STA442 Assignment 4

Yiwen Yang (#1004244800)

Dec.3rd, 2019

1. Analysis of Smoking

Introduction

The dataset we analyzed is the R version of result from the smoking survey, named as 'smoking. R'. The objective is finding difference of age of first smoking between male and female, people live in rural and urban area as well as ethnicities. Further, to investigate the hypotheses where the variation between US state is greater schools'; and whether two non-smoking children have same probability of trying smoking given cofounders and random effects.

Model and Interpretation

We first fitted the interaction model to predict the age of first smoking. However, from the summary of the fitted interaction model, the 95% credible intervals of all interaction terms include zero, indicating they are not significant. Please refer to the summary in the appendix.

The type of censoring is left censoring. In specific, some children already had begun smoking, but we'd like to know exactly when they started. We fitted the model using INLA from the Weibull family to estimate the mean age of first try smoking. The model only includes sex, rural/urban and ethnicity as cofounders, without interaction terms and it will be used entirely to investigate the hypotheses.

Model Assumptions:

$$Y_{ijk} | U_i, V_{ij} \sim Weibull(\lambda_{ijk}, \alpha)$$

$$\lambda_{ijk} = exp(-\eta_{ijk}),\ \eta_{ijk} = X_{ijk}\beta + U_i + V_{ij}$$

$$U_i \sim N(0, \sigma^2_{U_i})$$

$$V_{ij} \sim N(0, \sigma^2_{V_{ij}})$$

Y_{ijk} is the age of first smoking for i^{th} state, j^{th} school, k^{th} person.

$X_{ijk}\beta$ is the charateristics of the person such as rural/urban, sex and ethnicity.

U_i is the state random effect.

V_{ij} is the school random effect.

λ_{ijk} is the scale parameter and α is the shape parameter.

Priors:

The collaborating scientists have provided several prior information. They believe that some states have 2 to 3 times faster compare to individuals in other states. Mathematically, $\exp(U_i)$ = 2 or 3, but unlikely (5%) to exceed 10. We will set 10 as the upper bound, knowing that U_i = log (10) ≈ 2; since U_i follows normal distribution, according to the 68-95-99.7 rule, the 0.975 quantile has the value of 2. Thus, $2\sigma_U = 2 \Leftrightarrow \sigma_V = 1$. This is equivalent to the statement given by the scientists, where seeing the variation larger than 1 is unlikely.

Within a given state, the variation between schools is at most 50% in terms of first try smoking. Thus, $\exp(V_{ij})$ = 1.5 is the maximum random effect we can see. We know that V_{ij} = log (1.5) = 0.41; since V_{ij} follows normal distribution, applying the 68-95-99.7 rule, the 0.975 quantile has the value of 0.41. Thus, $2\sigma_V = 0.41 \Leftrightarrow \sigma_V$ = 0.2. This is equivalent to seeing the variation larger than 0.2 has a probability of 0.05.

Last but not least, the prior on Weibull distribution shape parameter should allow for a 1, but not larger than 4 or 5. A log-normal (log (1), 2/3) prior seems reasonable.

Thus, the three priors for standard deviation of states, standard deviation of school and shape parameter are as follows.

