University of Toronto

Smoking & Fatality on roads Analysis

YiFei Gu

1 Smoking

The report below aim to provide insights on the mean age children first try cigarettes, two hypothesis are to be examined:
1. Geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools.
2. 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, ethnicity) and random effects (school and state) are identical.

1.1 Methods

In the analysis, Weibull distribution was used, since the age of children first try cigarettes is modeled, note that k is the Weibull shape parameter that follows normal distribution with its own hyperparameters:

$$Z_{ijk}|Y_{ijk}, A_{ijk}, U_i, V_{ij} = \min(Y_{ij}, A_{ijk})$$

$$E_{ijk}|Y_{ijk}, A_{ijk}, U_i, V_{ij} = I(Y_{ijk} < A_{ijk})$$

$$Y_{ijk}|U_i, V_{ij} \sim \text{Weibull}(\rho_{ijk}, k)$$

Note that $X_{ijk}\beta$ is the gender, ethnicity and where they are from, either rural or urban:

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

$$\nu_{ijk} = X_{ijk}\beta + U_i + V_{ij}$$

And U_i is the state random effect, V_{ij} is the school random effect:

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

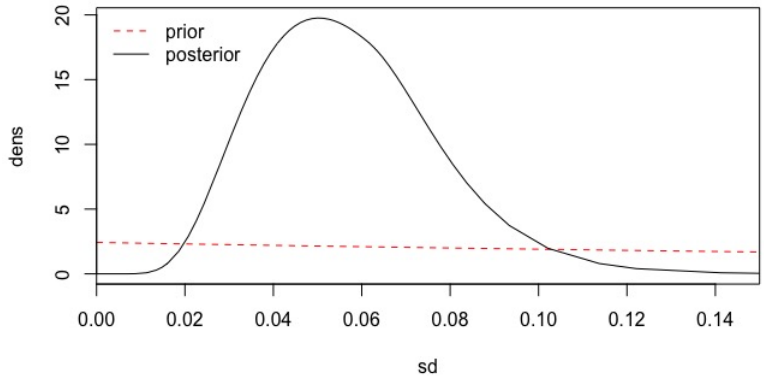
$$V_i \sim N(0, \sigma_V^2)$$

1.2 Results

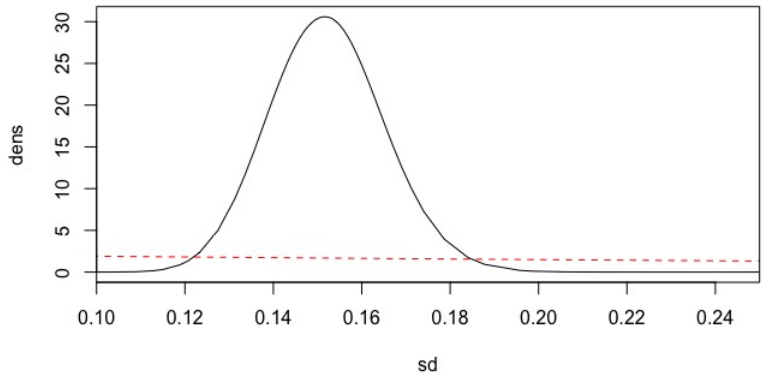
The table shown below indicates the variation amongst schools in the mean age children first try cigarettes and the variation amongst states in the mean age children first try cigarettes. Note that since the variation acroos schools is greater than the variation across the states, the first hypothesis is rejected.

	mean	0.025quant	0.975quant
SD for school	0.152106071	0.127475093	0.17863
SD for state	0.057060029	0.024364908	0.102150

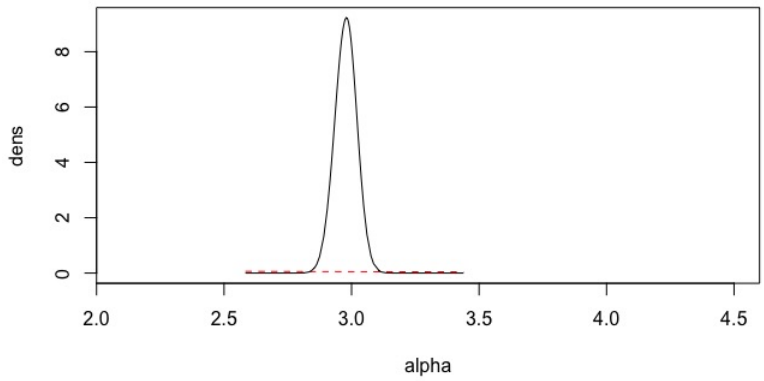
Here the penalized complexity prior for school used is 1.9, as $P(\sigma > 1.9) = 0.01$ sufficed for the assumption, given that $\exp(V_{ij}) = 1.5$. The graph illustrated the prior density of standard deviations for the school:



