STA442_HW4

Zeyang Zhang

29/11/2019

Smoking Report

Introduction

The data set is from the 2014 American National Youth Tobacco Survey. From the description, we can say that is likely that significant variation amongst the US states exists, and that there is variation from one school to the next. There are two hypotheses to be investigated:

- 1. Geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools.
- 2. First cigarette smoking has a flat hazard function, or in other words is a first order Markov process. Two non-smoking children have the same probability of trying cigarettes within the next month.

Method

First we convert the original time data into the decade scale, from 0.4(age 8 or already a smoker at age 8) to 1.5(individual at age 19) and use left censoring to encode the event into 3 category: 1,observed event, the individual is a smoker; 0,right censored event, the individual is not a smoker when he response to the survey; 2,left censored event, the individual is a smoker before 8 years old. Based on the hypothesis and the following prior information provided by collaborating scientists:

- The variability in the rate of smoking initiation between states substantial, with some states having double or triple the rate of smoking update compared other states for comparable individuals. Set U_i as the random effect for state i, we might see $\exp(U_i) = 2$ or 3 but unlikely to see at 10.
- Within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' schools or $\exp(V_{ij}) = 1.5$ for a school-level random effect is about the largest we'd see.
- A flat hazard function is expected, so the prior on the Weibull shape parameter should allow for a 1 but it is not believed that shape parameter is 4 or 5.

The model should be written as ;

$$
\begin{aligned}
&f(y;\lambda,\alpha) = \alpha y^{\alpha-1}\lambda^{\alpha}e^{-(\lambda y)^{\alpha}} \\
&Y_{ijk} \sim f(y;\lambda,\alpha) \\
&\lambda_{ijk} = e^{-\eta_{ijk}} \\
&\eta_{ijk} = \beta_0 + \boldsymbol{BX}_{ijk} + U_i + U_{ij} \\
&U_i \overset{ind}{\sim} N(0,\sigma_1^2) \\
&U_{ij} \overset{ind}{\sim} N(0,\sigma_2^2)
\end{aligned}
$$

$f(y;\lambda,\alpha)$ is the Weibull distribution defined in INLA with scale parameter λ and shape parameter α.

Y_{ijk} is the surv.inla data for the **k**th individual from **i**th state **j**th school.

η_{ijk} is the linear function for the **k**th individual from **i**th state **j**th school.

β_0 is the intercept of the model

$\boldsymbol{B}$ is the vector of the posterior means.

$\boldsymbol{X}_{ijk}$ is a vector that includes RuralUrban GENDER, RACE infomation of the **k**th individual from **i**th state **j**th school.

U_i is the random effect of **i**th state.

U_{ij} is the random effect of **j**th school in **i**th state.

Prior:

- P($\sigma_1 > \log(3)/2$)= 0.05, this makes $2\sigma_1 \leq log(3)$ under 95% CI. Then $exp(U_i) = 3$ can be the maximum we can expected. It corresponds with the prior information that the variability in the rate of smoking initiation between states substantial, with some states having double or triple the rate of smoking update compared other states for comparable individuals.
- P($\sigma_2 > \log(1.5)/2$)= 0.005, this makes $2\sigma_2 \leq log(1.5)$ under 99.5% CI. Then $exp(U_{ij}) = 1.5$ can be the largest we can expected. It corresponds with the prior information that within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' school.
- The prior gives the expectation of the shape parameter is 1, with precision $(0.7)^{-2}$. This is calculated with $qlnorm(0.99, log(1), 0.7) > 5$. When $\sigma = 0.7$, over 99% chance that the shape parameter can be less than 5. It corresponds with the prior information that the prior on the Weibull shape parameter should allow for a 1 but it is not believed that shape parameter is 4 or 5.

