STA442 Homework4

Hanyu Zhou

1003110858

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:

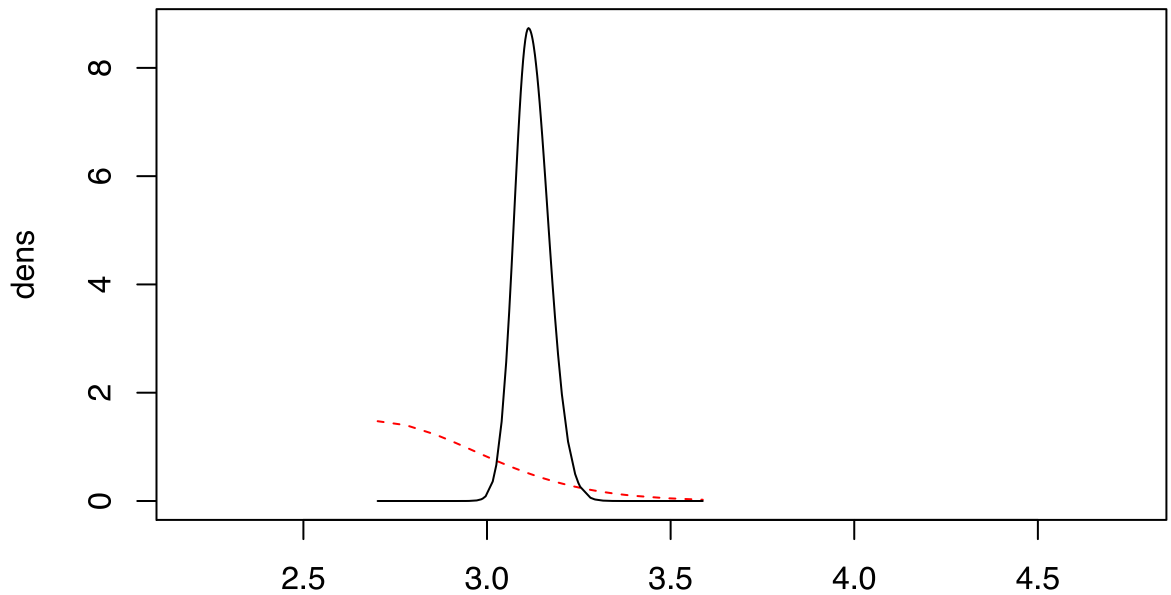
$$
\begin{aligned}
\kappa &\sim N(1,\ \ 0.1) \\
P(\sigma_U > 1.4) &= 0.01 \\
P(\sigma_V > 0.7) &= 0.01
\end{aligned}
$$

