Factors Affecting Children's First Smoking

Xinqi Shen

1 Introduction

The dataset is coming from the 2014 American National Youth Tobacco Survey, it provides the data related to the situation of smoking among young Americans. In this report, we will focus on addressing two hypotheses. First of all, whether geographic variation(between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. Secondly, whether two non-smoking children have the same probability of trying cigarettes within the next month, given the same confounders and random effects.

2 Method

We will fit a Weibull survival model to analyze the hypotheses given the Weibull distribution is approximately a good fit in our model. Then, we add school and state as random effects in our model in order to analyze the first hypothesis. In addition, some confounding variables should be added and treated as fixed effects, such as sex, rural/urban and ethnicity. Given the following model:

$$Z_{ijk}|Y_{ijk}, Aijk, School_{ij}, State_i = min(Y_{ijk}, A_{ijk})$$
$$E_{ijk}, A_{ijk}, School_{ij}, State_i = I(Y_{ijk} < A_{ijk})$$
$$Y_{ijk}|School_{ij}, state_i \sim Weibull(\lambda_{ijk}, \alpha)$$
$$\lambda_{ijk} = exp(-\eta_{ijk})$$
$$\eta_{ijk} = \beta_0 + \beta_1 I_{Rural,ijk} + \beta_2 I_{Female,ijk} + \beta_3 I_{Black,ijk} + \beta_4 I_{Spanic,ijk} + \beta_5 I_{Asian,ijk}$$
$$+\beta_6 I_{Native,ijk} + \beta_7 I_{Pacific,ijk} + School_{ij} + State_i$$
$$State_i \sim N(0, \sigma_u^2)$$
$$School_{ij}\ simN(0, \sigma_v^2)$$

Noted A_{ijk} is individual's age. When $E_{ijk} = 1$, then $Y_{ijk} = Z_{ijk}$. When $E_{ijk} = 0$, then $Z_{ijk} < Yijk < \infty$

Where Y_{ijk} is the age of child k first tried cigarette in state i and school j. $X_{ijk}\beta/(\beta_0 + ... + \beta_7 I_{Pacific,ijk})$ contains an intercept, and all the fixed effects, including sex, rural/urban and ethnicity. $State_i$ and $School_{ij}$ are both random effects.

0.025quant	mean	0.975quant
0.271	1	3.694

Table 1: Mean and 95%CI of shape prior

In terms of prior distribution, we firstly put Log-Normal(log(1), 2/3) for shape(α) parameter, it is consistent with our prior assumption based on Table 1. Since Weibull shape parameter that allows for the mean value 1 can make a flat hazard function, also, the shape parameter can not be 4 or 5, because they are not in the 95% CI. In addition, we use penalized complexity prior that follows the exponential distribution, for State σ_u^2 and School σ_v^2, with $P(\sigma_u > 0.7) = 0.01$ and $P(\sigma_v > 0.13) = 0.01$. (Mathematically, 0.7 is computed by log(10)/3 and 0.13 is computed by log(1.5)/3 since 3 standard deviations account for about 99% data, then $\sigma_u \sim exp(6.6)$ and $\sigma_v \sim exp(35.4)$) . They satisfy two assumptions: the variability in the rate of smoking initiation between states with some states having double or triple the rate of smoking update compared other states but unlikely to see at 10; the 'worst' schools are

expected to have at most 50% greater rate than the 'healthiest' schools. Based on the Figure 1 shown below, our prior chosen looks appropriate.

