STA442 A4

Tuoyue Huang Student number:1003906712

Thursday 28 November 2019

"Report of Smoking Data"

Summary

We analysis the 2014 American National Youth Tobacco Survey(NYTS2014) to study the age of children first try to smoke. We found that children in spcific schools could start smoke earlier, while the states they are living does not have much effects to the event. We also conclude that non-smoking children with higher ages are more likely to start smoking compare with younger ones with the same confounders, indicating non-flat hazard function.

Introduction

As smoking is a major health concern and is popular among youth, so we analyzed the NYTS2014, which is about smoking problems amongst American school children, through the R version of the dataset smoke.RData, which is accessible from pbrown.ca/teaching/appliedstats/data page. It was collected by FDA and CDC from Schools in the U.S by using multi-stage cluster sampling.

We have two hypothesis to investigate:
1. Geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools.
2. Two non-smoking children have the same probability of trying cigarettes within the next month, irrespective of their ages but provided the known confounders (sex, rural/urban, etnicity) and random effects (school and state) are identical.

Methods

The response variable interested here is survival time, which is the first time of a children to smoke cigarettes. We also have multiple measurement in the same school and state, thus we need include random effects. Therefore, a hierarchical survival model from the Weibull distribution family is what we need here and we could use the bayesian inference methodology with the INLA algorithm to analysis the results. The model assumption here we have:

$$Y_{ijk} \sim Weibull(\rho_{ijk}, \kappa) \quad \rho_{ijk} = exp(-\eta_{ijk}) \quad \eta_{ijk} = X_{ijk}\beta + U_i + V_{ij} \quad U_i \sim N(0, \sigma_u^2) \quad V_{ij} \sim N(0, \sigma_v^2)$$

where Y_{ijk} is the first time smoke for $individual_k$ in $school_j$ of $state_i$, and ρ_{ijk} and κ are the scale and shape parameter of the Weibull distribution. $X_{ijk}\beta$ contains covariates gender, ethnicity and studying in rural or urban school. U_i and V_{ij} are the random effects for $state_i$ and $school_{ij}$ in $state_i$
We also set the following priors, according to the information from collaborating scientists, which is the red cruve In Figure 1:

$$\sigma_u \sim Exponential(-log(0.05)/1.15) \quad \sigma_v \sim Exponential(-log(0.05)/0.203) \quad \kappa \sim lognormal(log(1), 0.64)$$

The penalized complexity prior, which is a exponential prior, we set on σ_v(school) means the P($\sigma_v > 0.203$) = 0.05. From our model assumption, V_{ij} should follow normal distribution with mean 0 and σ_v. We also know from the scientist that $\exp(V_{ij}) = 1.5$ for a school-level random effect is about the largest we would see, thus we could calculate $V_{ij} = \ln(1.5) = 0.406$ and this should be at the 95% quantiles which is $2\sigma_v$ from the mean. Therefore, we believe that the probability for σ_v larger than 0.406/2 is really small which is defined as 0.05 by us.

The other exponential prior set on state is according to $\exp(U_i) = 2.5$ but unlikely to see at 10 from scientists. Through the same method, we work out that the maximum value of V should be $\ln(10) = 2.3$, which is also $2\sigma_u$ from the mean. Hence, we could believe that $P(\sigma_u > 1.15) = 0.05$ for the variance of state random effect.

The last prior is a lognormal distribution set on the shape parameter of Weibull and a flat hazard function is expected from the scientist, so κ should allow for a 1 instead of 4 or 5. As a result, we expected the mean should be at log(1) and standard deviation shoud be around 0.64 by calculating using exponentiated qnorm function. κ would be (0.285, 1, 3.5) at (0.025, 0.5, 0.975) quantiles accordingly in this case, which is consistent with the scientist information.

We exclude the interactions between the confounders in our model since everyone prefers a simpler model if they give similar results, which is true in our scenario. To justify the two hypothesis, we will look at SD for school and state, graphs for prior and posterior densities of model parameters and also the cumulative hazard plot in the result part.

Results

Table 1: Table 1: Exponentiated Posterior Distribution for Model Parameters

	mean	0.025quant	0.975quant
(Intercept)	1.866	1.971	1.765
RuralUrbanRural	0.892	0.947	0.841
SexF	1.051	1.073	1.030
Raceblack	1.058	1.094	1.023
Racehispanic	0.967	0.994	0.941
Raceasian	1.214	1.301	1.137
Racenative	0.912	0.990	0.844
Racepacific	0.882	1.019	0.774
SD for school	0.149	0.125	0.176
SD for state	0.060	0.027	0.105

After take eˆ(-1*β), we could simply compare the posterior mean of these parameters with respect to 1. If the coefficients is smaller than 1, then it means the scale is smaller and the clock runs quicker, which indicates children start smoking in ealier ages, vice versa. Take RuralUrban as an example, the mean is less than 1, thus children study in rural areas tend to smoke earlier.

