140.623.01 - Statistical Methods in Public Health III

Assignment 4: Survival in Framingham Heart Study

Martin Skarzynski

March 13, 2018

Learning Objectives:

Students who successfully complete this section will be able to: - Analyze the relationship between grouped survival time data and baseline covariates of interest using log-linear Poisson regression models. - Check the assumptions for Poisson regression and use other models (such as negative binomial) as appropriate. - Summarize the findings in a brief fashion for public health readers. - Document and archive the steps of the statistical analysis.

Data Set: The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. Individuals were followed for 24 years. These data are binned into 5- year intervals (1876 days each) and stratified by gender and baseline current smoking, age category, BMI category, diabetes, and blood pressure medications (see Coding Description on the next page). The data are stored in the csv data set FraminghamPS4bin.csv which may be downloaded from the course website.

Methods: Use the data set described above and the appropriate statistical analyses to address the specific learning objectives. Hints: The hints shown below are based on a dataset with the name framData, read in with the following code. In the following list of commands, if you want to look at differences by other variables than drug, you should change the variable name! Create a new .R file to type/run your commands so that you will have a record of your analysis.

```r
library(readr)
framData = read_csv("FraminghamPS4bin.csv")
```

```
## Parsed with column specification:
## cols(
##   gender = col_integer(),
##   cursmoke = col_integer(),
##   diabetes = col_integer(),
##   bpmeds = col_integer(),
##   bmicat = col_integer(),
##   agecat = col_integer(),
##   tbin = col_integer(),
##   D = col_integer(),
##   Y = col_integer(),
##   Rate = col_double(),
##   Lower = col_double(),
##   Upper = col_double(),
##   L = col_integer()
## )
```

a. Explore the data using descriptive statistics:

- table()
- prop.table()
- summary() etc

```r
dim(framData)
```

```
## [1] 641  13
```

```r
str(framData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    641 obs. of  13 variables:
##  $ gender  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cursmoke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ diabetes: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bpmeds  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmicat  : int  1 1 1 1 2 2 2 2 3 3 ...
##  $ agecat  : int  1 2 3 4 1 2 3 4 1 2 ...
##  $ tbin    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ D       : int  0 0 0 0 0 1 7 4 0 2 ...
##  $ Y       : int  10950 7300 5475 5475 191625 385409 333030 148320 98550 284552 ...
##  $ Rate    : num  0.00 0.00 0.00 0.00 0.00 2.59e-06 2.10e-05 2.70e-05 0.00 7.03e-06 ...
##  $ Lower   : num  NA NA NA NA NA 3.65e-07 1.00e-05 1.01e-05 NA 1.76e-06 ...
##  $ Upper   : num  NA NA NA NA NA 1.84e-05 4.41e-05 7.19e-05 NA 2.81e-05 ...
##  $ L       : int  1825 1825 1825 1825 1825 1825 1825 1825 1825 1825 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 13
##   .. ..$ gender  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ cursmoke: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ diabetes: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ bpmeds  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ bmicat  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ agecat  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ tbin    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ D       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Y       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Rate    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ Lower   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ Upper   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ L       : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```r
summary(framData)
```

```
##      gender          cursmoke         diabetes          bpmeds
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :0.5757   Mean   :0.4867   Mean   :0.2902   Mean   :0.2777  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##                                                                     
##      bmicat          agecat           tbin            D         
##  Min.   :1.000   Min.   :1.000   Min.   :   0   Min.   : 0.000  
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1825   1st Qu.: 0.000  
##  Median :3.000   Median :3.000   Median :3650   Median : 1.000  
##  Mean   :2.789   Mean   :2.643   Mean   :3425   Mean   : 2.348  
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5475   3rd Qu.: 2.000  
##  Max.   :4.000   Max.   :4.000   Max.   :7300   Max.   :24.000  
##                                                                 
##        Y               Rate               Lower           Upper        
##  Min.   :    47   Min.   :0.0000000   Min.   :0e+00   Min.   :0.00002  
##  1st Qu.:  1825   1st Qu.:0.0000000   1st Qu.:1e-05   1st Qu.:0.00010  
##  Median :  7300   Median :0.0000176   Median :3e-05   Median :0.00040  
##  Mean   : 51123   Mean   :0.0002589   Mean   :9e-05   Mean   :0.00283  
##  3rd Qu.: 56869   3rd Qu.:0.0001333   3rd Qu.:8e-05   3rd Qu.:0.00162  
##  Max.   :528539   Max.   :0.0212766   Max.   :3e-03   Max.   :0.15104  
##                                       NA's   :284     NA's   :284      
##        L       
##  Min.   :1825  
##  1st Qu.:1825  
##  Median :1825  
##  Mean   :1825  
##  3rd Qu.:1825  
##  Max.   :1825  
## 
```

```r
library(purrr, help)
map(framData, class)
```

```
## $gender
## [1] "integer"
## 
## $cursmoke
## [1] "integer"
## 
## $diabetes
## [1] "integer"
## 
## $bpmeds
## [1] "integer"
## 
## $bmicat
## [1] "integer"
## 
## $agecat
## [1] "integer"
## 
## $tbin
## [1] "integer"
```

```
##
## $D
## [1] "integer"
##
## $Y
## [1] "integer"
##
## $Rate
## [1] "numeric"
##
## $Lower
## [1] "numeric"
##
## $Upper
## [1] "numeric"
##
## $L
## [1] "integer"
```

