0.14) = 1.15$ times as long as white rural males to begin smoking. The posterior density plot of the rate parameter for white urban males is given below.

Table 2: Parameter Estimates of Weibull GLMM for First Smoking Times with Covariates Rural/Urban, Sex, Race as well as School and State Level Random Effects and Full Interactions

	mean	0.025quant	0.975quant
(Intercept)	-0.619	-0.678	-0.560
SexF	-0.023	-0.069	0.022
Raceblack	-0.010	-0.072	0.052
Racehispanic	0.043	-0.006	0.092
Raceasian	-0.161	-0.264	-0.064

	mean	0.025quant	0.975quant
Racenative	0.038	-0.157	0.209
Racepacific	0.123	-0.122	0.333
RuralUrbanRural	0.142	0.075	0.209
SexF:Raceblack	-0.024	-0.105	0.058
SexF:Racehispanic	0.019	-0.046	0.083
SexF:Raceasian	-0.084	-0.237	0.067
SexF:Racenative	0.098	-0.146	0.345
SexF:Racepacific	-0.233	-0.785	0.220
SexF:RuralUrbanRural	-0.041	-0.097	0.016
Raceblack:RuralUrbanRural	-0.062	-0.144	0.021
Racehispanic:RuralUrbanRural	-0.020	-0.087	0.047
Raceasian:RuralUrbanRural	-0.046	-0.255	0.147
Racenative:RuralUrbanRural	0.107	-0.102	0.331
Racepacific:RuralUrbanRural	0.109	-0.194	0.420
SexF:Raceblack:RuralUrbanRural	0.011	-0.101	0.121
SexF:Racehispanic:RuralUrbanRural	-0.018	-0.108	0.073
SexF:Raceasian:RuralUrbanRural	0.221	-0.044	0.493
SexF:Racenative:RuralUrbanRural	-0.228	-0.544	0.082
SexF:Racepacific:RuralUrbanRural	0.097	-0.502	0.759
alpha for weibullsurv	3.100	3.004	3.186
sd for school	0.140	0.118	0.165
sd for state	0.058	0.026	0.100



Conclusions

Based on the fact that school level variation in the mean age children first try smoking is at least 1.5 times as big as state level variation, youth smoking prevention programs should target specific problematic schools as opposed to targeting states with the earliest smoking ages. Moreover, children become more and more susceptible to smoking as they age (in fact their hazard grows quadratically). As a result, youth smoking prevention programs should try and target the youngest students in order that their programs be as effective as possible.

Appendix

```
#Prior Densities and Their Plots
library("INLA")
par(mfrow=c(2,2))
#prior for alpha
xseq=seq(-3, 3, len=10000)
plot(inla.tmarginal(exp, cbind(xseq, dnorm(xseq, mean = 0.85, sd=sqrt(0.75)))), col="red",
     type="l", main="Prior for Alpha", xlab="x", ylab="Density Value")

#prior for sigma state
xseq=seq(0.02, 0.14, len=12000)
plot(xseq, dexp(xseq, rate=4.605), col="red", type="l", xlab="x", ylab="Density Value",
     main="Prior for State Sigma")

#prior for sigma school
xseq=seq(0.02, 0.14, len=12000)
plot(xseq, dexp(xseq, rate=23.026), col="red", type="l", xlab="x", ylab="Density Value",
     main="Prior for School Sigma")
par(mfrow=c(1,1))

#Setting up Data for Survivial Model
dataDir="C:/Users/amu/Documents/AppStats2"
smokeFile=file.path(dataDir, "smokeDownload.Rdata")
#smokeFile
if (!file.exists(smokeFile)){
  download.file("http://pbrown.ca/teaching/astwo/data/smoke.RData", smokeFile)
}
load(smokeFile)

forInla= smoke[, c("Age", "Age_first_tried_cigt_smkg", "Sex", "Race", "state",
                   "school", "RuralUrban")]
forInla=na.omit(forInla)
forInla$school<- as.factor(forInla$school)

forInla=as.list(forInla)

#library("INLA")

forSurv=data.frame(time=(pmin(forInla$Age_first_tried_cigt_smkg, forInla$Age)-4)/10,
                   event=forInla$Age_first_tried_cigt_smkg<=forInla$Age)
forSurv[forInla$Age_first_tried_cigt_smkg==8, "event"]=2
forInla$y=inla.surv(forSurv$time, forSurv$event)

#Model without Interactions
fitS2=inla(y~ RuralUrban + Sex + Race+
             f(school, model= "iid", hyper=list(prec=list(prior="pc.prec", param=c(0.1, 0.1))))+
           f(state, model="iid",
             hyper=list(prec=list(prior="pc.prec", param=c(0.5,0.1)))),
           control.family=list(variant=1,
                               hyper=list(alpha=list(prior="normal",
                                                     param=c(1, (0.75)^(-2))))),
           data=forInla, family="weibullsurv")
#summary(fitS2)
```

```r
library(Pmisc)
coeff=rbind(fitS2$summary.fixed[, c('mean', "0.025quant", "0.975quant")],
      Pmisc::priorPost(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")])
#Coeffficient Table
knitr:: kable(coeff, digits=3, caption="Parameter Estimates of Weibull
GLMM for First Smoking Time
              with Covariates Rural/Urban, Sex, Race as
well as School and State Level Random Effects")

#Posterior Plots of Densities
par(mfrow=c(2,2))
#alpha
plot(fitS2$marginals.hyperpar$`alpha parameter for weibullsurv`,
     col="red", type="l", xlab="x",
     ylab="Density Value", main="Posterior for Alpha")
#sigma state
post_sigmas_state<- inla.tmarginal(function (x) 1/sqrt(x),
                                   fitS2$marginals.hyperpar$`Precision for state`)
plot(post_sigmas_state, col="red", type="l", xlab="x",
     ylab="Density Value", main="Posterior for State Sigma")

#sigma school
post_sigmas_school<- inla.tmarginal(function (x) 1/sqrt(x),
                                    fitS2$marginals.hyperpar$`Precision for school`)
plot(post_sigmas_school, col="red", type="l", xlab="x",
     ylab="Density Value", main="Posterior for School Sigma")
par(mfrow=c(1,1))

#model with all iteractions
fitinters=inla(y~ Sex*Race*RuralUrban+
                 f(school, model= "iid",
                   hyper=list(prec=list(prior="pc.prec", param=c(0.1, 0.1))))+
             f(state, model="iid", hyper=list(prec=list(prior="pc.prec", param=c(0.5,0.1)))),
           control.family=list(variant=1, hyper=list(alpha=list(prior="normal",
                                                             param=c(0.85, (0.75)^(-2))))),
           data=forInla, family="weibullsurv")
coeff2=rbind(fitinters$summary.fixed[, c('mean', "0.025quant", "0.975quant")],
      Pmisc::priorPost(fitinters)$summary[, c("mean", "0.025quant", "0.975quant")])

#coefficient table
knitr:: kable(coeff2, digits = 3, caption = "Parameter Estimates of Weibull GLMM for
              First Smoking Times with Covariates Rural/Urban, Sex, Race as well as
              School and State Level Random Effects and Full Interactions")

#Plot of rate for White Urban Male
interceptmarg<-fitinters$marginals.fixed$`(Intercept)`
#rate parameter which is proportional to the expected lifetime
rate_uwm<-inla.tmarginal(function(x) exp(-x), fitinters$marginals.fixed$`(Intercept)`)
#interceptmarg$X is now a rate
plot(rate_uwm, col="red", type="l", xlab="x", ylab="Density Value",
     main="Posterior Rate Parameter for White Urban Males")
```