Table 1: Posterior estimates

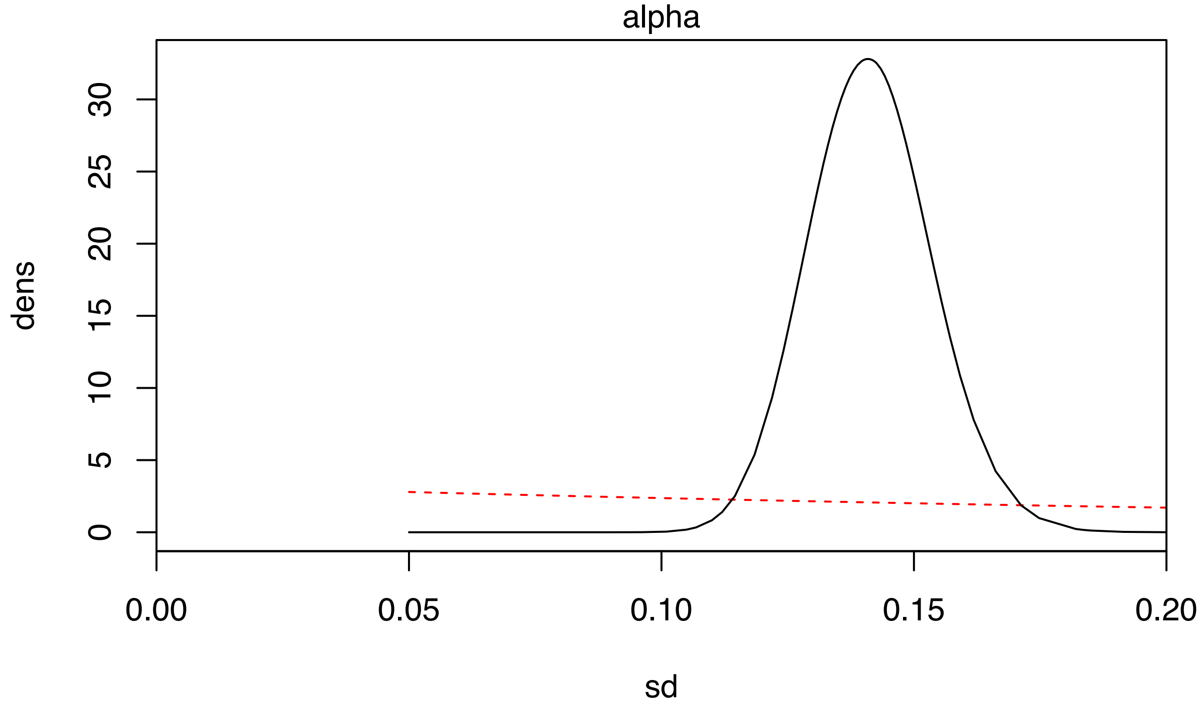
SD of School	SD of state
0.1421	0.0591

Result

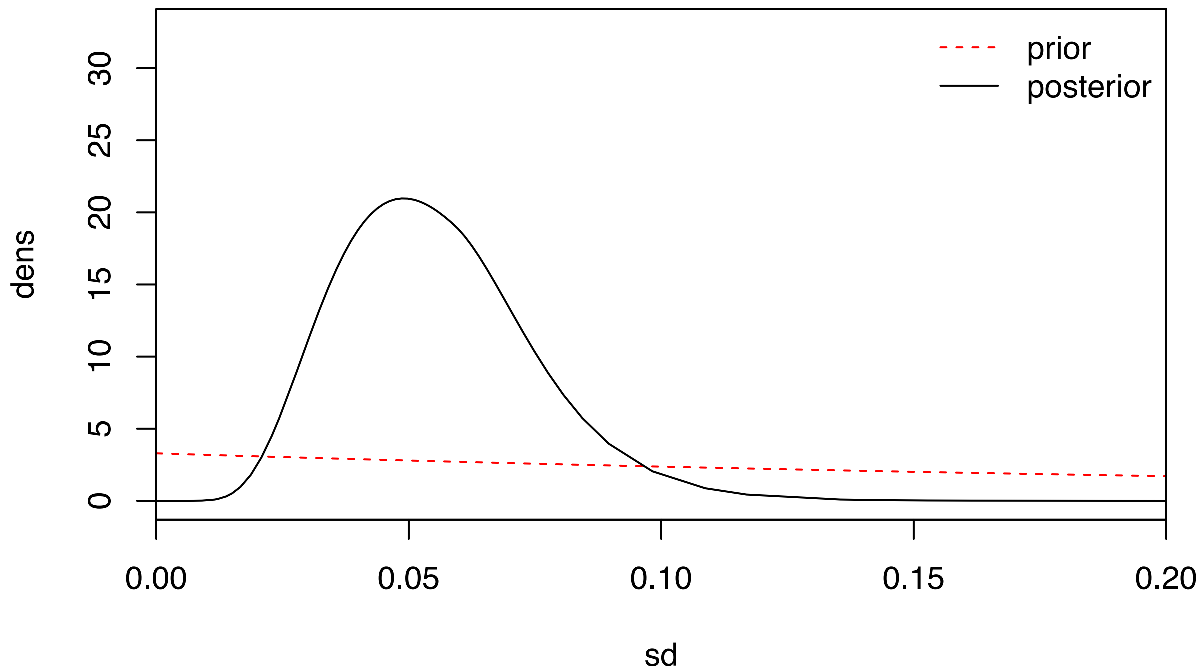
From the following 4 plots, it's clear that:
1, Geographic variation between states in the mean age children first try cigarettes is less than variation amongst schools. So tobacco control programs should actually target finding particular schools where smoking is a problem.
2, According to the plot 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.