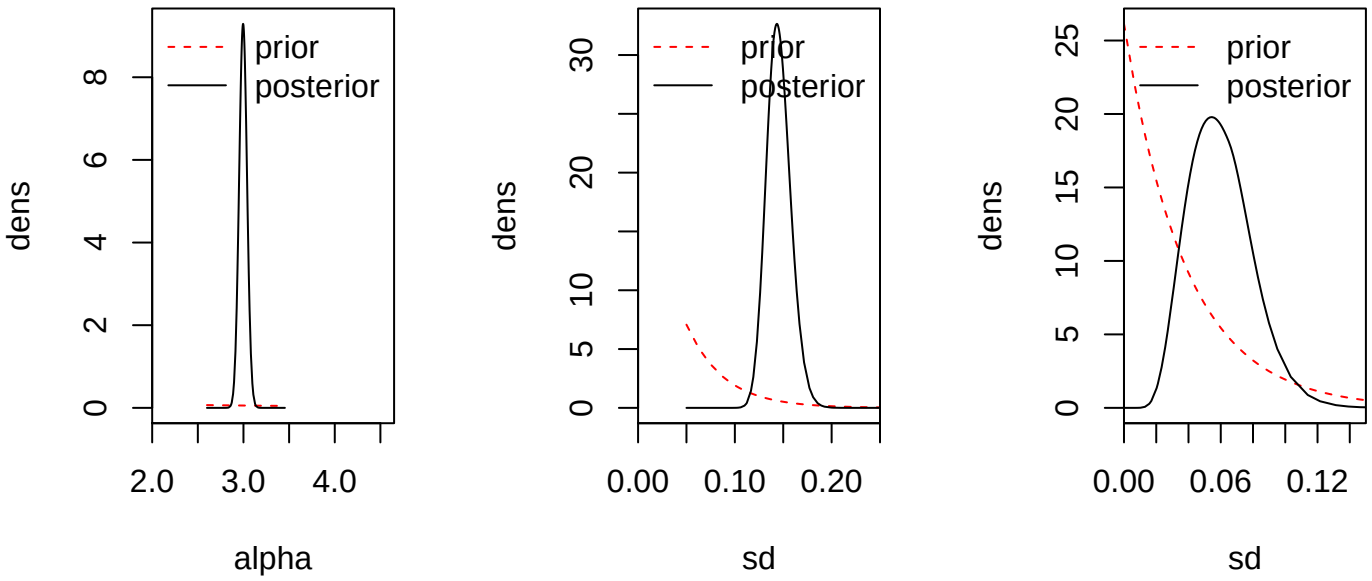
Result

The Summary of model coefficients (after exponential) in the model is given below. We can say that the model gives relatively precise result: most parameter 95% CI do not contain 1 and are significant. Higher exp(coef) gives lower λ (scale parameter) and the first time smoker age become smaller for that group, although the influence of each coefficient is not linear.

Table 1: Summary of exponential coefficients in the model

	mean	0.025quant	0.975quant
(Intercept)	0.539	0.511	0.570
RuralUrbanRural	1.119	1.056	1.186
SexF	0.952	0.933	0.971
Raceblack	0.946	0.914	0.978
Racehispanic	1.034	1.006	1.063
Raceasian	0.825	0.770	0.881
Racenative	1.097	1.011	1.184
Racepacific	1.133	0.982	1.289

The three graphs of prior and posterior densities of model parameters. The first is for shape parameter alpha, the mean is around 3. The second is standard deviation for random effect school, the mean is around 0.15. The third is standard deviation for random effect state, the mean is around 0.06. This result is also shown in other tables.

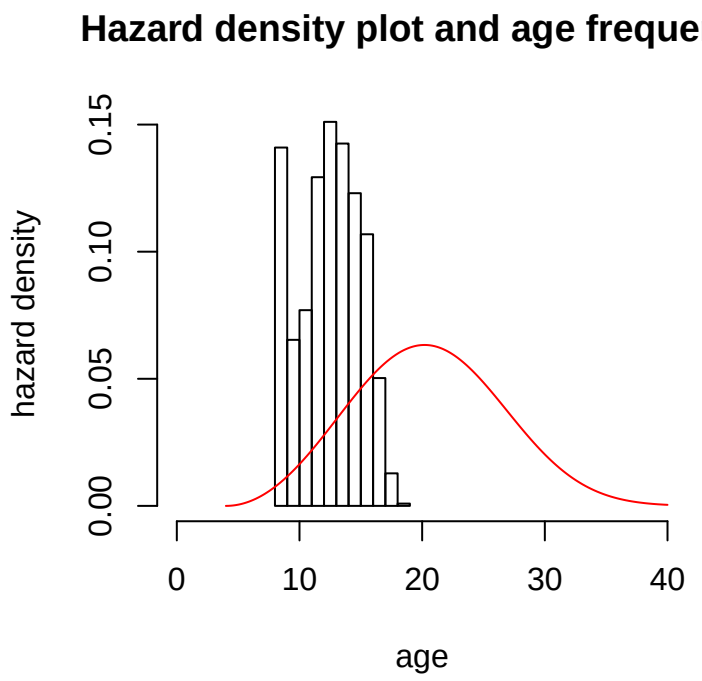


- The first hypothesis is wrong. The Geographic variation (between states) in the mean age children first try cigarettes is less than variation amongst schools. It can be seen from the table below and the postprior density plot above.