b. Explore several Poisson regression models using these grouped survival data and select between models:

```r
model1 = glm(D ~ gender, offset = log(Y), data =  framData, family=poisson(link="log"))
summary(model1)
```

```
##
## Call:
## glm(formula = D ~ gender, family = poisson(link = "log"), data = framData,
##     offset = log(Y))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8557  -0.7237  -0.3619   1.3158   5.4382
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -9.72601    0.03481 -279.363   <2e-16 ***
## gender      -0.50938    0.05179   -9.835   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1985.3  on 640  degrees of freedom
## Residual deviance: 1888.0  on 639  degrees of freedom
## AIC: 2915
##
## Number of Fisher Scoring iterations: 5
```

```r
AIC(model1)
```

```
## [1] 2914.99
```

c. Check the assumptions of your Poisson models; use other models as appropriate:

```r
# Pearson chi-square goodness-of-fit test (like poisgof in Stata)
X2 = sum(residuals(model1, type = "pearson")^2); X2
```

```
## [1] 5592.456
```

```r
df = model1$df.residual; df
```

```
## [1] 639
```

```r
pval = 1-pchisq(X2, df); pval
```

```
## [1] 0
```

```r
# Negative binomial regression
library(MASS)
model2 = glm.nb(D ~ gender + offset(log(Y)), data=framData)
summary(model2)
```

```
## 
## Call:
## glm.nb(formula = D ~ gender + offset(log(Y)), data = framData, 
##     init.theta = 0.9854264366, link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2750  -0.8745  -0.4765   0.3623   3.7012  
## 
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)    
## (Intercept) -9.14216    0.08451 -108.177  < 2e-16 ***
## gender      -0.48496    0.11733   -4.133 3.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(0.9854) family taken to be 1)
## 
##     Null deviance: 694.34  on 640  degrees of freedom
## Residual deviance: 680.23  on 639  degrees of freedom
## AIC: 2234.9
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  0.9854 
##           Std. Err.:  0.0975 
## 
##  2 x log-likelihood:  -2228.9260 
```

```r
AIC(model2)
```

```
## [1] 2234.926
```

d. Save your R script file that documents and archives the steps of your statistical analysis. This file will make your analysis "reproducible."
e. Summarize your findings in a brief report (less than two pages with at most one table and one figure) as if for a biomedical/public health journal.