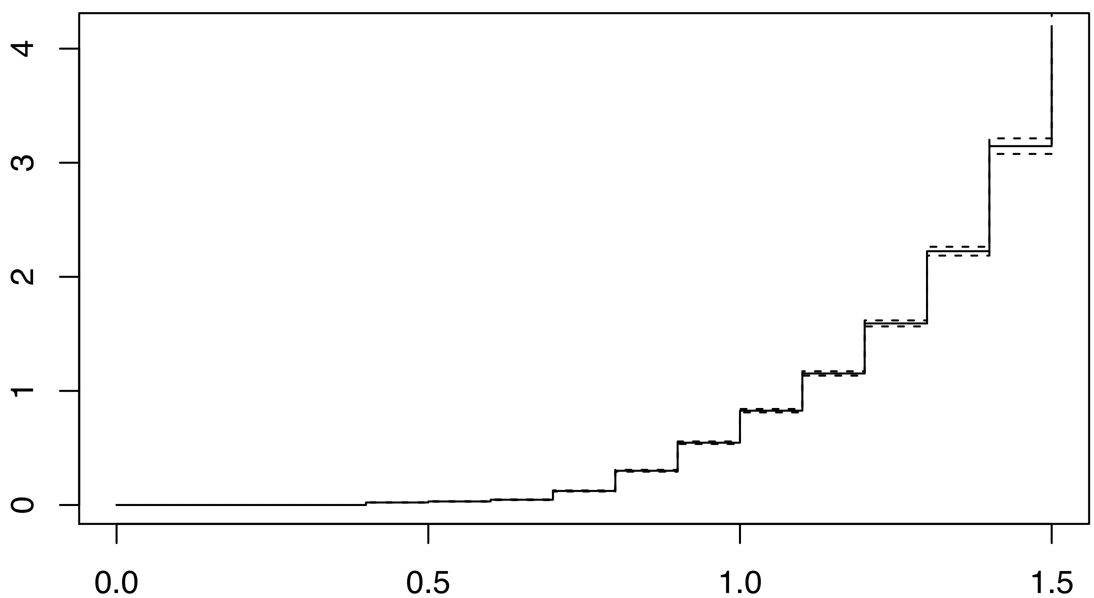


dens
0
2
4
6
8
2.5
3.0
3.5
4.0
4.5
alpha



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

Death on the roads

Introduction

I used the data from www.gov.uk/government/statistical-data-sets/ras30-reportedcasualties-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).

According to the hypothesis, we investigated that whether 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.

Method

First I fit a logistic regression model to examine the importance of chossing the lighting, weather and time conditions as strata.

Then I used conditional logistic regression to model the data.
We want

$$pr(Y_i = 1|X_i) = \lambda_i$$

$$\log[\lambda_i/(1-\lambda_i)] = \beta_0 + \sum_{p=1}^{P} X_{ip}\beta_p$$

We have

$$pr(Y_i = 1|X_i, Z_i = 1) = \lambda_i^*$$

$$\log[\lambda_i^*/(1-\lambda_i^*)] = \beta_0^* + \sum_{p=1}^{P} X_{ip}\beta_p^*$$

By the theorem we get:

$$\beta_p^* = \beta_0 + log[pr(Z_i = 1|Y_i = 1)/pr(Z_i = 1|Y_i = 0)] \text{ if } \ p = 0$$

$$\beta_p^* = \beta_p \text{ if } \ p \neq 0$$

where:

- $X_{ip}\beta$ is the subjects' gender, and age.
- Y_i is the status of casualty.
- Z_i is the strata of lightness, weather and time conditions.