The mean standard deviation(SD) for school is 0.149, where SD for state only has 0.06 which is less than half of the school's. Thus there is more variation between schools than states, and school effect is much more important to consider here. So, the first hypothesis is proved to be wrong because it states that geographic variation accounts for the most part and we should target specific states to deal with the early smoking problem. From result of the data, we know we should focus on schools with higher probability of smoking.

In Figure 1, we have the posterior SD for school and state, and we can clearly see that school has a higher mean than state, which gives the same conclusion from table 1.

The fisrt plot is for Weibull shape parameter, its posterior mean is around 2.9 while we are expecting a 1(flat harzard). If κ is larger than 1, then it means the harzard function is increasing. When children get older controling other confounders to be the same, they are more likely to start smoking cigarretes. We could also get the same results from the last plot in Figure 1. If the hazard function is 1, then the cumulative hazard plot should be a straight line. However, we can see it is a curve, which means we do not have a flat hazard function.

In conclusion, both hypotheses are rejected after we analyze the NYTS2014 data set.

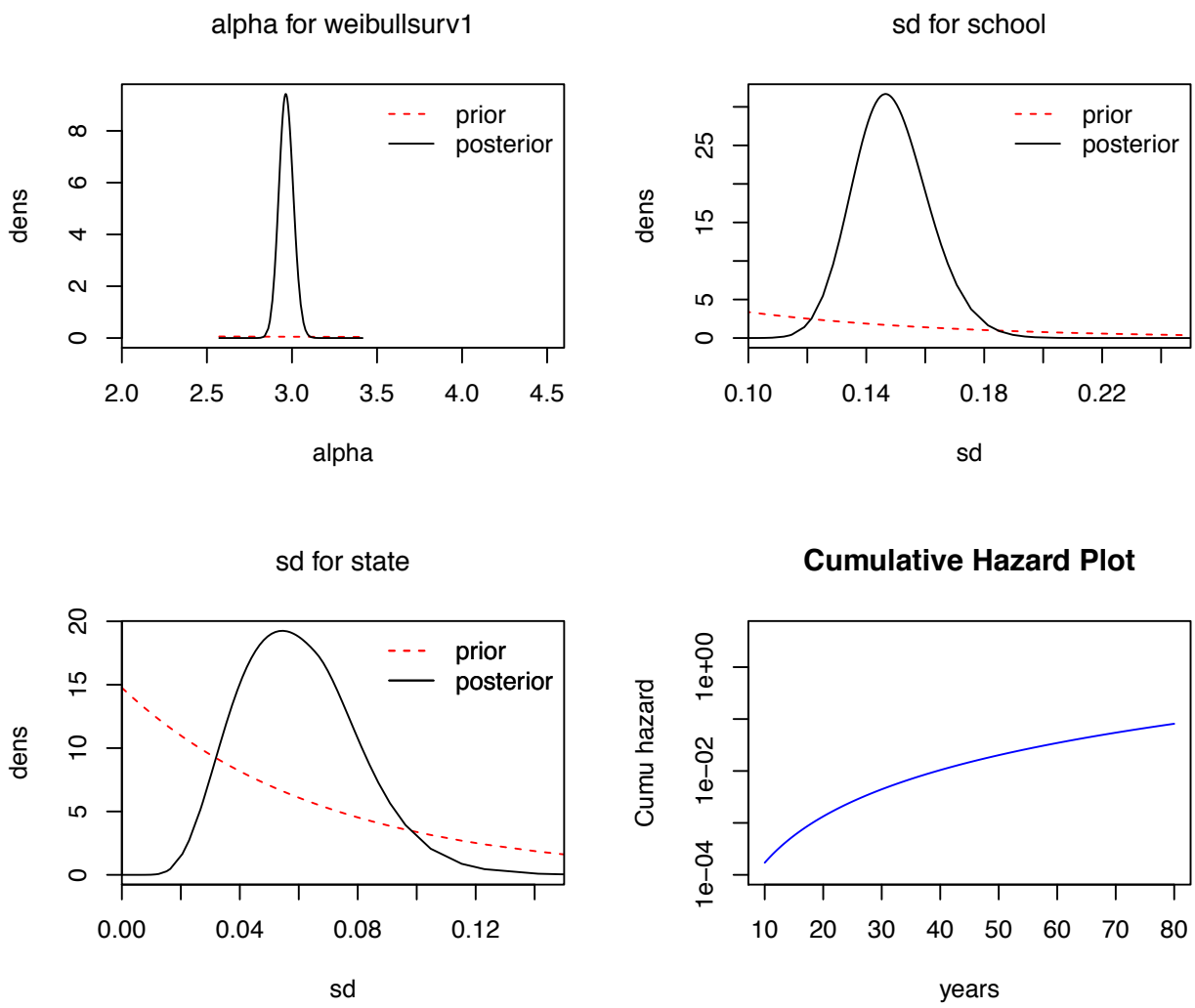

Figure 1: Prior and Posterior Plots and Cumulative Hazrd Plot

"Report of Death on the Roads"

Summary

Through analysis of data set, UK Road Accidents, by fitting a conditional logistic regression model with matched case control study, we could confirm that women is indeed safer as pedestrians than men on average, but particularly in age from 26 to 45 instead of teenager and early adulthood.

