PS3-Solution-2021

Elizabeth Colantuoni

I. Conditional Logistic Regression

Here you will be extending the analysis conducted and reviewed in Lecture9-Handout. Recall that the design was a matched case-control study conducted by Mack et al. (1976) to study the effect of exogenous estrogens on the risk of endometrial cancer. The data set is available on the Courseplus site, see Datasets folder. The dataset comprises 63 matched sets with one case and 4 controls per set. Controls were matched by being alive in the same community at the time of the case was diagnosed, having age within 1 year, same marital status and entering the community at roughly the same time. Controls could not have had a hysterectomy in which case they would not have been at risk of endometrial cancer. These data were made famous by the groundbreaking two volumes by Breslow and Day entitled Statistical Methods in Cancer Research. Chapters V and VI are excellent overviews of statistical methods for matched case-control studies.

The scientific questions of interest are:

- Are women who use estrogens, have a history of gall-bladder disease or hypertension at increased risk of endometrial cancer? Do these multiple risk factors may act synergistically?
- Does age or obesity modify the association between endometrial cancer and use of estrogens, history of gall-bladder disease or hypertension?

We explored Question A using only the first control in a 1-1 design, see Lecture9-Handout. You should repeat that analysis using all the available controls in the 1-4 design. Comment on how the strength of evidence changes with the addition of 3 additional controls per case. Then conduct an analysis to address Question B.

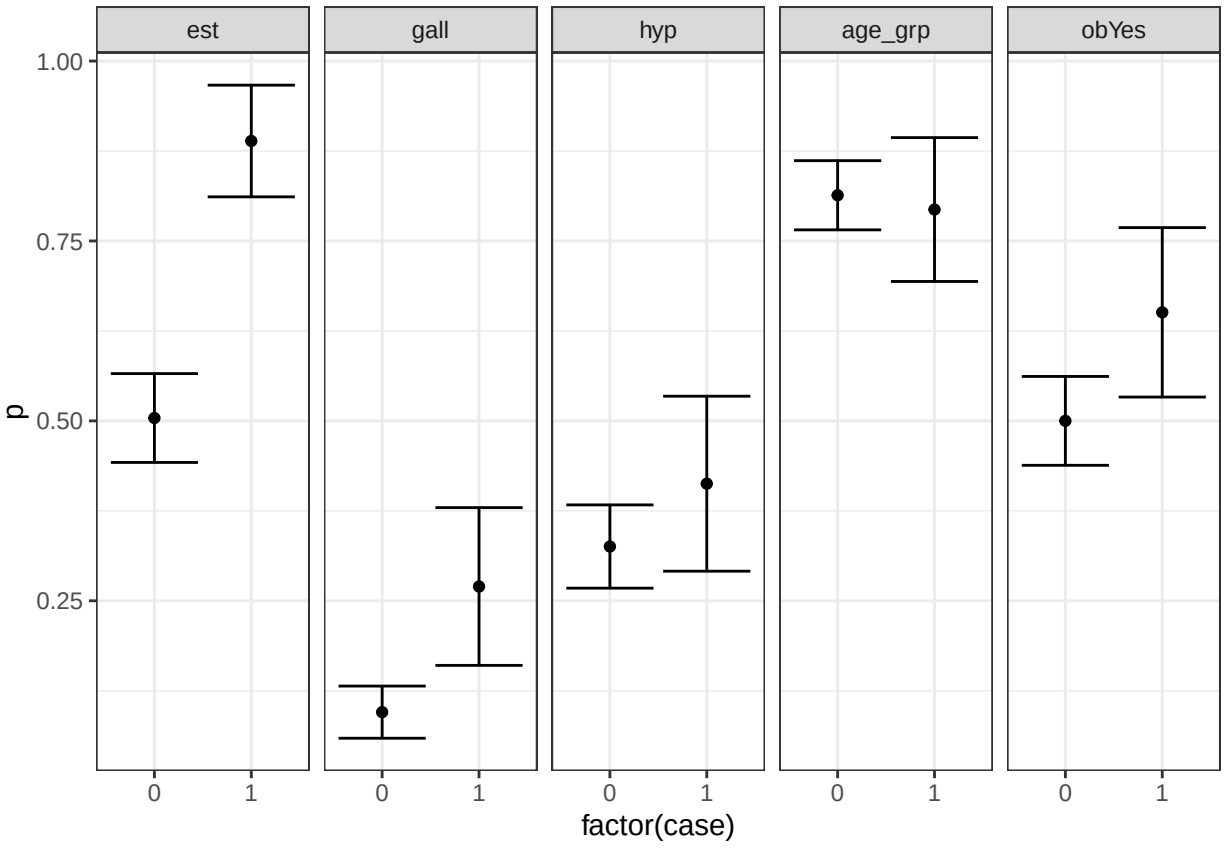
Prepare a one-page extended abstract plus one or two tables/figures that summarizes your work. State the questions. Describe key features of the data.