A suggested format is:

- Introduction – a few sentences about the research question(s)
- Data description – simple tabulations describing individual characteristics
- Results from multiple models that address question(s) (e.g., bivariate and multivariable)

- Graphical display that presents evidence in the data relevant to your scientific question.

```
model3 = glm(D ~ gender + cursmoke + diabetes + bpmeds + bmicat + agecat, offset = log(Y), data =  fram
summary(model3)
```

```
## 
## Call:
## glm(formula = D ~ gender + cursmoke + diabetes + bpmeds + bmicat + 
##     agecat, family = poisson(link = "log"), data = framData, 
##     offset = log(Y))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.6238  -0.9082  -0.4040   0.8807   4.1543  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -12.25845    0.14827 -82.675  < 2e-16 ***
## gender       -0.50352    0.05348  -9.415  < 2e-16 ***
## cursmoke      0.35391    0.05514   6.419 1.38e-10 ***
## diabetes      0.79385    0.11012   7.209 5.63e-13 ***
## bpmeds        0.64452    0.10893   5.917 3.28e-09 ***
## bmicat        0.12847    0.03718   3.455  0.00055 ***
## agecat        0.73529    0.03015  24.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 1985.3  on 640  degrees of freedom
## Residual deviance: 1095.5  on 634  degrees of freedom
## AIC: 2132.5
## 
## Number of Fisher Scoring iterations: 6
```

```
AIC(model3)
```

```
## [1] 2132.49
```

```
library(broom, help)
tidy(model3)
```

```
##          term    estimate  std.error  statistic       p.value
## 1 (Intercept) -12.2584464 0.14827319 -82.674733  0.000000e+00
## 2      gender  -0.5035228 0.05348142  -9.414912  4.734773e-21
## 3    cursmoke   0.3539143 0.05513926   6.418553  1.375756e-10
## 4    diabetes   0.7938520 0.11011576   7.209249  5.626121e-13
## 5      bpmeds   0.6445170 0.10892689   5.916968  3.279298e-09
## 6      bmicat   0.1284729 0.03718120   3.455319  5.496419e-04
## 7      agecat   0.7352874 0.03014957  24.387991 2.293420e-131
```

```
coef(model3)
```

```
## (Intercept)      gender    cursmoke    diabetes      bpmeds      bmicat 
## -12.2584464  -0.5035228   0.3539143   0.7938520   0.6445170   0.1284729 
##      agecat 
##   0.7352874 
```

```
tidy(model3)$p.value
```

```
## [1]  0.000000e+00  4.734773e-21  1.375756e-10  5.626121e-13  3.279298e-09
## [6]  5.496419e-04 2.293420e-131
```

```
df <- data.frame(adj_RR = round(exp(coef(model3))[-1], 4),
           lower_CI = round(exp(confint(model3)[-1,1]), 4),
           upper_CI = round(exp(confint(model3)[-1,2]), 4),
           p_value = round(tidy(model3)$p.value[-1], 5))
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

```
rownames(df) <- rownames(confint(model3))[-1]
```

```
## Waiting for profiling to be done...
```

```
df
```

```
##          adj_RR lower_CI upper_CI p_value
## gender   0.6044   0.5442   0.6711 0.00000
## cursmoke 1.4246   1.2787   1.5873 0.00000
## diabetes 2.2119   1.7705   2.7276 0.00000
## bpmeds   1.9051   1.5291   2.3447 0.00000
## bmicat   1.1371   1.0570   1.2229 0.00055
## agecat   2.0861   1.9667   2.2135 0.00000
```

