Survival Analysis

Smoking

Introduction

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.

Concerning the 2014 American National Youth Tobacco Survey and the dataset on the pbrown.ca/teaching/appliedstats/data, I intended to investigate the following hypotheses:

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.

Method

We need use survival analysis for this study since we intended to estimate the expected duration of time for children start smoking for the first time. And according to the shape of the data, I decided to use Weibull distribution to model it.

$$
\begin{aligned}
Y_{ijk} &\sim \text{Weibull}(\rho_{ijk}, \kappa) \\
\rho_{ijk} &= \exp(-\eta_{ijk}) \\
\eta_{ijk} &= X_{ijk}\beta + U_i + V_{ij} \\
U_i &\sim N(0, \sigma_U^2) \\
V_{ij} &\sim N(0, \sigma_V^2)
\end{aligned}
$$

where:

- $X_{ij}\beta$ is the subjects gender, ethnicity, whether they are from a rural or urban school
- U_i is the school random effect.
- V_{ij} is the state random effect.
- κ is the Weibull shape parameter.

From the prior information: $exp(U_i) = 2or3$ but unlikely to see at 10, I know the range of $exp(U_i)$ is <=7, so σ_U <= 1.7, and σ_V is less than the half of the σ_U
Therefore, I set the prior distributions as follows:

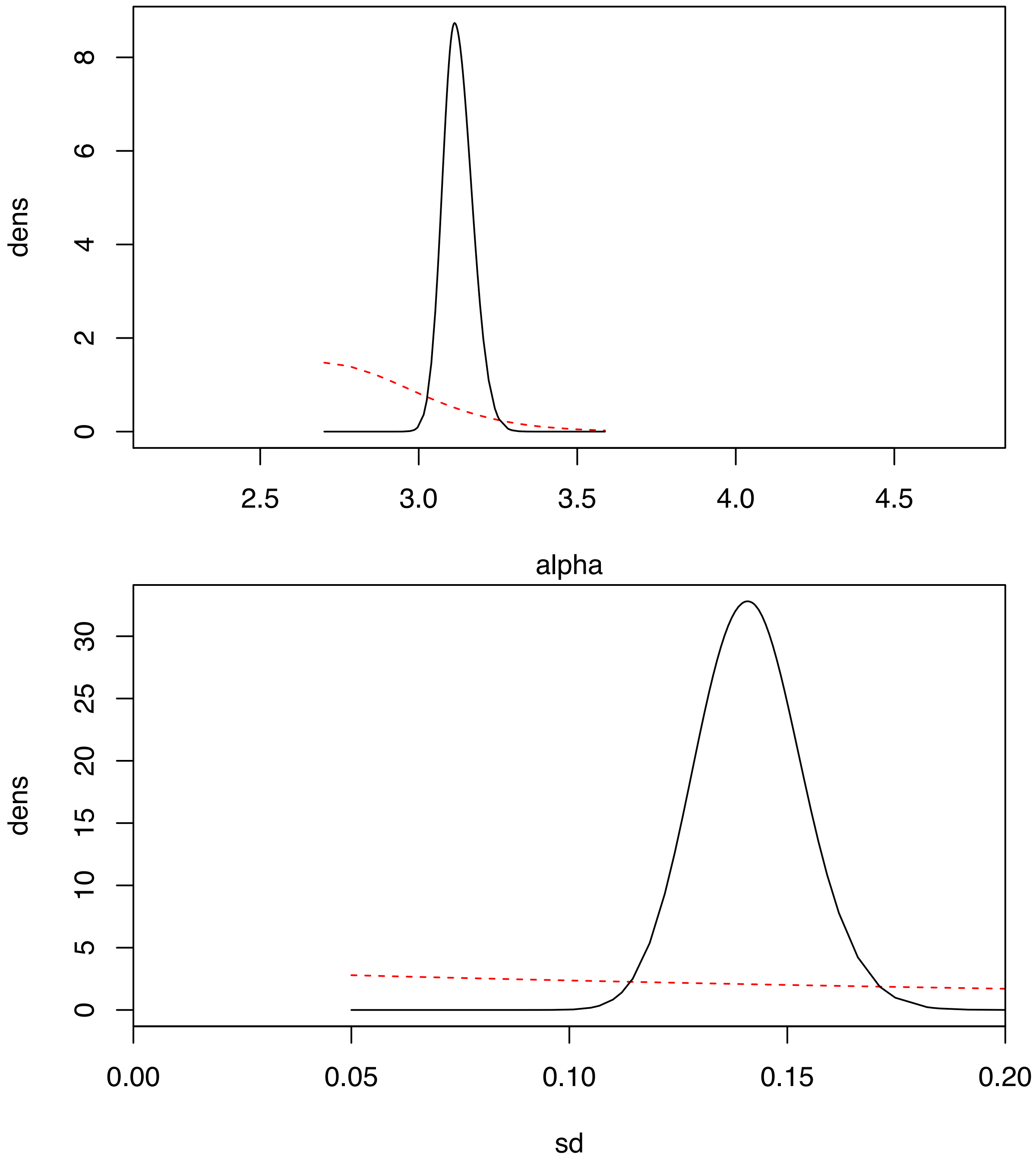
$$
\begin{aligned}
\kappa &\sim N(1,\ \ 0.1) \\
P(\sigma_U > 1.4) &= 0.01 \\
P(\sigma_V > 0.7) &= 0.01
\end{aligned}
$$

