QMSS GR5016 - Lab2

Mathieu Sauterey - UNI: mjs2364

October 31, 2018

Run a multiple variable survival analysis. You can perform the survival analysis either using discrete-time methods (i.e., event history analysis) or you can use Cox proportional hazards methods, either one is fine.

```
# Reads GSS Panel Data
pan = read.csv(file.choose())

# Selects and subsets demographic variables of interest
vars <- c("age", "year", "sex", "idnum", "panelwave", "childs", "educ", "region",
          "marital", "race", "attend")
sub <- pan[, vars]

# identify those who have at least one kid or not
sub$c = ifelse(sub$childs==0,0,1)

# identify those who have at least one kid or not
sub$m = ifelse(sub$marital==1,1,0)

# creates new variable for non valid wave 1 data (NA or already a.l 1 kid)
sub$drop = ifelse((sub$c==1 | is.na(sub$c)) &  sub$panelwave==1, 1,0)

# extends the variable above to all 3 waves of individuals where wave 1 is non valid
sub = merge(sub, aggregate(drop ~ idnum, data=sub, mean), by="idnum",
            suffixes=c("", ".all"))

# Subset only valid individuals
subpt = subset (sub, sub$drop.all==0)

# Transforms year to act as a cardinal variable
subpt$nyear = subpt$year-2006

# Identifies waves 2 where people became parents
subpt$ytwo = ifelse(subpt$c==1 & subpt$panelwave==2, 2, 0)

# Use the variable above to tag all 3 waves as "individual became parents in wave 2"
subpt = merge(subpt, aggregate(ytwo ~ idnum, data=subpt, max), by="idnum",
              suffixes=c("", ".two"))

# Identifies waves 3 where people became parents
subpt$ythree = ifelse(subpt$c==1 & subpt$panelwave==3, 3, 0)

# Use the variable above to tag all 3 waves as "individual became parents in wave 3"
subpt = merge(subpt, aggregate(ythree ~ idnum, data=subpt, max), by="idnum",
```

```
              suffixes=c("", ".three"))

# Creates a code that shows when people became parents
subpt$combo <- do.call(paste, c(subpt[c("ytwo.two", "ythree.three")], sep = ""))

# Prints table based on code
table(subpt$combo)
```

```
##
##   00   03   20   23
## 1440   87   48  102
```

```
# Considers that after people became parents, future is no longer of interest
subpt$c[subpt$combo=="20" & subpt$year==2010] <- NA
subpt$c[subpt$combo=="23" & subpt$year==2010] <- NA

# Analyzes how time affects becoming a parent
# summary(glm(c ~ as.factor(nyear), subpt, family="binomial", subset = subpt$year>2006))

# What demographic factors predict became a parent?
# summary(glm(c ~ as.factor(nyear) + sex + educ + age + as.factor(race) + attend, subpt,
#             family="binomial", subset = subpt$year>2006))

# Restricts study to individuals above 23
summary(glm(c ~ as.factor(nyear) + sex + educ + age + m + as.factor(race) + attend, subpt,
            family="binomial", subset = subpt$year>2006 & subpt$age>23))
```

```
##
## Call:
## glm(formula = c ~ as.factor(nyear) + sex + educ + age + m + as.factor(race) +
##     attend, family = "binomial", data = subpt, subset = subpt$year >
##     2006 & subpt$age > 23)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2163  -0.5035  -0.3545  -0.2424   2.8650
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.552722   0.870703  -0.635  0.52556
## as.factor(nyear)4 -0.278588   0.269572  -1.033  0.30140
## sex               -0.113371   0.271814  -0.417  0.67661
## educ              -0.054237   0.046820  -1.158  0.24669
## age               -0.032975   0.010999  -2.998  0.00272 **
## m                  1.688136   0.290118   5.819 5.93e-09 ***
## as.factor(race)2   0.807076   0.416978   1.936  0.05292 .
## as.factor(race)3   0.105514   0.458390   0.230  0.81795
## attend            -0.004431   0.052087  -0.085  0.93221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 450.43  on 640  degrees of freedom
```

```
## Residual deviance: 397.40  on 632  degrees of freedom
##   (409 observations deleted due to missingness)
## AIC: 415.4
##
## Number of Fisher Scoring iterations: 6
```

(a) State what your "failure" variable is and how you expect your independent variables to affect it.

Using discrete-time methods, I will be performing a survival analysis on becoming a parent, i.e individuals having their first child ever. To create this binary variable of becoming a parent, we assigned a value of 1 to all individuals with at least 1 child and 0 to those without any children.

I do not expect time or sex to affect the odds of becoming a parent because fertility rate does probably not change over the course of 4 years, and because we can reasonably assume that mothers and fathers become parent for the first time within the same couple. I also don't believe that race is a determinant factor for becoming a parent, it probably plays a role rather in the number of children that each parent has (which is not our concern here). Attending religious services is tricky: while we may expect it to have an effect on becoming a parent, I actually believe it will not be statistically significant as I think it only affects at what age people become parents. However, higher education could be statistically significant as more educated people may be more interested in their career than in parenting, thus reducing the odds of a first child. Moreover, I think that age (past the early twenties - this point is later discussed) reduces the odds of becoming a parent both for biological reasons (health problems such as infertility develop with age) and social reasons (people in the middle of their career are less likely to have kids than young professionals, for example). Finally, I expect marital status (recoded as a binary variable: "currently married" or not) to be strongly significant, with married respondent much more likely to become parents since they see the arrival of a first child as the next step in their couple.

(b) Explain how you determined the "risk window" (due to right truncation and left-censoring) and who is eligible for failure over the time you are studying.

The dataset only contains data from 3 waves: in 2006, 2008 and 2010. Thus right truncation is defined as being 2010, i.e the last year of data available. Left truncation is accounted for by column age which shows for how many years the respondents have been at "risk" of becoming parents. Only the patients who are more than 23 years old are eligible because we believe that the hazard function versus age increases until 23 years of age and then decreases. In other words, the probability of becoming a parent at any moment, given no children to date, increases until turning 23, and then decreases with time.

(c) Explain whether the results were consistent with your expectations, and do that by interpreting the coefficients from the models, model fit, and so on.

We see from the print above that the results are roughly consistent with our expectations. Indeed, all variables (sex, time, race, attend) are not statistically significant except age and marital status. Although education does show a negative coefficient estimate as expected, it is surprisingly not statistically significant. Interestingly, age reaches its lowest p-value (at the 99% confidence level) when we subset for individuals aged 23 and older. This explains why we believe that the hazard function versus age increases until 23 years old and then decreases. The coefficient estimate of age is -0.032975 which shows that the log odds of becoming a parent decreases by -0.032975 for each year after 23. We can better interpret this by calculating exp(-0.032975) = 0.968, which means that every additional year of age reduces the odds of becoming a parent by (1-0.966) = 0.032. Finally, there is no surprise that marital status is the most significant predictor at the 99.9% confidence level. Its coefficient estimate is 1.688136 which means that the log odds of becoming a parent

increases by 1.688136 for married respondents. We can better interpret this by calculating exp(-0.032975) = 5.41, which means that being married multiplies the odds of becoming a parent by a factor of x5.4