STA442 HW4

Xinyue Jiang

2019/11/30

Question 1

Introduction

We analyzed the age at which children first try cigarette smoking, using the data provided by the 2014 American National Youth Tobacco Survey. There are two hypotheses we wanted to investigate:
1. Geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. As a result, tobacco control programs should target the states with the earliest smoking ages and not concern themselves with finding particular schools where smoking is a problem.
2. First cigarette smoking has a flat hazard function, or in other words is a first order Markov process.This means 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.

Method

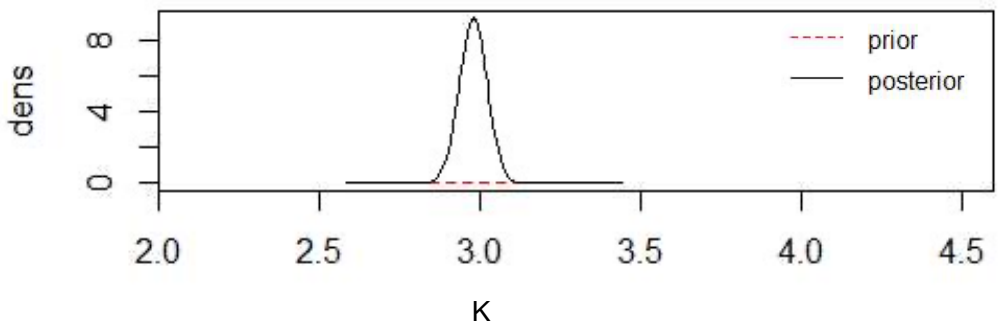
We used the Bayesian Inference and the survival mixed semi-parametric model.The age at which children first try cigarette smoking follows Weibull distribution(Y ~ Weibull(scale = λ, shape = k)). Sex + rural/urban + ethnicity + interaction between sex and ethnicity is the fixed effect, and is also the parametric part. State(U_i) + School(V_{ij}) is the random effect, and is also the nonparametric part. We used the Bayes inferance to smooth the nonparametric part,and estimate the shape parameter(k) of the Weibull deistribution.

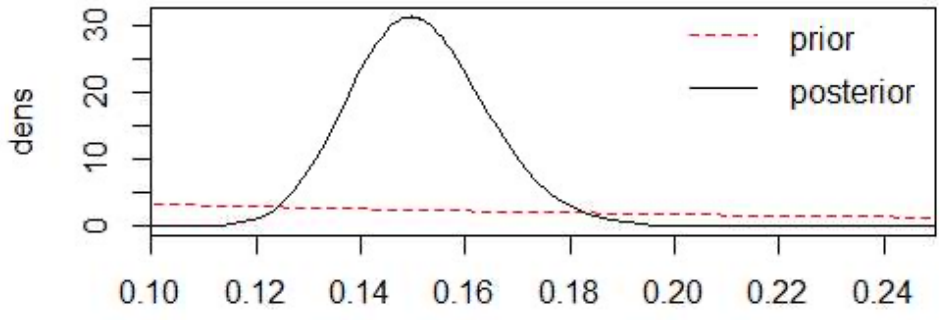
$Y_{ijk} \sim Weibull(\lambda_{ijk}, k)$
$\lambda_{ijk} = exp(-\eta_{ijk})$
$\eta_{ijk} = X_{ijk}\beta + U_i + V_{ij}$
$U_i \sim N(0, \sigma_U^2)$
$V_{ij} \sim N(0, \sigma_V^2)$

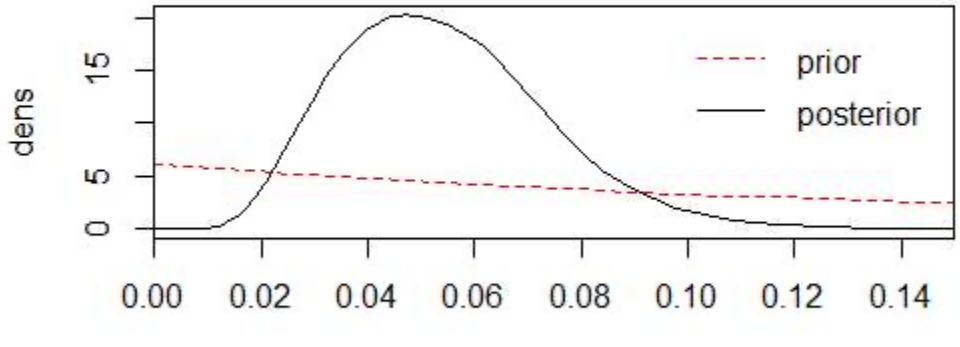
Y_{ijk}: the ages of children first try cigarettes
U_i: the state-level random effect
V_{ij}: the school-level random effect
X_{ijk}: the indicator variables for Sex, rural/urban, ethnicity, and the interaction between sex and ethnicity.

