STA442 Assignment 4

Xin Wei

23/11/2019

What Affects the Age Children First Try Cigars?

Summary

In our analysis of the age at which children first try cigarette smoking, we found that the variation in the mean age children first try cigarettes from one school to the next is more significant than that amongst the US states. And older children are more likely to start trying smoking cigarettes within the next month, provided the known confounders (sex, rural/urban, etnicity) and random effects (school and state) are identical.

Introduction

It is observed that an increasing number of teenagers have started smoking early, which causes lots of health issues. National Youth Tobacco Survey (NYTS) has helped provide the data and support state and national Tobacco prevention and control. We used the data available on the pbrown.ca page, where there is an R version of the 2014 dataset smoke.RData and a pdf documentation file 2014-Codebook. Our purpose was to explore whether the mean age at which children first try cigarettes varies more amongst different schools than US states. And we also wanted to know if first cigarette smoking is a first order Markov process and irrelevant to children's ages, given their sex, rural/urban, ethnicity, school and state identical.

Methods

Our dataset was released by Center for Disease Control and Prevetion. The study was conducted during spring of 2014, in order to support the estimation of tobacco-related knowledge, attitudes, and behaviors in a national population of public and private school students, by grade, age, school level, sex, and race/ethics. 22007 individuals voluntarily participated in the survey. The data was collected using a stratified, three-stage cluster sampling design. As an explanatory tool, we used a Weibull model as following:

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

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

$\eta_{ijk} = \sum_{l=1}^{3} X_{ijkl}\beta_l + U_i + V_{ij}$

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

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

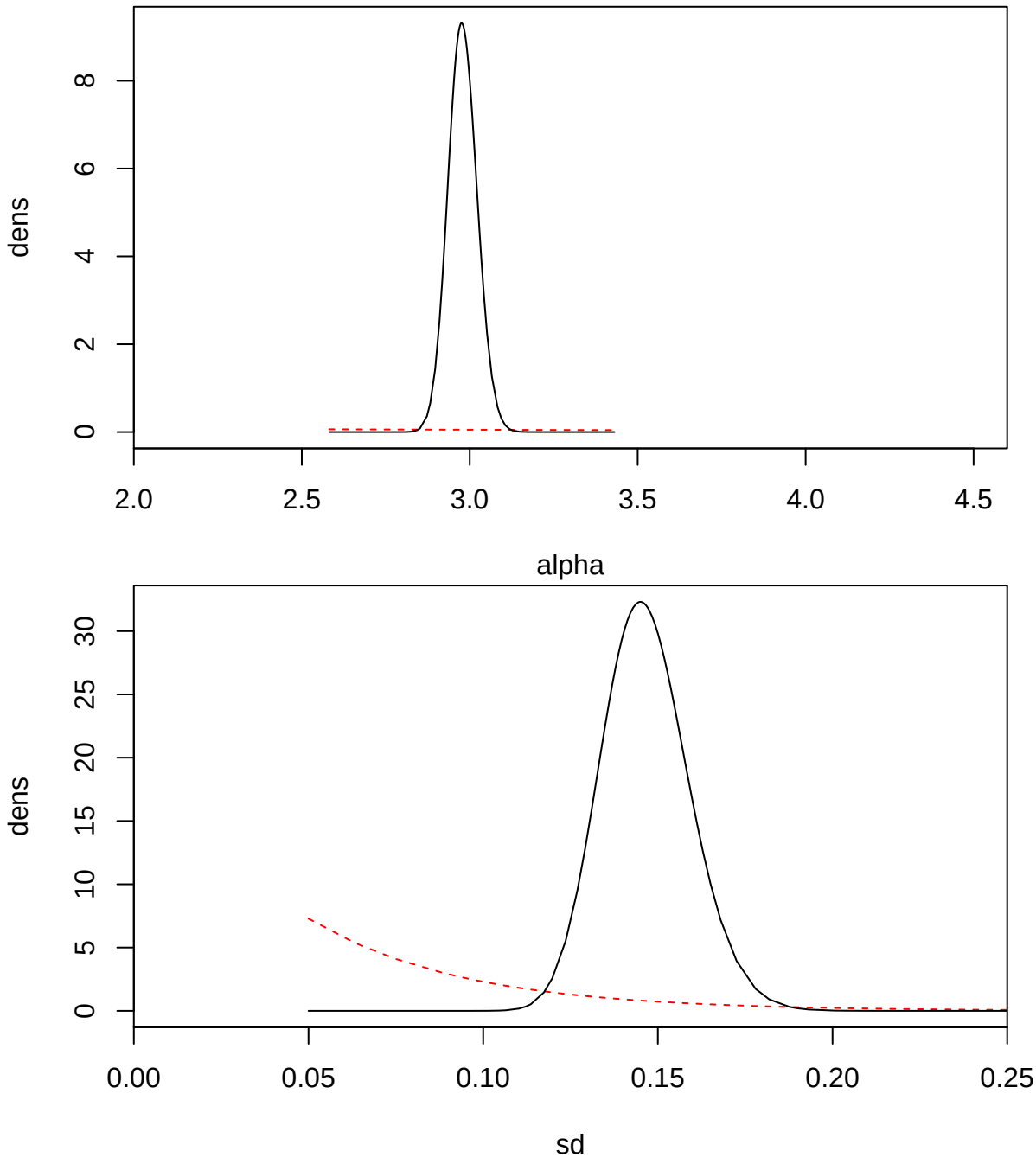
where: i for states, j for schools, and k for individuals. κ is the shape parameter, and ρ_{ijk} is the scale parameter. η_{ijk} is a sum of fixed effects(urban/rural, race, and gender) and random effects(U_i for state-to-state variations and V_{ij} for school-to-school variations). Firstly, based on historical data, we wanted to find the priors for the standard deviations of random variables. We wanted to use the penalized complexity prior here because it would be easier for us to adjust the paramters of our prior. Given the prior information that $\exp(U_i) = 2$ or 3 but unlikely to see at 10, hence U_i varies roughly from $\log(2) = 0.7$ to $\log(10) = 2.3$. Based on that, since U_i follows a Normal distribution with mean=0, we assumed the prior as $P(2\sigma_{U_i} > 2.2) = 0.01$, i.e., $P(\sigma_{U_i} > 1.1) = 0.01$ where $\sigma_{U_i} \sim exp(4.18)$. Similarly, $\exp(V_{ij})$ was given probably no greater than 1.5, thus, V_{ij} was probably no greater than $\log(1.5) = 0.405$. Then we assumed $P(\sigma_{V_{ij}} > 0.2) = 0.01$, where $\sigma_{V_{ij}} \sim exp(23)$. The shape parameter $\alpha \sim Lognormal(log(1), (\frac{2}{3})^{-2})$ where $(\frac{2}{3})^{-2}$ is the precision. The time (the age one