$$\boldsymbol{P(\sigma_U > 1) = 0.05 \Leftrightarrow \sigma_U \sim exp(3)}$$

$$\boldsymbol{P(\sigma_V > 0.2) = 0.05 \Leftrightarrow \sigma_V \sim exp(15)}$$

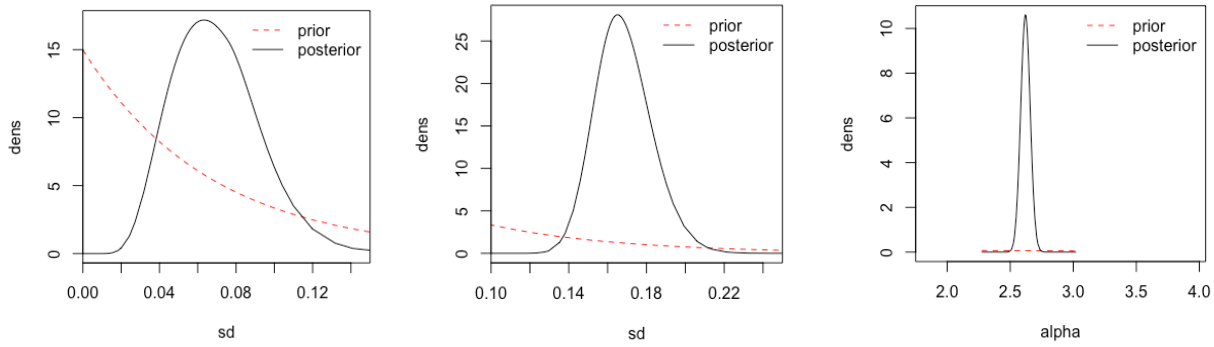
$$\boldsymbol{\alpha \sim log\text{-}Normal\left(log(1), \frac{2}{3}\right)}$$

Table 1 (Estimated exp (U_i), exp(V_{ij}), α at 0.025,0.5,0.975 Quantile)*

	0.025 quantile	quantile mean	0.975 quantile
exp(U_i)	0.135	1.105	7.389
exp(V_{ij})	0.670	1.020	1.492
shape parameter (α)	0.271	1.012	3.694

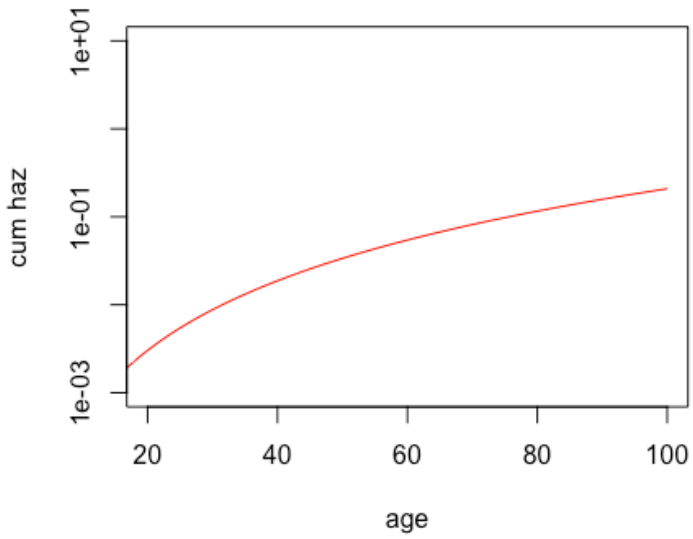
From the table summary, the prior we set is consistent with the assumption. The quantiles of exp(U_i) and exp(V_{ij}) can be calculated by ***exp (c (-2, 0.1, 2) * standard deviation)***. It is very likely to see exp(U_i) to be 1.1, but rarely to see the variation goes beyond 10. The 0.975 quantile for a school-level random effect, exp(V_{ij}), is 1.49, not exceeding the 50% greater rate. We set the 0.025 quantile to 0.27 for α, because the possibility that a child is less likely to smoke as they are getting older exists. A flat hazard is now allowed (CI includes 1), but unlikely to observe value of 4 or 5.

Plot 1 (Prior for σ_U, σ_V , α and Corresponding Posteriors from Left to Right Respectively)*



However, on one hand, the posterior suggests the standard deviation at state-level (σ_U) is likely to be 0.06, **less** than the standard deviation at school level(σ_V), 0.17. We need some extra result to fully reject the first hypothesis. On the other hand, the shape parameter is very likely to be 2.6 rather than 1 from the posterior distribution.

Plot 2 (Cumulative Hazard Function)*



We will have an **increasing hazard** function thus rejecting the second hypothesis where the hazard function is flat. This is because older people are likely to begin smoking, holding other constant.