We used penalized complexity prior as our prior distribution for σ_{U_i}. We were given the information:
$exp(U_i) = 2$ or 3, but unlikely to see at 10.
By the given information, we knew U_i is very likely to be 0.9, and unlikely to be 2.3. So the σ_{U_i} is very unlikely to be more than 1.4. We chose the parameters(0.75, 0.01), which means we thought $P(\sigma_{U_i} > 0.75) = 0.01$
We used penalized complexity prior as our prior distribution for $\sigma_{V_{ij}}$, and We were given the information: exp(V_{ij})= 1.5 for a school-level random effect is about the largest we'd see.
By the given information, we knew V_{ij} at most equals to 0.4, and $V_{ij} \geq 0$, so the $\sigma_{V_{ij}}$ is almost impossible to be more than 0.4, so we chose the parameter (0.4, 0.01),which means we thought $P(\sigma_{V_ij} > 0.4) = 0.01$
We used normal distribution as our prior for the Weibull shape parameter k. We were given the information that k should allow for 1,but it is not believed that k is 4 or 5, so we chose the parameter(log(1),(2/3)^(-2)), because we though the mean is 0, and the standard deviation is (2/3)^(-2), which allows k= 1, and k=4 or k=5 are not easy to happen.

We then used data Y to adjust the priors of σ_{U_i},σ_{V_ij} and k, got the posteriors for σ_{U_i},σ_{V_ij} and k, which are $[\sigma_{U_i}|Y]$,$[\sigma_{V_{ij}}|Y]$ and [k|Y].





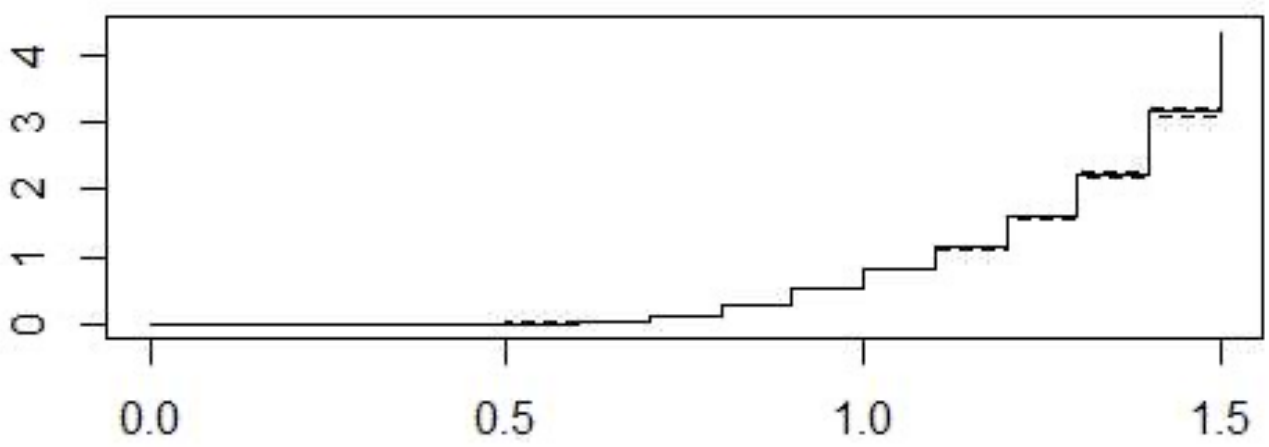


Result

The summary table shows that the standard error of school-level random effect is around 0.15, the standard error of state-level random effect is around 0.05, which means the variance explained by between school is much higher than the variance explained by between states. As a result, tobacco control programs should target the schools with the earliest smoking ages instead of states.