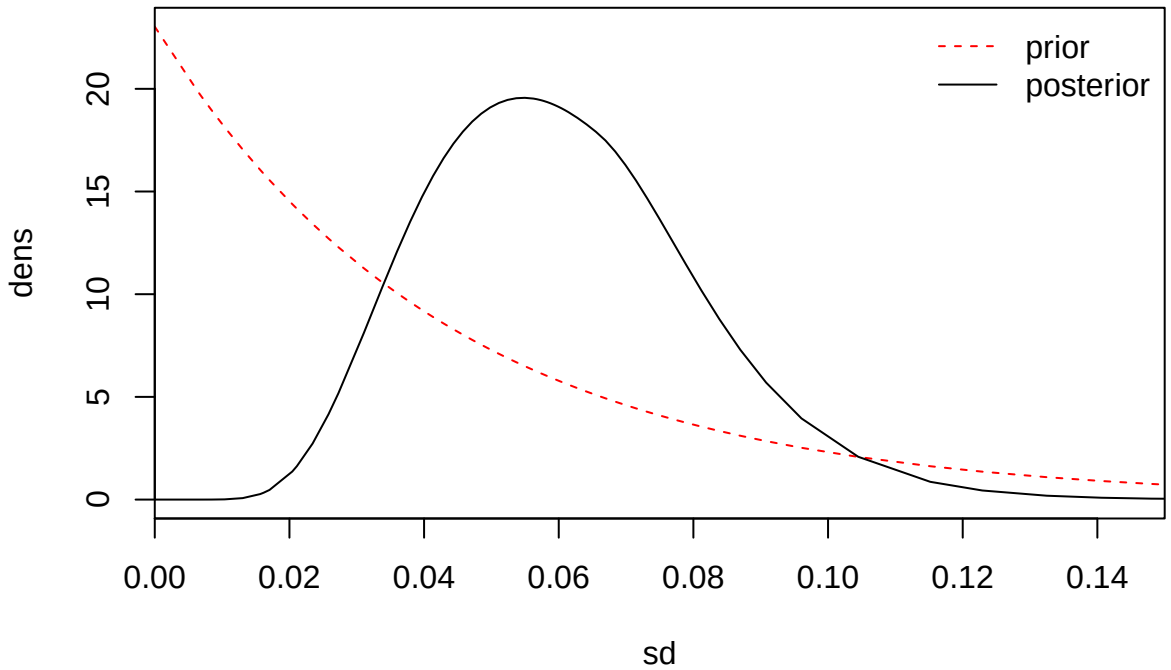
first tried smoking or the age one not smoked yet) has been standardized, by excluding children at or under 4 years old (assumed no one smokes before 5 years old), and divided by 10 (to show in a decade scale). As for the shape parameter α (greater than 0), we assumed a lognormal prior where the mean is 1 and the precision is $(\frac{2}{3})^{-2}$.