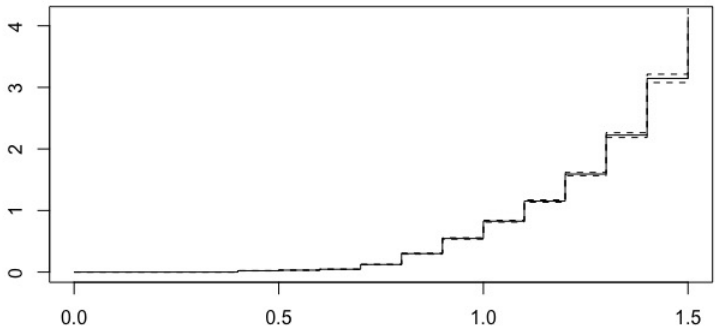
The penalized complexity prior for states is 1, with $P(\sigma > 1) = 0.01$, as illustrated in the below graph, the prior density for the standard deviations for the states:



The posterior density of Weibull function is shown in the below graph:



As shown above, note that the mean for the density of Weibull function is around 3.0, suggesting it will produce an increasing hazard function instead of a linear slope.



Here given cumulative hazard function is non-linear, implying that the mean probability of two non-smoking children try smoking cigarettes within next month depends on their age, provided their confounders are identical, which rejects the second hypothesis.

1.3 Conclusion

In conclusion, the first hypothesis, that the geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools, does not meet the analysis above. In fact, the variation amongst schools are greater than geographic variations in US. For the second hypothesis, the age of children does affect the probability of them trying cigarettes in the future, provided the known confounders(sex, rural/urban, ethnicity) and random effects (school and state) are identical. Both hypothesis are rejected as a result.

2 Death on the roads

Below is 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.

2.1 Methods

The model used below is conditional logistic model:

$$\text{logit}[\Pr(Y_{ij} = 1)] = \alpha_i + X_{ij}\beta$$

$$\text{logit}[\Pr(Y_{ij} = 1)|Z_{ij} = 1] = \alpha_i^* + X_{ij}\beta$$

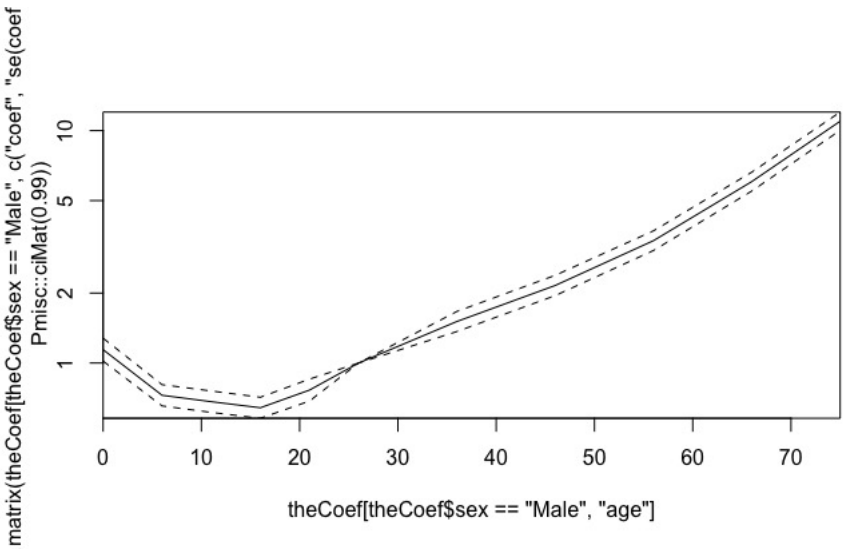
With

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

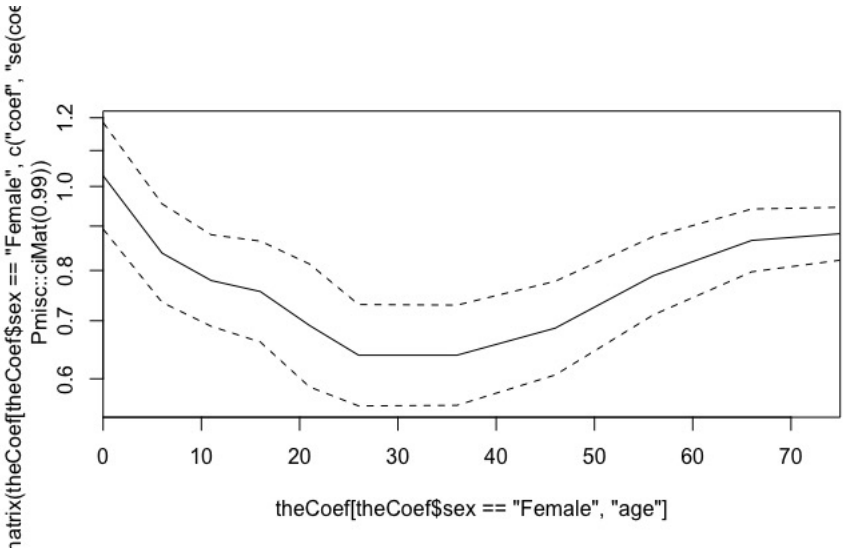
In the above model, i represents the strata, including light conditions and weather conditions. j represents each individuals. Y is the indicator for fatal injuries, and X_{ij} represents the interation of sex and age.