```
#install.packages("captioner")
library(captioner, help)
figs <- captioner(prefix="Figure")
tbls <- captioner(prefix="Table")
library(knitr)
```

Introduction

The Framingham Heart Study is a prospective study that followed study participants for 24 years in an attempt to better understand the etiology of cardiovascular disease. The study population is the community of Framingham, Massachusetts. The data that were obtain from the study were binned into 1875-day intervals (roughly 5 years). Additionally, categorical variables were created from the ages and body mass indices (BMI) of study participants.

Data Description

The research question that I will try to answer in this report is whether there is a relationship between grouped survival time data and baseline covariates of interest in the Framingham Heart Study. To answer this question, I will use a binned version of the Framingham data set and log-linear Poisson regression models. I hypothesize that smoking status, BMI and agecat will have a strong effect on the death rate. Specifically, I expect that male, obese, diabetic study participants that smoke and belong to the oldest age category will have a higher death rate. I will also assess whether anti-hypertensive medications provide any benefit.

Results

I calculated descriptive statistics and determined that the overall mean and median death rates are 0 and 0. Interestingly, the `Rate` variable is skewed highly to the right indicating that there are outliers with high death rates. The study population is 58% female; out of the total 641 study participants, 369 were women and 272 were men. Roughly 28% of the study participants took blood pressure medication. All of the variables in my log-linear model (gender, cursmoke, diabetes, bpmeds, bmicat, agecat) were statistically significant (based on an α value of 0.5). The results of the model of summarized in Table 1. The first column shows the death rate ratio, the second and third columns show the lower and upper confidence intervals (respectively) and the final column shows the first four subzero digits of the p-value. All of the death rate ratios were above 1 except for gender indication that being a current smoker, taking anti-hypertensive medication, being diabetic, being obese and being elderly were all associated with a higher death rate, while being female meant that participants were less likely to die in the Framingham study. The `diabetes` variable had the highest coefficient, but also the widest confidence interval.

Graphical Display

I decided to plot the log death rates per bin for every observation and color each of the variables of interest. The mean for each subgroup is shown as a horizontal bar. Interestingly, the diabetes (`diabetes=1`; bottom-left) and higher age categories (`agecat=3` and `4`; top-left) were consistently associated with a higher death rate. As for the bpmeds variable, I believe that the high coefficient associated with this variable would disappear if we controlled for blood pressure, as participants with the highest blood pressure would be most likely to be perscribed anti-hypertensive medicine and most likely to die of cardiovascular complications.

Conclusions

In conclusion, this analysis presents a multivariate log-linear model and univariate plots that highlight a potentially important link between various variables and the death rate in the Framingham Heart Study. Among the variables studied (gender, cursmoke, diabetes, bpmeds, bmicat, agecat), diabetes stood out as having the strongest association (highest coefficient) with the death rate. Further research is needed to improve our understanding of the interactions and etiologies of diabetes and cardiovascular disease. The analysis described herein also present the possibility that diabetes could be an important risk factor for cardiovascular disease. This work is only the beginning and more precise answers to the research questions discussed in the introduction will require further inspection with models more precisely adapted to each research question.

```r
knitr::kable(df, format = "markdown")
```

	adj_RR	lower_CI	upper_CI	p_value
gender	0.6044	0.5442	0.6711	0.00000
cursmoke	1.4246	1.2787	1.5873	0.00000
diabetes	2.2119	1.7705	2.7276	0.00000
bpmeds	1.9051	1.5291	2.3447	0.00000
bmicat	1.1371	1.0570	1.2229	0.00055
agecat	2.0861	1.9667	2.2135	0.00000

Table 1: Adjusted Rate Ratio Estimates of Death obtained from Log-Linear Regression.

```r
bins <- unique(framData$tbin)

par(mfrow=c(3,3), mar = c(0, 0, 0, 0), oma = c(4, 4, 0.1, 0.1))

plot(log(Rate) ~ jitter(tbin, 1),
     xaxt='n', ann=FALSE,
     data = framData, col = agecat)
ctgs <- unique(framData$agecat)
```

```r
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$agecat == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg, lwd = 3)
}}

legend("top",
       legend=paste0("agecat=", unique(framData$agecat)),
       pch = 1,
       col=1:length(unique(framData$agecat)),
       cex = 0.75)

plot(log(Rate) ~ jitter(tbin, 1),
     yaxt='n', xaxt='n', ann=FALSE,
     data = framData, col = bmicat)
ctgs <- unique(framData$bmicat)
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$bmicat == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg, lwd = 3)
}}
legend("top",
       legend=paste0("bmicat=", unique(framData$bmicat)),
       pch = 1,
       col=1:length(unique(framData$bmicat)),
       cex = 0.75)


