STA442 Assignment 4

Zhuojing Qiu

03/12/2019

Question 1. Smoking

Introduction

We analyzed the results of 2014 American National Youth Tobacco Survey using an R version of the 2014 dataset smoke.RData, which is available at pbrown.ca. Our main focus was to investigate whether the mean age of children first trying cigarettes depends more on the state that a child lives in or the school that he or she attends. We also investigated whether any two non-smoking children would be equally likely to try cigarettes within the next month, controlling for sex, demographics characteristics (Rural/Urban) and ethnicity.

Methods

Since our analysis focuses on how factors affect time (age) to some event (first tried smoking), we modelled the data using a Weibull distribution, as it is effective for such survival analysis data. To assess the fit of a Weibull distribution, we plotted the empirical distribution along with the our data distribution in Figure 1. This appears to be a reasonable fit.

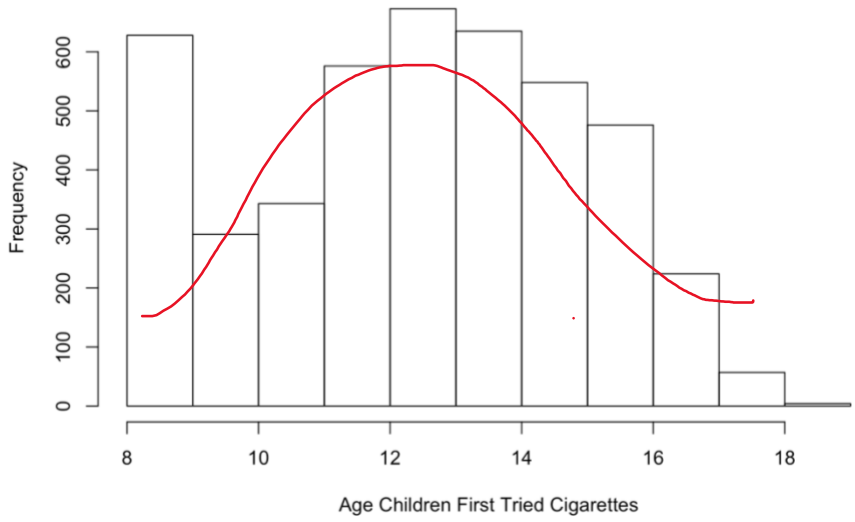

Figure 1. Assessing if the distribution can be fit to a Weibull distribution

So we used INLA with the interactions between sex and ethnicity in our model. The specific model is as follows:

$$Y_{ij} \mid U_i, V_i \sim Weibull(\lambda_{ij}, k)$$

$$-log(\lambda_{ij}) = \eta_{ij} = X_{ij}\beta + U_i + V_i$$

$$U_i \sim N(0, \sigma_U^2)$$

$$V_i \sim N(0, \sigma_V^2)$$

Where

- $X_{ij}\beta$ is the subjects gender, ethnicity, whether they are from a rural or urban school, and interaction between ethnicity and whether they are from a rural or urban school
- U_i is the school random effect
- V_i is the state random effect
- k is the Weibull shape parameter and is normally distributed with its own hyperparameters
- λ is the Weibull scale parameter.

	Mean/Alpha HyperParam	SD/Beta Hyperparam	0.025quant	0.975quant
Prior on Weibull shape	2.99	0.04	2.91	3.08
Prior on SD of School	0.15	0.13	0.18	0.86
Prior on SD of State	0.06	0.03	0.11	0.94

Table 1. Quantiles of Prior Distributions of Parameters

We selected the hyperparameters of the above model using the prior information provided by the collaborating scientists. They believe that some states should have update on children starting smoking 2-3 times faster than other states. However, of 10% probability or less, this rate of smoking update would exceed 10. Thus, we selected our prior to such that the 90th quantile was 5.

Schools were expected to have a less variability than states. Of 10% probability or less, the rate of smoking update would exceed 1.5. Thus, we chose our prior such that the 90th quantile was 1.5.

Finally, we expected the hazard function to be constant. Of 10% probability or less, the shape parameter would be greater than 4 or 5. Thus, we set our prior such that around 10 percent of the

time, the prior mean of alpha is 3, so the prior mean of log alpha is log 3. The prior on the shape parameter is a normal distribution.