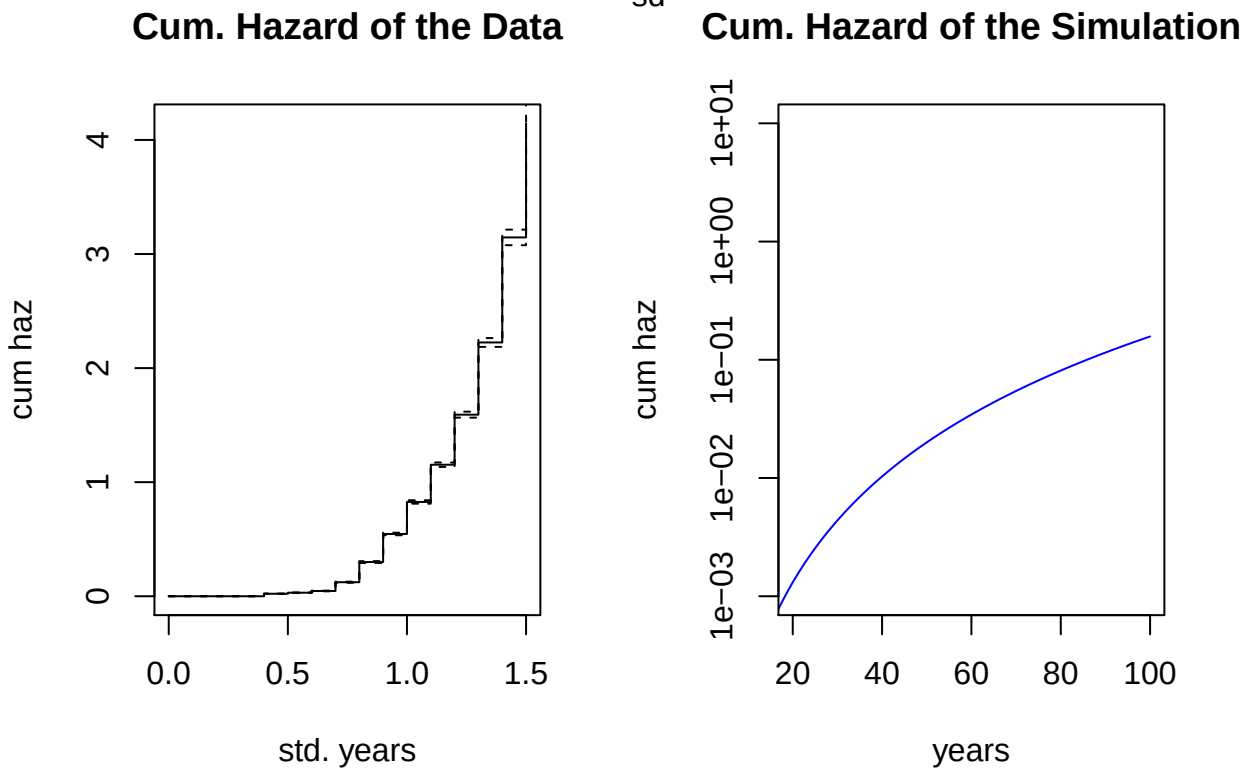
Results

Shown above are three graphs of prior and posterior distributions for parameters(alpha, sd for school, sd for state respectively). In the first distribution graph, we observe the estimate of α is nearly 3, which is greater than 1 and thus indicates an increasing hazard. In other words, the hazard function $h(t) = \frac{f(t)}{S(t)}$(S(t) as survival function) is not a constant function, but has some slope. It is confirmed in our cumulative hazard graphs. We know that a cdf of a constant pdf should be a linear function with some slope. But when we look at the cumulative hazard graph of our data, it has an increasing zigzag shape, which is far from a linear line. Also, the simulation plot shows a parabola instead of a linear shape. Hence, we have strong evidence to refuse the statement that first cigarette smoking has a flat hazard, instead, older children are more likely to try cigarette for the first time within the next month. After that, we take a look at the second and third prior-posterior graphs, and also the table of parameters of our model. we find the standard deviation for school is approximately 0.15, which is roughly five times the SD for state(0.05). In other words, the geometric variation among states in the mean age children first trying cigarettes is substantially smaller than the variation among schools. Hence, tobacco control programs should focus on specific schools where students started smoking earliest, instead of concerning themselves with some states, which are broad targets.

	mean	0.025quant	0.975quant
(Intercept)	-0.6206886	-0.6749550	-0.5655190
RuralUrbanRural	0.1134924	0.0546218	0.1719585
SexF	-0.0497342	-0.0698744	-0.0296955
Raceblack	-0.0559677	-0.0898431	-0.0224979
Racehispanic	0.0336575	0.0061832	0.0610378
Raceasian	-0.1934251	-0.2621386	-0.1277323
Racenative	0.0923675	0.0106972	0.1696136
Racepacific	0.1253913	-0.0187815	0.2552857
SD for school	0.1465533	0.1235968	0.1725616
SD for state	0.0605055	0.0272729	0.1044926


dens
0
2
4
6
8
2.0
2.5
3.0
3.5
4.0
4.5
alpha
dens
0
5
10
15
20
25
30
0.00
0.05
0.10
0.15
0.20
0.25
sd


prior
posterior
dens
0
5
10
15
20
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
sd


Cum. Hazard of the Data
Cum. Hazard of the Simulation
cum haz
