STA442 Homework 4, Survival

student number: 1003942326

Yulin WANG

2/12/2019

Question 1 Smoking

Introduction

We analyzed the 2014 American National Youth Tobacco Survey by using an R version of the dataset available at pbrown.ca. We had two main analysis purposes. Firstly, we wanted to analyze whether the mean age children first try cigarettes depends substantially more on the state the child lives in (geographic factor) or the school he or she goes to. Additionally, we also investigated whether two non-smoking children are equally as likely to try cigarettes within the next month, irrespective of their ages but provided identical confounders (sex, rural/urban, etnicity) and random effects (school and state).

Methods

Since the event that first trying cigarette smoking happens only once for each child, so we chose a Weibull distribution to model the dataset. The specific model for state i, school j, and individual k, is as follows:

$$Y_{ijk} \sim \text{Weibull}(\lambda_{ijk}, \alpha)$$

$$\lambda_{ijk} = \exp(-\eta_{ijk})$$

$$\eta_{ijk} = X_{ijk}\beta + U_i + V_{ij}$$

$$U_i \overset{iid}{\sim} N(0, \sigma_U^2)$$

$$V_{ij} \overset{iid}{\sim} N(0, \sigma_V^2)$$

where
Y_{ijk} is the age that the child first try cigarette smoking;
$X_{ijk}\beta$ represents the subjects gender, ethnicity and whether they are from rural or urban areas;
U_i is the state random effect and V_{ij} is the school random effect;
α is the Weibull shape parameter, which follows a log-normal distribution.

Our model did not include any interactions between gender, ethnicity and rural/urban confounders, since we fitted different models with different interactions and found that the interactions were not statistically significant.

Given prior information provided by collaborating scientists, we selected the hyperparameters of above model as follows:
For state random effect U_i:
since given that we might see $exp(U_i) = 2$ or 3 but unlikely to see at 10, so we assumed that $exp(U_i) \leq 9$, then the maximum of log scale of U_i we can take is $log(9) = 2.2$, so the standard deviation is $2.2/2 = 1.1$ with 95% confidence. Thus, we have $P(\sigma_U \leq 1.1) = 0.95$, then we should use penalized complexity precision: $P(\sigma_U > 1.1) = 0.05$. This can be checked by exp(c(-2, 2) * 1.1).

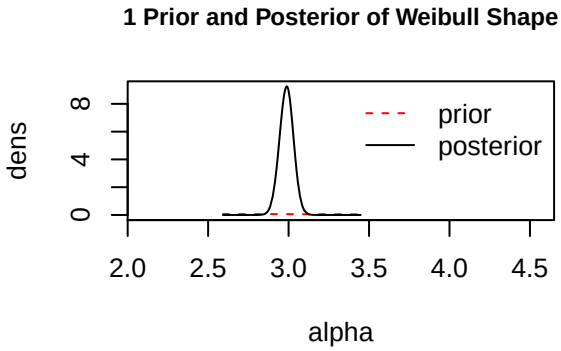
For school random effect V_{ij}:
since given that $exp(V_{ij}) = 1.5$ for a school-level random effect is about the largest we'd see. That is, the maximum of log scale of V_{ij} we can take is $log(1.5) = 0.4$, so the standard deviation is $0.4/2 = 0.2$ with

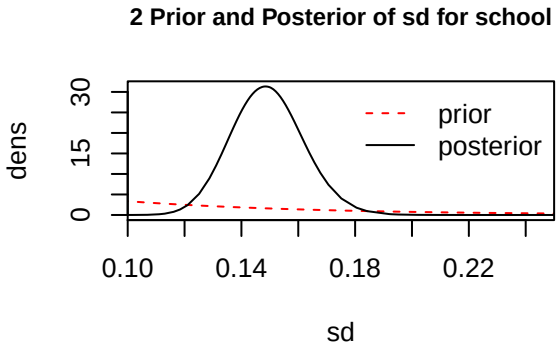
95% confidence. Thus, we have $P(\sigma_V \leq 0.2) = 0.95$, then we should use penalized complexity precision: $P(\sigma_V > 0.2) = 0.05$. This can be checked by exp(c(-2, 2) * 0.2).

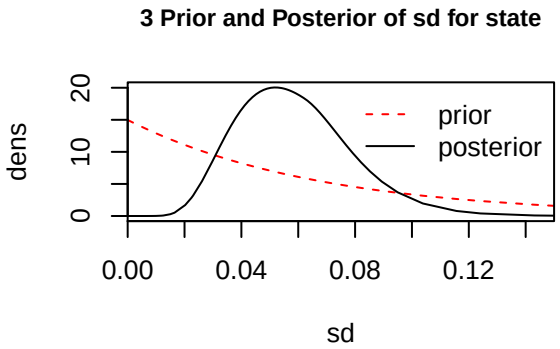
For the Weibull shape parameter α:
since given that a flat hazard function is expected, which means the prior on the Weibull shape parameter should allow for a 1 but it is not believed that shape parameter is 4 or 5. Thus, we assumed that $log(\alpha)$ follows a normal distribution that with mean of log(1.5) and the standard deviation of 2/3, which guarantee the Weibull shape parameter could be 1 and not much greater than 5. This can be checked by exp(qnorm(c(0.025, 0.5, 0.975), mean = log(1.5), sd = 2/3)).

