Refer to survival slide Pg 60

Homework 4, Survival

STA442 Methods of Applied Statistics

Due 3 Dec 2019

1 Smoking (20 marks)

This task concerns the 2014 American National Youth Tobacco Survey. On the pbrown.ca/teaching/appliedstats/data page there is an R version of the 2014 dataset `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`.

The age at which children first try cigarette smoking is known to be earlier for males than females, earlier in rural areas than urban areas, and to vary by ethnicity. It is likely that significant variation amongst the US states exists, and that there is variation from one school to the next.

The hypotheses to be investigated are:

1. 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.
2. 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.

The collaborating scientists have provided the following prior information

- The variability in the rate of smoking initiation between states substantial, with some states having double or triple the rate of smoking update compared other states for comparable individuals. If U_i is a random effect for state i, we might see $\exp(U_i) = 2$ or 3 but unlikely to see at 10. e.g. 3 referring to the rate tripled
- Within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' schools or $\exp(V_{ij}) = 1.5$ for a school-level random effect is about the largest we'd see.
- A flat hazard function is expected, so the prior on the Weibull shape parameter should allow for a 1 but it is not believed that shape parameter is 4 or 5.

Write a short consulting report addressing these hypotheses. Some additional notes:

- Show graphs of prior and posterior densities of model parameters related to the research questions.
- Interpret your model parameters in the context of the smoking problem, transforming model parameters to a more 'natural' scale as necessary.
- It is important to state precisely what your prior distributions are (i.e. a Gamma(0.4, 3.1) distribution for the log of the intercept parameter), but also show how these distributions are consistent with the prior assumptions by showing quantiles or means or tail probabilities.
- You're given three confounders (sex, rural/urban, ethnicity), it's up to you if you'd like to include interactions.

- You might want to fit more than one model, either as exploratory work or sensitivity assessments, but you should use a single 'best' model to answer the research questions. Fitting two models and selecting

one of them with a fairly *ad hoc* explanation is fine, comparing 10 models without some sort of formal assessment (a topic we haven't covered) wouldn't be.

```
smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/smoke.RData")
```

```
Loading required namespace: 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, 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$event)
fitS2 = inla(smokeResponse ~ RuralUrban + Sex * Race +
  f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",
    param = c(0.5, 0.05)))) + f(state, model = "iid",
  hyper = list(prec = list(prior = "pc.prec", param = c(0.5,
    0.05)))), control.family = list(variant = 1,
  hyper = list(alpha = list(prior = "normal", param = c(log(4),
    (2/3)^(-2))))), control.mode = list(theta = c(8,
  2, 5), restart = TRUE), data = forInla, family = "weibullsurv",
  verbose = TRUE)
rbind(fitS2$summary.fixed[, c("mean", "0.025quant",
  "0.975quant")], Pmisc::priorPostSd(fitS2)$summary[,
  c("mean", "0.025quant", "0.975quant")])
```

0 non-smoking & 1 smoking, ,2 means start to smoke at or before 8 years old

	mean	0.025quant	0.975quant
(Intercept)	-0.618123774	-0.673217580	-0.562381805
RuralUrbanRural	0.114219840	0.054982509	0.173126077
SexF	-0.050079551	-0.078482659	-0.021834950
Raceblack	-0.048030138	-0.090878915	-0.005850117
Racehispanic	0.025707830	-0.008877249	0.060088014
Raceasian	-0.194755194	-0.286897781	-0.108234511
Racenative	0.110090096	0.004857814	0.207960713
Racepacific	0.175543344	0.008688396	0.324061510
SexF:Raceblack	-0.016889846	-0.073916811	0.039994596
SexF:Racehispanic	0.016228302	-0.029712823	0.062147624
SexF:Raceasian	0.005526784	-0.121718791	0.131914982
SexF:Racenative	-0.043646334	-0.200293351	0.109720191
SexF:Racepacific	-0.169554281	-0.499950675	0.123079122
SD for school	0.150218268	0.125831434	0.176461812
SD for state	0.056808074	0.024521813	0.101231683

2 Death on the roads (20 marks)

The dataset below is a subset of the data from www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents, with all of the road traffic accidents in the UK from 1979 to 2015. The data below

consist of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestrians with moderate injuries have been removed).

```
dim(pedestrians)
```

```
[1] 1159453       6
```

```
pedestrians[1:3, ]
```

```
                  time     age  sex Casualty_Severity      Light_Conditions
54 1979-01-01 22:40:00 26 - 35 Male            Slight Darkness - lights lit
65 1979-01-02 10:40:00 26 - 35 Male            Slight               Daylight
79 1979-01-02 14:25:00 46 - 55 Male            Slight               Daylight
      Weather_Conditions
54 Snowing no high winds
65 Raining no high winds
79 Raining no high winds
```

```
table(pedestrians$Casualty_Severity, pedestrians$sex)
```

```
           Male Female
  Slight 637977 481832
  Fatal   24432  15212
```

Seems like male are more likely to have fatal injury

```
range(pedestrians$time)
```

```
[1] "1979-01-01 01:00:00 EST" "2015-12-31 23:35:00 EST"
```

Notice that men are involved in accidents more than women, and the proportion of accidents which are fatal is higher for men than for women. 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.

Write a short report assessing 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. Treat fatal accidents as cases and slight injuries as controls, and use a conditional logistic regression to adjust for time of day, lighting conditions, and weather. Make your report self-contained so it can be read by someone who has not seen this homework sheet. Some (but not all) of the code below could be helpful. Explain clearly how you have stratified the data (you could use a different stratification than I did if you wish).

Fit a glm

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

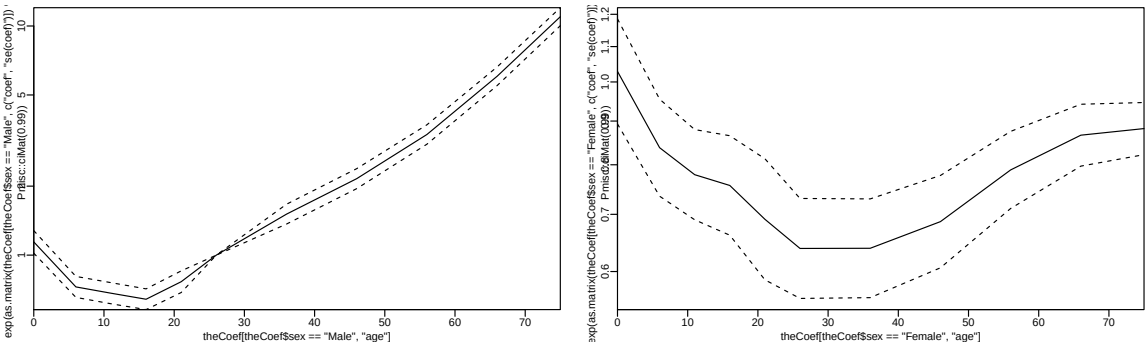
```
               Estimate Std. Error     z value      Pr(>|z|)
(Intercept) -3.2507678 0.02370768 -137.118754  0.000000e+00
sexFemale   -0.2988120 0.01245826  -23.985048 3.983187e-127
age0 - 5     0.1124659 0.03448000    3.261772  1.107180e-03
age6 - 10   -0.4369786 0.03241323  -13.481489  2.010085e-41
```

fit a conditional logistic model

```
library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata),
  data = x)
```

Some results

download data



(a) males (b) females

Figure 1: some results without adequate explanation

```
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)
```

Created strata for time, light conditions and weather conditions

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, ]
```

Get rid of strata with only one data

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