Result

Table 2: The coeficients of conditional logistic regression

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	sex	age
age0 - 5:sexFemale	0.0284229	1.0288306	0.0549522	0.5172285	0.6049967	Female	0
age6 - 10:sexFemale	-0.1771162	0.8376825	0.0507565	-3.4895264	0.0004839	Female	6
age11 - 15:sexFemale	-0.2498614	0.7789087	0.0471857	-5.2952744	0.0000001	Female	11
age16 - 20:sexFemale	-0.2791322	0.7564399	0.0520402	-5.3637766	0.0000001	Female	16
age21 - 25:sexFemale	-0.3691252	0.6913389	0.0633358	-5.8280613	0.0000000	Female	21
age26 - 35:sexFemale	-0.4482120	0.6387693	0.0522815	-8.5730476	0.0000000	Female	26
age36 - 45:sexFemale	-0.4482308	0.6387573	0.0516433	-8.6793515	0.0000000	Female	36
age46 - 55:sexFemale	-0.3763107	0.6863891	0.0482955	-7.7918406	0.0000000	Female	46
age56 - 65:sexFemale	-0.2370677	0.7889379	0.0403324	-5.8778460	0.0000000	Female	56
age66 - 75:sexFemale	-0.1433569	0.8664448	0.0323676	-4.4290313	0.0000095	Female	66
ageOver 75:sexFemale	-0.1256106	0.8819582	0.0272702	-4.6061492	0.0000041	Female	75
age0 - 5	0.1324083	1.1415744	0.0440170	3.0081179	0.0026287	Male	0
age6 - 10	-0.3196593	0.7263965	0.0408650	-7.8223298	0.0000000	Male	6
age11 - 15	-0.3829384	0.6818549	0.0411527	-9.3053109	0.0000000	Male	11
age16 - 20	-0.4432109	0.6419718	0.0404473	-10.9577480	0.0000000	Male	16
age21 - 25	-0.2680862	0.7648419	0.0421849	-6.3550264	0.0000000	Male	21
age 26 - 35	0.0000000	1.0000000	0.0000000	NA	NA	Male	26
age36 - 45	0.4115311	1.5091267	0.0386489	10.6479477	0.0000000	Male	36
age46 - 55	0.7682289	2.1559445	0.0389790	19.7087971	0.0000000	Male	46
age56 - 65	1.2120970	3.3605244	0.0378511	32.0227837	0.0000000	Male	56
age66 - 75	1.7972504	6.0330360	0.0363472	49.4467189	0.0000000	Male	66
ageOver 75	2.3957024	10.9759044	0.0351665	68.1244757	0.0000000	Male	75