Results

	mean	0.025quant	0.975quant	Exp.(-Estimate)
(Intercept)	-0.62	-0.68	-0.56	1.86
RuralUrbanRural	0.11	0.05	0.17	0.89
SexF	-0.05	-0.08	-0.02	1.05
Raceblack	-0.05	-0.09	-0.01	1.05
Racehispanic	0.03	-0.01	0.06	0.97
Raceasian	-0.20	-0.29	-0.11	1.22
Racenative	0.11	0.00	0.21	0.90
Racepacific	0.18	0.01	0.32	0.84
SexF:Raceblack	-0.02	-0.07	0.04	1.02
SexF:Racehispanic	0.02	-0.03	0.06	0.98
SexF:Raceasian	0.01	-0.12	0.13	0.99
SexF:Racenative	-0.04	-0.20	0.11	1.04
SexF:Racepacific	-0.17	-0.50	0.12	1.18
SD for school	0.15	0.13	0.18	-
SD for state	0.06	0.03	0.10	-

Table 2. Posterior Estimates of Hyperparameters

According to the posterior estimates of hyperparameters from our model, we found that surprisingly, there was less variation in the rate of smoking update between states than variation among schools. We can also compare the odds of first trying smoking by reading off the last column Exp.(-Estimate). For example, exponentiated negative estimate of RuralUrbanRural tells us that a non-smoking person in urban area are 11% less likely to try smoking than in rural area.

The graphs of prior and posterior densities of SD of school and SD of state are presented below.

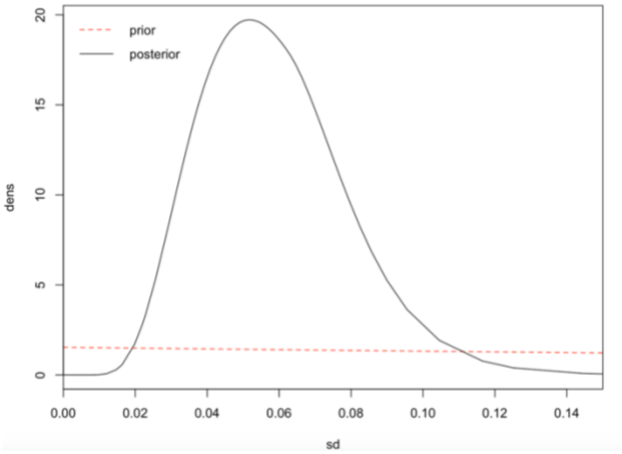

Figure 2. Prior and Posterior density of SD of state

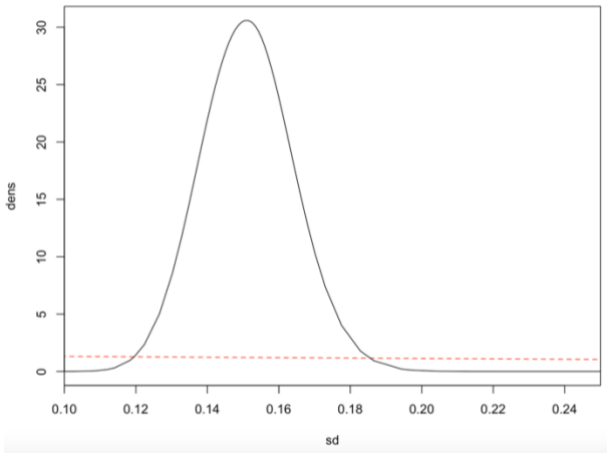

Figure 3. Prior and Posterior density of SD of school

According to Figure 2 and 3, SD of state centered around 0.06, whereas SD of school centered around 0.15. This result supports the posterior estimates of hyperparameters from our model graphically. Therefore, our model disagreed with the first researchers' first hypotheses, as tobacco control programs should target the schools with the earliest smoking ages irregardless of states states where smoking is a problem.

Finally, we plotted the prior and posterior density of Weibull Shape as shown below.