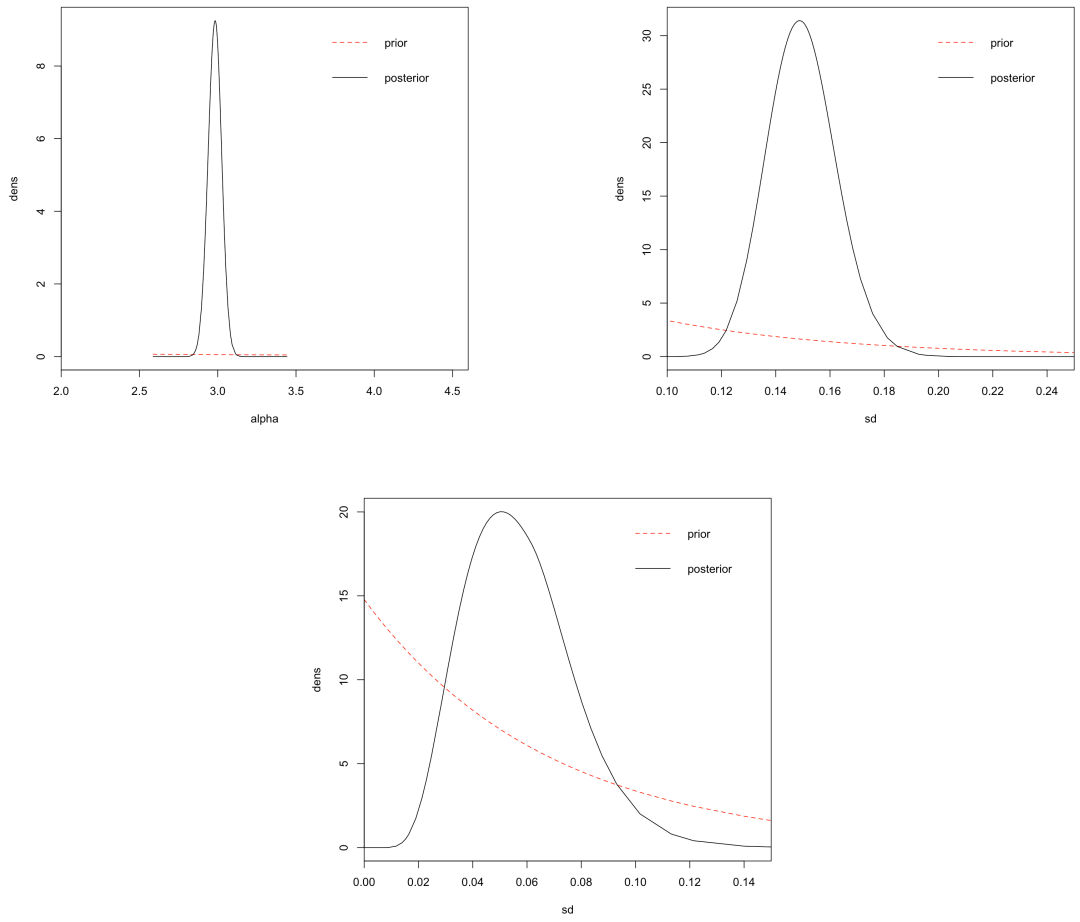

Figure 1: Posterior and Prior Distribution for Shape, School and State

3 Result

	mean	0.025quant	0.975quant
(Intercept)	1.857	1.960	1.758
RuralUrbanRural	0.893	0.947	0.843
SexF	1.051	1.072	1.030
Raceblack	1.057	1.094	1.022
Racehispanic	0.967	0.994	0.941
Raceasian	1.213	1.299	1.136
Racenative	0.912	0.989	0.844
Racepacific	0.882	1.019	0.775
SD for school	0.144	0.122	0.173
SD for state	0.059	0.027	0.103
alpha parameter for weibullsurv	2.985	2.901	3.070

Table 2: Posterior Means and Quantiles for Model Parameters.

The Table 2 provides the natural scale of the parameters of fixed effects. Based on our model shown above, we simply take the negative and exponential of original estimators since $\lambda_{ijk} = exp(-\eta_{ijk})$.

3.1 Hypothesis 1: variation between state and school

To addressing the first hypothesis, we can focus on the estimators of standard deviation(SD) for school and state shown in the Table 2. First of all, both 95% CI of SD do not included 0, which indicates state and school as random effects are significant. In other words, the age children first try cigarettes differ from both state and school. If we want to compare the variation, we found that the mean of standard deviation for school is more than twice the mean of standard deviation for state given their value are 0.144 and 0.059 respectively. Thus, tobacco control programs should target the schools with the earliest smoking ages and not concern themselves with finding particular states where smoking is a problem.

