STA442 Homework 4

Depeng Ye 1002079500

22/11/2019

Smoke

Background

We analyzed the youth smoking data set of 2014 American National Youth Tobacco Survey to try to assess two hypothesis in our report. The dataset was collected and created by Center of Disease. The first is to whether or not tobacco control programs should target the states with the earliest smoking ages instead of finding particular schools where smoking is a problem. The second hypothesis is that any non-smoking children, with the same known confounders and random effects, have the same probability to begin smoking within the next month regardless of their age. Our conclusions of analysis are that both hypothesis are wrong. The reasonings are shown as follows.

Model

A Weibull Survival model is fitted with school and state as Random Effect, and sex, urban/rural, ethnicity as Known Confounders. Model we used is:

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

$$\rho_{ij} = \exp(-\eta_{ij\kappa})$$
$$\eta_{ij\kappa} = X_{ij\kappa}\beta + U_i + V_{ij}$$
$$U_i \sim N(0, \sigma_U^2)$$
$$V_{ij} \sim N(0, \sigma_V^2)$$

where i is state, j is school and κ is individual person.

We use pc.prec method and the given parameters as our prior and posteriors of State, school, and K. The prior and posterior are ploted in the following appendix. We analysis the prior and posterior using parameters calculated in the following paragraph.

Prior & Posterior for States: $\exp(U_i) = 2 \Rightarrow U_i = 0.693$, $1 - e^{-\lambda t} = 0.01,\ t = 1 \Rightarrow \lambda = 0.010$.
Prior & Posterior for Schools: $\exp(V_{ij}) = 1.5 \Rightarrow V_{ij} = 0.405$, $1 - e^{-\lambda t} = 0.01,\ t = 0.7 \Rightarrow \lambda = 0.014$.

With the fitted model considered, the prior of our model follows Gamma(0.4, 3.1) for the log intercept parameter. Hzazard function were plotted as well.

Results

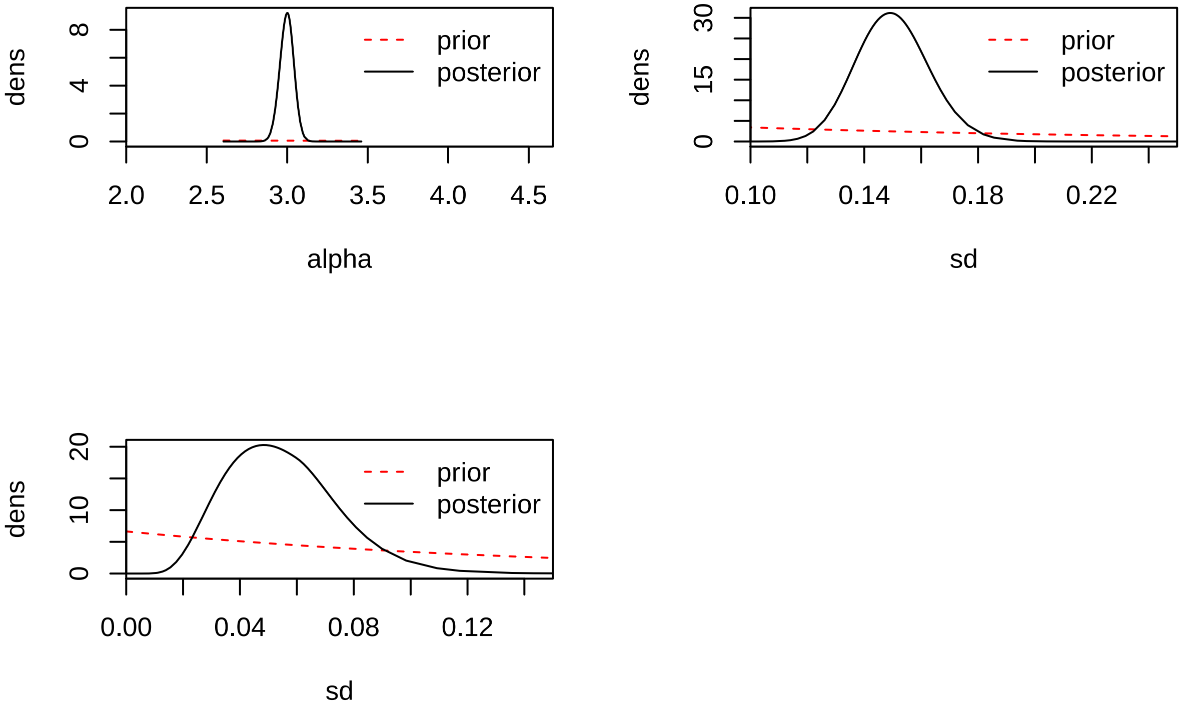
Notice in the table of model outcomes we have that the mean of School is a lot higher than State. This result shows that our first hypothesis is wrong, which means that it is not a proper solution to deal with the tobacco consumptions based on states instead of schools where smoking is a problem. According to the cumulative hazard function plot, we can notice that the cumulative hazard function is not a linear curve. Hence, the hazard function of this data set is not constant. Therefore, our second hypothesis is wrong as well. Because only when hazard function is constant can the hypothesis hold. As a result, when hold all known confounders and random effects identical, two non-smoking children will still have different probability of starting to take tobacco in the next month.

