Q1 Smoking

Introduction

We used R to analyze the 2014 American National Youth Tobacco Survey data in detail. The data was available from http://pbrown.ca/teaching/appliedstats/data/smoke.RData. The data was collected from students from different schools across the United States. We focus on the parameters that could affect the earliest smoking ages and smoking behaviors. For the first question, we wanted to explore whether the age children first try cigarettes depends more on the state the student lives in or the school he goes to. For the second question, controlling for known confounders (sex, rural/urban, ethnicity) and random effects (school and state), two non-smoking children have the same chance of trying cigarettes within a given period.

Methods

For analysis, we used the **Weibull distribution** to fits the data since children "first time start smoking" could be considered as a survival analysis. In this case, consider the following model:

$$Y_{ijk} | U_{ijk}, V_{ijk} \sim Weibull(\lambda_{ijk}, \alpha)$$

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

$$\eta_{ijk} = X_{ijk}\beta + V_{ijk} + U_{ijk}$$

$$U_{ijk} \sim N(0, \sigma_u^2)$$

$$V_{ijk} \sim N(0, \sigma_v^2)$$

Where

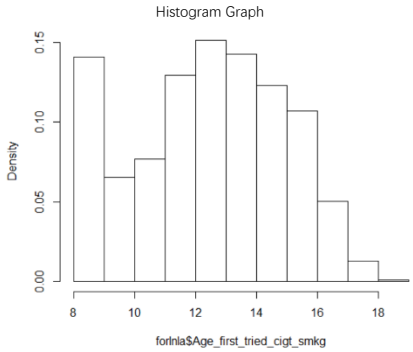
$$state\ i, school\ j, individual\ k$$

$X_{ijk}\beta$ is the fixed effect from subjects gender, ethnicity, whether they are from a rural or urban

$$U_{ijk}\ is\ state\ random\ effect$$

$$V_{ijk}\ is\ school\ random\ effect$$

The variance of Ui and Vi are hyperparameters following Penalized Complexity Prior(PC Prior)



We used histogram of data to check if the data fits the model well. The data and fitted line followed the same pattern (looks Weibull) which was good.

Quantiles of Prior Distributions of Parameters

```
Model hyperparameters:
                                    mean      sd 0.025quant 0.5quant 0.975quant   mode
alpha parameter for weibullsurv     2.98   0.043       2.89     2.98       3.06   2.97
Precision for school               46.95   8.020      32.98    46.35      64.46  45.23
Precision for state               390.21 319.800      84.37   299.53    1234.75 188.62
```

Because we used Bayesian here, we need a prior for all unknown quantities. We already modeled U and V as Gaussian. Since we don't know σ_v^2 *and* σ_u^2, we chose **penalized complexity prior** (PC Prior) which put an exponential prior on the standard deviation σ_u *and* σ_v. Therefore, $\boldsymbol{\sigma_u \sim \exp(\lambda_u)}$ ***and*** $\boldsymbol{\sigma_v \sim \exp(\lambda_v)}$.We selected the hyperparameters of the above model using information from the collaborating scientists. To calibrate the scaling of the random effects prior, we set A and p so that $P(\sigma > A) = p$.

We want $\exp(U_i) = 2$ *or* 3 but unlikely to see at 10, so $U_i = \ln 2 \approx 0.7$. We set **ln10 = 2. 3 at 0.975 quantile of U**, $\mu + 2\sigma_u = 2\sigma_u = 2.3$ $(\mu = 0\ since\ U \sim N(0, \sigma_u^2))$ *and* $0.7 < 2\sigma_u < 2.3$. We wanted the probability $\sigma_u > \frac{2.3}{2} = 1.15$ as small as possible, for precision, setting the prior to satisfy $P(\sigma_u > 1) = 0.05$.