0
1
2
3
4
0.0
0.5
1.0
1.5
std. years
cum haz
1e−03
1e−02
1e−01
1e+00
1e+01
20
40
60
80
100
years

Males More in Danger as Pedestrians Than Females?

Summary

In our analysis of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries in the UK from 1979 to 2015, we found that female in their teenage years are slightly more likely to get involved in severe accidents than teenage males. However, males in their older years are much more likely (above 11 times) to get involved in fatal accidents than the female conterparts.

Introduction

It has been an increasingly concerned issue that road accidents caused a large number of deaths each year. And the probabilities of such fatal accidents happening are believed to vary based on different light conditions, weather, and time of day. We used the dataset Casualties involved in reported road accidents (RAS30), which was provided by Department for Transport at the site of GOV.UK. Most of the statistics are based on road accidents reported to the police (Stats19). Our purpose is to explore how the probabilities of getting involved in severe road accidents differ between males and females in different age groups.

Methods

The Stats19 data we are using here are a set of numeric and alphabetic data. These data are collected by each police force in Great Britain. Every personal injury road traffic collision that is reported to the police is recorded in an administrative system that includes an element of statistical reporting. The information collected as part of Stats19 is revised every five years. The latest version of the form was introduced form the beginning of 2011.

As an explanatory tool, we firstly use a logistic model to estimate our coefficients (β's), then, since a matched case-control study is conducted, we want to fit a conditional logistic regression to our data, where fatal accidents were treated as cases and slight injuries as controls. The models are shown as follows:

$logit[P(Y_{ik} = 1)] = \alpha_i + X_{iage}\beta_{age} + X_{isex}\beta_{sex} + X_{ilight}\beta_{light} + X_{iweather}\beta_{weather}$,

$logit[P(Y_{ijk} = 1)|Z_{ijk} = 1] = \alpha_i^* + X_{ijage}\beta_{age} + X_{ijage*sex}\beta_{age*sex}$,

where $\alpha_i^* = \alpha_i + log[P(Z_{ijk} = 1|Y_{ijk} = 1)/P(Z_{ijk} = 1|Y_{ijk} = 0)]$.

