STA442 Homework4

SongQi Wang 1003439442

2019-12-03

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.

Base on the *2014 American National Youth Tobacco Survey* (pbrown.ca/teaching/appliedstats/data), we would like 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

Children start smoking for the first time once, therefore we chose Weibull distribution to model the data, which is good for survival analysis data.

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

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.

We setthe following as the prior distrobutions:

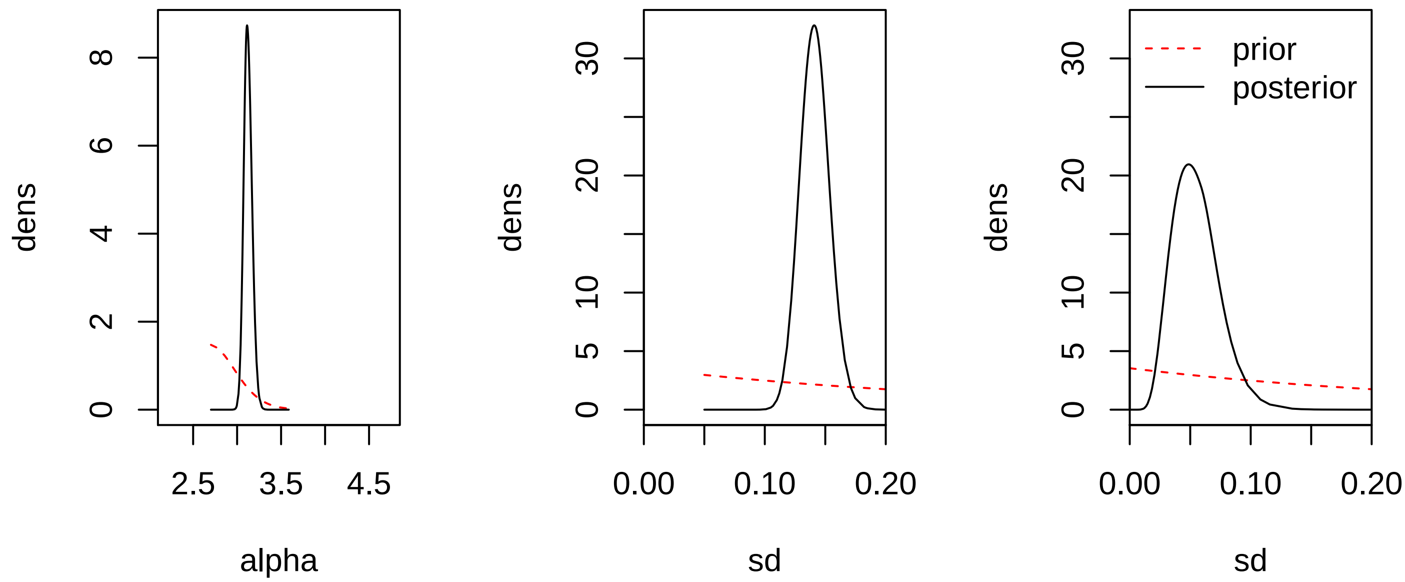
$$
\begin{aligned}
\kappa &\sim N(1,\ \ 0.1) \\
P(\sigma_U > 1.3) &= 1\% \\
P(\sigma_V > 0.6) &= 1\%
\end{aligned}
$$