3.2 Hypothesis 2: probability of trying cigarettes for non-smoking children

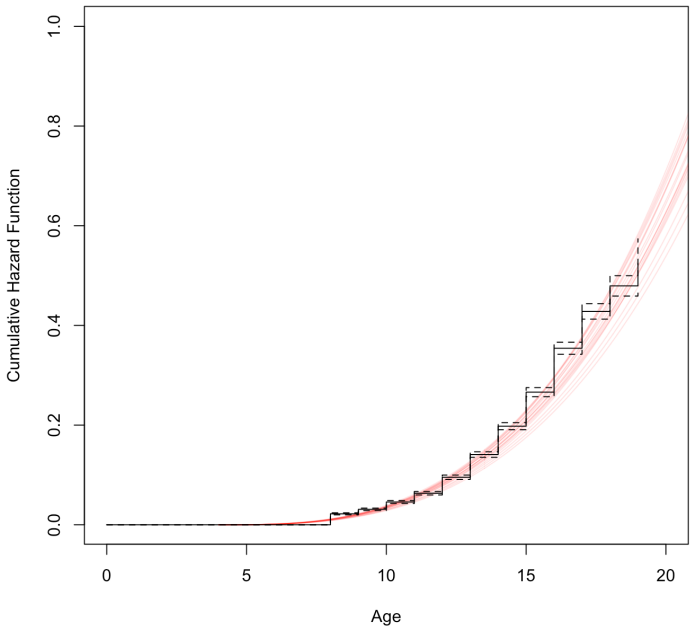

Figure 2: The Cumulative Hazard Function

To addressing the second hypothesis, we can first look at the estimated alpha(α)/shape parameter in the Table 2. If two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages but provided the same confounders(sex, rural/urban, ethnicity) and random effects(school and state), then the cigarette smoking must have a flat hazard function. In other words, the shape parameter should be approximately 1. We find that the estimated alpha parameter is very far from 1 and not even inside the 95% CI. Another way to check whether the hazard function is flat can look at the cumulative hazard function(Figure 2), we find that the plot shows a non-linear pattern. It supports the hazard function is not flat. Also, we notice that the cumulative hazard function increases more quickly when age getting older. In summary, age will affect the probability of non-smoking children trying cigarettes within the next month, and older the age, higher the probability.

4 Conclusion

Based on the analysis of the result of the dataset from the 2014 American National Youth Tobacco Survey, we have addressed two hypothesis. First of all, variation among schools in the mean age children first try cigarettes is substantially greater than geographic variation(between states). Thus, tobacco control programs should target the particular school instead of the state. Secondly, non-smoking children have the different probability of trying cigarettes within the next month given their different ages, even though they have the same sex, ethnicity and etc. More specific, the older children are more likely to smoke in the next month. In general, tobacco control programs should pay more attention to older children in particular schools.

5 Appendix

```
smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/
teaching/appliedstats/data/smoke.RData")
load(smokeFile)
smoke = smoke[smoke$Age > 9, ]
forInla = smoke[, c("Age", "Age_first_tried_cigt_smkg",
                    "Sex", "Race", "state", "school", "RuralUrban")]
forInla = na.omit(forInla)
forInla$school = factor(forInla$school)
library("INLA")
forSurv = data.frame(time = (pmin(forInla$Age_first_tried_cigt_smkg,
                                  forInla$Age) - 4)/10,
                                  event = forInla$Age_first_tried_cigt_smkg <=
                                  forInla$Age)
# left censoring
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2
smokeResponse = inla.surv(forSurv$time, forSurv$event)
fitS2 = inla(smokeResponse ~ RuralUrban + Sex + Race +
    f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",
    param = c(0.13, 0.01)))) +
    f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec",
    param = c(0.7,0.01)))),
    control.family = list(variant = 1,hyper = list(alpha = list(prior = "normal",
    param = c(log(1),(2/3)^(-2))))),
    control.mode = list(theta = c(8,2, 5), restart = TRUE),
    data = forInla, family = "weibullsurv",verbose = TRUE,
    control.compute=list(config = TRUE))

library(xtable)
alpha = rbind(c("0.025quant","mean","0.975quant"),
exp(qnorm(c(0.025, 0.5, 0.975), mean = log(1), sd = 2/3)))
print(xtable(alpha,digits = 3))

res = rbind(exp(-fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")]),
      Pmisc::priorPostSd(fitS2)$summary[,c("mean", "0.025quant", "0.975quant")],
      fitS2$summary.hyper[1,c("mean", "0.025quant", "0.975quant")])

print(xtable(res,digits = 3))

library(survival)
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters) {
```

```
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
  do.call(legend, fitS2$priorPost$legend)}

forSurv1 = data.frame(time = (pmin(
  forInla$Age_first_tried_cigt_smkg,forInla$Age) - 4)/10,
  event = forInla$Age_first_tried_cigt_smkg <= forInla$Age)
forSurv1$event = as.numeric(forSurv1$event)

hazEst = survfit(Surv(time*10+4, event) ~ 1, data=forSurv1)
xSeqNatural = seq(4, 100, len=1000)
xSeqTrans = (xSeqNatural-4)/10
densHaz = Pmisc::sampleDensHaz(fit = fitS2, x = xSeqTrans, n = 20)
matplot(xSeqNatural, densHaz[, "cumhaz", ], type = "l",
        lty = 1, col = "#FF000020", ylim = c(0.001,1), xlim = c(0,20),
        xlab = "Age",ylab = "Cumulative Hazard Function")
lines(hazEst, fun = "cumhaz")
```