For interpretation, there was a 5% chance that between subject variability $\sigma_u > 1$.

$$P(\sigma_u > 1) = 0.05 \quad \sigma_u \sim \exp(\lambda_u)$$

$$1 - [1 - F(\sigma_u; \lambda_u)] = 1 - [1 - e^{-\lambda\sigma_u}] = 1 - [1 - e^{-\lambda}] = 0.05$$

$$\lambda_u = 2.99 \quad \boldsymbol{\sigma_u \sim \exp(2.99)}$$

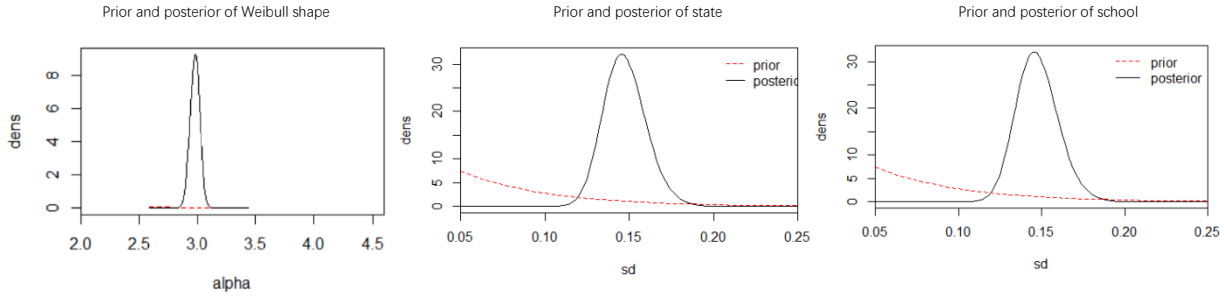
Within a given state, the 'worst' schools were expected to have at most 50% greater rate than the 'healthiest' schools or $exp(Vij) = 1.5$ for a school-level random effect was about the largest we'd see. We set **ln1.5=0.405 at 0.975 quantile of V**, $\mu + 2\sigma_v = 2\sigma_v = 0.405$ $(\mu = 0\ since\ V \sim N(0, \sigma_v^2))$ and $2\sigma_v < 0.405$. We wanted the probability $\sigma_v > \frac{0.405}{2} = 0.2$ as small as possible, for precision, setting the prior to satisfy $P(\sigma_u > 0.2) = 0.02$. For interpretation, there was a 2% chance that between subject variability $\sigma_u > 0.2$.

$$P(\sigma_v > 0.2) = 0.02 \quad \sigma_v \sim \exp(\lambda_v)$$

$$1 - [1 - F(\sigma_v;\ \lambda_u)] = 1 - [1 - e^{-\lambda\sigma_v}] = 1 - [1 - e^{-0.2\lambda}] = 0.02$$

$$\lambda_v = 19.6 \quad \sigma_v \sim \exp(19.6)$$

We also don't know the shape parameter (α) of Weibull. Since the prior on the Weibull shape parameter should allow for a 1, we set a normal distribution (log(1), (2/3)^(-2)) for the α of Weibull since exp(log(1))=1.



Result

Table 1

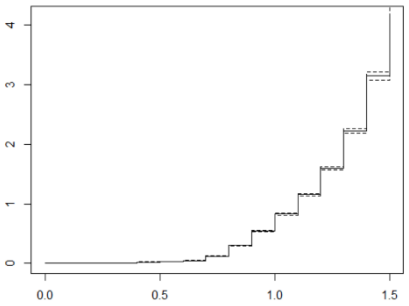
	mean	sd	q0.025	q0.5	q0.975	mode
alpha parameter for weibullsurv	2.97640098	0.04297350	2.89439055	2.97532733	3.0627624	2.97260770
SD for school	0.14752501	0.01252514	0.12472241	0.14685607	0.1738674	0.14546193
SD for state	0.06049408	0.02057544	0.02855085	0.05775116	0.1084676	0.05234072

The posterior mean random effect standard deviation for school was about 0.15 and a 95% credible interval was (0.124, 0.174). The posterior mean random effect standard deviation for state was 0.06 and a 95% credible interval was (0.029, 0.108). The posterior mean of α parameter for weibullsurv was 2.97 and a 95% credible interval was (2.89, 3.06).

Table 2

	mean	0.025quant	0.975quant
(Intercept)	-0.622564367	-0.678237345	-0.566190594
RuralUrbanRural	0.115118321	0.055248805	0.174619736
SexF	-0.050453055	-0.079060207	-0.022006660
Raceblack	-0.048306382	-0.091467650	-0.005818811
Racehispanic	0.025850805	-0.008986276	0.060482026
Raceasian	-0.195915076	-0.288719654	-0.108775679
Racenative	0.110525439	0.004534502	0.209105095
Racepacific	0.176525571	0.008473491	0.326112515
SexF:Raceblack	-0.016975589	-0.074410472	0.040315446
SexF:Racehispanic	0.016354777	-0.029914799	0.062602724
SexF:Raceasian	0.005507132	-0.122643508	0.132794519
SexF:Racenative	-0.043953958	-0.201723309	0.110510700
SexF:Racepacific	-0.170589608	-0.503344810	0.124136015
SD for school	0.151794442	0.127081632	0.178301464
SD for state	0.057946764	0.025210128	0.103239102