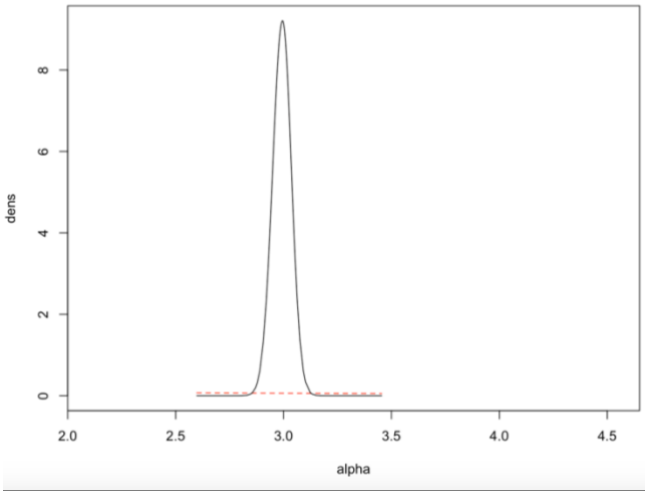

Figure 4. Prior and Posterior density of Weibull Shape

We can see from figure 4 that cigarette smoking does not have a flat hazard function. Therefore, two non-smoking children do not have the same probability of trying cigarette within the next month, irrespective of their age but provided all else being identical. The fact that older kids are more likely to start smoking might be the reason.

Conclusion

We carried out an analysis on the results of the 2014 American National Youth Tobacco Survey to help determine whether tobacco control programs should target the states with the ealiest smoking ages or concern themselves with finding particular schools where smoking is a problem.

From our results, we found that the school a child goes to has more impact on when the child will smoke their first cigarette than the state that he or she lives in. We also found that two non-smoking children do not have the same probability of first trying smoking within a given period of time. As such we should allocate the resources to focus towards schools rather than states to better control tobacco use for teenagers.

Question 2. Death on the roads

Introduction

We analyzed the subset of casualties-in-road-accidents data from www.gov.uk/government/statistical-data-sets/ras30. The original dataset included all the pedestrians involved in motor vehicle accidents in the UK from 1979 to 2015. Our main focus was to investigate whether men are involved in accidents more than women. Specifically, we would like to find out the whether the proportion of accidents which are fatal is higher for men than for women, at different age-stage.

Method

We fitted a conditional logistic regression model to perform a matched case-control studies, where we treated fatal accidents as cases and slight injuries as controls. Since there was interaction between sex and age, we used stratification by matching on time of day, whether condition is a lightning condition, and weather. Below is our fitted model:

$$logit[pr(Y_{ij} = 1)] = X_{ij}\beta + R_i\alpha$$

Where

- $X_{ij}\beta$ is the subjects age and interaction between age and gender
- $R_i\alpha$ is the binary indicator variables for each strata (matched sets)
- Y_{i1} is case i, Y_{ij} with j >1 are controls.

Results

	coef	exp(coef)
age0 - 5	0.1324083	1.1415744
age6 - 10	-0.3196593	0.7263965
age11 - 15	-0.3829384	0.6818549
age16 - 20	-0.4432109	0.6419718
age21 - 25	-0.2680862	0.7648419
age36 - 45	0.4115311	1.5091267

age46 - 55	0.7682289	2.1559445
age56 - 65	1.2120970	3.3605244
age66 - 75	1.7972504	6.0330360
ageOver 75	2.3957024	10.9759044
age26 - 35:sexFemale	-0.4482120	0.6387693
age0 - 5:sexFemale	0.0284229	1.0288306
age6 - 10:sexFemale	-0.1771162	0.8376825
age11 - 15:sexFemale	-0.2498614	0.7789087
age16 - 20:sexFemale	-0.2791322	0.7564399
age21 - 25:sexFemale	-0.3691252	0.6913389
age36 - 45:sexFemale	-0.4482308	0.6387573
age46 - 55:sexFemale	-0.3763107	0.6863891
age56 - 65:sexFemale	-0.2370677	0.7889379
age66 - 75:sexFemale	-0.1433569	0.8664448
ageOver 75:sexFemale	-0.1256106	0.8819582

Table 3.Estimated Coefficient of Parameters

Since we chose a baseline group of 25- 30, the coefficients for all the other age groups are presented relative to age group 25-30. For example, the negative coefficient of age group 21-25 indicates that age group 21-25 is less likely to get into a fatal accident than age group 25-30. Therefore, we can infer that men as teenagers tend to be safer as pedestrians than men in early adulthood, for the increasing coefficient from age group 6-10 to age group over 75. Notice that age group 0-5 is a special case that did not follow the pattern. This might be caused by the fact that small kids often like to work and play on the road, which increases the risks of unintentional injury.

