STA442 Assignment 4

Hongbo Du
1003568089
hongbo.du@mail.utoronto.ca

Question 1

Introduction

This report analysis the data from the 2014 American National Youth Tobacco Survey which is available on `http://pbrown.ca/teaching/appliedstats/data/`. The goal is to test two hypothesis. One hypothesis is that 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. the other hypothesis is whether 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.

Methods

The model used in this analysis is the Weibull distribution with more than one random effect, which responses with the states level random effect and school level random effect. Mathematically, it follows

$$Y_{ijk} \sim \text{Weibull}(\rho_{ijk}, \kappa)$$

$$\rho_{ijk} = \exp(-\eta_{ijk})$$

$$\eta_{ijk} = X_{ijk}\beta + U_i + Vij$$

$$U_i \sim \mathcal{N}(0, \sigma_U^2)$$

$$V_ij \sim \mathcal{N}(0, \sigma_V^2)$$

where Y_{ijk} represents each individual in school nested in states, U_i represents the states level random effect, and V_{ij} represents the school level random effect.

Results

Figure 1.1 is the prior and posterior density graph of `weibullsurv`. The red dashed line is the prior distribution and the black line is the posterior distribution. The prior on the Weibull shape parameter is log(1) and $1/\sqrt{2/3}$. The log(1) is due to the expectation of a flat hazard function, and $1/\sqrt{2/3}$ is the precision. So the distribution is a log-normal distribution, mathematically, it is

$$\log \mathrm{N}(\log(1), 2/3)$$

. Notice that the posterior has a mean around 3.0 and it is not a flat line.