Introduction

In this report, we are intersted in analysis the pedestrians' safety for men and women. The hypothesis we got is the following: women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

The road accidents data set from UK contains all of the road traffic accidents in the UK from 1979 to 2015, which can be accessed from www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents. We only used the subset of it, which consists of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestrians with moderate injuries have been removed), to analysis the hypothesis.

Methods

Here we have a matched case control study, where we treat fatal accidents as cases and slight injuries as controls, and use a conditional logistic regression to adjust for time of day, lighting conditions, and weather. Thus the conditional logistic regression model we used is the following:

$$Want \quad logit[pr(Y_{ij} = 1)] = \alpha_i + X_{ij}\beta$$

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

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

For each case i, we will find a number of similar controls, where Y_{i1} is case i and Y_{ij} with j>1 are controls. If Y = 1, it means the accident is fatal X_{ij} are covariates not used in matching and i is our strata. Z_{ij} represents the weather conditions, light conditions and also the happening time of accidents.

Since its matched case-control study, we need to build a strata for our model. The strata is builted with light and weather conditions and together with the time that accidents happened. We paste the three covariates together to form the strata in r. Each different combinations of the three variables is a $strata_i$ in this method. For expamle, we have strata like "Daylight Raining no high winds 1979_Mar_Sat_h11", which is consist of light, weather and time.

Results

Figure 2:Male Odds Ratio Relative to 1

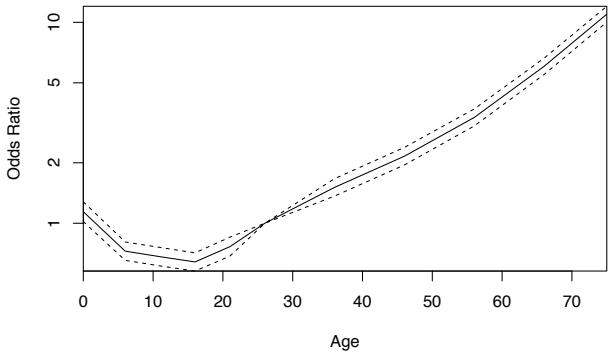

Figure 3:Female Odds Ratio Relative to Male at Same Age Group

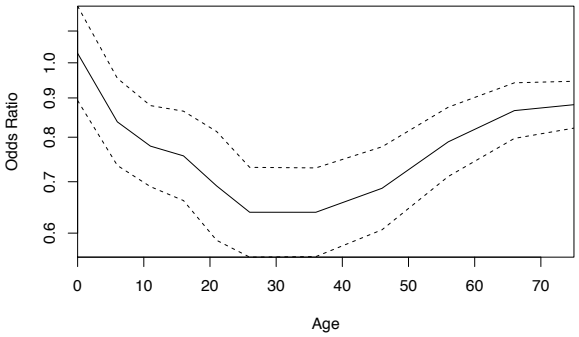

Table 2: Summary table for Model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	sex	age
age0 - 5:sexFemale	0.03	1.03	0.05	0.52	0.6	Female	0
age6 - 10:sexFemale	-0.18	0.84	0.05	-3.49	0.0	Female	6
age11 - 15:sexFemale	-0.25	0.78	0.05	-5.30	0.0	Female	11
age16 - 20:sexFemale	-0.28	0.76	0.05	-5.36	0.0	Female	16
age21 - 25:sexFemale	-0.37	0.69	0.06	-5.83	0.0	Female	21
age26 - 35:sexFemale	-0.45	0.64	0.05	-8.57	0.0	Female	26
age36 - 45:sexFemale	-0.45	0.64	0.05	-8.68	0.0	Female	36
age46 - 55:sexFemale	-0.38	0.69	0.05	-7.79	0.0	Female	46
age56 - 65:sexFemale	-0.24	0.79	0.04	-5.88	0.0	Female	56
age66 - 75:sexFemale	-0.14	0.87	0.03	-4.43	0.0	Female	66
ageOver 75:sexFemale	-0.13	0.88	0.03	-4.61	0.0	Female	75
age0 - 5	0.13	1.14	0.04	3.01	0.0	Male	0
age6 - 10	-0.32	0.73	0.04	-7.82	0.0	Male	6
age11 - 15	-0.38	0.68	0.04	-9.31	0.0	Male	11
age16 - 20	-0.44	0.64	0.04	-10.96	0.0	Male	16
age21 - 25	-0.27	0.76	0.04	-6.36	0.0	Male	21
age 26 - 35	0.00	1.00	0.00	NA	NA	Male	26
age36 - 45	0.41	1.51	0.04	10.65	0.0	Male	36
age46 - 55	0.77	2.16	0.04	19.71	0.0	Male	46
age56 - 65	1.21	3.36	0.04	32.02	0.0	Male	56
age66 - 75	1.80	6.03	0.04	49.45	0.0	Male	66
ageOver 75	2.40	10.98	0.04	68.12	0.0	Male	75