Table 2: Summary of random effect in the model

	mean	0.025quant	0.975quant
SD for school	0.145	0.123	0.172
SD for state	0.060	0.027	0.103

- The second hypothesis is also wrong. First cigarette smoking does not have a flat hazard function because the postprior density plot above gives the shape parameter should be around 3. Two non-smoking children have the increasing probability of trying cigarettes within the next month as he gets older. An example of hazard function and age frequency (for intercept group) is given below, we can see the hazards function is not flat when the shape is larger than 1.



Discussion

We fit the model based on the prior assumptions and the hypothesis. The two hypotheses are both wrong so the model might not be perfect. However, the model itself can reject the hypothesis and that should be enough.

Report about the Death on the roads

Introduction

The dataset is a subset of the data from all of the road traffic accidents in the UK from 1979 to 2015. The data consist of all pedestrians involved in motor vehicle accidents with either fatal or slight injuries. The goal is to investigate whether women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood.

Method

First, we notice men are involved in accidents more than 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.

We treat fatal accidents as cases and slight injuries as controls, and use a conditional logistic regression with strata adjust for time of day, lighting conditions, and weather. For each case i we have similar controls from 1 to n. The model we fit can be written as:

$$
\begin{aligned}
&logit[Pr(Y_{ij}=1)] = \alpha_i + X_{ij}\beta \\
&logit[Pr(Y_{ij}=1)|Z_{ij}=1] = \alpha_i^* + X_{ij}\beta \\
&\alpha_i^* = \alpha_i + log[\frac{pr(Z_{ij}=1|Y_{ij}=1)}{pr(Z_{ij}=1|Y_{ij}=0)}]
\end{aligned}
$$

- Y_{ij} is 1 when the ith starta jth control is fatal and is 0 when the ith starta jth control is not fatal.
- Z_{ij} is 1 when the ith starta jth control is selected into the study and is 0 when the ith starta jth control is not selected.
- α_i is the intercept of the model.
- β is the vector of the parameter.
- X_{ij} is a covariate vector of risk factor that includes gender and age infomation of the ith starta jth accident.

Note $\alpha_i^* = \alpha_i + log[\frac{pr(Z_{ij}=1|Y_{ij}=1)}{pr(Z_{ij}=1|Y_{ij}=0)}]$ is not known then the estimate of α would be not available. But clogit fucntion from survival package will give us the relative effect of the covariates.

Result

For male, the age 26-35 data is set as the intercept and all other groups are compared with this group. For female, each age strata is compared with the male strata at the same age.

The coefficient table below shows the proportion of accidents which are fatal is lower for women than for men (higher for men than for women). Except for age 0-6, female has significant lower fatal odd ratio at each age.

Table 3: Coefficient Table

	coef	exp(coef)	se(coef)	Pr(>|z|)
age0 - 5:sexFemale	0.028	1.029	0.055	0.605
age6 - 10:sexFemale	-0.177	0.838	0.051	0.000
age11 - 15:sexFemale	-0.250	0.779	0.047	0.000
age16 - 20:sexFemale	-0.279	0.756	0.052	0.000
age21 - 25:sexFemale	-0.369	0.691	0.063	0.000
age26 - 35:sexFemale	-0.448	0.639	0.052	0.000
age36 - 45:sexFemale	-0.448	0.639	0.052	0.000

	coef	exp(coef)	se(coef)	Pr(>|z|)
age46 - 55:sexFemale	-0.376	0.686	0.048	0.000
age56 - 65:sexFemale	-0.237	0.789	0.040	0.000
age66 - 75:sexFemale	-0.143	0.866	0.032	0.000
ageOver 75:sexFemale	-0.126	0.882	0.027	0.000
age0 - 5	0.132	1.142	0.044	0.003
age6 - 10	-0.320	0.726	0.041	0.000
age11 - 15	-0.383	0.682	0.041	0.000
age16 - 20	-0.443	0.642	0.040	0.000
age21 - 25	-0.268	0.765	0.042	0.000
age 26 - 35	0.000	1.000	0.000	NA
age36 - 45	0.412	1.509	0.039	0.000
age46 - 55	0.768	2.156	0.039	0.000
age56 - 65	1.212	3.361	0.038	0.000
age66 - 75	1.797	6.033	0.036	0.000
ageOver 75	2.396	10.976	0.035	0.000