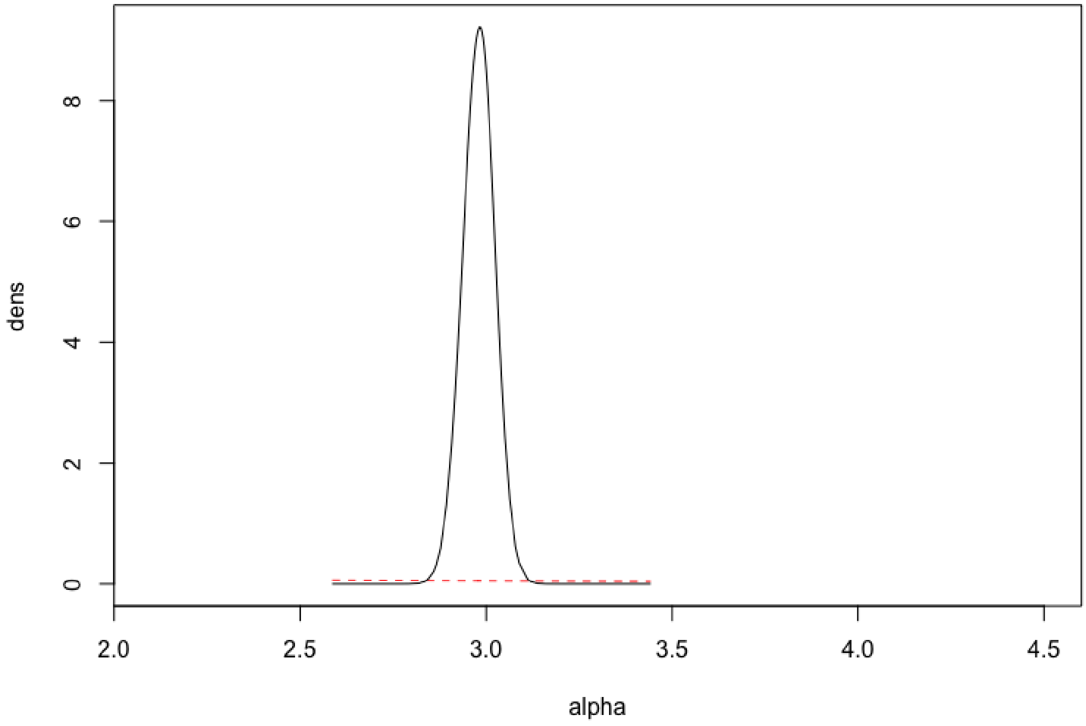

Figure 1.1

Figure 1.2 is the prior and posterior density graph of the random effect `School`. The red dashed line is the prior distribution and the black line is the posterior distribution. We set the PC prior with parameters 0.5 and 0.01. The reason is that according to the prior information, the $\exp(V_{ij}) = 1.5$ for a school-level random effect is the largest we would see. Thus we must have $V_{ij} \approx 0.4$ and so set $\mathrm{P}(> 0.5) = 0.01$.

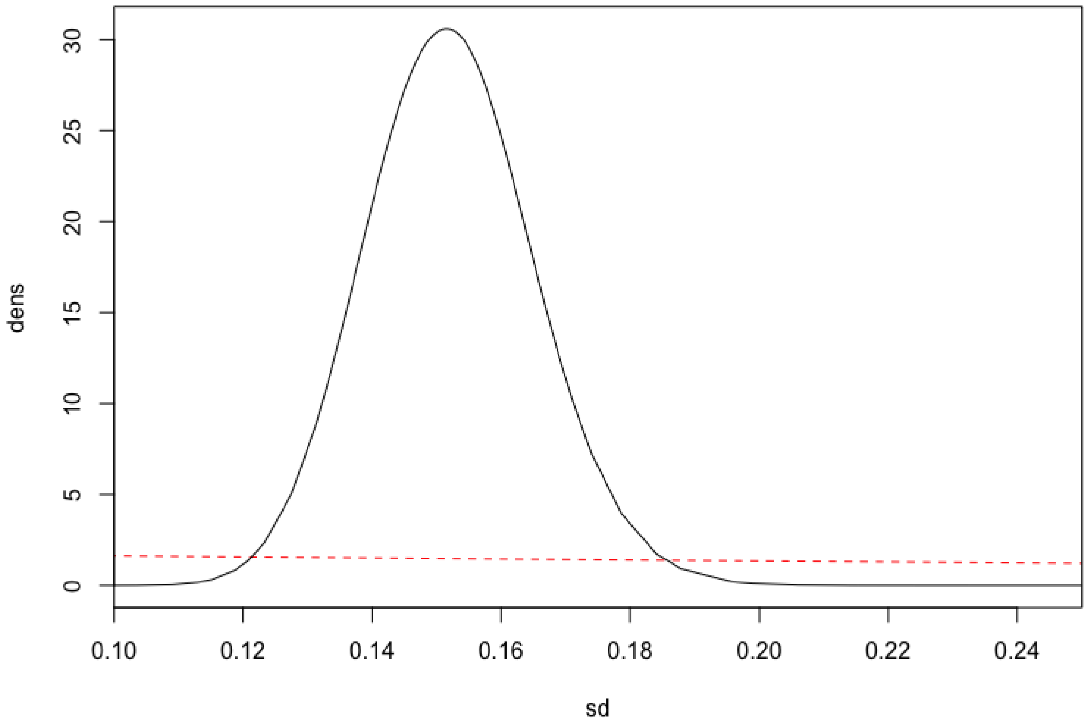

Figure 1.2

Figure 1.3 is the prior and posterior density graph of the random effect `State`. The red dashed line is the prior distribution and the black line is the posterior distribution. We set the PC prior with parameters 2.3 and 0.01. The reason is that according to the prior information, the $\exp(U_i) = 10$ for a state-level random effect is barely to see. Thus we must have $U_i \approx 2.3$ and so set $\mathrm{P}(> 2.3) = 0.01$.