	mean	0.025quant	0.975quant
(Intercept)	-0.62296427	-0.677868206	-0.567337294
RuralUrbanRural	0.11619906	0.056839998	0.175232621
SexF	-0.05045014	-0.079053050	-0.022005952
Raceblack	-0.04808544	-0.091226078	-0.005616732
Racehispanic	0.02586575	-0.008956012	0.060480323
Raceasian	-0.19599541	-0.288791959	-0.108862280
Racenative	0.11064923	0.004674943	0.209207720
Racepacific	0.17667058	0.008640868	0.326226866
SexF:Raceblack	-0.01697447	-0.074406011	0.040313288
SexF:Racehispanic	0.01633434	-0.029932655	0.062578751
SexF:Raceasian	0.00548459	-0.122668515	0.132773462
SexF:Racenative	-0.04386520	-0.201623888	0.110588553
SexF:Racepacific	-0.17078138	-0.503519461	0.123922446
SD for school	0.15116209	0.127361889	0.177979255
SD for state	0.05421607	0.022470260	0.097869057

We checked the posterior distribution of the Weibull shape parameter k, and found that k is more likely to lie around 3 (from 2.8 to 3.2), and is very unlikely to be 1, so the hazard function is not flat. We also checked the figure of the cumulative hazard. We let λ be the scale parameter of the Weibull distribution, if the shape parameter k = 1, then $H(y) = \frac{1}{\lambda^k} y^k$, changes to $H(y) = \frac{1}{\lambda} y$, becomes a linear function; however, our plot does not show a linear relationship, which means $k \neq 1$, the hazard function is not flat.



Question 2

Introduction

We analyzed the data of the road traffic accidents in the UK from 1979 to 2015. We wanted to know whether or not women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

Method

We used logistic model first to get β_0 and β_p. Then used β_0, β_p for the logistic case-control model to fit the data. Fatal or slight injuries is the response variable, follows bernoulli distribution(y ~ Bernoulli(λ), with fatal injuries be 1, slight injuries be 0).
X_{ip} are indicator variables for sex, age, Light Conditions and Weather Conditions.
Z_i is the strata, used to control different cases.

The original logistic model:
$\text{pr}(Y_i{=}1|X_i) = \lambda_i$
$\log[\lambda_i/(1-\lambda_i)]{=}\beta_0 + \sum_{p=1}^{P} X_{ip}\beta_p$

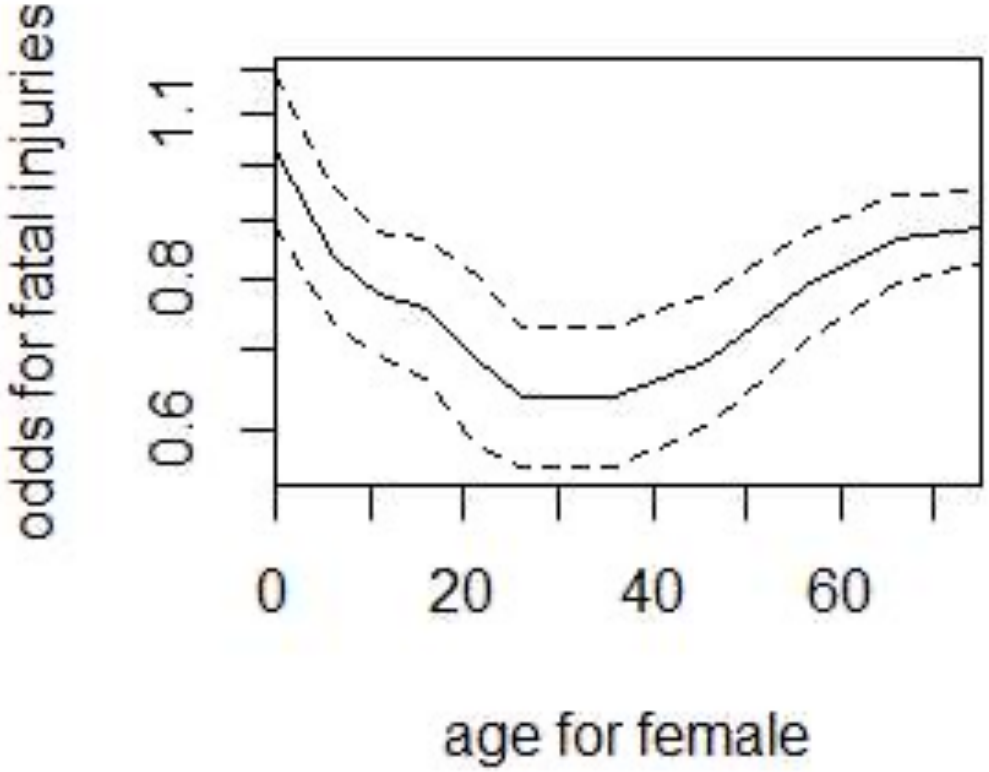
Transform β_0, β_p to β_0^*, β_p^*:
$\beta_p^* = \beta_0 + \log[\text{pr}(Z_i = 1|Y_i{=}1)/\text{pr}(Z_i = 1|Y_i{=}0)]$, when p=0
$\beta_p^* = \beta_p$, when $p \neq 0$