Results

	mean	0.025quant	0.975quant
(Intercept)	-0.620	-0.674	-0.565
RuralUrbanRural	0.114	0.055	0.173
SexF	-0.050	-0.070	-0.030
Raceblack	-0.056	-0.090	-0.023
Racehispanic	0.033	0.006	0.061
Raceasian	-0.193	-0.262	-0.127
Racenative	0.092	0.010	0.169
Racepacific	0.125	-0.019	0.254
SD for school	0.149	0.125	0.175
SD for state	0.059	0.026	0.104







According to the above table, we can find that males are more likely to start smoking than females, and children in rural areas tend to try first cigarette earlier than those in urban areas when they have same other features. Additionally, most of races are statistically significant to affect the the mean age children first try cigarettes except the pacific.
It is clear from figure 1 that cigarette smoking does not have a flat hazard function, which does not support the hypothesis. Therefore, two non-smoking children do not have the same probability of trying cigarettes within the next month, holding all other factors identical. Additionally, it shows that older children are more likely to first try smoking.
From figure 2 and 3, geographic variation (between states) in the mean age children first try cigarettes is substantially smaller than the variation amongst schools, which does not support the hypothesis. Therefore, it is not recommended for tobacco control programs to target states with the earliest smoking ages and not concern themselves with finding particular schools where smoking is a problem.

Conclusion

Firstly, geographic variation (between states) in the mean age children first try cigarettes is smaller than the variation amongst schools. As a result, we recommend that tobacco control programs should not only target the states with the earliest smoking ages, but also concern themselves with finding particular schools where smoking is a problem.
Secondly, two non-smoking children do not have the same probability of trying cigarettes within the next month, holding all other factors identical, and older children are more likely to first try smoking.

Question 2 Death on the roads

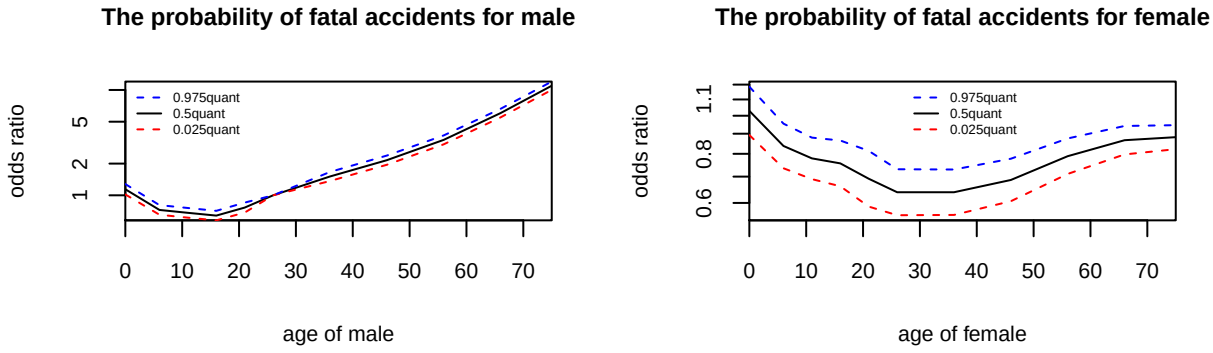
Introduction

We analyzed a subset of the data from www.gov.uk/government/statistical-data-sets/ras30-reportedcasualties-in-road-accidents, with all of the road traffic accidents in the UK from 1979 to 2015. We investigated all pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestrians with moderate injuries have been removed). We wanted to analyze whether the UK road accident data are consistent with the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

Methods

Since our purpose is to analyze how the gender affect the probability of fatal accidents for pedestrians, and there are three confounders: time of day, lighting and weather conditions, so we need to control these confounders identical for each case in order to investigate the risk of fatal accidents for both males and females when they are in the same situation. For instance, if it is rainy with no high winds and under normal daylight right now, under these conditions, the chance of having fatal accident as pedestrians for men and women may be different.
We stratified the data by accident time, light condition and weather condition. Each strata should at least have one case (fatal accident) and one control (slight injury).
Thus, we treated fatal accidents as cases and slight injuries as controls, removed stratas with no cases or no controls, and used a conditional logistic regression to adjust for time of day, lighting conditions, and weather.

Results



Note: These two plots treat 26-year-old men as the baseline.
According to the figure for men, it is expected that they become less likely to have fatal accidents as pedestrians until about 18 years old, and then the chance of fatal accidents tends to increase highly as they become older, and the odds ratio compared to the 26 year-old men is expected to reach 10 at 70-year-old.
When it comes to women, the odds ratio seems to fluctuate relatively stablely between 0.6 and 1.1, with decrease from birth date till about 30-year-old then increase gently. Additionally, women tend to have the lowest risk of fatal accidents when they are teenagers and in early adulthood.
Thus, men of most ages are substantially more likely to experience fatal accidents as pedestrians than women in UK.