Table 2 (Summary of Weibull Regression Model)*

	mean	0.025 quantile	0.975 quantile
Intercept	-0.596	-0.658	-0.533
RuralUrbanRural	0.130	0.0628	0.196
SexF	-0.0565	-0.0793	-0.0338
Raceblack	-0.0626	-0.101	-0.0247
Racehispanic	0.0382	0.00703	0.0692
Raceasian	-0.218	-0.296	-0.144
Racenative	0.104	0.0111	0.191
Racepacific	0.142	0.0219	0.289
Standard deviation for school	0.168	0.143	0.199
Standard deviation for state	0.0667	0.0298	0.117

We set **urban area**, **male** and race **White** as baselines. The λ ratio between rural area and urban area is $e^{-0.13} = 0.88 < 1$, which means rural area people are **12% earlier** to begin smoking. For sex comparison, the λ ratio between female and male is $e^{-(-0.0565)} = 1.06 > 1$, suggesting female are **6% later** to first try cigarettes comparing to male. Most ethnicity cofounders are significant because the credible

intervals exclude zero, especially for Asians as the 0.025 and 0.975 quantile has the value relatively far away from zero. The mean for ethnicity groups fluctuates between -0.218 to 0.142. Pacific children are 14% earlier to begin smoking whereas Asians are 24% later to smoke comparing to White people.

The within group variability is 0.168/ (0.168+0.0667) = 0.72, and the between group variability is 0.0667/ (0.168+0.0667) = 0.28. The higher value of within-group variability leads to the result, where the variation in mean age of children first try smoking between school is considerably greater than variation amongst American states. We will reject the first hypothesis.

Conclusion

The smoking result of 2014 American National Youth Tobacco Survey is as follows. Males are 6% earlier to smoke than females, rural dwellers start smoking 12% earlier than urban populations; the mean age of first try cigarette varies between ethnicity. The variation between US **schools** is **larger** than variation within each **state**, thus we recommend the tobacco control programs to target school-wise rather than state-wise. We have evidence that first cigarette smoking has an increasing hazard function, in another word, two non-smoking children does not have same chance of trying cigarettes within the next month; and the probability of smoking increases as they are getting older.

Appendix

Please refer to the last several pages of the PDF file.

2. Analysis of Death on the Roads

Introduction

We have analyzed the UK road accident data from 1979 to the end of 2015. The R version of the road accident data is named as 'pedestrians.rds', which can be downloaded on http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds page. Our primary focus is to investigate the hypothesis whether women pedestrians are safer compared to men; teenagers and people in early adulthood are safer than older population in terms of road crossing.

Methods and Interpretation

The unconditional logistic model may be inappropriate to use in this case because it is possible that women are not likely to go outdoor during poor weather or late at night but rather stay at home. For men, their routine may less likely be affected by the weather condition. To determine the proportion of fatal accidents between men and women, what we are interested in is the predictor sex, age and their interactions. However, sex will have interactions between time, weather and light condition as well. The result from this model will be biased.

Here is the unconditional generalized logistic model.

$$Pr\,(Y_{ij} = 1 | X_{ij}) = \lambda_{ij}$$

$$log\,[\lambda_{ij} / (1 - \lambda_{ij})] = \beta_0 + \sum_{p=1}^{2} X_{ip}\beta_p$$

Instead, to control covariate time, weather and light, we can stratify the data to several groups with case and control within each group. Each group has the same time of day, weather condition and light condition (daylight, darkness, etc.). The control is people having slight injuries whereas the case is for fatal accidents. To have a matched case-control design, we want the number of fatal accidents and slight injuries to be similar.

Down below is the fitted conditional logistic model after stratification. This model will be used for investigating the hypotheses. Some parameters are from the unconditional logistic model.

$$Pr\,(Y_{ij} = 1 | X_{ij},\, Z_{ij} = 1) = \lambda^*_{ij}$$

$$log\,[\lambda^*_{ij}/\,(1-\lambda^*_{ij})] = \beta^*_0 + \sum_{p=1}^{2} X_{ip}\beta^*_p$$

$$\beta^*_p = \begin{cases} \beta_0 + log\left[\dfrac{Pr(Z_{ij}=1\,|Y_{ij}=1)}{Pr(Z_{ij}=1|Y_{ij}=0)}\right] & p=0 \\ \beta_p & p\neq 0 \end{cases}$$

Y_{i1} is case *i* for fatal accident, Y_{i2} is control *i* for slight injuries.