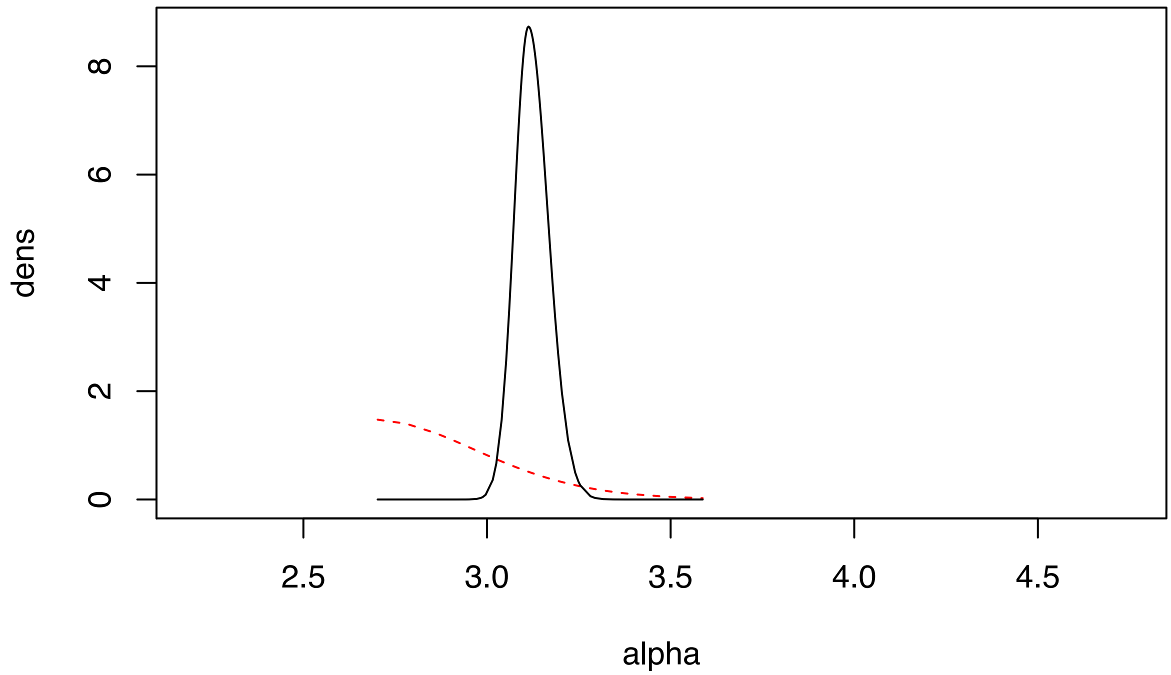
```
##                                              Estimate Std. Error   z value
## Light_ConditionsDarkness - lights lit       0.8544535 0.01366099 62.546952
## Light_ConditionsDarkness - lights unlit     2.4841936 0.17373844 14.298469
## Light_ConditionsDarkness - no lighting      2.7540798 0.03870802 71.150116
## Light_ConditionsDarkness - lighting unknown 2.4524585 0.17511200 14.005085
## Weather_ConditionsRaining no high winds     0.6907296 0.02127153 32.472019
## Weather_ConditionsSnowing no high winds     1.6384235 0.18251005  8.977168
## Weather_ConditionsFine + high winds         1.6771798 0.06946173 24.145377
## Weather_ConditionsRaining + high winds      1.2071754 0.07272304 16.599629
## Weather_ConditionsSnowing + high winds      2.0276695 0.43013292  4.714053
## Weather_ConditionsFog or mist               1.9549337 0.18625296 10.496122
##                                                 Pr(>|z|)
## Light_ConditionsDarkness - lights lit        0.000000e+00
## Light_ConditionsDarkness - lights unlit      2.236761e-46
## Light_ConditionsDarkness - no lighting       0.000000e+00
## Light_ConditionsDarkness - lighting unknown  1.451051e-44
## Weather_ConditionsRaining no high winds     2.648603e-231
## Weather_ConditionsSnowing no high winds      2.778250e-19
## Weather_ConditionsFine + high winds         8.350045e-129
## Weather_ConditionsRaining + high winds       7.012350e-62
## Weather_ConditionsSnowing + high winds       2.428370e-06
## Weather_ConditionsFog or mist                9.000212e-26
```

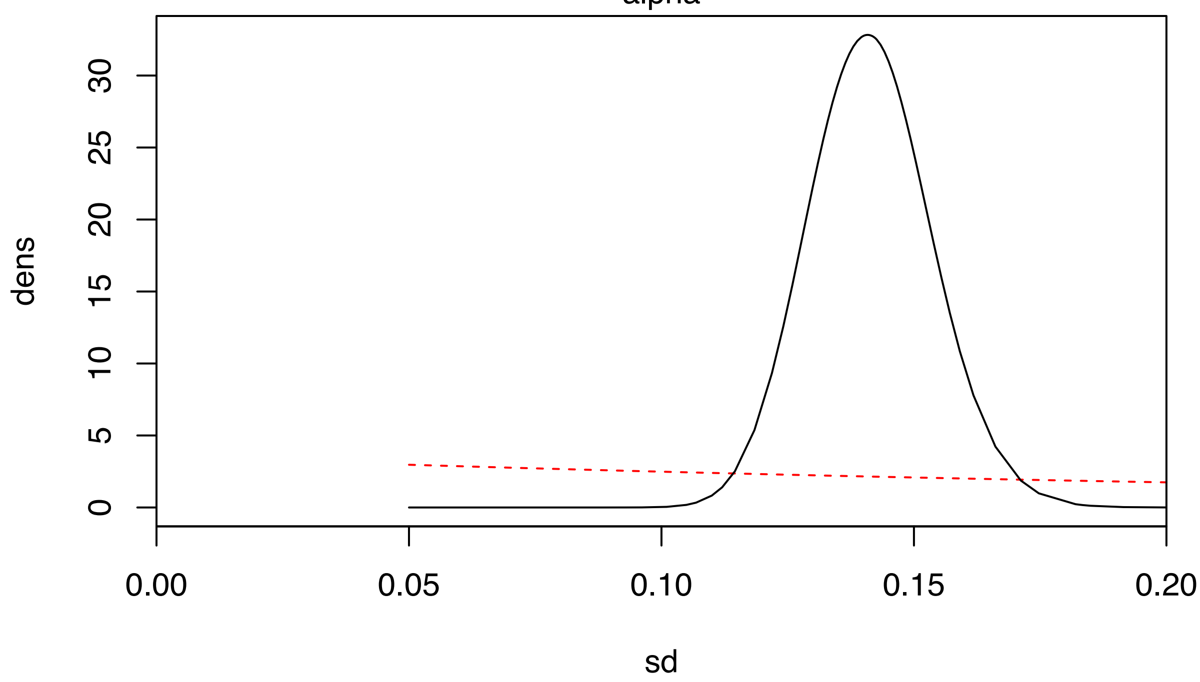
1,The summary of coefficients of logistics regression model is listed as above, we can see the p-value of lighting and weather conditions are extremely small, it means the lighting and weather conditions are both statistically significant, so they have notable influence to the outcome. And the time condition have relationship with lighting condition(darkness at night). Therefore I set them as strata.
2,The coefficients of conditional logistic regression are summarized in the table. The reference group is the male with age from 26 to 35(coef = 0). It is clear that men are more likely to involved in accidents than women in the rough. After age 35, the proportion of accidents which are fatal is higher for men than for women, but the proportion is pretty much the same from age 0 to 35. Also it's notable that wamen tend to be safer particularly as middle-aged(26-45), rather than as teenagers and in early adulthood.

