STA442 Homework4-Survival

Yingzhu Zhang

03/12/2019

1. Smoking

Introduction

We analyzed the 2014 American National Youth Tobacco Survey using an R version of the dataset. The original dataset was released by the Center for Disease Control. The data was collected from a survey administered to 258 Schools across the United States. We want to investigate whether geographic variation (between states) in the mean age children first try cigarettes is substantially greater than variation amongst schools. Also, 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.

Method

Since children only start smoking for the first time once, we modelled the data using a Weibull distribution, as is convention for such survival analysis data. The specific model we used is as follows:

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

Where:

- state i, school j, individual k
- $X_{ij}\beta$ is the fixed effect for the subjects gender, ethnicity, whether they are from a rural or urban school
- U_i is the state random effect
- V_{ij} is the school random effect
- The variance of U_i and V_{ij} are themselves hyperparameters following a Normal distribution
- κ is the Weibull shape parameter and is normally distributed with its own hyperparameters.

We selected the hyperparameters of the above model using information from the collaborating scientists. For state parameter, we expected $exp(U_i) = 2$ or 3, but unlikely to see at 10. When $exp(U_i) = 2$, we got $U_i = 0.69$; when $exp(U_i) = 10$, we got $U_i = 2.30$. By unlikely we interpreted it as with less then 1% probability, and the rate ratio is between $exp(U_i \pm 2\sigma)$ and $exp(U_i)$. As such we chose our prior such that $U_i = 0.69$ is very likely to be included within $2\sigma = 2$ range, and $U_i = 2.30$ is very unlikely to see within the range. Therefore, the prior for state is to have the standard deviation exceeding 1 around 1% of the time.

For school parameter, within a given state, the 'worst' schools are expected to have at most 50% greater rate than the 'healthiest' schools or $exp(V_{ij}) = 1.5$, which means we'd like to include $V_{ij} = 0.4$ when we set our prior. Same reasoning as state parameter, we chose our prior for school to have the standard deviation exceeding 0.4 around 1% of the time.

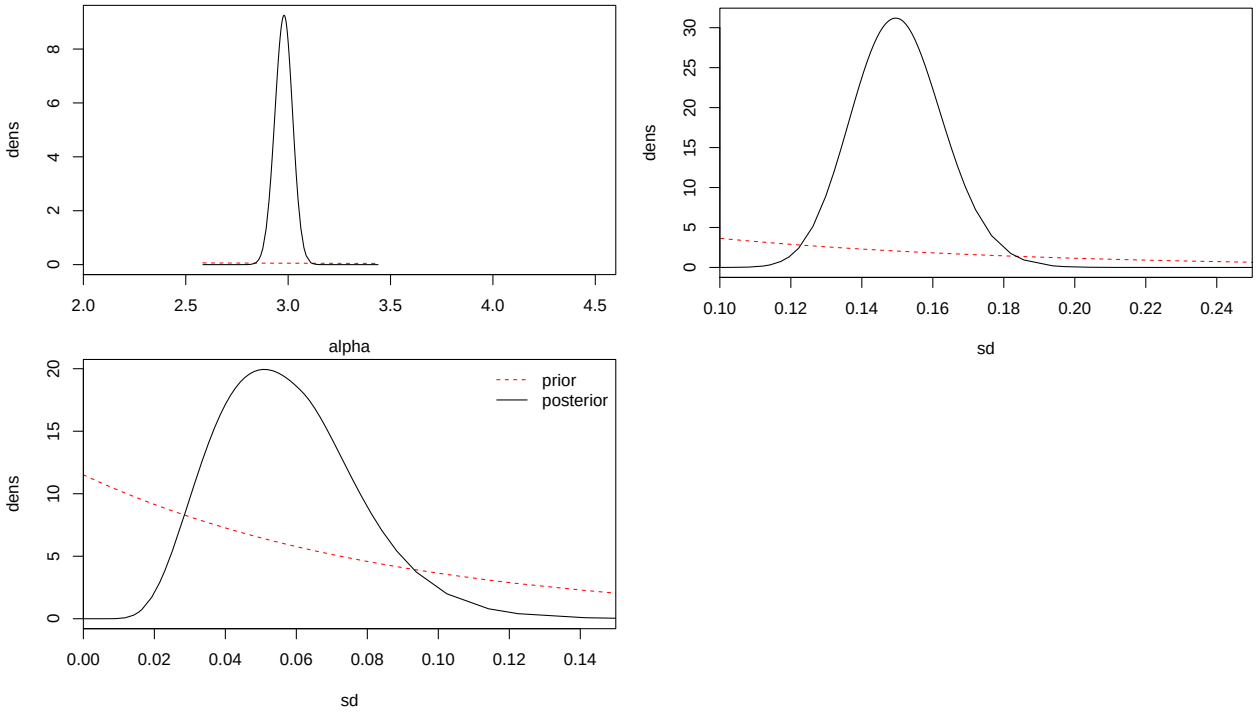
For prior on the Weibull shape parameter, it is allowed for a 1 but it is not believed that shape parameter is 4 or 5. In our model, the distribution of the prior is log-normal. Hence, $Log - Normal(log(1), 2/3)$ prior seems reasonable, with mean=log(1), standard deviation=2/3. The 95% confidence interval of the prior is (0.27, 3.69) as shown above.

