STA442 Homework 4

Kevin Dang

December 01, 2019

Smoking

Introduction

The 2014 American National Youth Tobacco Survey is a rich data source that will allow us to answer various research hypotheses. The first hypothesis that we investigated is:

"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."

The second hypothesis that we investigated is:

"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."

After our investigation, we found that both hypotheses are not correct.

Methods

A Weibull distribution is fit to the response variable which is a survival object. State and school are used as random effects in the model. The normal prior for alpha with a mean of log(1) allows for a Weibull shape parameter of 1 since a flat hazard function is expected. PC priors are used to penalize model complexity for both the school and state random effects. The standard deviation of the PC prior for school (0.25) was chosen to be smaller than the PC prior for state (0.5) because the first hypothesis expects the variation between states to be larger. The model for state i, school j, individual k can be expressed as:

$$
\begin{aligned}
Y_{ijk} &\sim Weibull(\rho_{ijk}, \kappa) \\
\kappa &\sim N(\log(1), (3/2)^2) \\
\rho_{ijk} &= exp(-\eta_{ijk}) \\
\eta_{ijk} &= X_{ijk}\beta + U_i + V_{ij} \\
U_i &\sim N(0, \sigma_u^2) \\
\sigma_u^2 &\sim PC(0.5, 0.05) \\
V_{ij} &\sim N(0, \sigma_v^2) \\
\sigma_v^2 &\sim PC(0.25, 0.05)
\end{aligned}
$$

Results

In Table 1, the standard deviation for the school random effect is 0.1502, with a 95% credible interval of [0.1262,0.1765]. The standard deviation for the state random effect is 0.0571, with a 95% credible interval of [0.0247,0.1016]. Since the lower bound of the standard deviation for school (0.1262) is larger than the upper bound of the standard deviation for state (0.1016). This strongly suggests that the variation between states is significantly smaller than the variation amongst schools.

Table 1: Weibull GLM for Smoke data including Random effects

	mean	0.025quant	0.975quant
(Intercept)	-0.6220357	-0.6772820	-0.5660167
RuralUrbanRural	0.1149248	0.0555165	0.1739811
SexF	-0.0504601	-0.0790618	-0.0220176
Raceblack	-0.0481423	-0.0912823	-0.0056742
Racehispanic	0.0259263	-0.0088973	0.0605428
Raceasian	-0.1959019	-0.2886925	-0.1087757
Racenative	0.1106609	0.0046959	0.2092110
Racepacific	0.1767657	0.0087469	0.3263137
SexF:Raceblack	-0.0169491	-0.0743771	0.0403351
SexF:Racehispanic	0.0163584	-0.0299059	0.0626003
SexF:Raceasian	0.0055102	-0.1226357	0.1327920
SexF:Racenative	-0.0437959	-0.2015459	0.1106493
SexF:Racepacific	-0.1708109	-0.5035316	0.1238761
SD for school	0.1502480	0.1262369	0.1764638
SD for state	0.0570653	0.0246568	0.1015851

Figure 1 below contains plots for the prior and posterior of the shape parameter alpha as part of the Weibull survival model, as well as plots for the prior and posterior densities for school and state random effects that were used in the above model. The plot of alpha is centered around about 3, which already indicates to us that the hazard function will not be flat as alpha is not centered at 1. This will be verified in a plot of the hazard function.