Are female pedestrians safer than male in UK?

Xinqi Shen

1 Introduction

The dataset is coming from the department of transport in the UK, it consists of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries(pedestrians with moderate injuries have been removed). In this report, we will addressing the hypothesis that whether women trend to be, on average, safer as pedestrians than men, particular as teenagers and in early adulthood.

2 Method

We will fit a conditional logistic regression model. First of all, we treat fatal accidents as cases and slight injuries as controls. Then, in order to adjust for time of day, lighting conditions, and weather, we split our data into different strata based on the same conditions. Last but not least, we include sex and the interaction between sex and age in our model for analyzing our hypothesis. Given the following model:

$$logit[pr(Y_{ij} = 1)] = \alpha_i + X_{ij}\beta$$
$$logit[pr(Y_{ij} = 1)|Z_{ij} = 1] = \alpha_i^* + X_{ij}\beta$$
$$\alpha_i^* = \alpha_i + log[pr(Z_{ij} = 1|Y_{ij} = 1)/pr(Z_{ij} = 1|Y_{ij} = 0)]$$

Where Y_{ij} is either 1 or 0 representing fatal or slight injuries. Noted that i is strata, then Y_{i1} is case i with j > 1 are controls. α_i^* is some constant value for strata i. $X_{ij}\beta$ contains age and the interaction term between sex and age. Z_{ij} is 'in the study' indicators.

3 Result

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age0 - 5	0.132	1.142	0.044	3.008	0.003
age6 - 10	-0.320	0.726	0.041	-7.822	0.000
age11 - 15	-0.383	0.682	0.041	-9.305	0.000
age16 - 20	-0.443	0.642	0.040	-10.958	0.000
age21 - 25	-0.268	0.765	0.042	-6.355	0.000
age36 - 45	0.412	1.509	0.039	10.648	0.000
age46 - 55	0.768	2.156	0.039	19.709	0.000
age56 - 65	1.212	3.361	0.038	32.023	0.000
age66 - 75	1.797	6.033	0.036	49.447	0.000
ageOver 75	2.396	10.976	0.035	68.124	0.000
age26 - 35:sexFemale	-0.448	0.639	0.052	-8.573	0.000
age0 - 5:sexFemale	0.028	1.029	0.055	0.517	0.605
age6 - 10:sexFemale	-0.177	0.838	0.051	-3.490	0.000
age11 - 15:sexFemale	-0.250	0.779	0.047	-5.295	0.000
age16 - 20:sexFemale	-0.279	0.756	0.052	-5.364	0.000
age21 - 25:sexFemale	-0.369	0.691	0.063	-5.828	0.000
age36 - 45:sexFemale	-0.448	0.639	0.052	-8.679	0.000
age46 - 55:sexFemale	-0.376	0.686	0.048	-7.792	0.000
age56 - 65:sexFemale	-0.237	0.789	0.040	-5.878	0.000
age66 - 75:sexFemale	-0.143	0.866	0.032	-4.429	0.000
ageOver 75:sexFemale	-0.126	0.882	0.027	-4.606	0.000