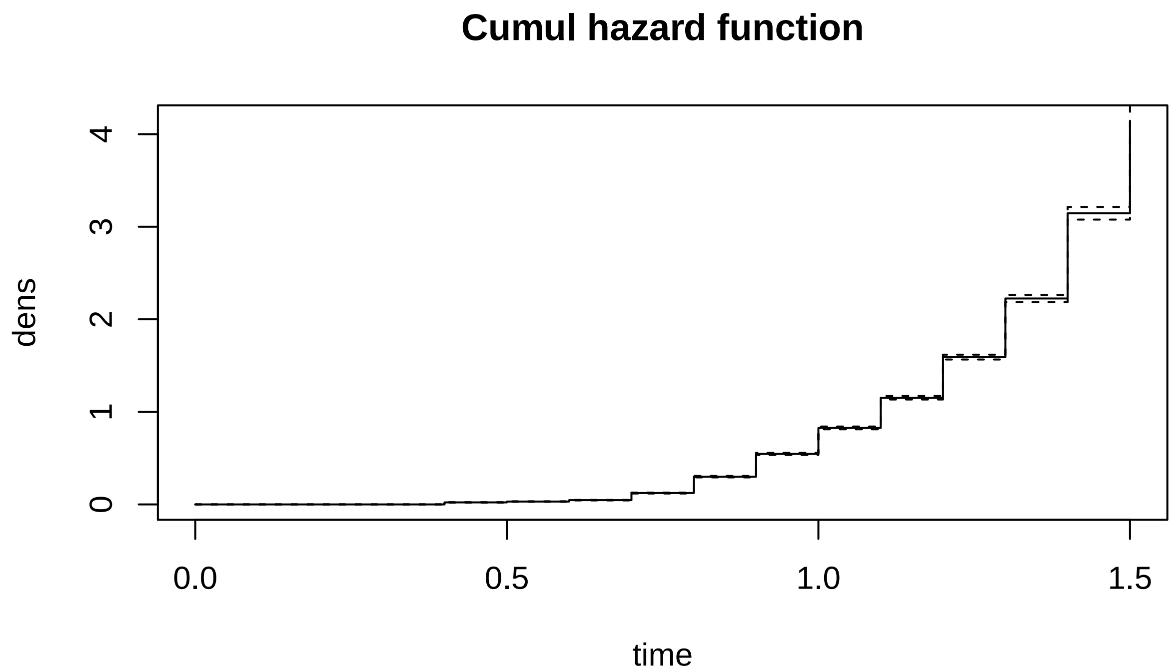
Result

Table 1: Posterior estimates

SD of School	SD of state
0.142	0.059

Since $\sigma_U = 0.142$ and $\sigma_V = 0.059$, geographic variation (between states) in the mean age children first try cigarettes is less than variation amongst schools. Tobacco control programs should actually target particular schools where smoking is a problem.





Base on the plots of prior distribution and the plot of hazard fuction, we can tell the first cigarette smoking does not have a flat hazard function. The non-smoking children with higher age have the higher probability of trying cigarettes within the next month.

Death on the roads

Introduction

We used the data from www.gov.uk/government/statistical-data-sets/ras30-reportedcasualties-in-road-accidents, the difference in accidents between the male and female, 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).

we would like to investigate the following hypotheses:

1. Men are involved in accidents more than women
2. 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

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

$$pr(Y_i = 1|X_i) = \lambda_i$$
$$\log(\frac{\lambda_i}{1-\lambda_i}) = \beta_0 + \sum_{p=1}^{P} X_{ip}\beta_p$$

We have

$$pr(Y_i = 1|X_i, Z_i = 1) = \lambda_i^*$$
$$\log(\frac{\lambda_i^*}{1-\lambda_i^*}) = \beta_0^* + \sum_{p=1}^{P} X_{ip}\beta_p^*$$

Then we finally get:

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

where:

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

Result

Table 2: The coeficients of conditional logistic regression

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

The coeficients of conditional logistic regression are summarized in the table. The reference group is the male with age from 26 to 35. It is easy to see that generally men are involved in accidents more than women. 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.