The plot below shows how much the fatal odd ratio is for female comparing with male at each age. The ratio 1 is marked in black line. Most of 95% Confidence Interval lies below that line, which means the fatal accident odd ratio is lower for women than men. The plot illustrates that women are safer as pedestrians than men, especially as teenagers and in early adulthood. Then the hypothesis is correct.

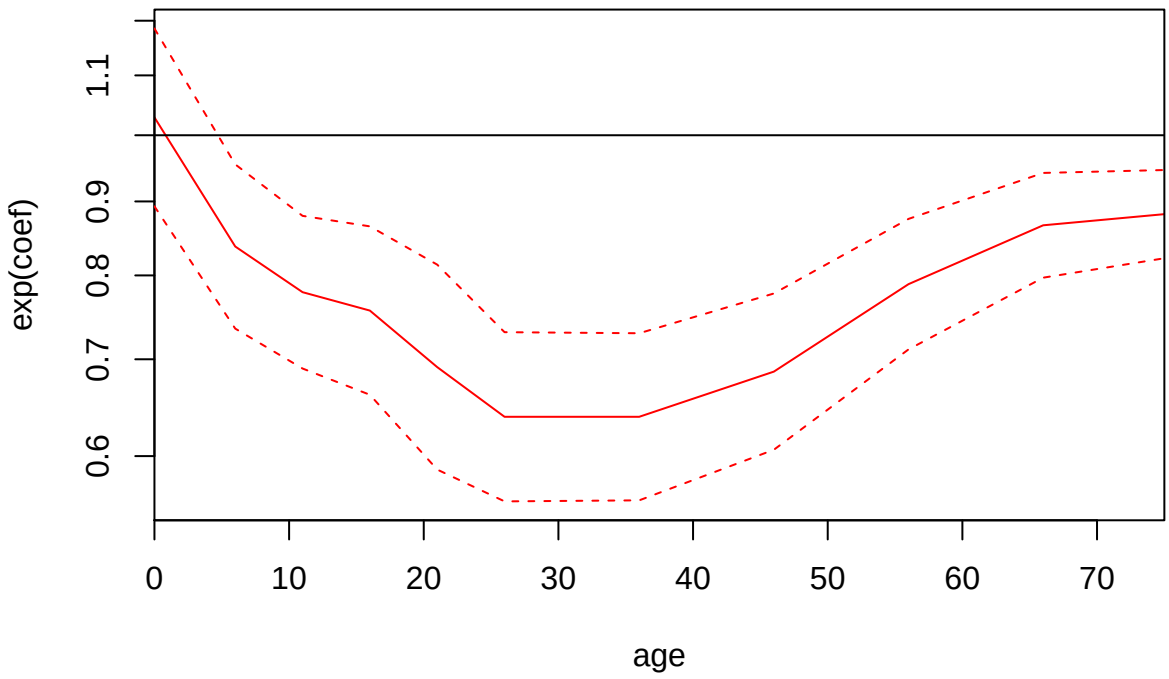


Figure 1: 95% Confidence Interval of the fatal accident odd ratio for female comparing with male

Discussion

The model treats fatal accidents as cases and slight injuries as controls so the the conclusion should better mention that women tend to be safer as pedestrians than men when they are invloved in accidents on roads. Since we do not have other data set. This is the best conclusion we can get.