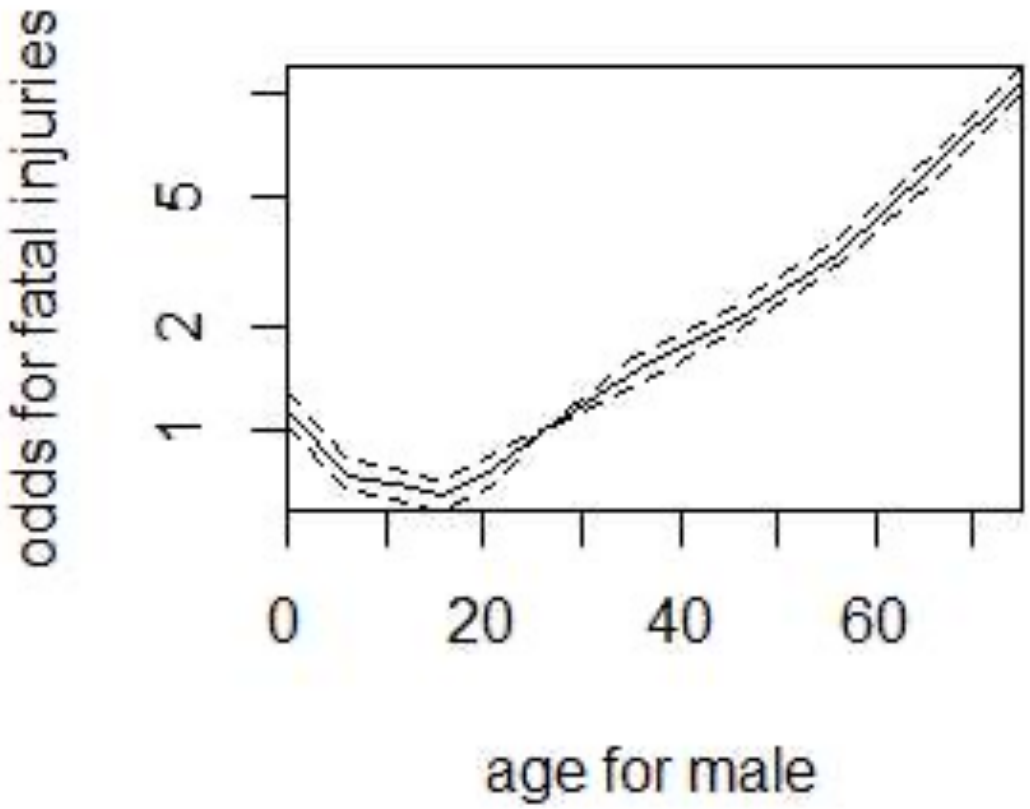
We use β_0^* and β_p^* to fit the logistic case-control model:
$\text{pr}(Y_i = 1|X_i, Z_i{=}1) = \lambda_i^*$
$\log[\lambda_i^*/(1\text{-}\lambda_i^*)] = \beta_0^* + \sum_{p=1}^{P} X_{ip}\beta_p^*$