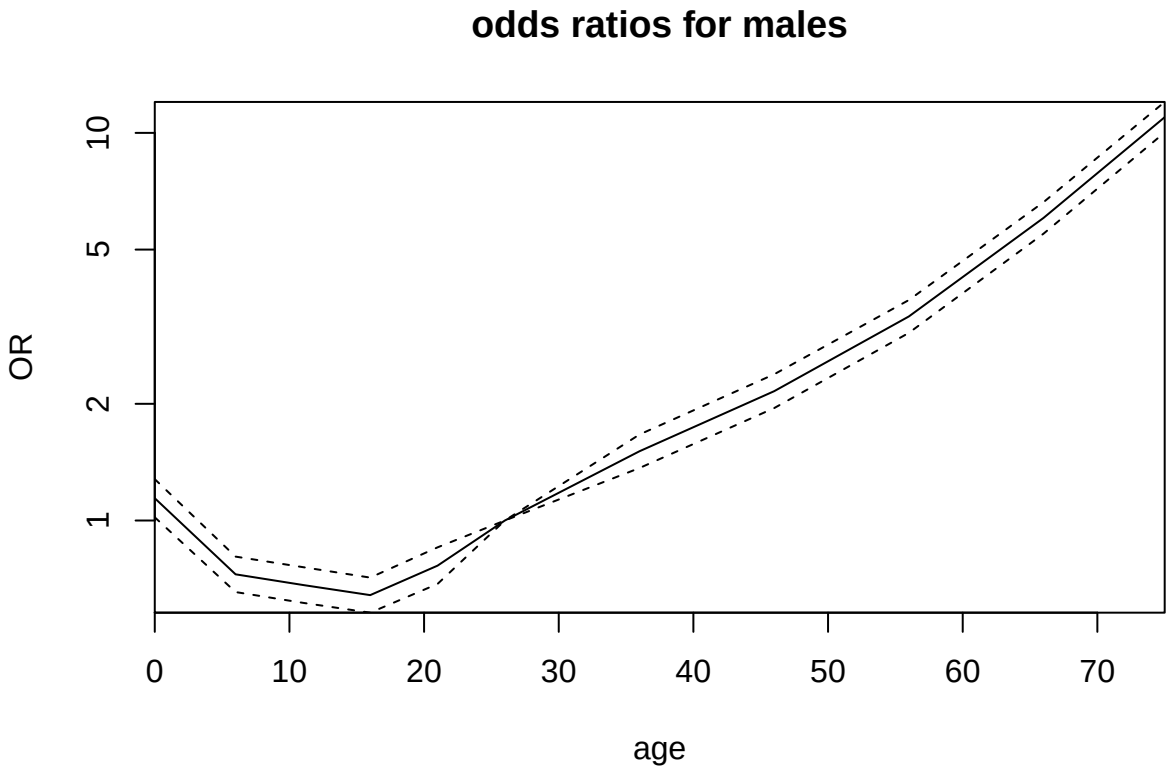
We treat a fatal accident(j=1) as case i, and slight injuries as control(j=2). Covariates X_{ijk} (k represents age & sex interacting age) are variables not used in matching, and α_i's are stratum constants, where stratifications are made based on time of the day, lighting conditions(daylight, darkness, etc.), and weather conditions(raining no high winds, fine + high winds, etc.), In other words, groups of controls are matched to subgroups of cases, all of which have the same values of the confounding variables(time of the day, lighting conditions, and weather). (Stratification is defined as the process of partitioning data into distinct or nonoverlapping groups.) Furthermore, the stratum effects were applied to each matching case-control pair. The group of male aged from 26 to 35 was treated as the reference group.