X_{ij} are covariates not used in matching (age and sex).

Z_{ij} is the given time of day, weather and light conditions.

λ^*_{ij} is the probability of fatal accident for case *i, and j = 1,2.*

i is the strata (group), for each case *i,* we want to find similar number of controls.

Table 1 (Summary of Conditional Logistic Model Interaction with Females)*

Female Age Groups	exp(coefficient)	p-values
age0-5:sexFemale	1.0288306	6.049967e-01
age6-10:sexFemale	0.8376825	4.838772e-04
age11-15:sexFemale	0.7789087	1.188378e-07
age16-20:sexFemale	0.7564399	8.149970e-08
age21-25:sexFemale	0.6913389	5.607500e-09
age26-35:sexFemale	0.6387693	1.007818e-17
age36-45:sexFemale	0.6387573	3.980319e-18
age46-55:sexFemale	0.6863891	6.604004e-15
age56-65:sexFemale	0.7889379	4.156393e-09
age66-75:sexFemale	0.8664448	9.465729e-06
ageOver75:sexFemale	0.8819582	4.101941e-06

From the summary above, take a look of at the exponentiated coefficient. It represents the ratio of odds of female getting car accident to same aged male's odds. The interaction terms with female all has value less than one except for a new born female baby. Thus, females are safer pedestrians generally. Women at their mid-age (36-45) obtain the lowest ratio, but the ratio gradually rises after 46 years old. The p-value suggests theses interaction covariates are important.

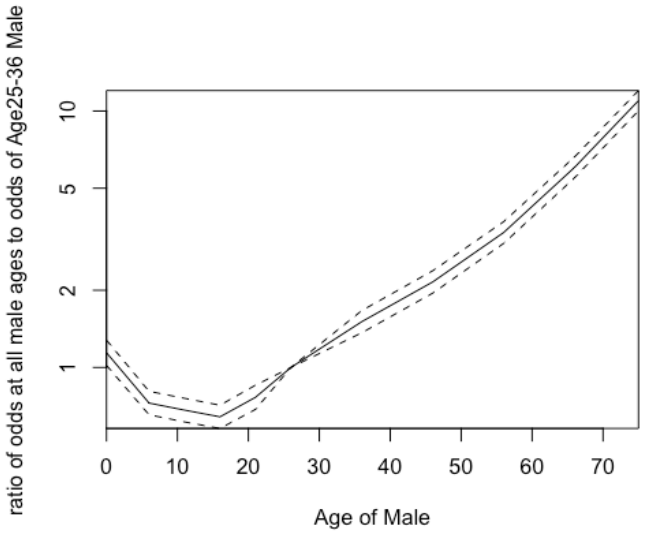
Table 2 (Summary of Conditional Logistic Model Males Comparison)*

Male Age Groups	exp(coefficient)	p-values
age 0-5	1.1415744	2.628711e-03
age 6-10	0.7263965	5.185454e-15
age 11-15	0.6818549	1.336009e-20
age 16-20	0.6419718	6.099840e-28
age 21-25	0.7648419	2.083908e-10
age 26-35	1.0000000	NA
age 36-45	1.5091267	1.782514e-26
age 46-55	2.1559445	1.812237e-86
age 56-65	3.3605244	5.254565e-225
age 66-75	6.0330360	0
age Over75	10.9759044	0

We have set male aged from 26-35 as baseline. The exponentiated coefficient table is ratio of odds at the age group to odds of baseline. A new born male baby is riskier than a grownup male. The ratios are less than one before the age of 26-35, meaning the male teenagers or in early adulthood are safer. However, the ratio rises from age of 36, until the age of 75, the ratio has a surprising value of 11. It means elders are much more likely to engage risky behaviors. The p-values suggests all these factors are significant.