1. Analysis

```
endo=read.table('./endometrial.txt')
names(endo)=c('set','case','age','ageg','est','gall','hyp','ob','non')
#make binary variables 0/1 (instead of 1/2)
endo[,c('est','gall','hyp','ob','non')]=endo[,c('est','gall','hyp','ob','non')] - 1
#make obesity a factor, since 0=no, 1=yes and 2=unknown
endo$ob=factor(endo$ob)
# Create binary age variable
endo$age_grp=(endo$age >= 65)
endo$obYes=(endo$ob==1)
```

Recall some of the prior descriptive analyses:

```
## `summarise()` regrouping output by 'case' (override with `.groups` argument)
```



Cases have substantially higher rates of estrogen usage (statistically significant), somewhat higher rates of gallbladder disease (statistically significant), and marginally higher rates of hypertension (not statistically significant).

In our prior analyses, we found that hypertension was not associated with risk of endometrial cancer so we dropped this variable from consideration.

When using the 1:1 design, we found the following:

- The estimated odds of being a case for subjects with only estrogren use are 14.5 (95% CI: 3.1 to 71.4) times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.
- The estimated odds of being a case for subjects with only a history of gall bladder disease are 9.9 (95% CI: 0.95 to 104.8) times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.
- Finally, the estimated odds of being a case for subjects with both estrogen use and gall bladder disease are 16.8 (95% CI: 2.9 to 99.0) times the odds of being a case for subjects with neither estogren use or history of gallbladder disease. This is approximately double the odds ratio from either risk factor alone.

Next, we used the 1:4 design and refit the model that includes main effects for estrogen use and history of gallbladder disease and a model that included the interaction of estrogen use and history of gallbladder disease to explore whether estrogen use and gallbladder disease may act synergistically.

```
fit1=clogit(case ~ est + gall + strata(set), data=endo)
coefficients(summary(fit1))
```

```
##       coef exp(coef) se(coef)     z   Pr(>|z|)
## est  2.115     8.288   0.4398 4.809 0.00000152
## gall 1.275     3.577   0.4109 3.102 0.00191993
```

```
fit1.out = lincom(fit1,c("est","gall"),eform=TRUE,digits=2)
fit1.out
```

```
##      Estimate 2.5 % 97.5 % Chisq Pr(>Chisq)
## est  8.288    3.5   19.62  23.12 0.00000152
## gall 3.577    1.599 8.004  9.625 0.00192
```

Estrogen use and history of gallbladder disease both appear to have a statistically significant association with higher risk of endometrial cancer.