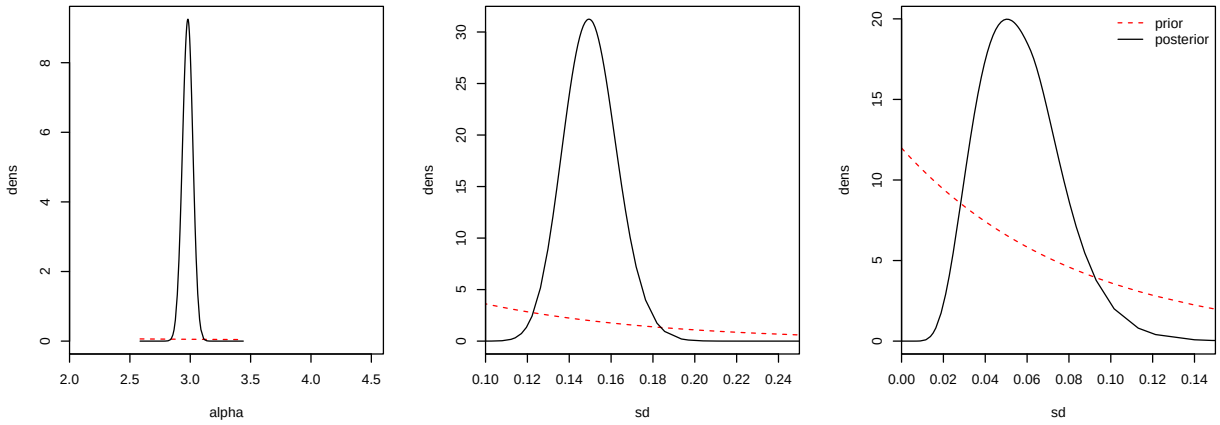

Figure 1: alpha for weibullsurv, sd for school, sd for state

Figure 2 contains a plot of the cumulative hazard function. If a hazard function is flat, then the cumulative hazard function should be a straight line with a constant slope. That is not the case here, where we can see the function increasing and the slope is increasing. This means that first cigarette smoking does not have a flat hazard function. In other words, two non-smoking children have the same probability of trying cigarettes within the next month if they both have the same sex, ethnicity, geographic location and if they attend the same school.

