Question 1: Smoking

Summary

This report analyzes the 2014 American National Youth Tobacco survey using a survival model with a Weibull distribution, modelling the age of children first trying cigarettes. We first investigate the hypothesis about whether there is a larger variation between states than schools in the mean age of children first trying cigarettes, and we have found it to be false. It seems like the variation between schools is substantially larger than the variation between states, hence the tobacco control program should be targeting particular schools rather than states. The second hypothesis we investigate states that whether two identical non-smoking children of different ages have the same probability of trying cigarettes within the next month, and we have found it to be false as well. It seems like as the age of the non-smoking children increases, the probability of children starting to smoke increases, given they have identical sex, residency (urban/rural) and ethnicity.

Introduction

In this report, we will be analyzing the 2014 American National Youth Tobacco survey, to investigate the hypotheses about whether variation between states and schools in the mean age of children first trying cigarettes is larger and whether two identical non-smoking children have the same probability of trying cigarettes within the next month. In the first hypothesis, it is stated that the variation between states in the mean age is substantially greater than that between schools, and hence tobacco control programs should target the states with the earliest smoking ages, not the schools. In the second hypothesis, it is stated that two identical non-smoking children have the same probability of trying cigarettes within the next month, irrespective of the ages.

Methods

Model Selection

The survival model with a Weibull distribution is used to analyze the data, as we want to model the age of children first trying cigarettes. For the variables in the model, three confounders (sex, urban/urban, ethnicity) and two random effects (school and states) are included. No interaction terms are added into the model, as after testing multiple models with interaction terms, there isn't much improvement in the model fits.

Prior Distribution Selection

I have chosen to use the penalized complexity prior, which puts an exponential prior to the standard deviation σ_u^2,σ_e^2 and requires me to calibrate the scaling of random effect priors.

For the random effect for states, the variability in the rate of smoking initiation in some states has doubled or tripled compared to other states, but that rate won't exceed ten times.
Calculation:

$$\exp(U_i) = 2$$
$$U_i = \ln(2) = 0.693$$
$$\exp(U_i) = 10$$
$$U_i = \ln(10) = 2.303$$

Given that Ui can be selected within the range of 0.7 - 2.3, I have chosen the value Ui = 1, additionally, I set p = 0.01, meaning that I think there is a 99% chance that the between state variability is greater than 1.2.

For the random effect for schools, the largest variability between the two schools is approximately 1.5 times, meaning that the 'worst' schools are expected to have at most 50% at a greater rate than the 'healthiest' school.
Calculation:

$$\exp(Vij) = 1.5$$
$$Vij = \ln(1.5) = 0.405$$

Given that the largest Vij value can be 0.4, U have choosen Vij = 0.3, additionally I set p = 0.0, meaning that I think there's a 99% chance that the between school variability is greater than 0.3.