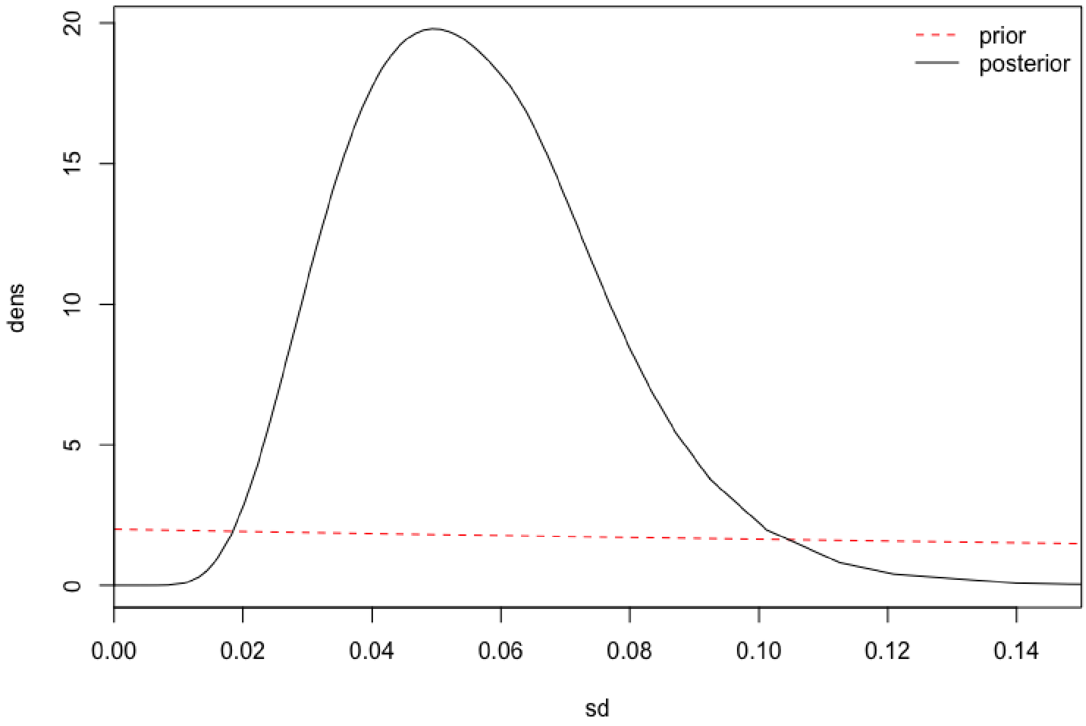

Figure 1.3

Table 1.1 shows that the SD for school has a mean of 0.15 and the SD for state has a mean of 0.06.

	mean	0.025quant	0.975quant
(Intercept)	-0.62	-0.68	-0.57
RuralUrbanRural	0.12	0.06	0.18
SexF	-0.05	-0.07	-0.03
Raceblack	-0.06	-0.09	-0.02
Racehispanic	0.03	0.01	0.06
Raceasian	-0.19	-0.26	-0.13
Racenative	0.09	0.01	0.17
Racepacific	0.12	-0.02	0.25
SD for school	0.15	0.13	0.18
SD for state	0.06	0.02	0.10

Table 1.1

Conclusion

We can conclude that school level has a higher mean than state level, according to table 1.1 and figure 1.2 and 1.3. Therefore we do not conclude the hypothesis that 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 schools with the earliest smoking ages. Secondly, according to figure 1.1, the graph appears to have a mean of 3.0, which implies the shape parameter of weibull is around 3.0 instead of 1. Hence we cannot say that the first cigarette smoking has a flat hazard function. This means two non-smoking children do not 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.

Question 2

Introduction

This report is to analysis the data of the road traffic accidents in the UK from 1979 to 2015. The goal is to assess the hypothesis that whether women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood in UK in road accidents.

Methods

This report uses the case-control method. We set the slight injuries as control, and the fatal injuries as cases. So we use logistic models for case control. Mathematically, it follows that

$$\text{we have } \Pr(Y_i = 1 | X_i, Z_i = 1) = \lambda_i^*$$

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

$$\text{we want } \Pr(Y_i = 1 | X_i) = \lambda_i$$

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

where β is the odds ratio of the model.