Appendix

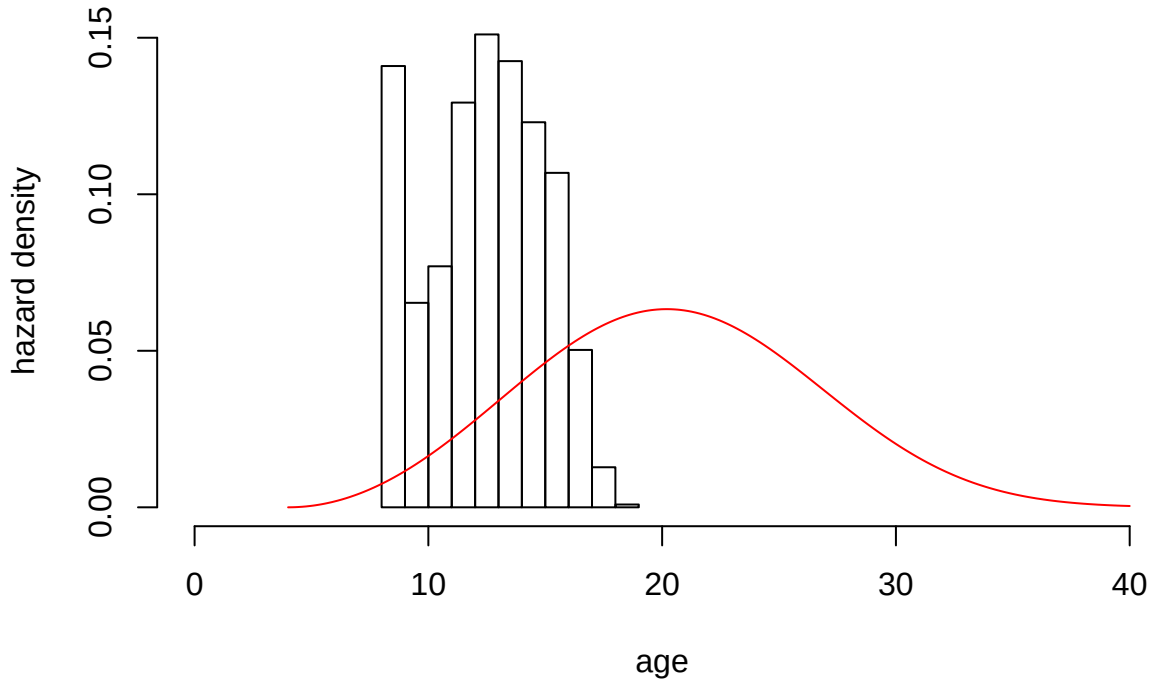
```
# Smoke Report
smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/smoke.RData")
load(smokeFile)
smoke = smoke[smoke$Age > 9, ]
forInla = smoke[, c("Age", "Age_first_tried_cigt_smkg", "Sex", "Race", "state", "school", "RuralUrban")]
forInla = na.omit(forInla)
forInla$school = 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(log(1.5)/2, 0.005)))) +
f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(log(3)/2, 0.05)))),
control.family = list(variant = 1,hyper = list(alpha = list(prior = "normal", param = c(log(1), (0.7)^(-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)     -0.61717748 -0.67102652 -0.56249355
## RuralUrbanRural  0.11274802  0.05438607  0.17072343
## SexF            -0.04945090 -0.06947809 -0.02952459
## Raceblack       -0.05572380 -0.08940363 -0.02244729
## Racehispanic     0.03350156  0.00618553  0.06072409
## Raceasian       -0.19250092 -0.26082518 -0.12718019
## Racenative       0.09212944  0.01092411  0.16893500
## Racepacific      0.12485866 -0.01850427  0.25402256
## SD for school    0.14541787  0.12278731  0.17152728
## SD for state     0.05997851  0.02710544  0.10340331
```