Table 3 also shows us the interaction effect between sex and female, which allows us to calculate the mean effect for females by adding the interaction coefficient and age coefficient in the same age group. For example, the mean effect for female in age group over 75 is 2.396 + (-0.126) = 2.27. Therefore, if the interaction coefficient is negative, the odds of female getting into a fatal accident would be lower than males. Clearly, we can see that the interaction coefficients from age group 6-10 to age group over 75 are all negative. That is, women tend to be, on average, safer as pedestrians.

We plotted the odds of having a fatal accident for males and the estimated effect for females relative to males. Both plots are presented below.

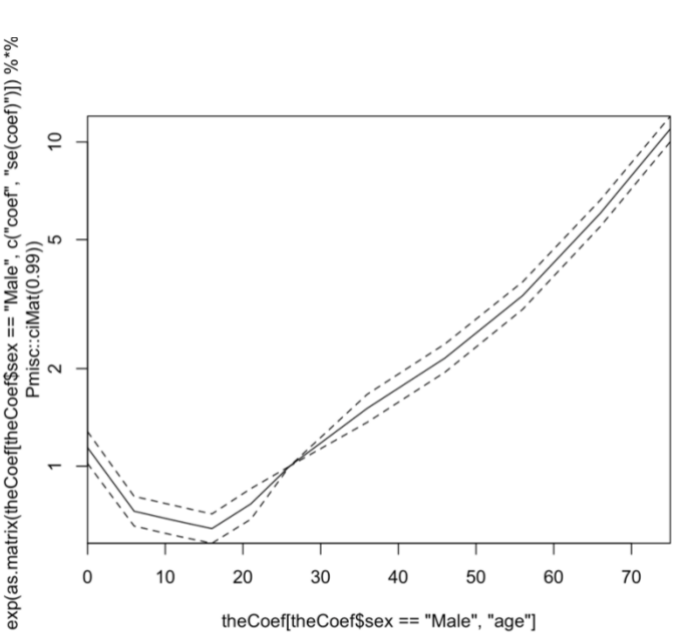

Figure 5. The Odds of Having a Fatal Accident for Males

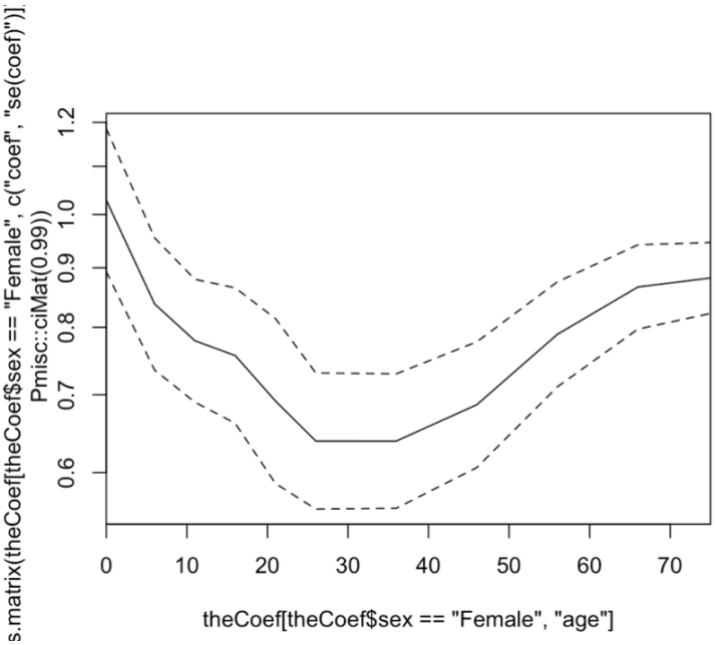

Figure 6. Estimated Effects for Females Relative to Males

According to Figure 5, age group 25-30 has an exponentiated estimates set to 1 as a baseline group. We can see that as age of men increases, the odds of having a fatal accident increases as well. This agrees with our previous analysis.