Table 1: Prior for Weibull shape with 95% Confidence Interval

2.5% quantile	mean	97.5% quantile
0.2707275	1	3.693752

Result

We plotted the prior and posterior of each parameters in figures below. Two in the first row are the Weibull shape and school, the bottom one is state. From the Weibull shape figure, we can see the highest posterior (λ) after fitting the model is around 3, which is greater than 1, indicating an increasing hazard funtion. From school and state figure, fistly, we can see an approximately exponential shape of the prior graph for both parameters; secondly, the posterior for school is much more greater than it for state. Provided visual evidence for greater variation among shcools rather than among states.

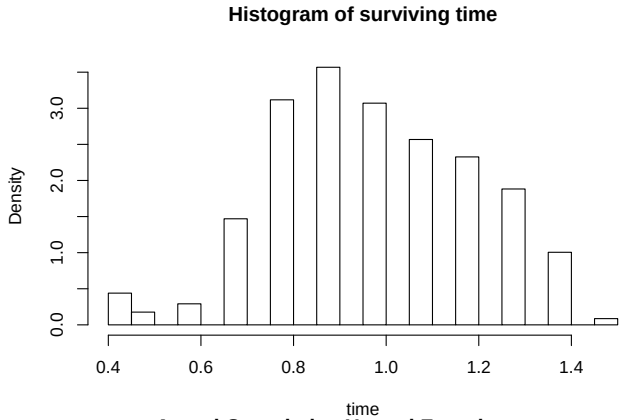


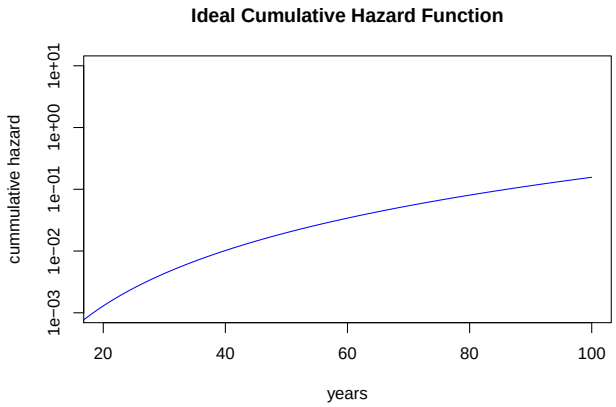
Moreover, the table below shown the Standard Deviation for School and State with 95% Confidence Interval. Geographic variation (between states) in the mean age children first try cigarettes is 0.058, and the school variation is 0.150. Hence, the hypothesis of geographic variation is substantially greater than variation amongst schools is surprisingly incorrect. As a result, tobacco control programs should target the schools with the earliest smoking ages, also slightly concern themselves with finding particular state where smoking is a problem.

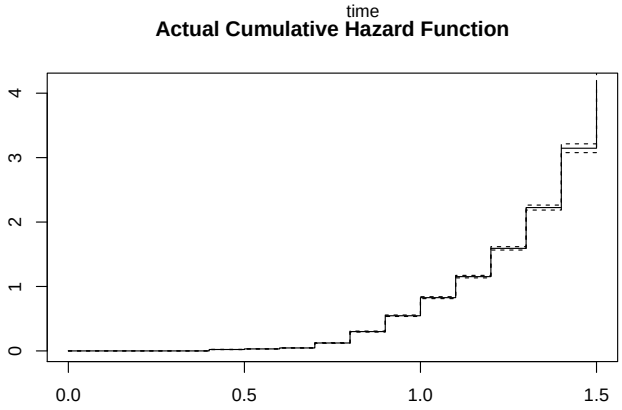
Table 2: Standard Deviation for School and State with 95% Confidence Interval

	mean	0.025quant	0.975quant
SD for school	0.1503061	0.1262391	0.1765229
SD for state	0.0576209	0.0250717	0.1024117

The figure below shows the histogram of the time before a subject smoking for the first time, and two cumulative hazard funtion figures. We want the cumulative hazard function to be smooth linear as shown in Ideal Cumulative Hazard Function figure, which indicates that two non-smoking children have the same probability of trying cigarettes within the next month. However, the Actual Cumulative Hazard Function figure shows that the hazard function in increasing, moreover, as mentioned previously, the highest posterior (λ) after fitting the model is around 3. Both visually and mathematically we can say that older children have higher probablity of trying cigarettes within the next month than relatively younger children.