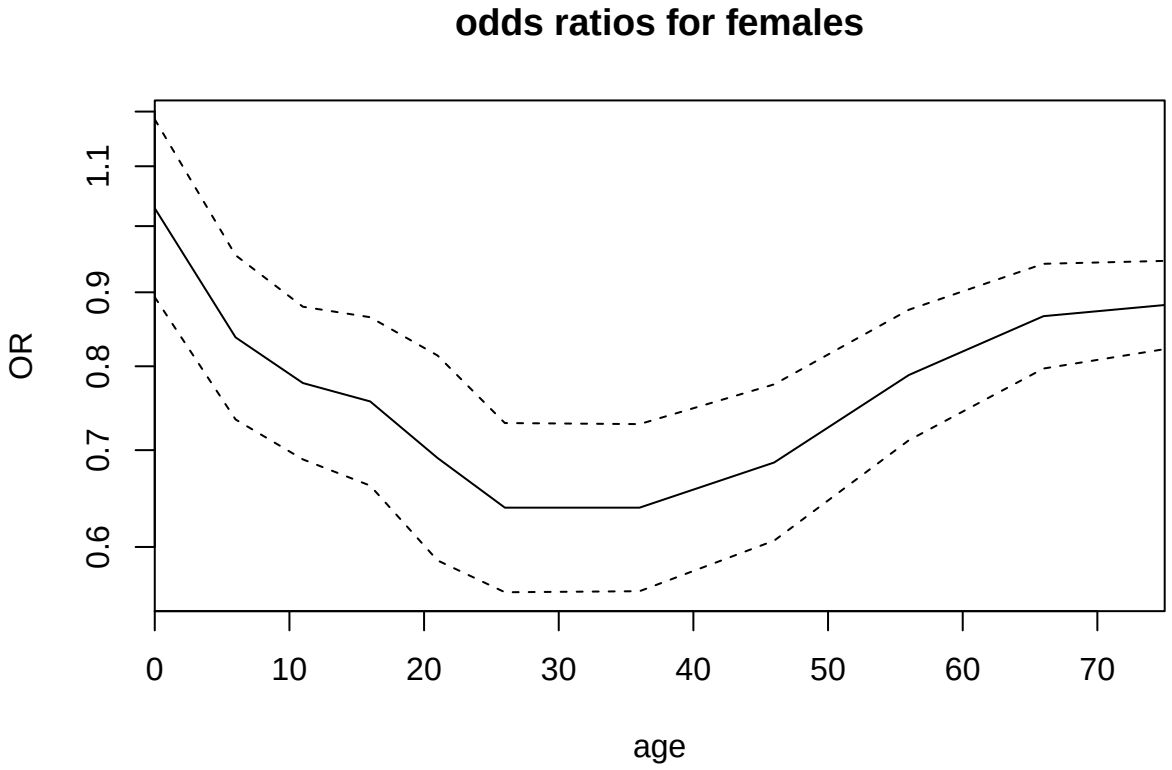
```
## Call:
## coxph(formula = Surv(rep(1, 445302L), y) ~ age + age:sex + strata(strata),
##     data = x, method = "exact")
##
##   n= 445302, number of events= 34299
##
##                          coef exp(coef) se(coef)       z Pr(>|z|)
## age0 - 5              0.13241   1.14157  0.04402   3.008 0.002629 **
## age6 - 10            -0.31966   0.72640  0.04086  -7.822 5.19e-15 ***
## age11 - 15           -0.38294   0.68185  0.04115  -9.305  < 2e-16 ***
```

```
## age16 - 20           -0.44321   0.64197  0.04045 -10.958  < 2e-16 ***
## age21 - 25           -0.26809   0.76484  0.04218  -6.355 2.08e-10 ***
## age36 - 45            0.41153   1.50913  0.03865  10.648  < 2e-16 ***
## age46 - 55            0.76823   2.15594  0.03898  19.709  < 2e-16 ***
## age56 - 65            1.21210   3.36052  0.03785  32.023  < 2e-16 ***
## age66 - 75            1.79725   6.03304  0.03635  49.447  < 2e-16 ***
## ageOver 75            2.39570  10.97590  0.03517  68.124  < 2e-16 ***
## age26 - 35:sexFemale -0.44821   0.63877  0.05228  -8.573  < 2e-16 ***
## age0 - 5:sexFemale    0.02842   1.02883  0.05495   0.517 0.604997
## age6 - 10:sexFemale  -0.17712   0.83768  0.05076  -3.490 0.000484 ***
## age11 - 15:sexFemale -0.24986   0.77891  0.04719  -5.295 1.19e-07 ***
## age16 - 20:sexFemale -0.27913   0.75644  0.05204  -5.364 8.15e-08 ***
## age21 - 25:sexFemale -0.36913   0.69134  0.06334  -5.828 5.61e-09 ***
## age36 - 45:sexFemale -0.44823   0.63876  0.05164  -8.679  < 2e-16 ***
## age46 - 55:sexFemale -0.37631   0.68639  0.04830  -7.792 6.60e-15 ***
## age56 - 65:sexFemale -0.23707   0.78894  0.04033  -5.878 4.16e-09 ***
## age66 - 75:sexFemale -0.14336   0.86644  0.03237  -4.429 9.47e-06 ***
## ageOver 75:sexFemale -0.12561   0.88196  0.02727  -4.606 4.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      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
##
## Concordance= 0.739  (se = 0.002 )
## Likelihood ratio test= 25992  on 21 df,   p=<2e-16
## Wald test            = 23970  on 21 df,   p=<2e-16
## Score (logrank) test = 30867  on 21 df,   p=<2e-16
```