On the other hand, figure 6 shows us the exponentiated coefficient for the interaction effect. To get the odds of getting into a fatal accident, we now need to multiply the two point estimates in the two plots at the same age. Clearly, the point estimates in figure 6 are all smaller than 1, which again proves that the odds of women getting into a fatal accident are always lower than men.

Conclusion

We carried out an analysis on the road traffic accidents data in UK from 1979 to 2015 to determine whether women tend to be, on average, as pedestrians than men. From our results, we found that the odds of women getting into a fatal accident is actually lower than men. Particularly, men as teenagers are safer than men in adulthood. This could reflect that men are more likely to engage in risky behavior than women, especially men in adulthood.

Appendix

##Question1

```{r pressure, echo=FALSE}

#install.packages("R.utils")

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)

hist(forInla$Age_first_tried_cigt_smkg, xlab="Age Children First Tried Cigarettes", main = '')

```

##Methods

Since our analysis focuses on how factors affect time (age) to some event (first tried smoking), we modelled the data using a Weibull distribution, as it is effective for such survival analysis data. To assess the fit of a Weibull distribution, we plotted the empirical distribution along with the our data distribution in Figure 1. This appears to be a reasonable fit.

$$Y_{ij} \mid U_i, V_i \sim Weibull(\lambda_{ij}, k)$$

$$-log(\lambda_{ij}) = \eta_{ij} = X_{ij}\beta + U_i + V_i$$

$$U_i \sim N(0 , \sigma^2_{U}) $$

$$V_i \sim N(0 , \sigma^2_{V}) $$

Where

- $X_{ij}\beta$ is the subjects gender, ethnicity, whether they are from a rural or urban school, and interaction between ethnicity and whether they are from a rural or urban school

- U_i is the school random effect

- V_i is the state random effect

- The variance of U_i and V_i are themselves hyperparameters following a loggamma distribution

- k is the Weibull shape parameter and is normally distributed with its own hyperparameters

- λ is the Weibull scale parameter.

```{r, echo=FALSE}

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)

# if event=2, it's a left censoring

# 0 means they havent smoked yet, 1 means they smoked after 8, 2 means they smoked before age 8

forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2

cbind(forInla$Age, forInla$Age_first_tried_cigt_smkg,forSurv)[1:10,]
```

```
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(1.5,0.1)))) +
               f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(5, 0.1)))),
             control.family = list(variant = 1,
                                   hyper = list(alpha = list(prior = "normal", param = c(log(3), (2/3)^(-2))))),
             control.mode = list(theta = c(8,2, 5), restart = TRUE), data = forInla, family = "weibullsurv",
             verbose = TRUE)
```

```{r, echo=FALSE}
Table1 = rbind(fitS2$summary.fixed[, c("mean", "0.025quant",
                          "0.975quant")], Pmisc::priorPostSd(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")])
Ratio1 = exp(-Table1['mean'])
colnames(Ratio1) = 'Exp.(-Estimate)'
Table2 <-cbind(Table1, Ratio1)
Table2
knitr::kable(Table2,digits=2)
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters) {
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
}
fitS2$priorPost$legend$x = "topleft"
do.call(legend, fitS2$priorPost$legend)
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters) {do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)}
do.call(legend, fitS2$priorPost$legend)
```

##Question2

```
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)
#remove strata with no cases or no controls

theTable = table(pedestrians$strata, pedestrians$y)
onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)]
x = pedestrians[!pedestrians$strata %in% onlyOne, ]

library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
summary(theClogit)$coef[,1:2]

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$age), ]
theCoef
matplot(theCoef[theCoef$sex == "Male", "age"], exp(as.matrix(theCoef[theCoef$sex == "Male", c("coef", "se(coef)")]) %*%
                                        Pmisc::ciMat(0.99)),
      log = "y", type = "l", col = "black", lty = c(1,2, 2), xaxs = "i", yaxs = "i")
matplot(theCoef[theCoef$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")
```