Result

We checked the the summary table and the two figures of the odds of fatal injuries for male and female, found that in general, women tend to be, on average, safer as pedestrians than men; however, the phenomenon is not obvious in the period from age 15 to age 25. The odds of fatal injuries for male is much higher than female after age 46, which means, after 46 years old, women on average are much safer as pedestrians than men.

	exp(coef)	exp(-coef)	lower .95	upper .95
age0 - 5	1.1416	0.87598	1.0472	1.2444
age6 - 10	0.7264	1.37666	0.6705	0.7870
age11 - 15	0.6819	1.46659	0.6290	0.7391
age16 - 20	0.6420	1.55770	0.5930	0.6949
age21 - 25	0.7648	1.30746	0.7041	0.8308
age36 - 45	1.5091	0.66263	1.3990	1.6279
age46 - 55	2.1559	0.46383	1.9974	2.3271
age56 - 65	3.3605	0.29757	3.1202	3.6193
age66 - 75	6.0330	0.16575	5.6182	6.4785
ageOver 75	10.9759	0.09111	10.2449	11.7591
age26 - 35:sexFemale	0.6388	1.56551	0.5766	0.7077
age0 - 5:sexFemale	1.0288	0.97198	0.9238	1.1458
age6 - 10:sexFemale	0.8377	1.19377	0.7584	0.9253
age11 - 15:sexFemale	0.7789	1.28385	0.7101	0.8544
age16 - 20:sexFemale	0.7564	1.32198	0.6831	0.8377
age21 - 25:sexFemale	0.6913	1.44647	0.6106	0.7827
age36 - 45:sexFemale	0.6388	1.56554	0.5773	0.7068
age46 - 55:sexFemale	0.6864	1.45690	0.6244	0.7545
age56 - 65:sexFemale	0.7889	1.26753	0.7290	0.8538
age66 - 75:sexFemale	0.8664	1.15414	0.8132	0.9232
ageOver 75:sexFemale	0.8820	1.13384	0.8361	0.9304


odds for fatal injuries
0.6
0.8
1.1
0
20
40
60
age for female


odds for fatal injuries
1
2
5
0
20
40
60
age for male

Appendix

```
library("survival")
# Question1

smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/
                                 appliedstats/data/smoke.RData")
#Loading required namespace: R.utils
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)


#dont run this line
fitS2 = inla(smokeResponse ~ RuralUrban + Sex * Race +
               f(school, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec",param = c(0.75, 0.01))))
             + f(state, model = "iid",hyper =
                   list(prec = list(prior = "pc.prec", param = c(0.4, 0.01)))),
             control.family = list(variant = 1,
                  hyper = list(alpha = list(prior = "normal", param = c(log(1),

             control.mode = list(theta = c(8,  2, 5), restart = TRUE),
             data = forInla, family = "weibullsurv",
             verbose = TRUE)

rbind(fitS2$summary.fixed[, c("mean", "0.025quant", "0.975quant")], Pmisc::priorPostSd(fitS2)$summary[,
                                                                    c("mean", "0.025quant",

fitS2$priorPost = Pmisc::priorPost(fitS2)

for (Dparam in fitS2$priorPost$parameters) {
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
  }
do.call(legend, fitS2$priorPost$legend)

forSurv$one = 1
hazEst = survfit(Surv(time,one)~1,data=forSurv)
plot(hazEst,fun='cumhaz')
```

```
### Question 2

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$timeCat)
dim(pedestrians)
pedestrians[1:3, ]
table(pedestrians$Casualty_Severity, pedestrians$sex)
range(pedestrians$time)
summary(glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
            data = x, family = "binomial"))$coef[1:4, ]

library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata),
                   data = x)
summary(theClogit)

theTable = table(pedestrians$strata, pedestrians$y)
onlyOne = rownames(theTable)[which(theTable[, 1] ==
                                     0 | theTable[, 2] == 0)]
x = pedestrians[!pedestrians$strata %in% onlyOne, ]
theCoef = rbind(as.data.frame(summary(theClogit)$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",
        ylab = 'odds for fatal injuries', xlab = 'age for 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 for fatal injuries', xlab = 'age for female')
```