2.2 Results

The below figure is the odds ration of male involved in incidents for different ages, the credible interval converges at age 26, as age 26 is the reference group. As male gets older, the odds of involving fatal injuries increase as shown.



The below figure shows the odds ratio of involving in injuries for female. In comparison to male, the odds of getting injured for different ages has less variation for female.



	coef	exp(coef)	se(coef)	z	Pr(>|z|)	sex	age
age0 - 5:sexFemale	0.03	1.03	0.05	0.52	0.60	Female	0.00
age6 - 10:sexFemale	-0.18	0.84	0.05	-3.49	0.00	Female	6.00
age11 - 15:sexFemale	-0.25	0.78	0.05	-5.30	0.00	Female	11.00
age16 - 20:sexFemale	-0.28	0.76	0.05	-5.36	0.00	Female	16.00
age21 - 25:sexFemale	-0.37	0.69	0.06	-5.83	0.00	Female	21.00
age26 - 35:sexFemale	-0.45	0.64	0.05	-8.57	0.00	Female	26.00
age36 - 45:sexFemale	-0.45	0.64	0.05	-8.68	0.00	Female	36.00
age46 - 55:sexFemale	-0.38	0.69	0.05	-7.79	0.00	Female	46.00
age56 - 65:sexFemale	-0.24	0.79	0.04	-5.88	0.00	Female	56.00
age66 - 75:sexFemale	-0.14	0.87	0.03	-4.43	0.00	Female	66.00
ageOver 75:sexFemale	-0.13	0.88	0.03	-4.61	0.00	Female	75.00
age0 - 5	0.13	1.14	0.04	3.01	0.00	Male	0.00
age6 - 10	-0.32	0.73	0.04	-7.82	0.00	Male	6.00
age11 - 15	-0.38	0.68	0.04	-9.31	0.00	Male	11.00
age16 - 20	-0.44	0.64	0.04	-10.96	0.00	Male	16.00
age21 - 25	-0.27	0.76	0.04	-6.36	0.00	Male	21.00
age 26 - 35	0.00	1.00	0.00			Male	26.00
age36 - 45	0.41	1.51	0.04	10.65	0.00	Male	36.00
age46 - 55	0.77	2.16	0.04	19.71	0.00	Male	46.00
age56 - 65	1.21	3.36	0.04	32.02	0.00	Male	56.00
age66 - 75	1.80	6.03	0.04	49.45	0.00	Male	66.00
ageOver 75	2.40	10.98	0.04	68.12	0.00	Male	75.00

2.3 Conclusion

Note that the early adulthood and the age of teenagers, as pedestrian for both sexes are shown in the above table, note that the exp(coef) of female and male, during the age of 16 - 20, and 21-25, are both low, indicating that there are not significant evidence that women tend to be safer as pedestrians than men. However, the odds ratio across different ages between female and male tends to be lesser than 1, implying that women tend to be safer as pedestrians than men especially when they are older, with age above 35.

3 Appendix

```
# Q1
> 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(1.9, 0.01)))) +
+                  f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec",
↪  param = c(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)
> rbind(fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")],
↪  Pmisc::priorPostSd(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")])
                         mean   0.025quant   0.975quant
(Intercept)       -0.622825696 -0.678441514 -0.566416951
RuralUrbanRural    0.115378078  0.055459970  0.174945234
SexF              -0.050456356 -0.079067640 -0.022005179
Raceblack         -0.048305723 -0.091472523 -0.005811282
Racehispanic       0.025838645 -0.009003629  0.060474376
Raceasian         -0.195950359 -0.288777156 -0.108790960
Racenative         0.110524127  0.004513611  0.209120502
Racepacific        0.176512622  0.008435974  0.326118074
SexF:Raceblack    -0.016982715 -0.074427974  0.040318551
SexF:Racehispanic  0.016351868 -0.029925973  0.062607614
SexF:Raceasian     0.005501409 -0.122680750  0.132819125
SexF:Racenative   -0.043977182 -0.201774721  0.110514373
SexF:Racepacific  -0.170591811 -0.503412809  0.124183855
SD for school      0.152106071  0.127475093  0.178636039
SD for state       0.057060029  0.024364908  0.102150936
> forSurv$ones=1
> xSeq=seq(5, 100, len = 1000)
> kappa = fitS2$summary.hyper['alpha','mode']
> lambda = exp(-fitS2$summary.fixed['(Intercept)', 'mode'])
> plot(xSeq, (xSeq / (100*lambda))^kappa, col = 'blue', type = 'l', log='y',
↪  ylim=c(0.001, 10), xlim = c(20, 100), xlab= 'years', ylab = 'cum haz')
> hazEst = survfit(Surv(time, ones) ~ 1, data = forSurv)
> plot(hazEst, fun = 'cumhaz')
> fitS2$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$legend)
# Q2
> pedestrainFile =
↪  Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
> pedestrians = readRDS(pedestrainFile)
> pedestrians = pedestrians[!is.na(pedestrians$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, ]
```

```
> 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")
```