Smoke Question Appendix

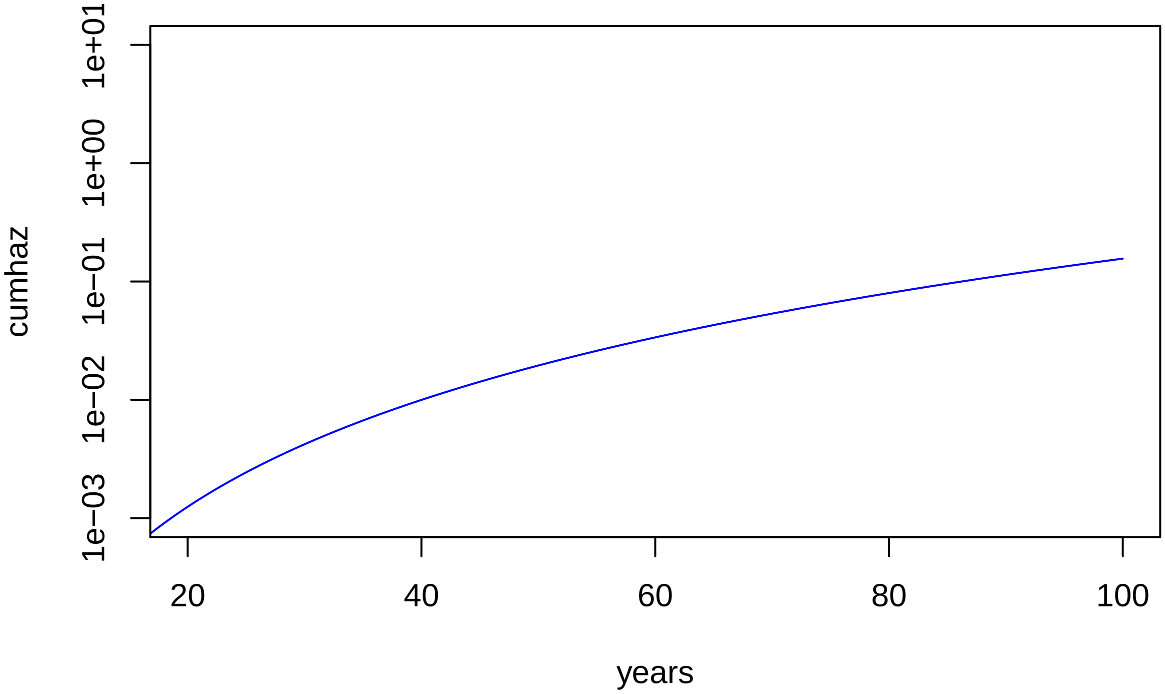
```r
smokeFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/smoke.RData")
load(smokeFile)
smoke = smoke[smoke$Age > 9, ]
forInla = smoke[, c("Age", "Age_first_tried_cigt_smkg",
"Sex", "Race", "state", "school", "RuralUrban")]
forInla = na.omit(forInla)
forInla$school = factor(forInla$school)
library("INLA")

# create data frame of data and situation
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$event)
fitS2 = inla(smokeResponse ~ RuralUrban + Sex * Race +
               f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",
                                                                 param = c(0.693, 0.01)))) +
               f(state, model = "iid", hyper = list(prec = list(prior = "pc.prec",
                                                                param = c(0.405, 0.014)))),
             control.family = list(variant = 1,
                                   hyper =
                                     list(alpha =
                                            list(prior = "normal",
                                                 param = c(log(4), (2/3)^(-2))))),
             control.mode = list(theta = c(8, 2, 5),
                                 restart = TRUE), data = forInla,
             family = "weibullsurv", verbose = TRUE, control.compute = list(config = T))
knitr::kable(
  rbind(fitS2$summary.fixed[,c("mean", "0.025quant","0.975quant")],
        Pmisc::priorPostSd(fitS2)$
          summary[,c("mean", "0.025quant", "0.975quant")]), digits = 3)
```