Appendix

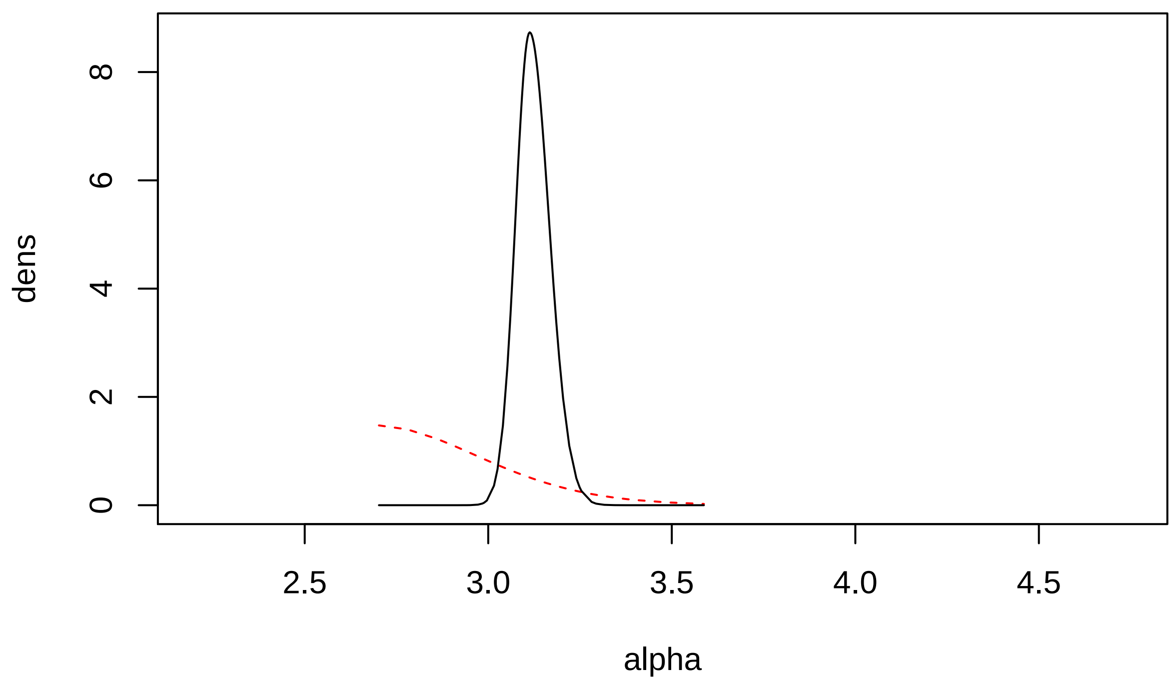
```
### Model Code ###

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(1.3,0.01)))) +
  f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec", param = c(0.6,0.01))))


prof_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))
```

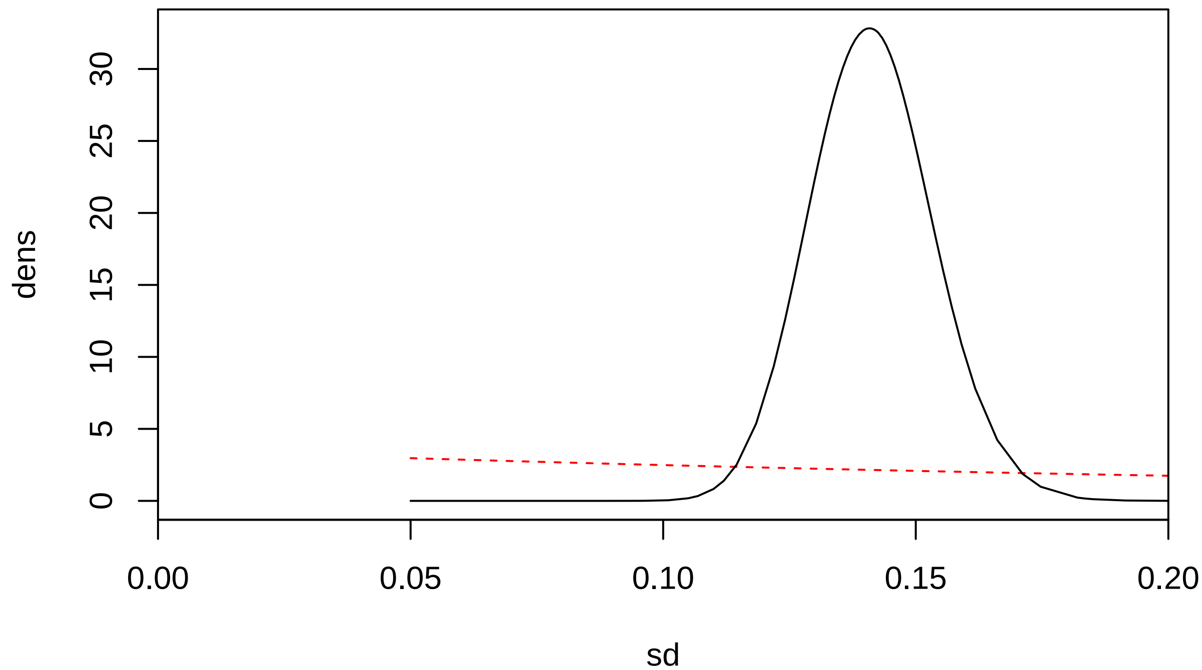
```
### model para ###

# rbind(prof_model$summary.fixed[, c("mean", "0.025quant",
# "0.975quant")],Pmisc::priorPostSd(prof_model)$summary[,
# c("mean", "0.025quant", "0.975quant")])

post.dat=round(exp(prof_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") %>%
  kable_styling(latex_options = "hold_position")
```