Appendix

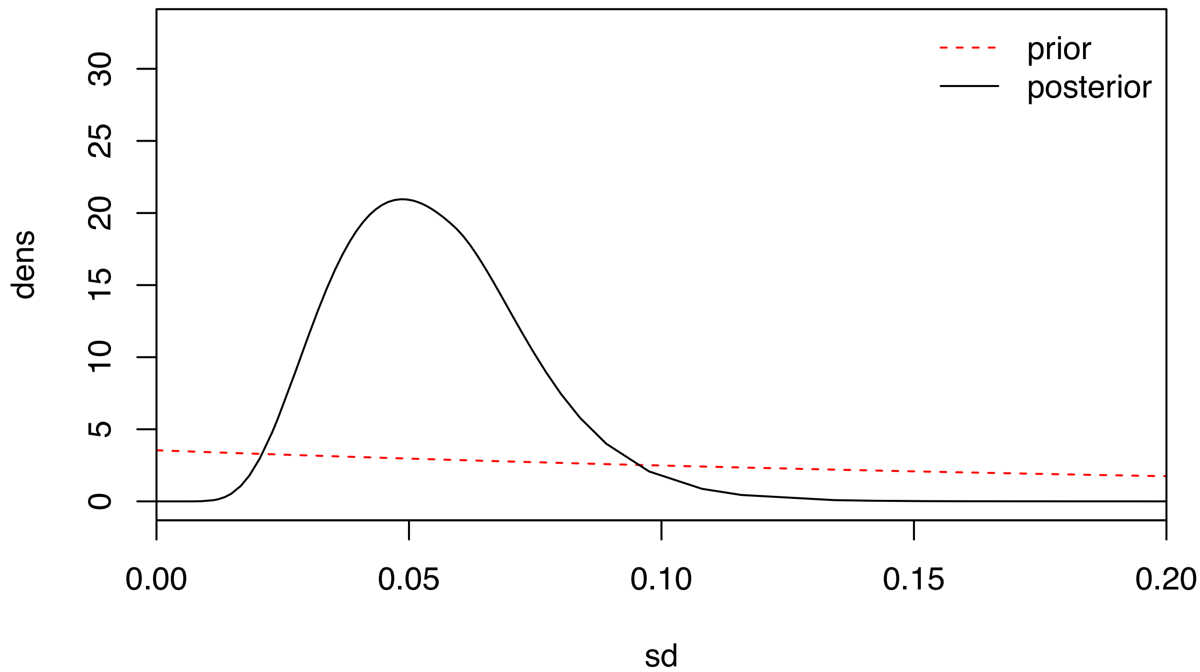
```r
# Smoking

### 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)
}
```