Results

To begin with, we can take a look at the second column of our summary table, which are odds ratios (ORs), that is, odds of males or females involved in fatal accidents in different age divided by odds of males involved in fatal accidents aged between 26-35(reference group, OR=1). We do not discuss the group of females aged from 0 to because the estimate is not statistically significant. For children aged 6-10, and young adults 11-20(including two age groups), the ORs of males are slightly smaller than that of females, which is not consistent with the statement that women are safer as pedestrians than men as teenagers and young adults. However, older groups (over 20) of men are much more likely to be involved in fatal accidents than women. As seen in the table, except the age group of 21-25, ORs of males are all above 1, and the rates demonstrate a sharp accelerated growth along the ages, especially for men above 75 years old, their odds is nearly 11 times that of men age 26-35. On the contrary, ORs of women above 20 years old do not change much, and those rates are all below that of our reference group. Our findings from the table can be confirmed in the two graphs below. Hence, we are safe to conclude that though teenage females are slightly more involved in fatal accidents than males, males are much more likely to be in such danger in their older years. And on average, women tend to be safer as pedestrians than men. This might be due in part to women being more reluctant than men to walk outdoors late at night or in poor weather, and could also reflect men being on average more likely to engage in risky behaviour than women.




odds ratios for females
OR
0.6
0.7
0.8
0.9
1.1
0
10
20
30
40
50
60
70
age

Appendix

```
#Question1
smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/smoke.RData")
library(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,
                     #for each decade
                     event = forInla$Age_first_tried_cigt_smkg
                     <=forInla$Age) # left censoring: a data point is below a certain value but it is u
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2
# 1 means no censoring, 0 for right censoring and 2 for left censoring.
smokeResponse = inla.surv(forSurv$time, forSurv$event)

fitS2 = inla(smokeResponse ~ RuralUrban + Sex + Race +
               f(school, model = "iid",
                 hyper = list(prec = list(prior="pc.prec",
                                          param = c(0.2, 0.01))))+
               f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec",
                                          param =c(1.1, 0.01)))),
             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)

#exp(qnorm(c(0.025, 0.5, 0.975), mean=log(1), sd=2/3))

tab1 <- rbind(fitS2$summary.fixed[, c("mean",
                                      "0.025quant","0.975quant")],
              Pmisc::priorPostSd(fitS2)$summary[,c("mean","0.025quant","0.975quant")])
knitr::kable(tab1)

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
```

```
xSeq = seq(5,100,len=1000)
kappa = fitS2$summary.hyper['alpha', 'mode']
lambda = exp(-fitS2$summary.fixed['(Intercept)', 'mode'])
#plot(xSeq,dweibull(xSeq/100, shape = kappa, scale = lambda)/100,
# col='blue')

hazEst = survfit(Surv(time, one) ~ 1, data=forSurv)
plot(hazEst, fun='cumhaz')
plot(xSeq, (xSeq / (100*lambda))^kappa, col='blue', type='l', log='y',
     ylim=c(0.001, 10), xlim = c(20,100), xlab='years', ylab = 'cum haz')

#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)

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

summary<-summary(glm(y ~ sex + age + Light_Conditions +
                       Weather_Conditions,
                     data = x,
                     family=binomial))$coef[1:4,]
library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
summary(theClogit)
theCoef = rbind(as.data.frame(summary(theClogit)$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",
  xlab="age", ylab="OR", main="odds ratios for males")

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="OR", main="odds ratios for females")
```