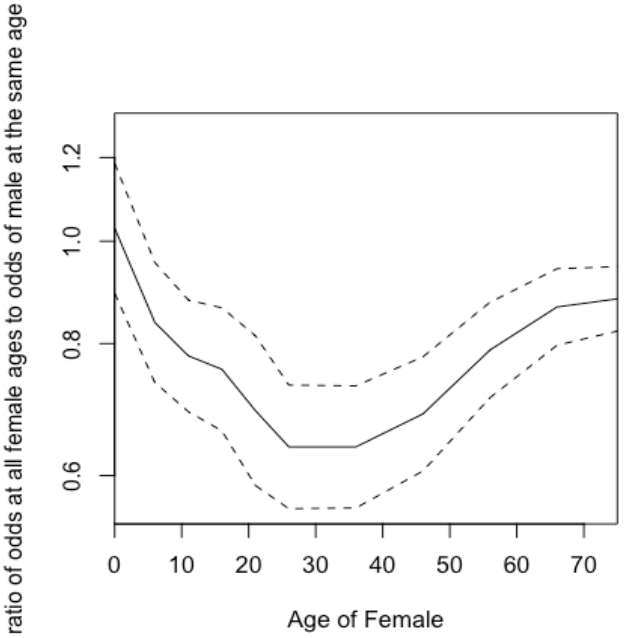
Analyzing Plots

Plot 1 (ratio of odds at all male ages to odds of Age 25-36 Male)*



The black line is the mean of prediction, and the dotted line indicates 0.005 quantile and 0.995 quantiles. As we can see from the plot, a new born baby has a ratio of 1.2, meaning he is slightly riskier than a 26 to 35-year-old male (reference group). The ratio decreases to around 0.3 at the age of 16, the odds of getting involved in fatal accident for a 16-year-old teenager is only 30% to the odds of 35-year-old male. The ratio increases greatly as the male gets older. The odds of getting fatal injury for a 70-year-old man is 10 times to the odds of our reference group's! Since male with age 26 to 35 is our baseline, the ratio to itself has a value of one with no lower and upper quantile. This explains the disappearance of dotted line between age of 25 to 30 in the plot.

Plot 2 (ratio of odds at all female ages to odds of male at the same age)*



A new born female baby has the same odds of getting severely injured in a car accident to male baby. Until the age of 25, female behaves safer, and the ratio decreases to 0.65. They continue to engage safer behaviors until 36-years-old. Then the ratio gradually rises to 0.9 as the age of 70. The odds of getting fatal car crash for 70-year old women is 90% of odds of same aged male pioneers. Overall, for older males and females, both gender seems to engage risky behaviors. Note that the credible interval is wider because the scale of y-axis is smaller. The width of credible interval should be the same for plot 1&2 if under the same scale.

Conclusion

The odds of getting a fatal injury for men increases steeply from the age of 18. The interesting thing is, teenager males or in the early adulthood are particularly safer compared to all other ages of male. For women pedestrians, the ratio of the odds to same aged man's is less than 1, thus women are generally safer to men. This is due to the fact females are reluctant to go outdoors late at night and under bad weathers. However, even though female teenagers or in their early adulthood is less likely to be involved in a serious road accident, it is not the age they engage the safest behavior. The lowest odds ratio is 0.65 at around their 30's, meaning the odds to get fatal accident for female is only about 65% to a same aged male. Overall, we **fail to reject** the hypothesis where women are safer than men, but do **have evidence to reject** teenager females or in early adulthood is the safest age.

Reference

1. http://pbrown.ca/teaching/appliedstats/slides/survival.pdf
2. http://pbrown.ca/teaching/appliedstats/slides/casecontrol.pdf