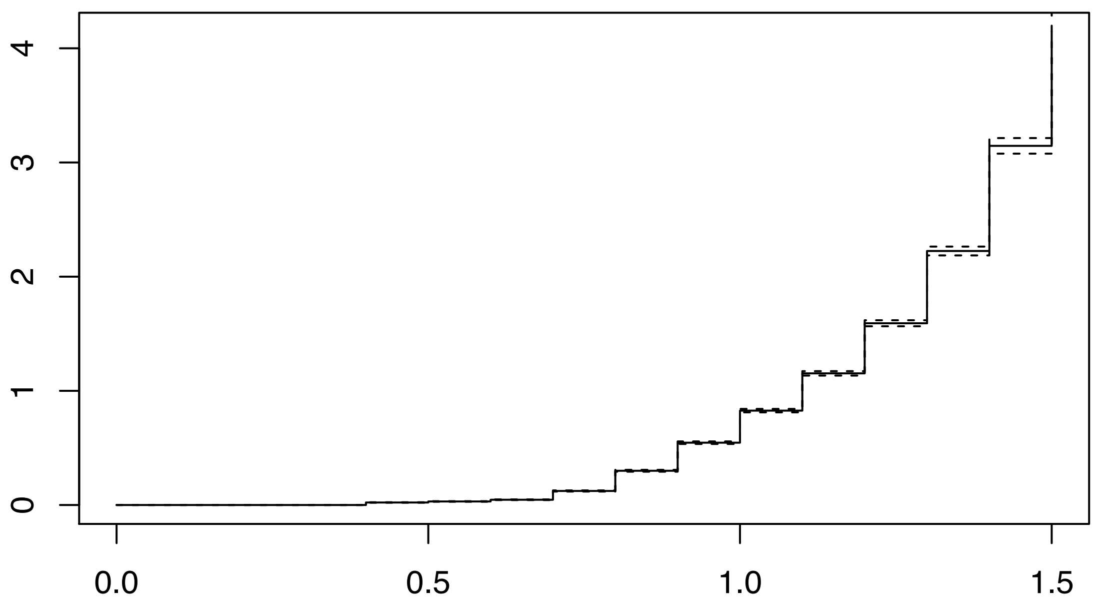





```
do.call(legend,  seco_model$priorPost$legend)
```




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




```
# Death
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,
```

```r
                              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,]

theTable = table(pedestrians$strata, pedestrians$y)

onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)]
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
model2 = glm(y ~ sex + age + Light_Conditions + Weather_Conditions, data = x, family = "binomial")
summary(model2)$coef
```

```
##                                               Estimate Std. Error
## (Intercept)                                 -3.2505406 0.02370688
## sexFemale                                   -0.2987419 0.01245846
## age0 - 5                                     0.1122174 0.03447951
## age6 - 10                                   -0.4372269 0.03241273
## age11 - 15                                  -0.5189038 0.03167783
## age16 - 20                                  -0.3864282 0.03229939
## age21 - 25                                  -0.1901664 0.03457809
## age36 - 45                                   0.3543664 0.03122667
## age46 - 55                                   0.6565748 0.03084867
## age56 - 65                                   1.0814320 0.02906168
## age66 - 75                                   1.6574226 0.02710365
## ageOver 75                                   2.2402435 0.02604294
## Light_ConditionsDarkness - lights lit        0.8544535 0.01366099
## Light_ConditionsDarkness - lights unlit      2.4841936 0.17373844
## Light_ConditionsDarkness - no lighting       2.7540798 0.03870802
## Light_ConditionsDarkness - lighting unknown  2.4524585 0.17511200
## Weather_ConditionsRaining no high winds      0.6907296 0.02127153
## Weather_ConditionsSnowing no high winds      1.6384235 0.18251005
## Weather_ConditionsFine + high winds          1.6771798 0.06946173
## Weather_ConditionsRaining + high winds       1.2071754 0.07272304
## Weather_ConditionsSnowing + high winds       2.0276695 0.43013292
## Weather_ConditionsFog or mist                1.9549337 0.18625296
##                                                 z value      Pr(>|z|)
## (Intercept)                                 -137.113806  0.000000e+00
## sexFemale                                    -23.979039 4.601842e-127
## age0 - 5                                       3.254611  1.135479e-03
## age6 - 10                                    -13.489356  1.806722e-41
## age11 - 15                                   -16.380664  2.628586e-60
## age16 - 20                                   -11.963949  5.488823e-33
## age21 - 25                                    -5.499620  3.806105e-08
## age36 - 45                                    11.348198  7.570648e-30
## age46 - 55                                    21.283729 1.606435e-100
## age56 - 65                                    37.211614 4.428418e-303
## age66 - 75                                    61.151265  0.000000e+00
## ageOver 75                                    86.021148  0.000000e+00
## Light_ConditionsDarkness - lights lit         62.546952  0.000000e+00
## Light_ConditionsDarkness - lights unlit       14.298469  2.236761e-46
## Light_ConditionsDarkness - no lighting        71.150116  0.000000e+00
```