Results

The table 2.1 shows the proportion of slight accidents for men is 637919 and for women is 481811; and the the fatal accidents for men is 24429 and for women is 15212.

	Male	Female
Slight	637919	481811
Fatal	24429	15212

Table 2.1

According to the figure 2.1, it is noticeable that the coefficient for men is increasing by age. There is a point where the confidence interval shrink to it. The vertical scale is from 0 up to 10.

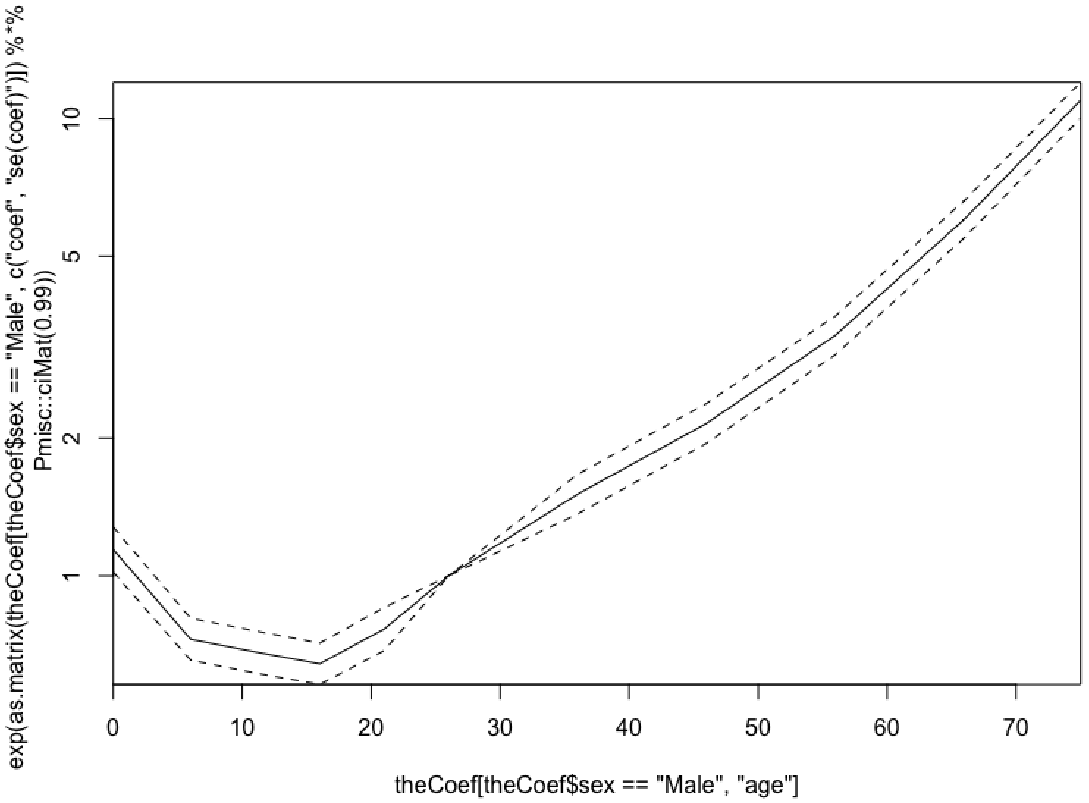

Figure 2.1

According to the figure 2.2, the exponential coefficient for women line is concave up with a minimum around age 30. The vertical scale is from 0 up to 1.2.