Results

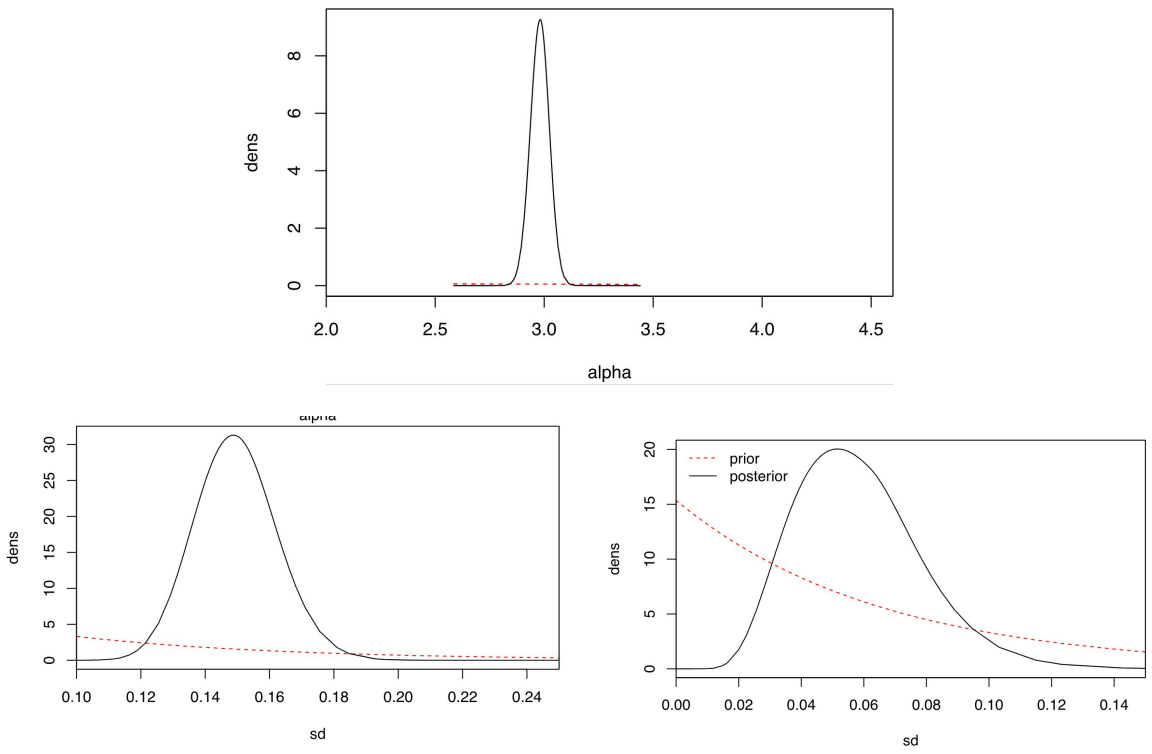


Figure 1. Prior and Posterior Distribution of Hyperparameters
(Top: Alpha parameter for Weibull, Bottom Left: SD for school, Bottom Right: SD for state)

	mean	0.025quant	0.975quant
alpha for weibullsurv	2.980	2.895	3.064
sd for school	0.150	0.126	0.176
sd for state	0.058	0.026	0.103

Figure 2. Posterior Estimates of Hyperparameters

Figure 2 gives us estimates of the variations between school and state. From this result, it seems like the variation between schools (0.150) is substantially larger than the variation between states (0.058), showing that the first hypothesis is false. In this case, it would be recommended that tobacco control programs should target particular schools with the earliest smoking ages, instead of targeting specific states with that problem.

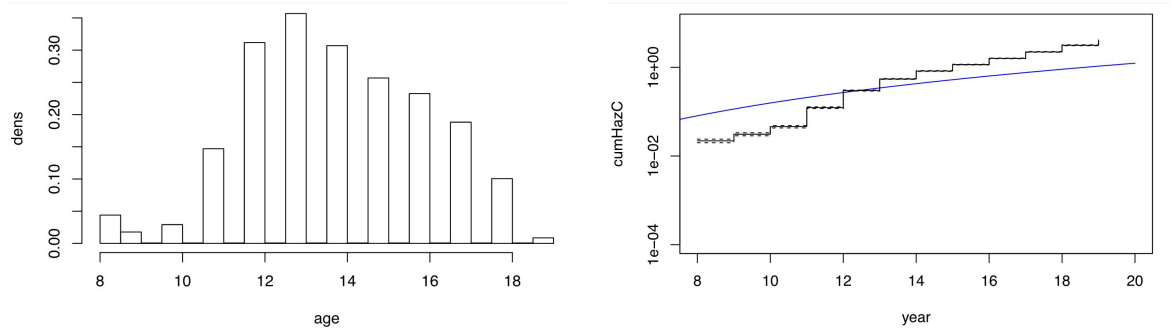


Figure 3. Hazard Density Function (Left), Cumulative Hazard (Right)

The second hypothesis proposes that first cigarette smoking has a flat hazard function, that is the shape parameter is equal to 1 and the probability of two identical non-smoking children starting to smoke in the next month is the same, irrespective of their ages. However, in Figure 2, we have found that the shape parameter (alpha for weibullsurv) is approximately 3; additionally, in Figure 3, the hazard function is not flat. Hence we can conclude that first cigarette smoking has an increasing hazard function, meaning that as the age of the non-smoking children increases, the probability of children starting to smoke increases, given identical cofounders (sex, urban/rural, ethnicity).

Question 2: Death on the roads

Summary

This report analyzes a subset of data concerning road traffic accidents in the UK from 1979 to 2015 using a conditional logistic regression model. We investigate the hypothesis stating that women tend to be safer as pedestrians than men in motorcycle accidents, especially for younger women, and we have found it to be half true. Women are indeed safer pedestrians comparing to their male counterparts, however, middle-aged women pedestrians are the least likely to encounter fatal accidents compared to the other age groups.

Introduction

In this report, we will be analyzing a subset of data with all of the road traffic accidents in the UK from 1979 to 2015, consisting of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries. Using this subset of data, we want to investigate the hypothesis stating that women tend to be safer as pedestrians than men, particularly as teenagers and in early adulthood.

Methods

There are two models that are fitted to answer the hypothesis. We have fitted a logistic regression model, to model the odds of fatality of pedestrians involved in motor vehicle accidents. For the variables, we have included variables such as sex, age, light conditions and weather conditions in the model. This model helps us identify potential variables that can be used for stratification, which are usually variables that are highly influential on the response variable and thus vary within at least one strata. Secondly, we have fitted a conditional logistic regression model, also to model the odds of fatality but with stratification using the variables time of day, lighting conditions and weather conditions. Fatal accidents (34299) are treated as cases, whereas accidents with sligh injuries (411003) are treated as controls. For the variables, we have included the variables of interest, which are sex and age of the pedestrians involved in these accidents.

Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.251	0.024	-137.114	0.000
sexFemale	-0.299	0.012	-23.979	0.000
age0 - 5	0.112	0.034	3.255	0.001
age6 - 10	-0.437	0.032	-13.489	0.000
age11 - 15	-0.519	0.032	-16.381	0.000
age16 - 20	-0.386	0.032	-11.964	0.000
age21 - 25	-0.190	0.035	-5.500	0.000
age36 - 45	0.354	0.031	11.348	0.000
age46 - 55	0.657	0.031	21.284	0.000
age56 - 65	1.081	0.029	37.212	0.000
age66 - 75	1.657	0.027	61.151	0.000
ageOver 75	2.240	0.026	86.021	0.000
Light_ConditionsDarkness - lights lit	0.854	0.014	62.547	0.000
Light_ConditionsDarkness - lights unlit	2.484	0.174	14.298	0.000
Light_ConditionsDarkness - no lighting	2.754	0.039	71.150	0.000
Light_ConditionsDarkness - lighting unknown	2.452	0.175	14.005	0.000
Weather_ConditionsRaining no high winds	0.691	0.021	32.472	0.000
Weather_ConditionsSnowing no high winds	1.638	0.183	8.977	0.000
Weather_ConditionsFine + high winds	1.677	0.069	24.145	0.000
Weather_ConditionsRaining + high winds	1.207	0.073	16.600	0.000
Weather_ConditionsSnowing + high winds	2.028	0.430	4.714	0.000
Weather_ConditionsFog or mist	1.955	0.186	10.496	0.000

Figure 1. Results from the logistic model

Figure 1 shows us the results from the logistic regression model we have fitted. It seems like the odds of fatality varies greatly under different lighting condition, the odds seem to be lower when the lights are lit (0.854) while it is much higher when the lights are unlit (2.484) or there is no lighting at all (2.754). The odds of fatality also vary under different weather conditions; the odds are the highest when it is snowing and there are high winds (2.028), while the odds are the lowest when it is raining and there are no high winds (0.691). Therefore both light conditions and weather conditions are suitable to be included in the stratification of the data, given their substantial effects. Another variable we are not able to fit into this model is the time of day, but we can potentially associate that with the light condition, for instance at night when it is darker, the odds of fatality must be significantly bigger than that in the morning or afternoon.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age0 - 5	0.132	1.142	0.044	3.008	0.003
age6 - 10	-0.320	0.726	0.041	-7.822	0.000
age11 - 15	-0.383	0.682	0.041	-9.305	0.000
age16 - 20	-0.443	0.642	0.040	-10.958	0.000
age21 - 25	-0.268	0.765	0.042	-6.355	0.000
age36 - 45	0.412	1.509	0.039	10.648	0.000
age46 - 55	0.768	2.156	0.039	19.709	0.000
age56 - 65	1.212	3.361	0.038	32.023	0.000
age66 - 75	1.797	6.033	0.036	49.447	0.000
ageOver 75	2.396	10.976	0.035	68.124	0.000
age26 - 35:sexFemale	-0.448	0.639	0.052	-8.573	0.000
age0 - 5:sexFemale	0.028	1.029	0.055	0.517	0.605
age6 - 10:sexFemale	-0.177	0.838	0.051	-3.490	0.000
age11 - 15:sexFemale	-0.250	0.779	0.047	-5.295	0.000
age16 - 20:sexFemale	-0.279	0.756	0.052	-5.364	0.000
age21 - 25:sexFemale	-0.369	0.691	0.063	-5.828	0.000
age36 - 45:sexFemale	-0.448	0.639	0.052	-8.679	0.000
age46 - 55:sexFemale	-0.376	0.686	0.048	-7.792	0.000
age56 - 65:sexFemale	-0.237	0.789	0.040	-5.878	0.000
age66 - 75:sexFemale	-0.143	0.866	0.032	-4.429	0.000
ageOver 75:sexFemale	-0.126	0.882	0.027	-4.606	0.000

Figure 2. Results from the conditional logistic model

Figure 2 shows the results from the conditional logistic model we have fitted. The results are supportive of the hypothesis, stating that women tend to be safer as pedestrians than men if we compare the estimates of regression coefficients. For men, especially after the age of 36-45, the estimates are above one and increasing, indicating that the older men get, the more likely they will encounter a fatal accident as a pedestrian. Whereas for women, the estimates remain less than one throughout most age groups, indicating that women pedestrians are indeed less likely to encounter a fatal accident.