	mean	0.025quant	0.975quant
(Intercept)	-0.619	-0.673	-0.563
RuralUrbanRural	0.115	0.056	0.174
SexF	-0.050	-0.078	-0.022
Raceblack	-0.048	-0.091	-0.006
Racehispanic	0.026	-0.009	0.060
Raceasian	-0.195	-0.287	-0.108
Racenative	0.110	0.005	0.208
Racepacific	0.176	0.009	0.324
SexF:Raceblack	-0.017	-0.074	0.040
SexF:Racehispanic	0.016	-0.030	0.062
SexF:Raceasian	0.006	-0.122	0.132
SexF:Racenative	-0.044	-0.200	0.110
SexF:Racepacific	-0.170	-0.500	0.123
SD for school	0.150	0.126	0.176
SD for state	0.055	0.023	0.098

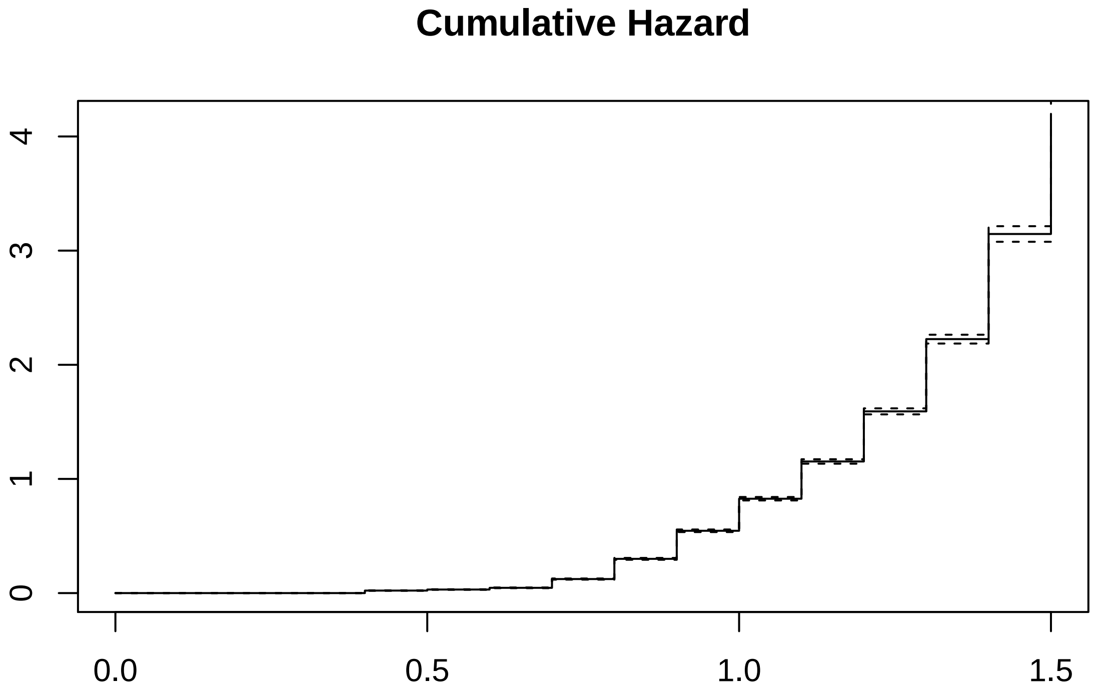
```r
# prior and posterior
par(mfrow = c(2,2))
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters) {
  do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)
    do.call(legend, fitS2$priorPost$legend)
    }

#hazard function
forSurv$ones = 1
xSeq = seq(5, 100, len = 1000)
par(mfrow = c(1, 1))
```



```r
kappa = fitS2$summary.hyper["alpha", "mode"]
lambda = exp(-fitS2$summary.fixed["(Intercept)", "mode"])
plot(xSeq, (xSeq / (100 * lambda)) ^ kappa, col = "blue",
     type = "l", log = 'y', ylim = c(0.001, 10),
     xlim = c(20, 100), xlab = "years", ylab = "cumhaz")
```



```
hazEst = survfit(Surv(time, forSurv$ones) ~ 1, data  = forSurv)
plot(hazEst, fun = "cumhaz", main = "Cumulative Hazard")
```


Cumulative Hazard

Death on the roads

Background

We are having a Casualties involved in reported road accidents dataset published by the British Government collecting all road traffic accident data from 1979 to 2015 in UK. The data segment we are using consist of all pedestrains involved in motor vehicle accidents with either fatal or slight injuries with the pedestrians with moderate injuries removed.