plot(log(Rate) ~ jitter(tbin, 1),
     yaxt='n', xaxt='n', ann=FALSE,
     data = framData, col = cursmoke + 1)
ctgs <- unique(framData$cursmoke)
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$cursmoke == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg+1, lwd = 3)
}}
legend("top",
       legend=paste0("cursmoke=", unique(framData$cursmoke)),
       pch = 1,
       col=1:length(unique(framData$cursmoke)),
       cex = 0.75)

plot(log(Rate) ~ jitter(tbin, 1), data = framData, col = diabetes + 1)
ctgs <- unique(framData$diabetes)
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$diabetes == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg+1, lwd = 3)
}}
legend("top",
       legend=paste0("diabetes=", unique(framData$diabetes)),
       pch = 1,
       col=1:length(unique(framData$diabetes)),
       cex = 0.75)


plot(log(Rate) ~ jitter(tbin, 1),
     yaxt='n', ann=FALSE,
     data = framData, col = bpmeds + 1)
ctgs <- unique(framData$bpmeds)
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$bpmeds == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg+1, lwd = 3)
}}
legend("top",
       legend=paste0("bpmeds=", unique(framData$bpmeds)),
       pch = 1,
       col=1:length(unique(framData$bpmeds)),
       cex = 0.75)

plot(log(Rate) ~ jitter(tbin, 1),
     yaxt='n', ann=FALSE,
     data = framData, col = gender + 1)
ctgs <- unique(framData$gender)
for(bin in bins){
    for(ctg in ctgs){
        avg <- log(mean(framData$Rate[framData$tbin == bin & framData$gender == ctg]))
        lines(c(bin-250, bin+250), (c(avg, avg)), col = ctg+1, lwd = 3)
}}
legend("top",
       legend=paste0("gender=", unique(framData$gender)),
       pch = 1,
       col=1:length(unique(framData$gender)),
       cex = 0.75)
```

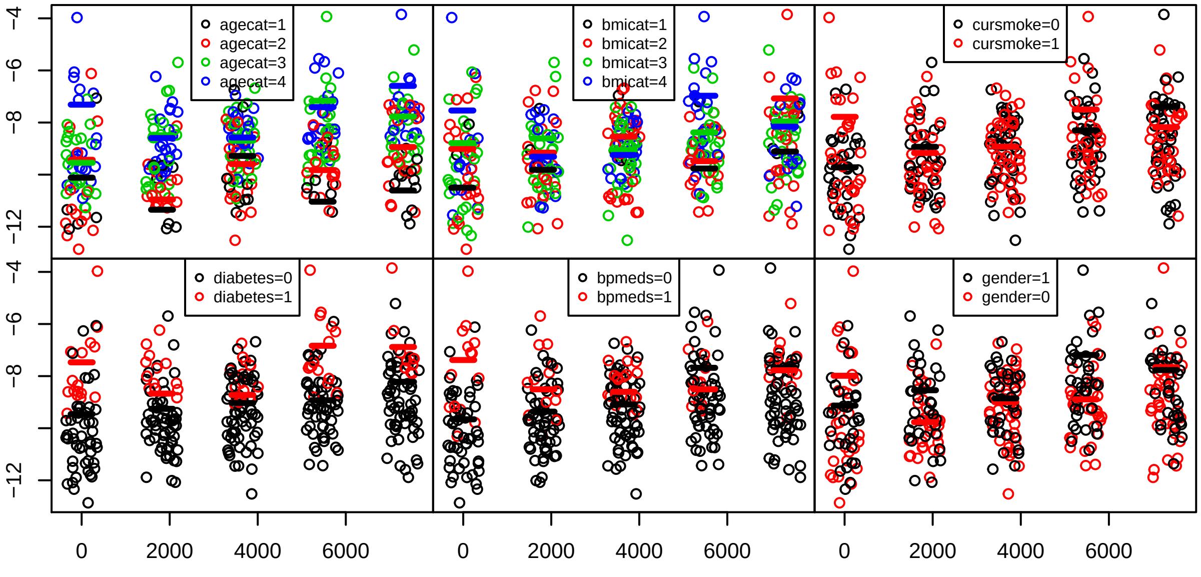

Figure 1: Death Rates per Time Bin in the Framingham Heart Study