Table 1: Estimated Model Parameters

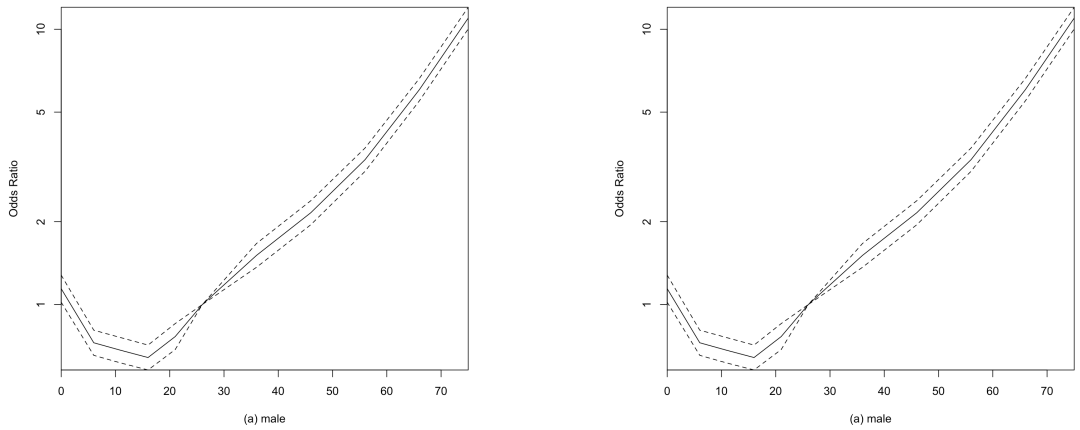

Figure 1: Estimated Model Parameters

Both the Table 1 and the Figure 1 provide the information about the odds ratio involved in motor vehicle accidents with fatal injuries for male and female. Noted that, we should look at exp(coef) column and the reference group here is male with age 26-35. Then, the exponential coefficients of age term provide the odds ratio for male between some age group and the reference age group. However, the exponential coefficients of interaction terms provide the odds ratio between female and male in the same age group. Thus, based on the results among interaction terms, we found that except women with age 0-5, all the odds ratio for women are less than 1. In other words, in general, women tend to be safer as pedestrians than men. Also, we noticed that the smallest odds ratio for women is in age 26-45, provided the value 0.639. Hence, women pedestrians are safer than men, particularly in middle-aged.

4 Conclusion

Based on the analysis of the result of the dataset from the department of transport in the UK, we have addressed our hypothesis. Thus, we can conclude that women tend to be, on average, safer as pedestrians than men, particular in middle-aged(26-45). Anyway, everyone needs to raise security awareness.

5 Appendix

```
library(Pmisc)
library("survival")
pedestrainFile = Pmisc::downloadIfOld(
  "http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time),]
pedestrians$y = pedestrians$Casualty_Severity == "Fatal"
pedestrians$timeCat = format(pedestrians$time, "%Y_%b_%a_h%H")
pedestrians$strata = paste(pedestrians$Light_Conditions,
            pedestrians$Weather_Conditions, pedestrians$timeCat)
theTable = table(pedestrians$strata, pedestrians$y)
onlyOne = rownames(theTable)[which(theTable[, 1] ==
                                     0 | theTable[, 2] == 0)]
x = pedestrians[!pedestrians$strata %in% onlyOne, ]
summary(glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
            data = x, family = "binomial"))$coef[1:4, ]
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
```

```
theCoef = rbind(as.data.frame(summary(theClogit)$coef),
                `age 26 - 35` = c(0, 1, 0, NA, NA))
theCoef$sex = c("Male", "Female")[1 + grepl("Female",
                                            rownames(theCoef))]
theCoef$age = as.numeric(gsub("age|Over| - [[:digit:]].*|[:].*",
                              "", rownames(theCoef)))
theCoef = theCoef[order(theCoef$sex, theCoef$age),]
res = summary(theClogit)$coef[,1:5]
library(xtable)
print(xtable(res,digits = 3))

matplot(theCoef[theCoef$sex == "Male", "age"],
exp(as.matrix(theCoef[theCoef$sex ==
                    "Male", c("coef", "se(coef)")]) %*% Pmisc::ciMat(0.99)),
        log = "y", type = "l", col = "black", lty = c(1,
                2, 2), xaxs = "i", yaxs = "i",ylab='Odds Ratio', xlab='(a) male')
matplot(theCoef[theCoef$sex == "Female", "age"],
exp(as.matrix(theCoef[theCoef$sex ==
                     "Female", c("coef", "se(coef)")]) %*% Pmisc::ciMat(0.99)),
        log = "y", type = "l", col = "black", lty = c(1,
                2, 2), xaxs = "i",ylab='Odds Ratio', xlab='(b) female')
```