A hypothesis has been made that men are involved in accidents more than women, and the proportion of accidents which are fatal is higher for men than for women, particularly as teenagers and in early adulthood. 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.

Model

In this question, we are not using the glm model. Instead, we are investigating in these hypothesis using conditional logistic model where cases are fatal accidents and controls are slight injuries. Also, the strata used in this model inlcudes Light conditions, weather conditions and time catagories. Mathematiclly, our model look like: 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:

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

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

Result

In the result of Clogit fit, we are interested in the summary of coefficients (Table 5). Take a look at the exp(coef) column. exp(coef) represents the odds of fatal accidents with the reference group being Male26-35 and the odds of this group being 1. We can see that, in general, men are more likely to experience a fatal accident, while women involved in an accident are more likely to be slightly injured. However, when investigated more in depth, we will notice that in the later adulthood instead of teenagers and early adulthood, the odds of fatal accidents are higher. This result implies the hypothesis of “women are especially safer than men as teenagers and in early adulthood” is improper. As we have used case controls on weather, light, and time conditions by putting these factors into the strata of out Conditional Logistic Model, the result of our mode should have already removed the influences of those factors and hence its reliable.

Death on Road Question Appendix

```
pedestrainFile =
  Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time),]

dim(pedestrians)
```

```
## [1] 1159371       6
```

```
knitr::kable(pedestrians[1:3, ])
```

	time	age	sex	Casualty_Severity	Light_Conditions	Weather_Conditions
54	1979-01-01 22:40:00	26 - 35	Male	Slight	Darkness - lights lit	Snowing no high winds
65	1979-01-02 10:40:00	26 - 35	Male	Slight	Daylight	Raining no high winds
79	1979-01-02 14:25:00	46 - 55	Male	Slight	Daylight	Raining no high winds

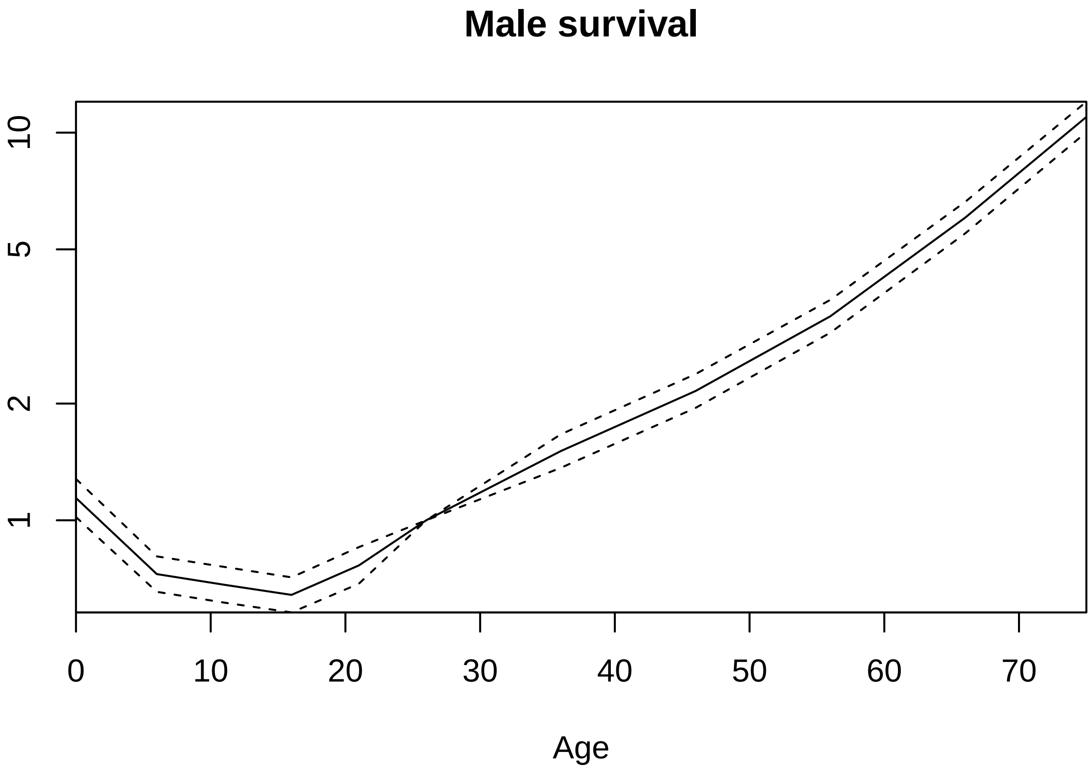
```
knitr::kable(table(pedestrians$Casualty_Severity, pedestrians$sex))
```