Appendix

Code for Analyzing Smoking

```
load('smoke.RData')
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)-5)/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$event)
fitS2 = inla(smokeResponse ~ RuralUrban * Sex * Race +f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",param = c(0.2, 0.05))))
        + f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(1,0.05)))),
        control.family = list(variant = 1,hyper = list(alpha = list(prior = "normal", param = c(log(1),(2/3)^(-2))))),
        control.mode = list(theta = c(8,2, 5), restart = TRUE), data = forInla, family = "weibullsurv",verbose = TRUE)
rbind(fitS2$summary.fixed[, c("mean", "0.025quant","0.975quant")], Pmisc::priorPostSd(fitS2)$summary[,c("mean", "0.025quant",
"0.975quant")])

alpha.prior = c(log(1),2/3) #assume flat hazard rate
exp(qnorm(c(0.025, 0.5, 0.975), mean = log(1), sd = 2/3))

#for state
exp(c(-2,0.1, 2) * 1)

#for school
exp(c(-2, 0.1,2) * 0.2)


#inla plots
```

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

```
fitS3 = inla(smokeResponse ~ RuralUrban + Sex + Race +f(school, model = "iid", hyper = list(prec = list(prior = "pc.prec",param = c(0.2, 0.05))))
        + f(state, model = "iid",hyper = list(prec = list(prior = "pc.prec", param = c(1,0.05)))),
        control.family = list(variant = 1,hyper = list(alpha = list(prior = "normal", param = c(log(1),(2/3)^(-2))))),#informative prior
        control.mode = list(theta = c(8,2, 5), restart = TRUE), data = forInla, family = "weibullsurv",verbose = TRUE)
rbind(fitS3$summary.fixed[, c("mean", "0.025quant","0.975quant")], Pmisc::priorPostSd(fitS3)$summary[,c("mean", "0.025quant", "0.975quant")])
```

Interaction Factor of Fitted Interaction Model

	mean	0.025 quantile	0.975 quantile
RuralUrbanRural:SexF	-0.0528	-0.120	0.0144
RuralUrbanRural:Raceblack	-0.0740	-0.172	0.0242
RuralUrbanRural:Racehispanic	-0.0265	-0.107	0.0533
RuralUrbanRural:Raceasian	-0.0882	-0.344	0.146
RuralUrbanRural:Racenative	0.124	-0.124	0.390
RuralUrbanRural:Racepacific	0.122	-0.237	0.490
SexF:Raceblack	-0.0342	-0.131	0.0626
SexF:Racehispanic	0.0160	-0.0606	0.0925
SexF:Raceasian	-0.103	-0.286	0.0749
SexF:Racenative	0.110	-0.179	0.403
SexF:Racepacific	-0.283	-0.937	0.254
RuralUrbanRural:SexF:Raceblack	0.0196	-0.112	0.151
RuralUrbanRural:SexF:Racehispanic	-0.0196	-0.127	0.0878
RuralUrbanRural:SexF:Raceasian	0.293	-0.0251	0.621
RuralUrbanRural:SexF:Racenative	-0.262	-0.636	0.106
RuralUrbanRural:SexF:Racepacific	0.125	-0.585	0.909

Code for Analyzing Death on Roads

```
library('R.utils')
pedestrainFile = Pmisc::downloadIfOld("http://pbrown.ca/teaching/appliedstats/data/pedestrians.rds")
pedestrians = readRDS(pedestrainFile)
pedestrians = pedestrians[!is.na(pedestrians$time),] #delete na in variable 'time',the reason is
pedestrians$y = pedestrians$Casualty_Severity == "Fatal"# create binary outcome
pedestrians$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$y)
onlyOne = rownames(theTable)[which(theTable[, 1] == 0 | theTable[, 2] == 0)] #remove unmatched observation(not balanced)
x = pedestrians[!pedestrians$strata %in% onlyOne, ] #select the data with matched case
dim(pedestrians)
dim(x)

#Unconditional logistic model
```

```
summary(glm(y ~ sex + age + Light_Conditions + Weather_Conditions,data = x, family = "binomial"))$coef

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

theCoef = rbind(as.data.frame(summary(theClogit)$coef),`sexMale:age26 - 35` = c(0, 1, 0, NA, NA))
rownames(theCoef)[1] <- 'sexFemale:age26 - 35'
theCoef$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",xlab = 'Age of Male',ylab= 'ratio of odds at all male
ages to odds of Age25-36 Male' )

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",xlab = 'Age of Female',ylab= 'ratio of odds at all
female ages to odds of male at the same age')
```