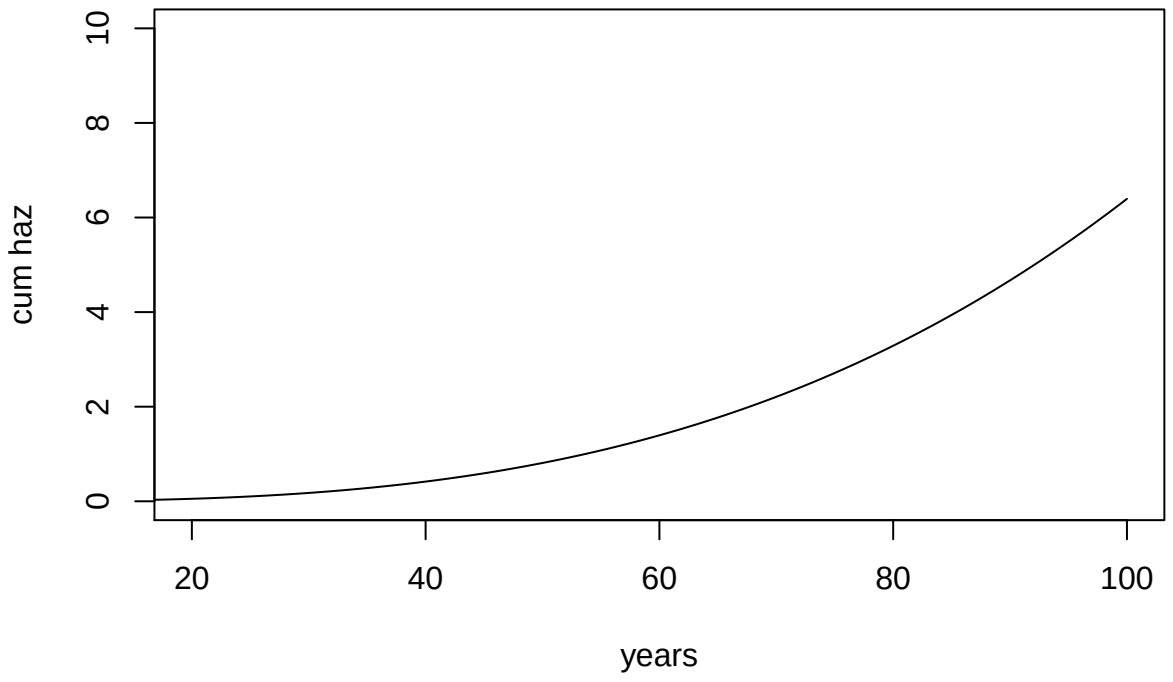

Figure 2: Cumulative Hazard Function

Conclusion

The variation between states is significantly smaller than the variation amongst schools, due to the 95% credible interval for the standard deviation of state lying below the 95% credible interval of the standard deviation for school. In addition, since the hazard function is not flat this means that two non-smoking children have the same probability of trying cigarettes within the next month if their confounding variables and random effects are the same. In conclusion, our results reject both hypotheses.

Appendix A: Smoking Code

```
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(0.25, 0.
#               f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(0.5, 0.05
#             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)

# Save and reload to save time on knitting
#saveRDS(fitS2, "fitS2.rds")
fitS2 <- readRDS("fitS2.rds")

# GLM output
knitr::kable(rbind(fitS2$summary.fixed[, c("mean", "0.025quant", "0.975quant")],
                   Pmisc::priorPostSd(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")]),
             caption = "\\label{tab:1} Weibull GLM for Smoke data including Random effects")

# Prior and Posterior plots
par(mfrow=c(1,3))

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

xSeq = seq(5,100,len=1000)
kappa = fitS2$summary.hyper['alpha', 'mode']
lambda = exp(fitS2$summary.fixed['(Intercept)', 'mode'])

# Cumulative hazard function
plot(xSeq, (xSeq / (100*lambda))^kappa, type='l', ylim=c(0.001, 10), xlim = c(20,100),
     xlab = 'years', ylab = 'cum haz')
```

Death on the roads

Introduction

We explore a dataset containing car accident information in the UK from 1979 to 2015. This dataset contains accidents that were either fatal or resulted in slight injuries. The hypothesis of interest to be explored is whether teenage girls and young women are safer than teenage boys and young men on average.

Methods

A Conditional Logistic Regression model is used to answer the research question with fatal accidents as the response variable, age and sex as the covariates. This model provides estimates of the regression coefficients of age and sex that vary within three strata. The data is stratified by light conditions (daylight, darkness with lighting, darkness), weather conditions (rain, snow, fog, mist, wind) and date (specifically by hour). The model is written as:

$$Y_i \sim Bin(N_i, \mu_i)$$
$$\log(\frac{\mu_i}{1-\mu_i}) = X_i\beta + z_j\alpha$$

where z_j are binary indicator variables for each strata and α's are the regression coefficients associated with the stratum indicator variables. The CLR algorithm will estimate the β's (not the α's), which can be used to analyze odds ratios of each covariate adjusted for the others.

Results

From Table 2, we can see that the odds-ratio for females aged 16-20, 21-25 and 25-36 are 0.756, 0.691 and 0.639 respectively. This means that females of those age groups (teenage years and early adulthood) are less likely to be involved in a fatal accident compared to males of the same age group.

Table 2: Conditional Logistic Regression model for pedestrian accident data

	coef	exp(coef)
age0 - 5	1.1415744	3.131695
age6 - 10	0.7263965	2.067616
age11 - 15	0.6818549	1.977542
age16 - 20	0.6419718	1.900224
age21 - 25	0.7648419	2.148655
age36 - 45	1.5091267	4.522779
age46 - 55	2.1559445	8.636043
age56 - 65	3.3605244	28.804292
age66 - 75	6.0330360	416.979063
ageOver 75	10.9759044	58448.681531
age26 - 35:sexFemale	0.6387693	1.894148
age0 - 5:sexFemale	1.0288306	2.797792
age6 - 10:sexFemale	0.8376825	2.311005
age11 - 15:sexFemale	0.7789087	2.179093
age16 - 20:sexFemale	0.7564399	2.130677
age21 - 25:sexFemale	0.6913389	1.996387
age36 - 45:sexFemale	0.6387573	1.894125
age46 - 55:sexFemale	0.6863891	1.986529
age56 - 65:sexFemale	0.7889379	2.201057
age66 - 75:sexFemale	0.8664448	2.378440
ageOver 75:sexFemale	0.8819582	2.415625

We can further expand on the results by looking at Figure 3 below. The odds ratio for females consistently remain below 1 starting from childhood. The only time the odds ratio is above 1 is during infancy. This shows that females are safer than men across almost all age groups. Specifically, the age group in which women have the smallest odds ratio is 26-35 years.