Table 3: Cumulative Hazard Function



The hypothesis stated that the school a child goes to affects less on their first smoke cigarette age than the state that they live in. From the summary table 2, it contained the mean and 95% credible interval of fixed effects and standard deviation of random effect. Noticed that $\sigma_v^{school} = 0.15$ *and* $\sigma_u^{state} = 0.057$. $\boldsymbol{\sigma_v^{school} > \sigma_u^{state}}$, we **rejected the hypothesis**. In other words, holding other variables constant, geographic variation (between states) in the **mean age children first try cigarettes was substantially smaller than variation amongst schools**. As a result, tobacco control programs should concern themselves with finding schools where smoking was a problem and not target the states with the earliest smoking ages.

The hazard function models which periods had the highest or lowest chances of an event. The Weibull density haea Hazard function $h(x; k, \lambda) = \frac{kx^{k-1}}{\lambda^k}$ (λ scale parameter, k shape parameter) and cumulative Hazard function $H(x; k, \lambda) = \left(\frac{x}{\lambda}\right)^k$. Depending on whether k was greater than or less than 1, the hazard can increase or decrease with increasing y. When shape = 1, we could get a constant hazard. From table 1, the posterior mean of α shape parameter for weibullsurv was 2.97 and a 95% credible interval was (2.89, 3.06). **1 was not in credible interval, thus the hazard function is not constant.** We had an increasing hazard and cumulative Hazard function had not linear line. Furthermore, from table 3, also noticed that cigarette smoking does not have a flat cumulative hazard function. Therefore, **two nonsmoking children do not have the same probability of trying cigarettes within the next month**, irrespective of their ages but provided the known confounders (sex, rural/urban, ethnicity) and random effects (school and state) are identical. The older children are more likely to begin smoking.

Q2 Death on the roads

Introduction

We used R to analyze all the road traffic accidents in the UK from 1979 to 2015 data in detail. The data was available from https://www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents. The information was collected from all pedestrians involved in motor vehicle accidents with either fatal or slight injuries. We focus on the parameters that could affect the safety of pedestrians. For the first question, we wanted to explore whether the women tend to be, on average, safer as pedestrians than men. For the second question, if women were safer than men, particularly as teenagers and in early adulthood.

The method

Conditional logistic regression is a specialized type of logistic regression usually used when case subjects with a particular condition. For this analysis, we used **the Conditional Logistic Regression** and consider the following model:

The generalized linear regression model:

$$\mathbf{pr}(\mathbf{Y_{ij}} = \mathbf{1}|\mathbf{X_{ij}}) = \boldsymbol{\lambda_{ij}}, \mathbf{log}\left[\frac{\boldsymbol{\lambda_{ij}}}{\mathbf{1}-\boldsymbol{\lambda_{ij}}}\right] = \boldsymbol{\beta_0} + \sum_{\boldsymbol{p=1}}^{\boldsymbol{p}} \boldsymbol{X_{ip}\beta_p}\ \boldsymbol{without\ grouping}$$

The Conditional Logistic Regression model:

$$\text{logit}\left[\mathbf{pr}(\boldsymbol{Y_{ij}} = \mathbf{1})\right] = \boldsymbol{\alpha_i} + \mathbf{X_{ij}}\boldsymbol{\beta}$$

$$\text{logit}\left[\frac{\text{pr}(\text{Y}_{\text{ij}} = 1)}{\text{Z}_{\text{ij}} = 1}\right] = \alpha_i^* + \text{X}_{\text{ij}}\beta$$

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

For each case i find a number of similar controls; Y_{i1} *is case of* i, Y_{ij} *with* $j > 1$ *are controls*

Z_{ij} *is* "in the study" *indicator*; $\text{Z}_{\text{ij}} = 1$ *means in study*

covariates X_{ij} *are variables not used in matching*

i is the strata: we had lots of group with different Light Conditions, Weather Conditions, time

β's are the population regression coefficients to be estimated

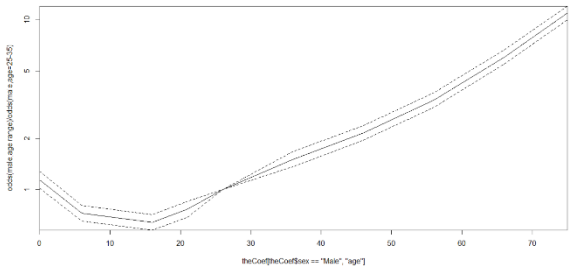
We set fatal accidents as cases and slight injuries as controls.