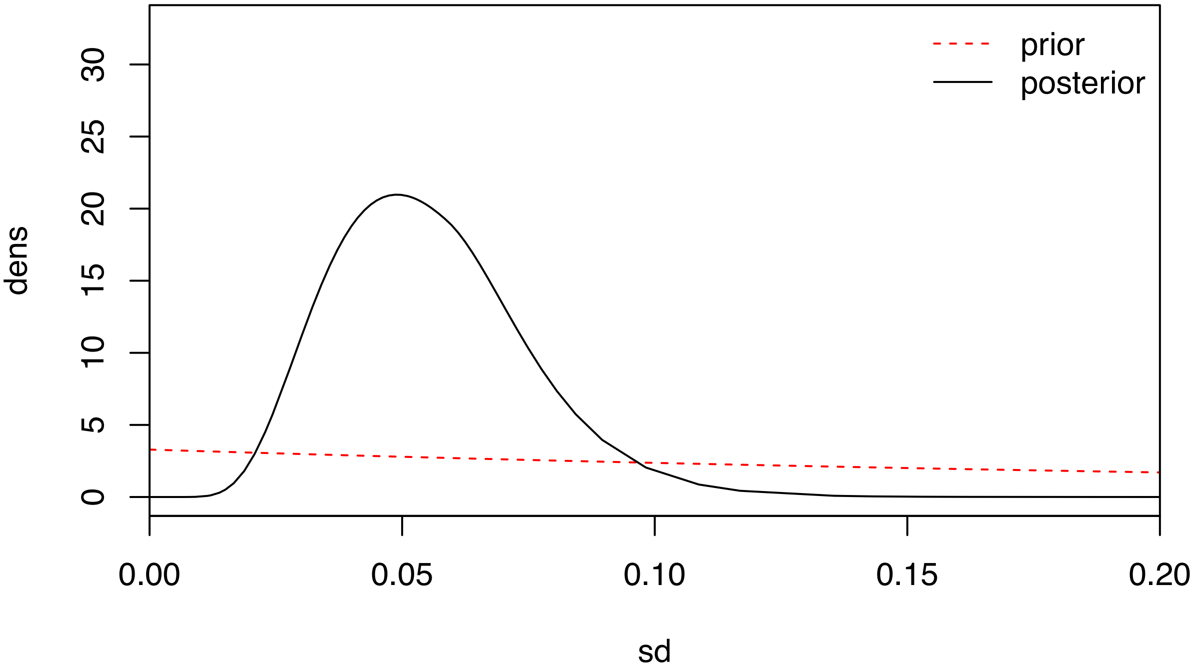
Table 1: Posterior estimates

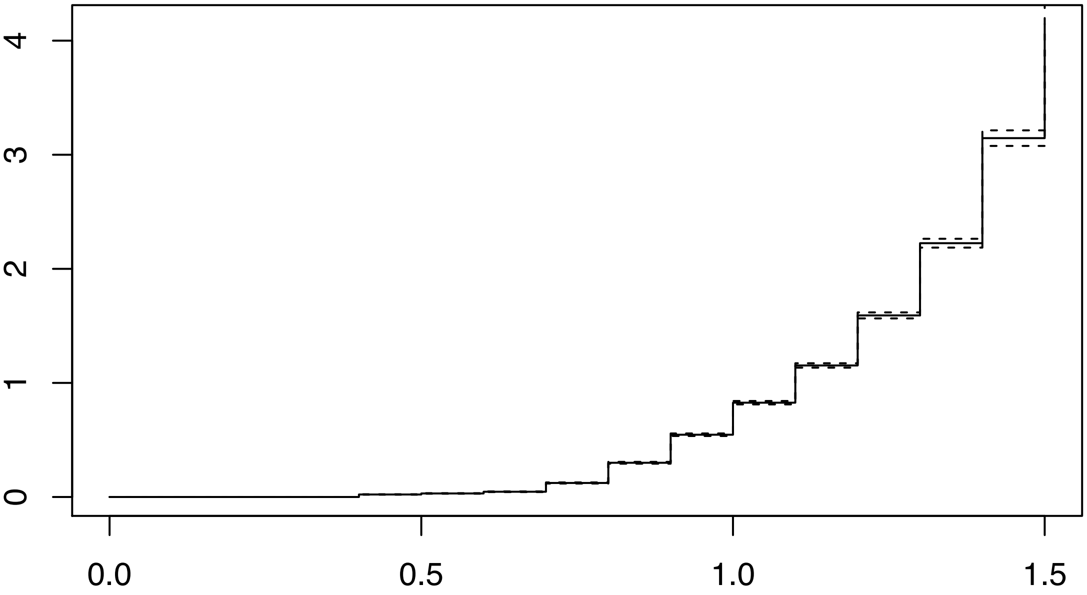
SD of School	SD of state
0.1421	0.0591

Result

From the following 4 plots, it's clear that:
1, Geographic variation between states in the mean age children first try cigarettes is less than variation amongst schools. So tobacco control programs should actually target finding particular schools where smoking is a problem.
2, According to the plot of prior distribution and the plot of hazard fuction, we can conclude the first cigarette smoking does not have a flat hazard function. And the non-smoking children with higher age have the higher probability of trying cigarettes within the next month provided the known confounders and random effects are identical.