In the summary table, we will mainly look at the exponentiated coefficient colunm. For sex as males, it represents the odds ratio relative to the male base line group, which is the male group aged from 26 to 35 with exp(coef) = 1. While for sex as females, all these coefficients are odds ratio relative to the males who are in the same age groups. We know odds and probability transformation is monotonic, so a higher odds means a relatively higher probability of having fatal accidents. Hence, if the odds ratio is less than 1 for females, then it means females having relatively lower probability of having fatal accidents compare with the same age group of the males.

After looking at the colunm, we found that females have all odds ratio less than one, except for the age group from 0 to 5, which is 1.03. But 1.03 is almost 1, which means they have the same probability of getting fatal accidents, and it should be resonable to claim that women on average tend to be safer as pedestrians than men.

For the second part of the hypothesis, it says women is more safer than men among the period from 10 to 40. It is true that women is much more safe from 26 to 40 with the lowest odds ratio of 0.64. However, for teenager age(10 to 20), the average odds ratio is about 0.77, which is higher than group of 46-55 and almost the same as the group of 56-65. Therefore, the second part of hypothesis is not completely correct. It should change to "particularly in age from 26 to 45".

We could also get the same results by looking at figure 2 and 3, since the two plots come from the exponentiated coefficients of the table together with a credible interval.

In conclusion, the first part of the hypothesis in correct, but the second part should be change to age from 26 to 35. The reason is that we found women have the lowest odds ratio(0.64) in that range through our analysis.

Apendix

```
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(0.203, 0.05)))) +
               f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec",
                                          param = c(1.15, 0.05)))),
             control.family = list(variant = 1,
                                   hyper = list(alpha = list(
                                     prior = "normal",
                                     param = c(log(1), (0.64)^(-2))))),
             control.mode = list(theta = c(8, 2, 5), restart = TRUE),
             data = forInla, family = "weibullsurv", verbose = TRUE)

table1 <- rbind(exp(-fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")]),
                Pmisc::priorPostSd(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")])
knitr::kable(table1, digits = 3,
             caption = "Table 1: Exponentiated Posterior Distribution for Model Parameters")
#exp(qnorm(c(0.025,0.5,0.975), mean = log(1), sd = 0.7))
par(mfrow=c(2,2))
## prior plot
old.par <- par(mfrow=c(2, 2))
fitS2$priorPost = Pmisc::priorPost(fitS2)
i = 1
for (Dparam in fitS2$priorPost$parameters) {
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
```

```
  do.call(legend, fitS2$priorPost$legend)
  title(main = list(fitS2$priorPost$parameters[i], cex = 1.1, font = 1))
  i = i + 1
}
do.call(legend, fitS2$priorPost$legend)

xSeq = seq(10,80,len=1000)
kappa = fitS2$summary.hyper['alpha', 'mode']
lambda = exp(-fitS2$summary.fixed['(Intercept)', 'mode'])
plot(xSeq, (xSeq / (100*lambda))^kappa, col='blue', type='l', log='y',
     ylim=c(0.0001, 5), xlim = c(10, 80), xlab='years', ylab = 'Cumu hazard')
title(main = "Cumulative Hazard Plot", cex = 1.1, font = 1 )

par(old.par)

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, ]

theClogit = clogit(y ~ age + age:sex + strata(strata),
                   data = x)

#glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
    #data = x, family = "binomial")

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), ]

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",
        xlab = "Age",ylab = "Odds Ratio" )
title(main = "Figure 2:Male Odds Ratio Relative to 1", cex = 1.1, font = 1 )

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", yaxs = "i",
        xlab = "Age",ylab = "Odds Ratio" )
title(main = "Figure 3:Female Odds Ratio Relative to Male at Same Age Group", cex = 1.1, font = 1 )

knitr::kable(theCoef, digits = 2,
             caption = "Table 2: Summary table for Model ")
```