Conclusion

In this analysis, we investigated two hypothesises, and the model actully did not conform to the expectations. We found that tobacco control programs should target the schools with the earliest smoking ages, also slightly concern themselves with finding particular state where smoking is a problem. Also, older children have higher probablity of trying cigarettes within the next month than relatively younger children, irrespective of their ages but provided the known confounders (sex, rural/urban, etnicity) and random effects (school and state) are identical. These surprising results imply that further studies should be done.

2. Death on the roads

Introduction

In this analysis, we analyzed the dataset below consist of with all of the road traffic accidents in the UK from 1979 to 2015. All pedestrians involved in motor vehicle accidents with either fatal or slight injuries (pedestriansn with moderate injuries have been removed). We assessed whether the UK road accident data are consistent with the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood. Treat fatal accidents as cases and slight injuries as controls, and use a conditional logistic regression to adjust for time of day, lighting conditions, and weather.

Method

Since we treated fatal accidents as cases and slight injuries as controls, and use a conditional logistic regression to adjust for time of day, lighting conditions, and weather, we have models as following:

We want:

$$pr(Y_i = 1|X_i) = \lambda_i$$

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

After setting case and control, and stratification, we have:

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

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

By the previous theorem, for β_p^*:

$$\beta_p^* = \beta_0 + log[pr(Z_i = 1|Y_i = 1)/pr(Z_i = 1|Y_i = 0)] \quad p = 0$$

$$\beta_p^* = \beta_p \quad p \neq 0$$

Where:

- $Y_i = 1$ is the event of death
- $Z_i = 1$ is the strata

We stratified the male and female sample population based on their casualty severity, light conditions and weather conditions. The specific stratifications are shown in the tables below:

Table 3: Casualty Severity

	Male	Female
Slight	637919	481811
Fatal	24429	15212

Table 4: Light Conditions

	Male	Female
Daylight	468340	389257
Darkness - lights lit	170461	97134
Darkness - lights unlit	4068	2506
Darkness - no lighting	13710	4763
Darkness - lighting unknown	5769	3363
Data missing or out of range	0	0
missing	0	0

Table 5: Weather Conditions

	Male	Female
Fine no high winds	561181	415151
Raining no high winds	74584	61333
Snowing no high winds	3722	2812
Fine + high winds	11100	8300
Raining + high winds	8332	7200
Snowing + high winds	803	567
Fog or mist	2626	1660
Other	0	0
Unknown	0	0
Data missing or out of range	0	0
missing	0	0

Result

The following table shown the summary of coefficents of conditional logistic regression model. By comparing exp(coef), which is the odds ratio of female or male death on the road, we can investigate the hypothesis. Female odds ratios are relative to males at each corresponding age. Female odds ratios are mostlt below 1. This implies that women tend to be, on average, safer as pedestrians than men, since female overall have siginificantly smaller odds of having accidents on the road.

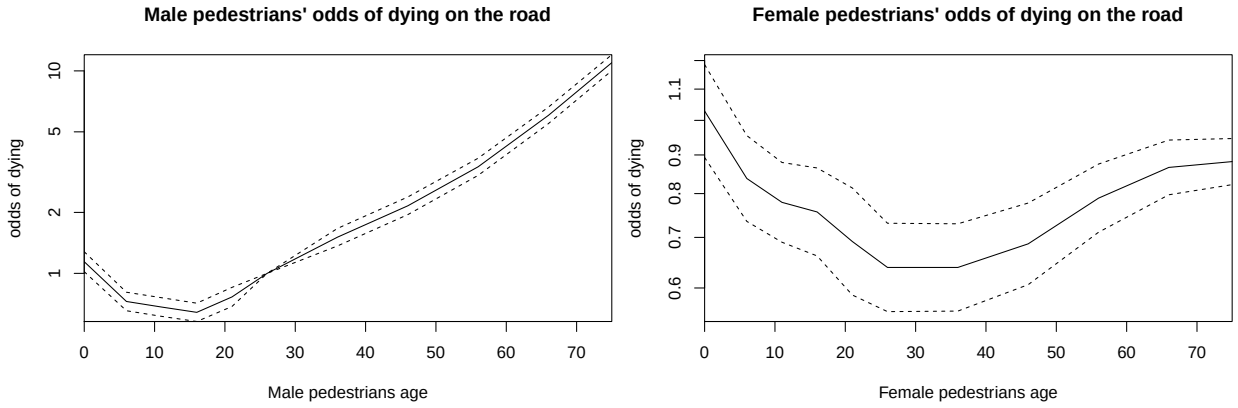
Table 6: Summary of Coefficents of Fitted Model