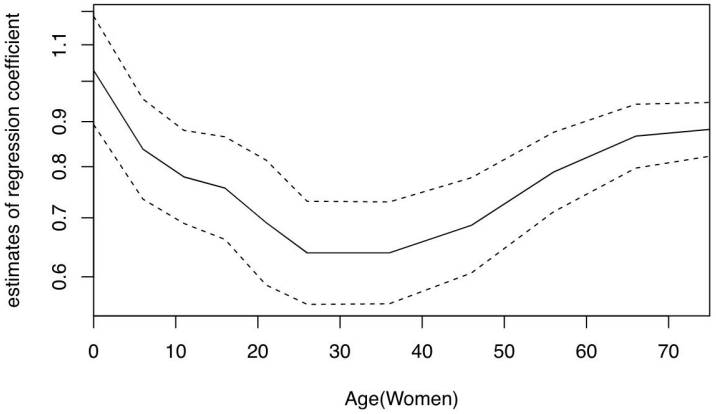


Figure 3. Graphical representation of results of the conditional logistic model for women

The second half of the hypothesis states that women are safer as pedestrians as teenagers or in early adulthood, which is false. Women are the least likely to encounter a fatal accident when they are in the age group 36-45 (0.639) and 46-55 (0.686), indicating that middle-aged women are the safest as pedestrians. This phenomenon can also be observed in Figure 3, which shows that women as teenagers or in early adulthood have relatively high estimates of regression coefficients compared to middle-aged women.

STA442 HW4 Q1

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

```
## Loading required namespace: R.utils
```

```
## Registered S3 method overwritten by 'R.oo':
##   method        from
##   throw.default R.methodsS3
```

```
load(smokeFile)
smoke = smoke[smoke$Age > 9, ]

# forInla
forInla = smoke[, c("Age", "Age_first_tried_cigt_smkg", "Sex", "Race", "state", "school", "RuralUrban")]
forInla = na.omit(forInla)
forInla$school = factor(forInla$school)
```

```
# forSurv
library("INLA")
```

```
## Loading required package: Matrix
```

```
## Loading required package: sp
```

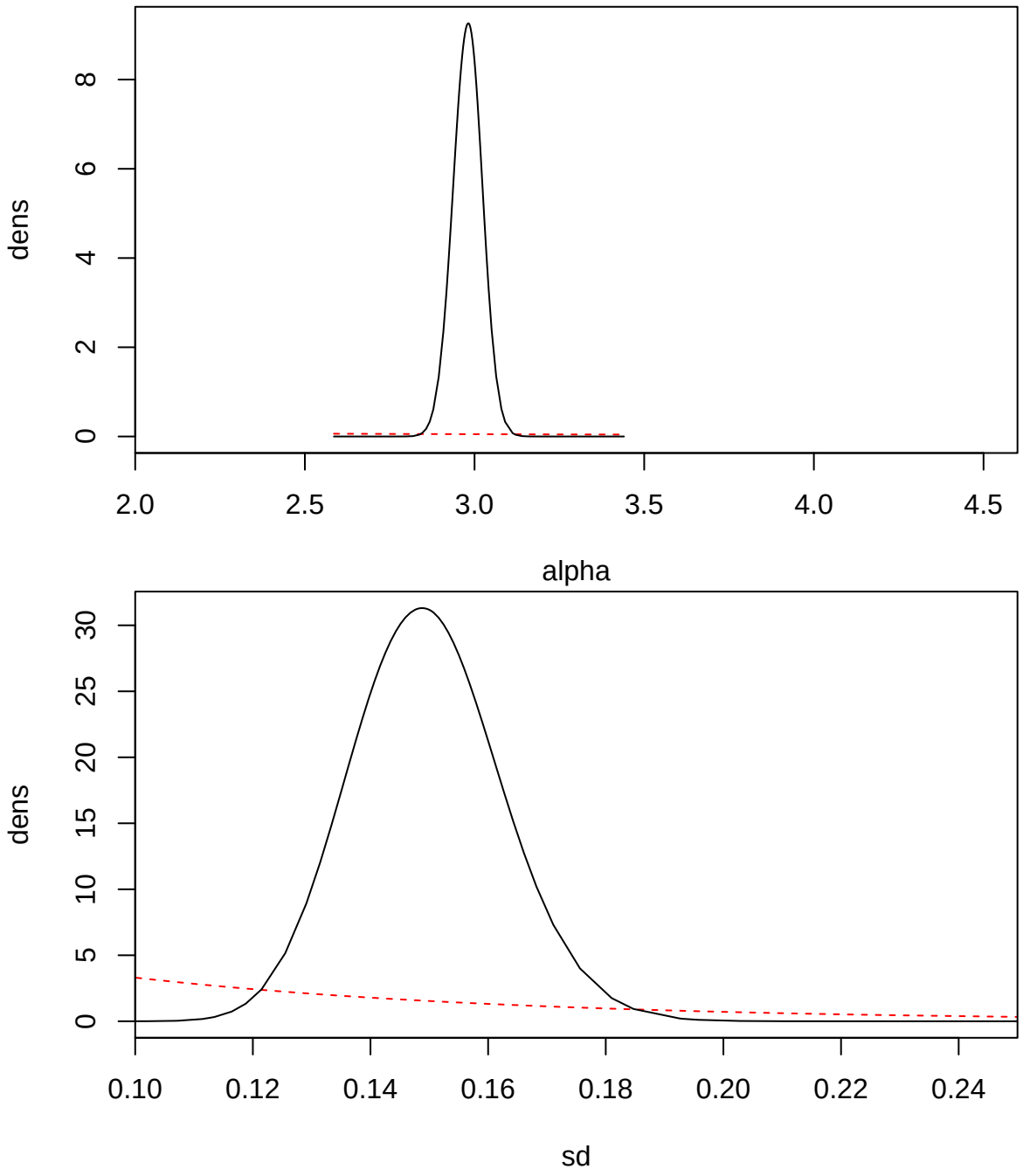
```
## Loading required package: parallel
```

```
## This is INLA_19.09.03 built 2019-09-03 09:07:31 UTC.
## See www.r-inla.org/contact-us for how to get help.
## To enable PARDISO sparse library; see inla.pardiso()
```