Conclusion

We analyzed a dataset of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries in UK from 1979 to 2015. We found that if given identical time of day, light conditions and weather in the UK, women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

Appendix

Question 1

```
#load and clean the data
load('smoke.RData')
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$Age)-4)/10,
                     #rescaling the time to event
                     event = forInla$Age_first_tried_cigt_smkg <=forInla$Age)
hist(smoke$Age_first_tried_cigt_smkg)
# left censoring
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2
smokeResponse = inla.surv(forSurv$time, forSurv$event)
```

```
#fit a model including all interaction terms
fitS2 = inla(smokeResponse ~ RuralUrban * Sex * Race +
               f(school, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(0.2, 0.05))))
             + f(state, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(1.1,0.05)))),
             control.family = list(variant = 1, hyper =
                                     list(alpha = list(prior = "normal",
                                                       param = c(log(1.5),(2/3)^(-2))))),
             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", "0.975quant")])
```

```
#Since the interactions are not significant, so we can go for a simpler model with no interactions.
fitS = inla(smokeResponse ~ RuralUrban + Sex + Race +
               f(school, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(0.2, 0.05))))
             + f(state, model = "iid", hyper =
                   list(prec = list(prior = "pc.prec", param = c(1.1,0.05)))),
             control.family = list(variant = 1, hyper =
                                     list(alpha = list(prior = "normal",
                                                       param = c(log(1.5),(2/3)^(-2))))),
             control.mode = list(theta = c(8, 2, 5), restart = TRUE),
             data = forInla, family = "weibullsurv", verbose = TRUE)

knitr::kable(rbind(fitS$summary.fixed[, c("mean", "0.025quant","0.975quant")],
      Pmisc::priorPostSd(fitS)$summary[,c("mean", "0.025quant", "0.975quant")]), digits = 3)
```

```
#Prior and posterior of hyperparameters plots
par(mfrow=c(2,2))
fitS$priorPost = Pmisc::priorPost(fitS)
mains <- c("Weibull Shape", "sd for school", "sd for state")
i <- 1
```

```
for (param in fitS$priorPost$parameters) {
  do.call(matplot, c(fitS$priorPost[[param]]$matplot))
  do.call(legend, fitS$priorPost$legend)
  title(main = paste(i, "Prior and Posterior of", mains[i]))
  i <- i + 1
}
```

Question 2

```
library('R.utils')
pedestrainFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
pedestrians = readRDS(pedestrainFile) #str(pedestrians) glimpse
pedestrians = pedestrians[!is.na(pedestrians$time),] #delete na in variable 'time'
pedestrians$y = pedestrians$Casualty_Severity == "Fatal" # binary outcome
pedestrians$timeCat = format(pedestrians$time, "%Y_%b_%a_h%H")
#paste the confounders together
pedestrians$strata = paste(pedestrians$Light_Conditions,
                           pedestrians$Weather_Conditions, pedestrians$timeCat)
# remove strata with no cases or no controls, one to one, case -> control
theTable = table(pedestrians$strata, pedestrians$y) #head(pedestrians$strata)
onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)]
#remove unmatched observation, only remain  have fatal and slight at the same time
x = pedestrians[!pedestrians$strata %in% onlyOne, ] #select the data with matched cases
```

```
#fit the conditional logistic model
library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
```

```
#add baseline in summary table
theCoef = rbind(as.data.frame(summary(theClogit)$coef),`age26 - 35` = c(0, 1, 0, NA, NA))
#create a new column sex
theCoef$sex = c("Male", "Female")[1 + grepl("Female", rownames(theCoef))]
#create a new column age by using regular expression
theCoef$age = as.numeric(gsub("age|Over| - [[:digit:]].*|[:].*", "", rownames(theCoef)))
theCoef = theCoef[order(theCoef$sex, theCoef$age),] #reorder
```

```
par(mfrow=c(2,2))
# create plot for male
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 = c("black", "red", "blue"), lty = c(1, 2, 2), xaxs = "i", yaxs = "i",
  xlab = "age of male", ylab = "odds ratio", main = "The probability of fatal injuries for male")
legend("topleft", bty = "n", lty = c(2, 1, 2), col = c("blue", "black", "red"),
       legend = c("0.975quant", "0.5quant", "0.025quant"), cex = 0.5)

# create plot for female
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 = c("black", "red", "blue"), lty = c(1, 2, 2), xaxs = "i",
  xlab = "age of female", ylab = "odds ratio", main = "The probability of fatal injuries for female")
legend("topright", bty = "n", lty = c(2, 1, 2), col = c("blue", "black", "red"),
       legend = c("0.975quant", "0.5quant", "0.025quant"), cex = 0.5)
```