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

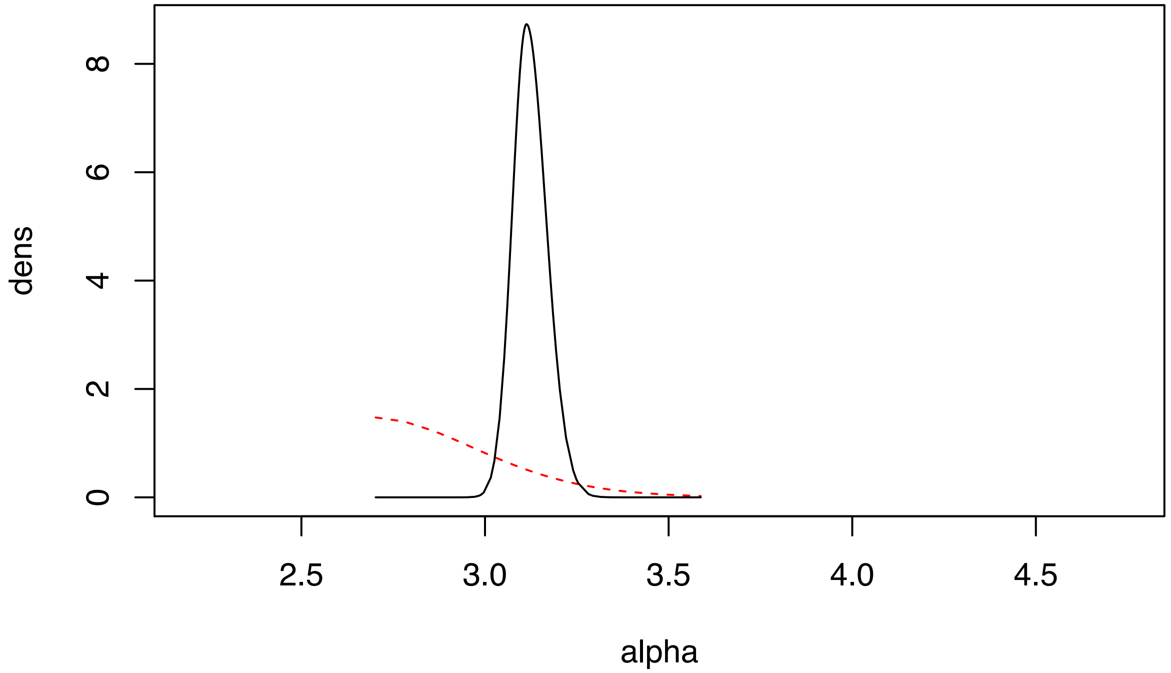

prior
posterior
dens
0
5
10
15
20
25
30
0.00
0.05
0.10
0.15
0.20
sd

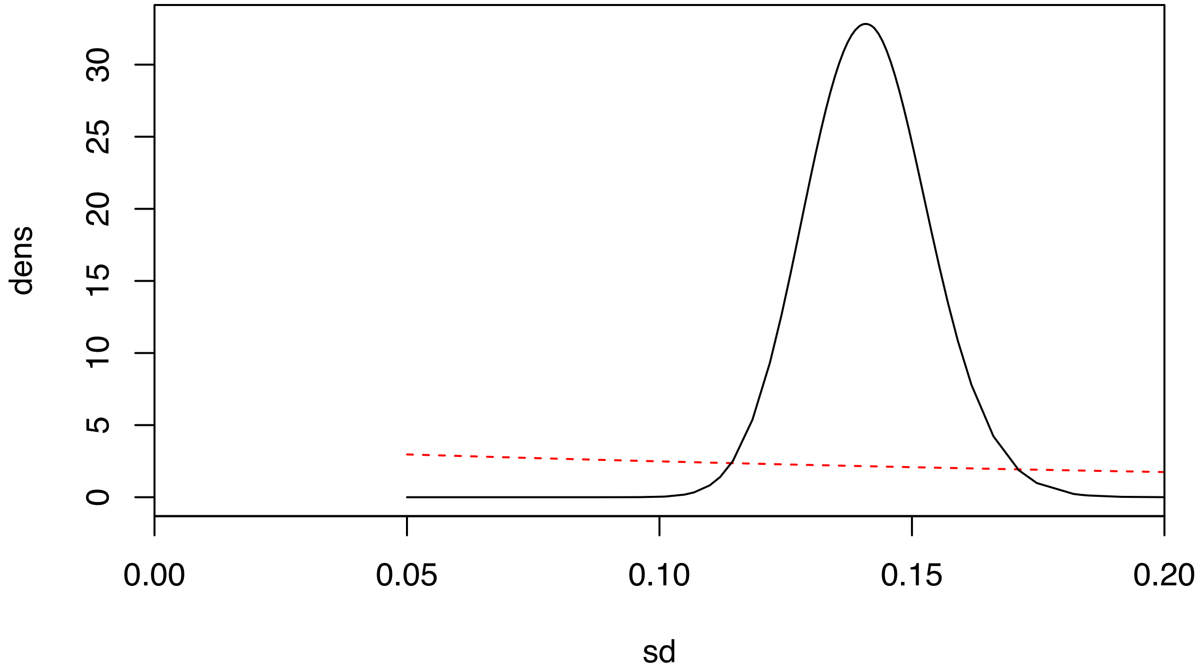

0
1
2
3
4
0.0
0.5
1.0
1.5

Appendix

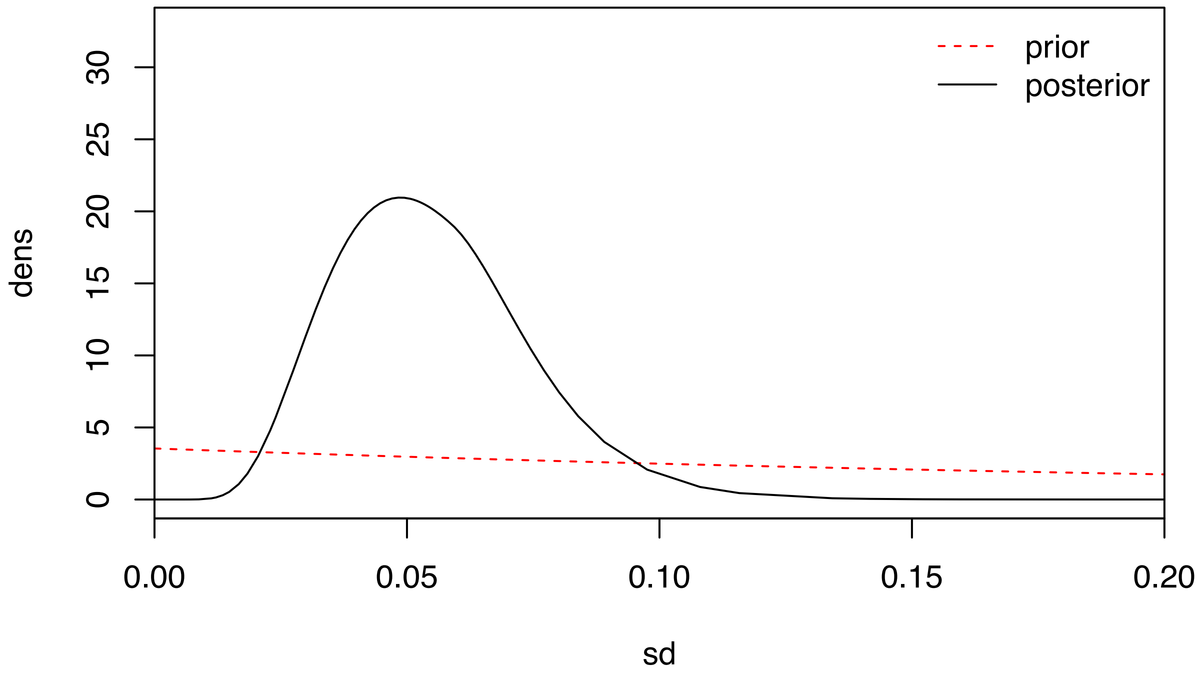
```
# Smoking

### Model 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)

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)

inla_formula = smokeResponse ~ Race + Sex + RuralUrban +
  f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(1.4,0.01)))) +