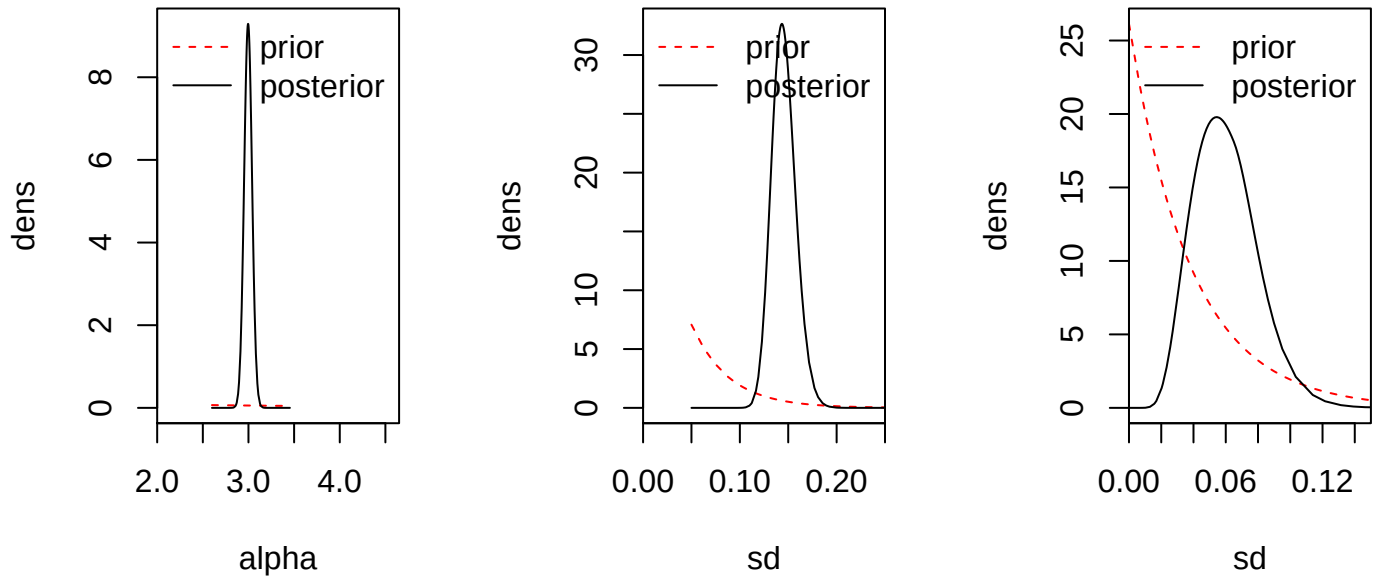
```
qlnorm(0.99,0,0.7)
```

```
## [1] 5.095937
```

```
xSeq=seq(4,40,len=1000)
hist(forInla$Age_first_tried_cigt_smkg, main='',
     xlim=c(0,40), xlab='age', ylab='hazard density', prob=TRUE)
kappa=fitS2$summary.hyperpar['alpha','mode']
lambda=exp(-fitS2$summary.fixed['(Intercept)','mode'])
lines(xSeq, dweibull((xSeq-4)/10, shape = kappa, scale = lambda)/10, col='red')
```




```
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters) {
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
  do.call(legend, fitS2$priorPost$legend)
}
```




```
# Report about the Death on the roads
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$timeC

theTable = table(pedestrians$strata, pedestrians$y)
onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)]
x = pedestrians[!pedestrians$strata %in% onlyOne, ]

library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)

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[,c(1,2,3,5)],digits = 3,caption = "Coefficient Table")
```

Table 4: Coefficient Table

	coef	exp(coef)	se(coef)	Pr(>|z|)
age0 - 5:sexFemale	0.028	1.029	0.055	0.605
age6 - 10:sexFemale	-0.177	0.838	0.051	0.000
age11 - 15:sexFemale	-0.250	0.779	0.047	0.000
age16 - 20:sexFemale	-0.279	0.756	0.052	0.000
age21 - 25:sexFemale	-0.369	0.691	0.063	0.000
age26 - 35:sexFemale	-0.448	0.639	0.052	0.000

	coef	exp(coef)	se(coef)	Pr(>|z|)
age36 - 45:sexFemale	-0.448	0.639	0.052	0.000
age46 - 55:sexFemale	-0.376	0.686	0.048	0.000
age56 - 65:sexFemale	-0.237	0.789	0.040	0.000
age66 - 75:sexFemale	-0.143	0.866	0.032	0.000
ageOver 75:sexFemale	-0.126	0.882	0.027	0.000
age0 - 5	0.132	1.142	0.044	0.003
age6 - 10	-0.320	0.726	0.041	0.000
age11 - 15	-0.383	0.682	0.041	0.000
age16 - 20	-0.443	0.642	0.040	0.000
age21 - 25	-0.268	0.765	0.042	0.000
age 26 - 35	0.000	1.000	0.000	NA
age36 - 45	0.412	1.509	0.039	0.000
age46 - 55	0.768	2.156	0.039	0.000
age56 - 65	1.212	3.361	0.038	0.000
age66 - 75	1.797	6.033	0.036	0.000
ageOver 75	2.396	10.976	0.035	0.000

```r
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 = "red",
        lty = c(1, 2, 2), xaxs = "i",
        xlab="age", ylab="exp(coef)" , main="95% Confidence Interval for female comparing with male")
abline(h=1, col="black")
```