Table 4: The coeficients of conditional logistic regression

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	sex	age
age0 - 5:sexFemale	0.0284229	1.0288306	0.0549522	0.5172285	0.6049967	Female	0
age6 - 10:sexFemale	-0.1771162	0.8376825	0.0507565	-3.4895264	0.0004839	Female	6
age11 - 15:sexFemale	-0.2498614	0.7789087	0.0471857	-5.2952744	0.0000001	Female	11
age16 - 20:sexFemale	-0.2791322	0.7564399	0.0520402	-5.3637766	0.0000001	Female	16
age21 - 25:sexFemale	-0.3691252	0.6913389	0.0633358	-5.8280613	0.0000000	Female	21
age26 - 35:sexFemale	-0.4482120	0.6387693	0.0522815	-8.5730476	0.0000000	Female	26
age36 - 45:sexFemale	-0.4482308	0.6387573	0.0516433	-8.6793515	0.0000000	Female	36
age46 - 55:sexFemale	-0.3763107	0.6863891	0.0482955	-7.7918406	0.0000000	Female	46
age56 - 65:sexFemale	-0.2370677	0.7889379	0.0403324	-5.8778460	0.0000000	Female	56
age66 - 75:sexFemale	-0.1433569	0.8664448	0.0323676	-4.4290313	0.0000095	Female	66
ageOver 75:sexFemale	-0.1256106	0.8819582	0.0272702	-4.6061492	0.0000041	Female	75
age0 - 5	0.1324083	1.1415744	0.0440170	3.0081179	0.0026287	Male	0
age6 - 10	-0.3196593	0.7263965	0.0408650	-7.8223298	0.0000000	Male	6
age11 - 15	-0.3829384	0.6818549	0.0411527	-9.3053109	0.0000000	Male	11
age16 - 20	-0.4432109	0.6419718	0.0404473	-10.9577480	0.0000000	Male	16
age21 - 25	-0.2680862	0.7648419	0.0421849	-6.3550264	0.0000000	Male	21
age 26 - 35	0.0000000	1.0000000	0.0000000	NA	NA	Male	26
age36 - 45	0.4115311	1.5091267	0.0386489	10.6479477	0.0000000	Male	36
age46 - 55	0.7682289	2.1559445	0.0389790	19.7087971	0.0000000	Male	46
age56 - 65	1.2120970	3.3605244	0.0378511	32.0227837	0.0000000	Male	56
age66 - 75	1.7972504	6.0330360	0.0363472	49.4467189	0.0000000	Male	66
ageOver 75	2.3957024	10.9759044	0.0351665	68.1244757	0.0000000	Male	75

```
## Light_ConditionsDarkness - lighting unknown   14.005085  1.451051e-44
## Weather_ConditionsRaining no high winds       32.472019 2.648603e-231
## Weather_ConditionsSnowing no high winds        8.977168  2.778250e-19
## Weather_ConditionsFine + high winds           24.145377 8.350045e-129
## Weather_ConditionsRaining + high winds        16.599629  7.012350e-62
## Weather_ConditionsSnowing + high winds         4.714053  2.428370e-06
## Weather_ConditionsFog or mist                 10.496122  9.000212e-26
```

```r
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), ]

knitr::kable(theCoef, caption="The coeficients of conditional logistic regression")
```