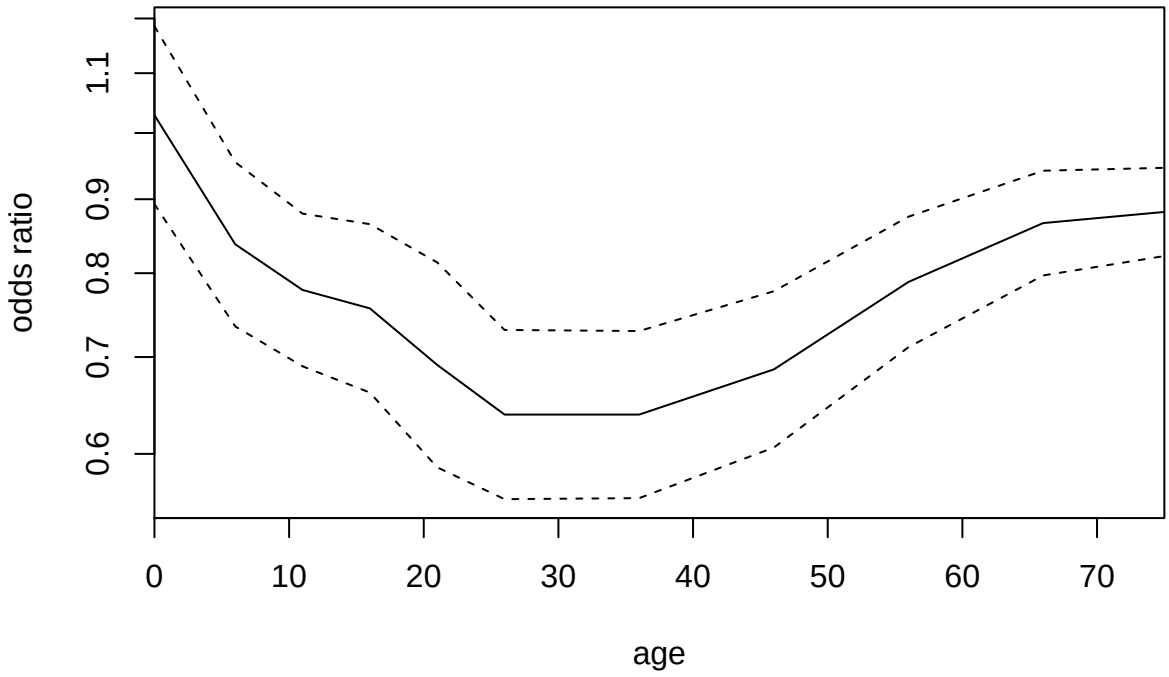

Figure 3: 95% Credible intervals for Female odds ratio by age

Conclusion

On average, women tend to be safer as pedestrians than men, particularly as teenagers and in early adulthood. This also holds true for just about every other age group, with the exception of infants and toddlers who are an anomaly because they lack the cognitive abilities and awareness to make decisions and avoid dangers.

Appendix B: Death Code

```
#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$t
#
#theTable = table(pedestrians$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)  # conditional logistic regression

# Save and reload to save time on knitting
#save(theClogit, file="theClogit.Rdata")
load("theClogit.Rdata")
knitr::kable(exp(summary(theClogit)$coef[,1:2]),
             caption = "\\label{tab:2} Conditional Logistic Regression model for pedestrian accident da

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), ]

# Odds ratio vs Age
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", xlab = "age", ylab = "odds
```