	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

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

Moreover, by comparing male and female's odds of death on the road during teenagers and in early adulthood (age 16 to 25), we found that the odds for young female pedestrians is 0.69 to 0.76. This finding indicates that the odds ratio for female death on the road is 0.69 to 0.76 times lower than male's. The hypothesis of young female are much safer than young male is correct.

Noticeably, as age goes up, the odds of male death is increasing dramatically. From age 66 and over, the odds of male pedestriansn death is 6 to 10 times higher than those in female pedestriansn. From the figure below, we can also see the odds for male pedestriansn is accelerating quickly as age goes up, whereas odds for female have less variation.



Conclusion

In this analysis, we investigate the hypothesis that women tend to be, on average, safer as pedestrians than men, particularly as teenagers and in early adulthood. By fitting conditional logistic regression model, we found female pedestrians are safer than male pedestrians overall, further, as age goes up, male elderly have much higher odds of death on the raod than female elderly. This is probably because female elderly pay much more caution on the road than male. However, no siginificant difference found between teenger male and female.

Appendix

```
CI = 2*confint.lm(fit)
knitr::kable(CI, caption = "95% Confidence Interval of Effects")
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$school)
library("INLA")
library(survival)
forSurv = data.frame(time = (pmin(forInla$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(0.4, 0.01))))
             + f(state, model = "iid",
                 hyper = list(prec = list(prior = "pc.prec", param = c(1,0.01)))),
             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)
ci = exp(qnorm(c(0.025, 0.5, 0.975), mean = log(1), sd = 2/3))
weibull = matrix(ci, ncol=3, byrow = TRUE)
colnames(weibull) = c("2.5% quantile", "mean", "97.5% quantile")
knitr::kable(weibull, caption = "Prior for Weibull shape with 95% Confidence Interval")
fitS2$priorPost = Pmisc::priorPost(fitS2)
for (Dparam in fitS2$priorPost$parameters)
  {do.call(matplot, fitS2$priorPost[[Dparam]]$matplot)}
do.call(legend, fitS2$priorPost$legend)
table = rbind(fitS2$summary.fixed[, c("mean", "0.025quant", "0.975quant")],
      Pmisc::priorPostSd(fitS2)$summary[,c("mean", "0.025quant", "0.975quant")])
knitr::kable(tail(table, n = 2),
             caption = "Standard Deviation for School and
             State with 95% Confidence Interval")
forSurv$one = 1
xSeq = seq(5, 100, len=  1000)
hist(forSurv$time, prob = TRUE, xlab="time", main = "Histogram of surviving time")
kappa = fitS2$summary.hyper['alpha', 'mode']
lambda = exp(-fitS2$summary.fixed['(Intercept)', 'mode'])
lines(xSeq, dweibull(xSeq/100, shape = kappa, scale = lambda)/100, col = 'blue')
plot(xSeq, (xSeq/(100*lambda))^kappa, col="blue", type = "l", log = "y",
     ylim = c(0.001, 10), xlim = c(20, 100),
     xlab = "years", ylab = "cummulative hazard", main ="Ideal Cumulative Hazard Function")
hazEst = survfit(Surv(time, one) ~ 1, data = forSurv)
```

```r
plot(hazEst, fun="cumhaz", main ="Actual Cumulative 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, ]
library("survival")
theClogit = clogit(y ~ age + age:sex + strata(strata), data = x)
knitr::kable(coefficients(summary(theClogit)),
             caption = "Summary of Coefficents of Fitted Model")
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$sex == "Male", c("coef", "se(coef)")]) %*% Pmisc::ciMat(0.99)),
        log = "y", type = "l", col = "black", lty = c(1, 2, 2), xaxs = "i", yaxs = "i",
  main = "Male pedestrians' odds of dying on the road", xlab = "Male pedestrians age",
  ylab = "odds of dying")
matplot(theCoef[theCoef$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",
  main = "Female pedestrians' odds of dying on the road", xlab = "Female pedestrians age",
  ylab = "odds of dying")
```