	Male	Female
Slight	637919	481811
Fatal	24429	15212

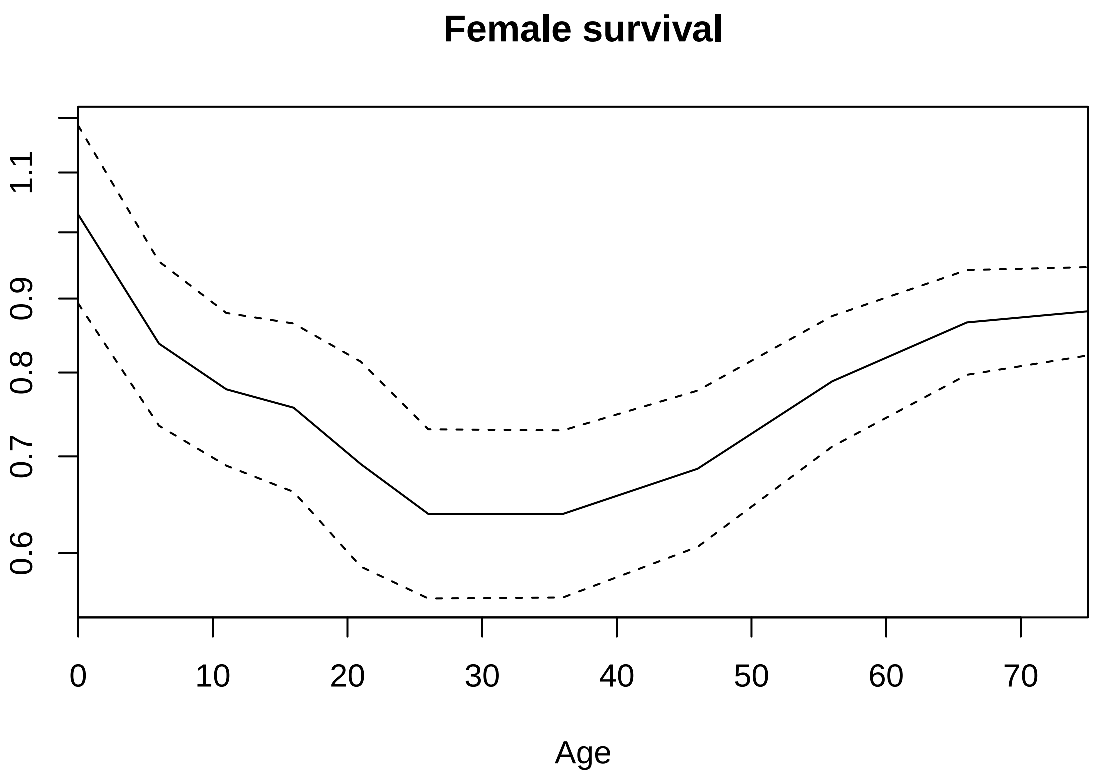
```
range(pedestrians$time)
```

```
## [1] "1979-01-01 01:00:00 EST" "2015-12-31 23:35:00 EST"
```

```
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)
# remove strata with no cases or no controls
theTable = table(pedestrians$strata, pedestrians$y)
onlyOne = rownames(theTable)[which(theTable[, 1] ==
0 | theTable[, 2] == 0)]
x = pedestrians[!pedestrians$strata %in% onlyOne, ]



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$age),
]
matplot(theCoef[theCoef$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 = "",
        xlab = "Age", main = "Male survival")
```



```
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 = "", xlab = "Age", main = "Female survival")
```



```
knitr::kable(summary(glm(y ~ sex + age + Light_Conditions +
                           Weather_Conditions, data = x, family = "binomial"))$
               coef[1:4,], digits = 3, caption = "Logistic fit")
```

Table 4: Logistic fit

	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

```
knitr::kable(summary(theClogit)$coef, digits = 3, caption = "Clogit Fit Coefficient")
```

Table 5: Clogit Fit Coefficient

	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

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
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