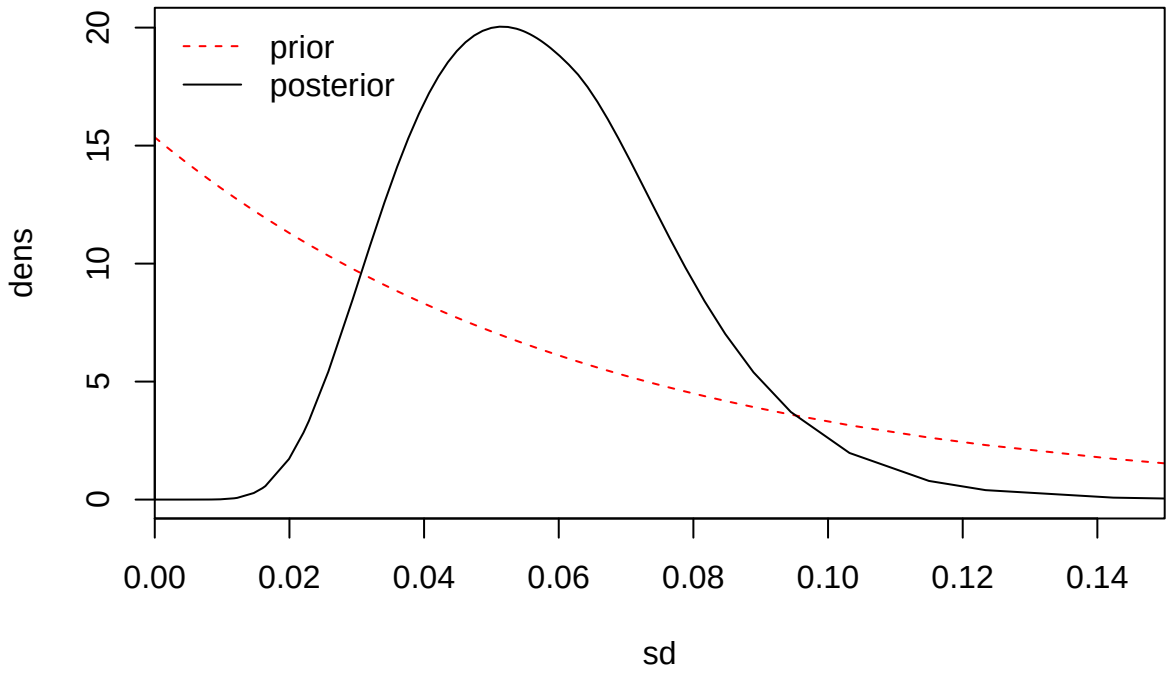
```
forSurv = data.frame(time = (pmin(forInla$Age_first_tried_cigt_smkg,forInla$Age) - 4)/10,
                     event = forInla$Age_first_tried_cigt_smkg <= forInla$Age)

# 0 means non-smoking, 1 means smoking, 2 means starts to smoke at or before 8 years old
forSurv[forInla$Age_first_tried_cigt_smkg == 8, "event"] = 2

smokeResponse = inla.surv(forSurv$time, forSurv$event)
```

```
# model fit
fitS2 = inla(smokeResponse ~ RuralUrban + Sex + Race +
               f(school, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(0.3, 0.01))))
             + f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(1.2, 0.01)))),
             control.family = list(variant = 1,
                                   hyper = list(alpha = list(prior = "normal",
                                                             param = c(log(0.9), (2/3)^(-2))))),
             control.mode = list(theta = c(8,2,5), restart = TRUE),
             data = forInla, family = "weibullsurv",verbose = TRUE)
```

```
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)
```




```
# Results on a more natural scale
modFix = exp(-fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")])
knitr::kable(modFix, digits = 3)
```