Results

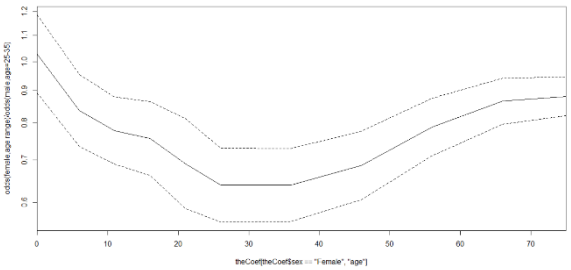
Graph 1

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

Graph 2: odds ratio graph for male



Graph 3: odds ratio graph for female



We set male from 26-35 years old as reference group. Noticed that $\exp(\text{coef}) = \frac{\text{odds}(sex, age\ range)}{odds(male,\ age{=}26{-}35)} = odds\ ratio$. Larger odds ratio represented larger probability of involved in fatal accidents. When we look at the table 1, we found that women and men had similar odds ratio around 0.6 to 0.76 from age 5-25. For example, $\frac{\text{odds}(male,\ age{=}11{-}15)}{odds(male,\ age{=}26{-}35)} = 0.68\ and\ \frac{\text{odds}(female,\ age{=}11{-}15)}{odds(male,\ age{=}26{-}35)} = 0.78$. The odds ratios were pretty close. Thus, we expected that males and females had a similar chance involved in fatal accidents when they were young. However, the table indicated a significant difference between men and women in odds ratios as their age grows. When males and females both over 75 years old, the odds ratio of the male was 10.98, which was ten times compared to females' odds ratio(0.88). The graph 2 and graph 3 illustrated that as age increased, males had a significant increase in odds rates with a range from 0.64 to 10.98, and females only had a small fluctuation in the odds ratios with a range from 0.64 to 1.03. Thus, **we could conclude that the chance of women involved in fatal accidents was less than men.** In general, women acted as safer pedestrians.

Furthermore, the hypothesis stated that women were safer as pedestrians than men particularly as teenagers and in early adulthood. We defined teenagers and early adulthood with age range 15-20. The odds ratio for male was $\frac{\text{odds}(male,\ age{=}15{-}20)}{odds(male,\ age{=}26{-}35)} = 0.64$ and for femal was $\frac{\text{odds}(female,\ age{=}15{-}20)}{odds(male,\ age{=}26{-}35)} = 0.76$. The difference on odds ratio was not significant. In fact, men performed as safer pedestrians as women when they were 15-20 years old. However, the odds ratios gap between male and female were enlarged as age increased. And women performed much well compared to men. Graph 2 and 3 indicated that over 46 years old, women's odds ratio range were 0.69 to 0.88 and men's odds ratio range were 2.16 to 10.98. Men had a larger probability on involved in fatal accidents. Therefore, the hypothesis was invalid. In conclusion, **women were safer as pedestrians than men particularly in elderly age.**

Q2- 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$timeCat)

theTable = table(pedestrians$strata, pedestrians$y)
# theTable[, 2]==0 ???TRUE
# theTable[, 1]==0 ???false

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

summary(glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
            data = x, family = "binomial"))$coef[1:4, ]

a=glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
      data = x, family = "binomial")
hist(a$coefficients,main = "",prob=TRUE)
summary(a)

install.packages("Publish")
library(Publish)
publish(a,intercept=TRUE)

library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata),
                   data = x)
summary(theClogit)
knitr::kable(summary(theClogit)$coef,digits=2)

publish(theClogit,intercept=TRUE)

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)), ylab="odds(male,age range)/odds(male,age=25-35)", 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)), ylab="odds(female,age range)/odds(male,age=25-35)",log = "y", type =
"l", col = "black", lty = c(1,2, 2), xaxs = "i")
```