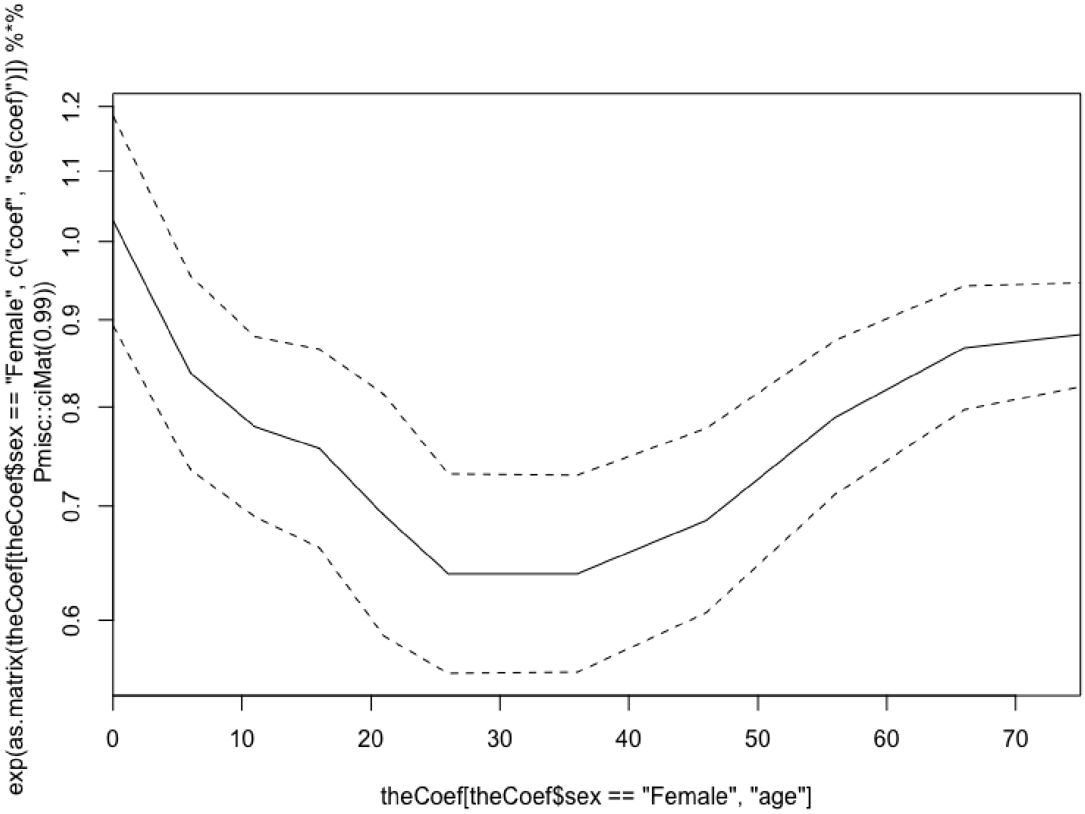

Figure 2.2

According to the table 2.2, we have the similar result obtain from the figures. Notice that female have the lowest exponential value 0.639 at age 26 to 45. Then the table shows that younger and older females have higher exponential coefficients. Female at age 0 has the exponential coefficient of 1.029 with the p-value of 0.605. Similarly, the table shows that males at ages of 16-20 have the lowest exponential coefficient, that is 0.642. Then the exponential coefficients of male are increasing along with the age. In addition, it is obvious that males have higher exponential coefficients after age 21. All of the p-values are around 0 except for female at age 0. Notice that women's standard errors are larger than men's standard errors.