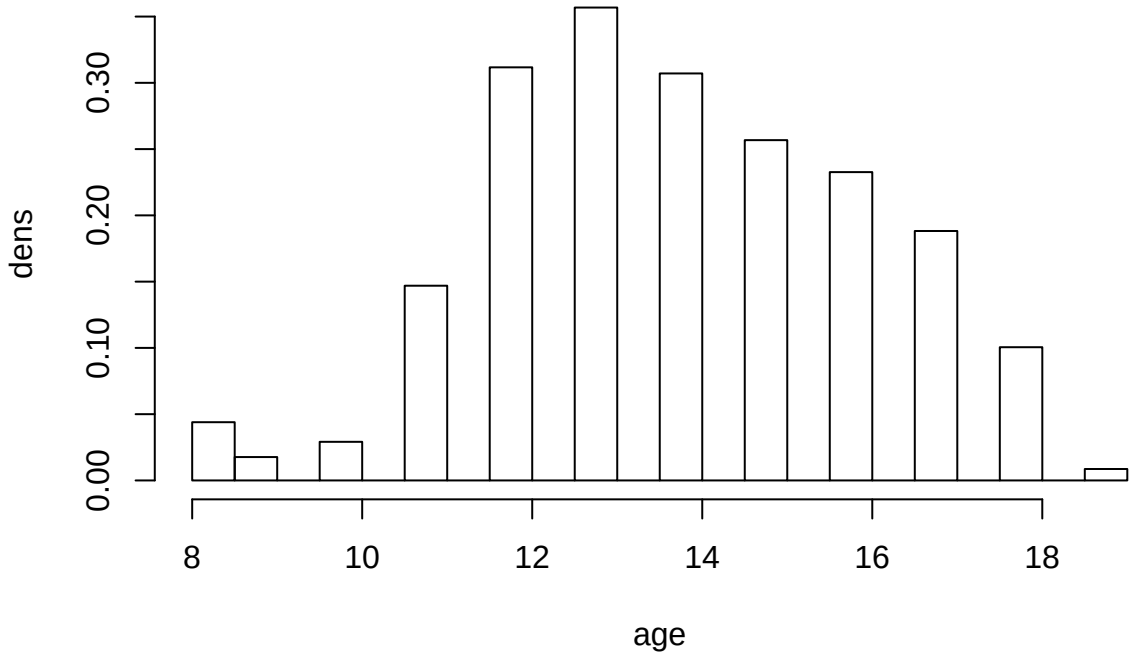
	mean	0.025quant	0.975quant
(Intercept)	1.862	1.966	1.762
RuralUrbanRural	0.892	0.946	0.841
SexF	1.051	1.072	1.030
Raceblack	1.058	1.094	1.023
Racehispanic	0.967	0.994	0.941
Raceasian	1.213	1.300	1.136
Racenative	0.912	0.990	0.844
Racepacific	0.882	1.019	0.775

```
# Posterior estimates of hyperparamters
modSd = Pmisc::priorPost(fitS2)$summary[, c("mean", "0.025quant", "0.975quant")]
knitr::kable(modSd, digits = 3)
```