Table 3: Posterior estimates

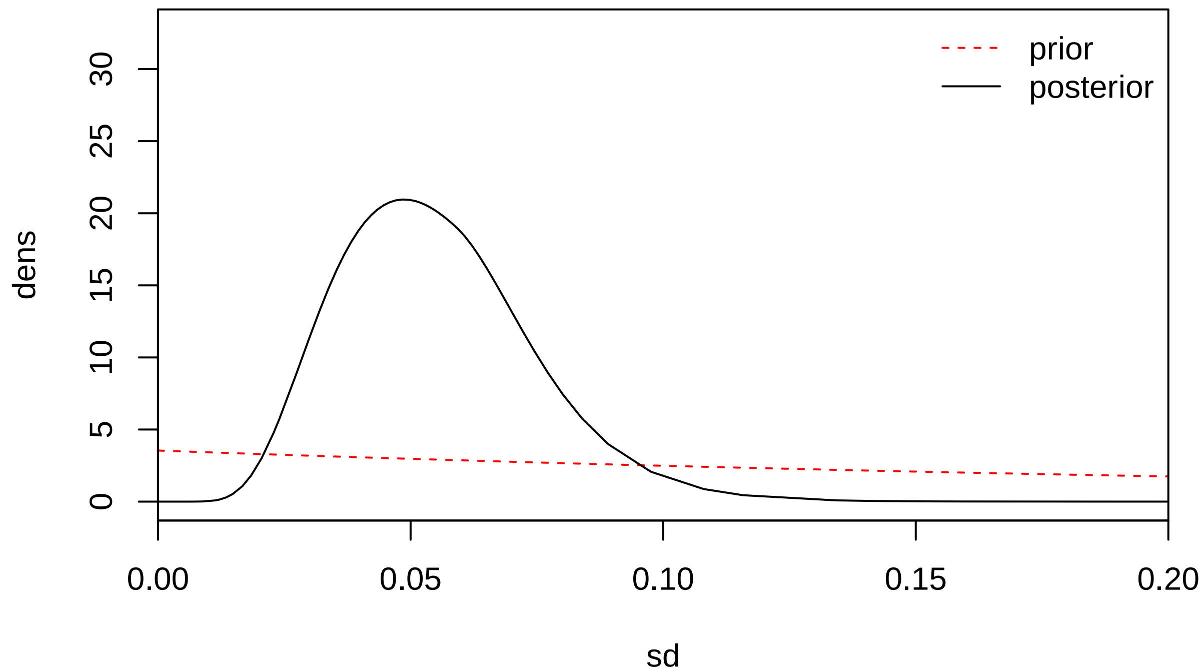
SD of School	SD of state
0.142	0.059

```r
### porio plot

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

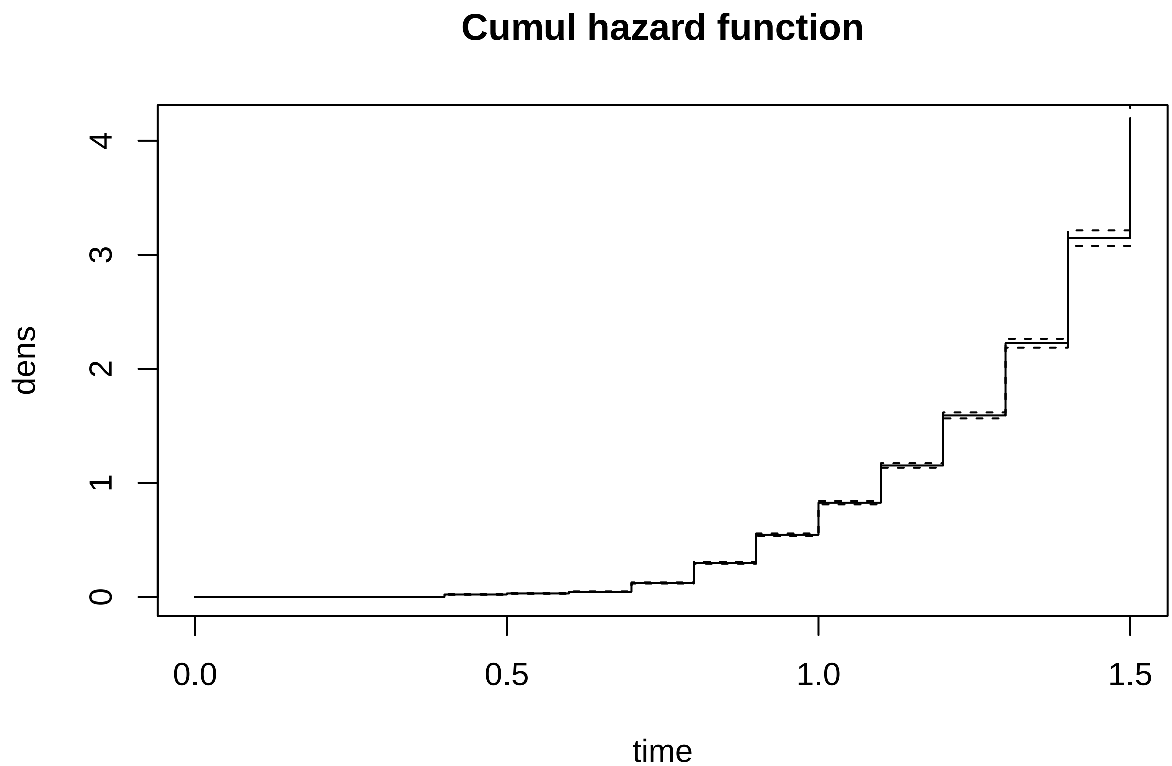





```r
do.call(legend, prof_model$priorPost$legend)
```




```
forSurv$one = 1
hazEst = survfit(Surv(time, one) ~ 1, data=forSurv)
plot(hazEst, fun='cumhaz', ylab = 'dens', xlab = 'time', main = "Cumul hazard function")
```




```
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,]

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

onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)]

# model2 = glm(y ~ sex + age + Light_Conditions + Weather_Conditions, data = x, family = "binomial")

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

```r
theCoef = as.data.frame(summary(theClogit)$coef)

knitr::kable(theCoef, caption="The coeficients of conditional logistic regression")%>%
  kable_styling(latex_options = "hold_position")
```

Table 4: The coeficients of conditional logistic regression

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