  f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(0.7,0.01))))


seco_model = inla(inla_formula,
                  control.family = list(variant = 1, hyper = list(alpha = list(prior = "normal", param =
                  control.mode = list(theta = c(8, 2, 5), restart = TRUE),
                  data = forInla, family = "weibullsurv", verbose = TRUE,
                  control.compute=list(config = TRUE))


### model para

post.dat=round(exp( seco_model$mode$theta),2)
table1 <- 1/ sqrt(post.dat[2:3])
table1<-round(table1,4)
table1<-matrix(table1,nrow=1)

colnames(table1)<-c("SD of School", "SD of state")

knitr::kable(table1, caption="Posterior estimates")
```

```
 seco_model$priorPost = Pmisc::priorPost( seco_model)
for (Dparam in  seco_model$priorPost$parameters) {
do.call(matplot   seco_model$priorPost[[Dparam]]$matplot)
}
```

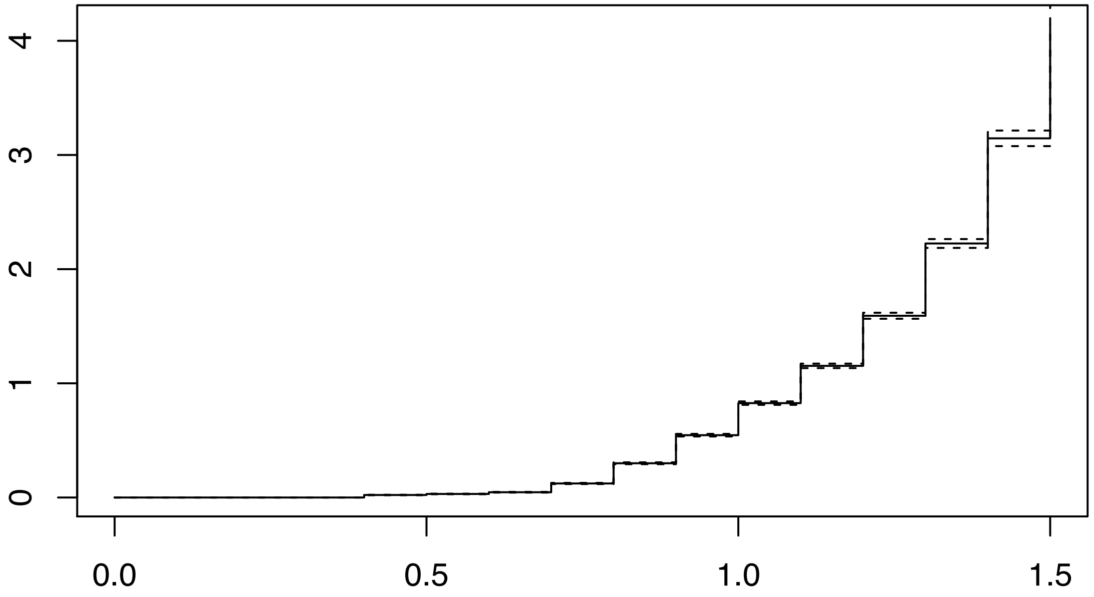




```
do.call(legend,  seco_model$priorPost$legend)
```



```r
forSurv$one = 1
hazEst = survfit(Surv(time, one) ~ 1, data=forSurv)
plot(hazEst, fun='cumhaz')
```


0
1
2
3
4
0.0
0.5
1.0
1.5