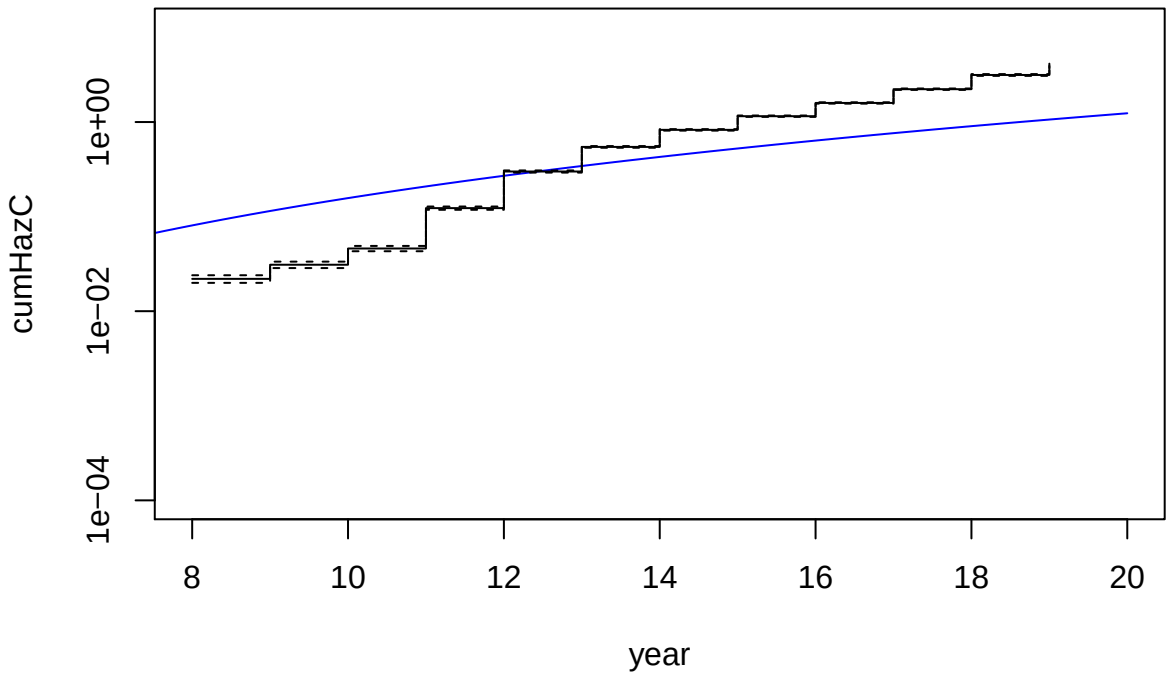
	mean	0.025quant	0.975quant
alpha for weibullsurv	2.980	2.895	3.064
sd for school	0.150	0.126	0.176
sd for state	0.058	0.026	0.103

```
library('survival')
# Density Function
forSurv$ones = 1
xSeq = seq(5,20,len=1000)
kappa = fitS2$summary.hyperpar['alpha','mean']
lambda = exp(-fitS2$summary.fixed['(Intercept)','mean'])
```

```r
hist((forSurv$time*10)+4, main='', xlab='age', ylab='dens', prob=TRUE)
```




```r
# Cumulative Hazard
cumHazC = ((xSeq/(10*lambda))^kappa)
plot(xSeq, cumHazC, col='blue',type='l',
     log='y', ylim=c(0.0001, 10), xlim = c(8,20),
     xlab = "year")
hazEst = survfit(Surv(time,ones) ~ 1, data=forSurv)
hazEst$time = (hazEst$time*10)+4
lines(hazEst,fun="cumhaz")
```



STA442 HW4 Q2

```
# download data
pedestrainFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
```

```
## Loading required namespace: R.utils
```

```
## Registered S3 method overwritten by 'R.oo':
##   method        from
##   throw.default R.methodsS3
```

```
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time),]
```

```
# create strata for time, light conditions and weather conditions
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)]
# remove strata with no cases and no controls
x = pedestrians[!pedestrians$strata %in% onlyOne, ]

# GLM
table(x$Casualty_Severity)
```

```
##
## Slight  Fatal
## 411003  34299
```

```
xSubM = as.data.frame(x[x$Casualty_Severity %in% c("Fatal", "Slight"),])

theLogit = glm(y ~ sex + age + Light_Conditions + Weather_Conditions, data = xSubM, family = "binomial")
knitr::kable(summary(theLogit)$coef,digit = 3)
```