	coef	exp(coef)	se(coef)	z	Pr(> |z|)	sex	age
age0 - 5:sexFemale	0.028	1.029	0.055	0.517	0.605	Female	0
age6 - 10:sexFemale	-0.177	0.838	0.051	-3.490	0.000	Female	6
age11 - 15:sexFemale	-0.250	0.779	0.047	-5.295	0.000	Female	11
age16 - 20:sexFemale	-0.279	0.756	0.052	-5.364	0.000	Female	16
age21 - 25:sexFemale	-0.369	0.691	0.063	-5.828	0.000	Female	21
age26 - 35:sexFemale	-0.448	0.639	0.052	-8.573	0.000	Female	26
age36 - 45:sexFemale	-0.448	0.639	0.052	-8.679	0.000	Female	36
age46 - 55:sexFemale	-0.376	0.686	0.048	-7.792	0.000	Female	46
age56 - 65:sexFemale	-0.237	0.789	0.040	-5.878	0.000	Female	56
age66 - 75:sexFemale	-0.143	0.866	0.032	-4.429	0.000	Female	66
ageOver 75:sexFemale	-0.126	0.882	0.027	-4.606	0.000	Female	75
age0 - 5	0.132	1.142	0.044	3.008	0.003	Male	0
age6 - 10	-0.320	0.726	0.041	-7.822	0.000	Male	6
age11 - 15	-0.383	0.682	0.041	-9.305	0.000	Male	11
age16 - 20	-0.443	0.642	0.040	-10.958	0.000	Male	16
age21 - 25	-0.268	0.765	0.042	-6.355	0.000	Male	21
age 26 - 35	0.000	1.000	0.000	NA	NA	Male	26
age36 - 45	0.412	1.509	0.039	10.648	0.000	Male	36
age46 - 55	0.768	2.156	0.039	19.709	0.000	Male	46
age56 - 65	1.212	3.361	0.038	32.023	0.000	Male	56
age66 - 75	1.797	6.033	0.036	49.447	0.000	Male	66
ageOver 75	2.396	10.976	0.035	68.124	0.000	Male	75

Table 2.2

Conclusion

First of all, the zero coefficient and the shrinkage at age 26 is because it is the reference group. By looking at the table 2.1, the number of men in injuries is higher than women's in both slight injuries and fatal accidents. Notice that for men after age 21, the exponential coefficients for men is higher than women at the same ages. The two vertical scale of the graphs show that men's scale is much greater than women's. In addition, the larger standard errors for women than men's is another result that makes the women's confidence interval look wider than men's. Therefore, from the result of table 2.1 and by comparing the exponential coefficients, we can conclude that men are involved in accidents more than women, and the proportion of accidents which are fatal is higher for men than for women, which is equivalently to say that women as pedestrians is safer than men.
Secondly, since both graph of exponential coefficient are concave up; and the results from table 2.1 shows that teenagers have more accidents than adults. Hence there is no evidence showing that as teenagers and in early adulthood is safer for both male and female.

Appendix

```
# Question 1

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)

# Weibull
fitS2 = inla(smokeResponse ~ RuralUrban + Sex + Race +
             f(school, model = "iid",
               hyper = list(prec = list(prior = "pc.prec", param = c(2.3,
                  0.01)))) +
             f(state, model = "iid", hyper = list(prec = list(prior = "pc.
                prec", param = c(0.5,0.01)))),
           control.family = list(variant = 1, hyper =
                                 list(alpha = list(prior = "normal", param
                                    = c(log(1),(2/3)^(-2))))),
           control.mode = list(theta = c(8, 2, 5), restart = TRUE),
           data = forInla,
           family = "weibullsurv",
           verbose = TRUE)


summ <- rbind(fitS2$summary.fixed[, c("mean", "0.025quant", "0.975quant")],
       Pmisc::priorPostSd(fitS2)$summary[,c("mean", "0.025quant", "0.975
          quant")])
```

```
knitr::kable(summ, digit = 2, "latex", booktabs = T)

plot(fitS2$marginals.fixed$SexF, type="l")
# females tend to smoke later
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)

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

# Question 2

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)

dim(pedestrians)
pedestrians[1:3, ]
knitr::kable(table(pedestrians$Casualty_Severity, pedestrians$sex), "latex",
    booktabs = "T", digit = 3)
range(pedestrians$time)
x = pedestrians[!pedestrians$strata %in% onlyOne, ]
summary(glm(y ~ sex + age + Light_Conditions + Weather_Conditions,
          data = x, family = "binomial"))$coef[1:4, ]

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

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

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


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, "latex", booktabs = "T", digit = 3)

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")
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")
```