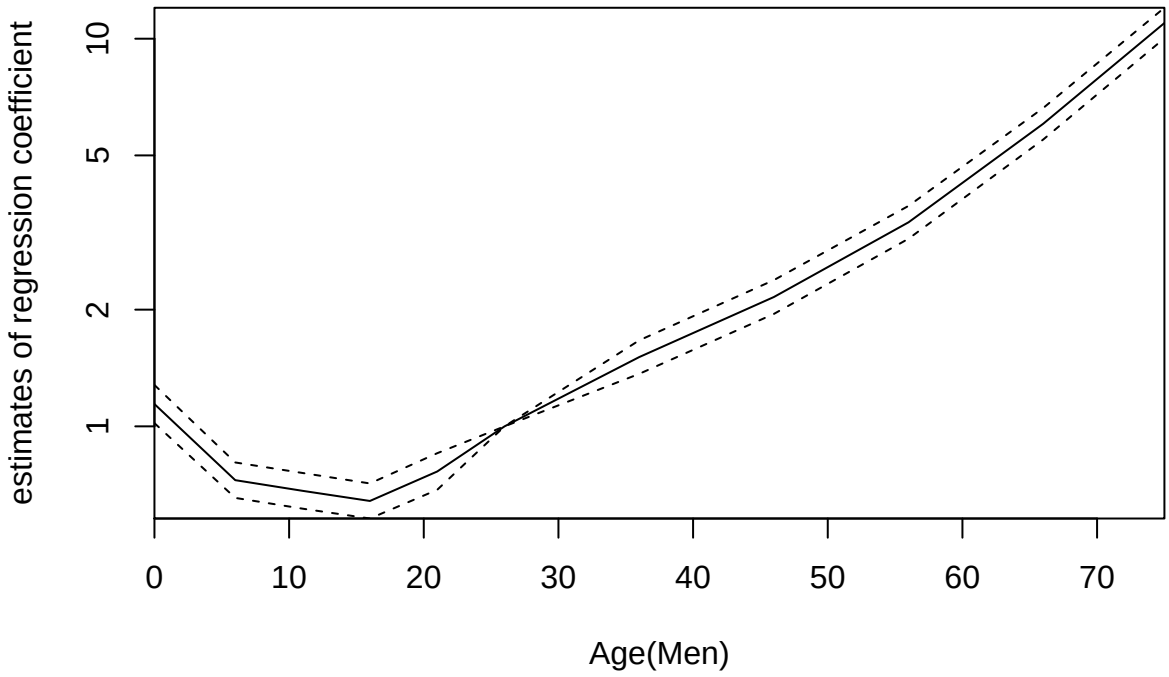
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.251	0.024	-137.114	0.000
sexFemale	-0.299	0.012	-23.979	0.000
age0 - 5	0.112	0.034	3.255	0.001
age6 - 10	-0.437	0.032	-13.489	0.000
age11 - 15	-0.519	0.032	-16.381	0.000
age16 - 20	-0.386	0.032	-11.964	0.000
age21 - 25	-0.190	0.035	-5.500	0.000
age36 - 45	0.354	0.031	11.348	0.000
age46 - 55	0.657	0.031	21.284	0.000
age56 - 65	1.081	0.029	37.212	0.000
age66 - 75	1.657	0.027	61.151	0.000
ageOver 75	2.240	0.026	86.021	0.000
Light_ConditionsDarkness - lights lit	0.854	0.014	62.547	0.000
Light_ConditionsDarkness - lights unlit	2.484	0.174	14.298	0.000
Light_ConditionsDarkness - no lighting	2.754	0.039	71.150	0.000

	Estimate	Std. Error	z value	Pr(>|z|)
Light_ConditionsDarkness - lighting unknown	2.452	0.175	14.005	0.000
Weather_ConditionsRaining no high winds	0.691	0.021	32.472	0.000
Weather_ConditionsSnowing no high winds	1.638	0.183	8.977	0.000
Weather_ConditionsFine + high winds	1.677	0.069	24.145	0.000
Weather_ConditionsRaining + high winds	1.207	0.073	16.600	0.000
Weather_ConditionsSnowing + high winds	2.028	0.430	4.714	0.000
Weather_ConditionsFog or mist	1.955	0.186	10.496	0.000

```
# conditional logistic model
library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata),data = x)
knitr::kable(summary(theClogit)$coef,digit = 3)
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age0 - 5	0.132	1.142	0.044	3.008	0.003
age6 - 10	-0.320	0.726	0.041	-7.822	0.000
age11 - 15	-0.383	0.682	0.041	-9.305	0.000
age16 - 20	-0.443	0.642	0.040	-10.958	0.000
age21 - 25	-0.268	0.765	0.042	-6.355	0.000
age36 - 45	0.412	1.509	0.039	10.648	0.000
age46 - 55	0.768	2.156	0.039	19.709	0.000
age56 - 65	1.212	3.361	0.038	32.023	0.000
age66 - 75	1.797	6.033	0.036	49.447	0.000
ageOver 75	2.396	10.976	0.035	68.124	0.000
age26 - 35:sexFemale	-0.448	0.639	0.052	-8.573	0.000
age0 - 5:sexFemale	0.028	1.029	0.055	0.517	0.605
age6 - 10:sexFemale	-0.177	0.838	0.051	-3.490	0.000
age11 - 15:sexFemale	-0.250	0.779	0.047	-5.295	0.000
age16 - 20:sexFemale	-0.279	0.756	0.052	-5.364	0.000
age21 - 25:sexFemale	-0.369	0.691	0.063	-5.828	0.000
age36 - 45:sexFemale	-0.448	0.639	0.052	-8.679	0.000
age46 - 55:sexFemale	-0.376	0.686	0.048	-7.792	0.000
age56 - 65:sexFemale	-0.237	0.789	0.040	-5.878	0.000
age66 - 75:sexFemale	-0.143	0.866	0.032	-4.429	0.000
ageOver 75:sexFemale	-0.126	0.882	0.027	-4.606	0.000

```
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",
        ylab = "estimates of regression coefficient", xlab = "Age(Men)")
```




```
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",
        ylab = "estimates of regression coefficient", xlab = "Age(Women)")
```