Next, We add interaction terms with age (indicator of age > 65) and obesity to the model with main terms of estogen use and history of gall bladder disease, and perform likelihood ratio tests to test for the significance of those effects.

```
fit1age=clogit(case ~ est + gall + est:age_grp + gall:age_grp + strata(set), data=endo)
coefficients(summary(fit1age))
```

```
##                      coef exp(coef) se(coef)       z Pr(>|z|)
## est                2.06812     7.910   0.7702 2.68526 0.007247
## gall               0.40801     1.504   0.9475 0.43060 0.666760
## est:age_grpTRUE    0.06918     1.072   0.7706 0.08977 0.928468
## gall:age_grpTRUE 1.07629     2.934   1.0553 1.01987 0.307788
```

```
fit1age.out = lincom(fit1age,c("est","est+est:age_grpTRUE","gall","gall+gall:age_grpTRUE"),eform=TRUE,d
fit1age.p = anova(fit1, fit1age, test='LRT')[2,4]
fit1age.p
```

```
## [1] 0.5838
```

```
fit1ob=clogit(case ~ est+gall+ob+est:ob+gall:ob+strata(set), data=endo)
coefficients(summary(fit1ob))
```

```
##              coef exp(coef) se(coef)       z Pr(>|z|)
## est       2.55664   12.8925   1.0908  2.3438  0.01909
## gall      1.28857    3.6276   0.7521  1.7133  0.08666
## ob1       1.30389    3.6836   1.1977  1.0887  0.27630
## ob2       0.74109    2.0982   1.3429  0.5518  0.58106
## est:ob1  -0.72376    0.4849   1.2250 -0.5908  0.55463
## est:ob2   0.47724    1.6116   1.5498  0.3079  0.75814
## gall:ob1  0.08346    1.0870   0.8879  0.0940  0.92511
## gall:ob2 -0.27068    0.7629   1.7141 -0.1579  0.87452
```

```
fit1ob.out = lincom(fit1ob,c("est","est+est:ob1","est+est:ob2","gall","gall+gall:ob1","gall+gall:ob2"),
fit1ob.p = anova(fit1, fit1ob, test='LRT')[2,4]
fit1ob.p
```

```
## [1] 0.6035
```

Age and obesity do not appear to moderate the association between the risk factors and risk of endometrial cancer.

Fit the model with the interaction between estrogen use and gallbladder disease.

```
fit4=clogit(case ~ est*gall + strata(set), data=endo)
coefficients(summary(fit4))
```

```
##            coef exp(coef) se(coef)      z   Pr(>|z|)
## est       2.700   14.8818   0.6118  4.414 0.00001016
## gall      2.894   18.0717   0.8831  3.278 0.00104673
```

```
## est:gall -2.053    0.1284   0.9950 -2.063 0.03910155
```

```
fit4.out = lincom(fit4,c("est","gall","est+gall+est:gall"),eform=TRUE,digits=2)
fit4.p = anova(fit1,fit4,test="LRT")[2,4]
coef=round(coefficients(fit4),2)
CI=round(confint(fit4),2)
#compute variance and CI of sum of coefficients
var_coefsum=c(1,1,1) %*% vcov(fit4) %*% c(1,1,1)
CI_coefsum=round(sum(coef) + c(-1,1)*1.96*sqrt(var_coefsum),2)
```

```
## Warning in c(-1, 1) * 1.96 * sqrt(var_coefsum): Recycling array of length 1 in vector-array arithmeti
##   Use c() or as.vector() instead.
```

The interaction between estrogen use and history of gallbladder disease is statistically significant and has a negative coefficient. This shows that the risks associated with estrogen use and gallbladder disease are not additive: the log odds of endometrial cancer for subjects with estrogen use (only) is 2.70 (95% CI: [1.50, 3.90]), for subjects with history of gallbladder disease (only) is 2.89 (95% CI: [1.16, 4.63]), and for subjects with both is 3.54 (95% CI: [2.12, 4.96]). We see that subjects with both risk factors have only slightly higher log odds of endometrial cancer than subjects with only a single risk factor.

Compared to the 1:1 design, we have gained precision to estimate the main effect of estogren use and gallbladder disease but wider confidence interval for the interaction between estrogen use and gallbladder disease.

2. One-page extended abstract

Objective: To study the effect of exogenous estrogens and history of gallbladder disease on the risk of endometrial cancer and determine whether age or obesity modifies these effects.

Design: A matched case-control study comprised of 63 matched sets with one case and 4 controls per set. Controls were matched by being alive in the same community at the time of the case was diagnosed, having age within 1 year, same marital status and entering the community at roughly the same time. Controls could not have had a hysterectomy in which case they would not have been at risk of endometrial cancer.

Methods: Descriptive statistics were computed to compare report of estrogen use and gallbladder disease among the cases and controls. Several conditional logistic regression models were constructed including a model with only main terms for report of estrogen use and history of gallbladder disease. This model was extended where each moderator (indicator of age > 64 and indicators for obesity: yes, unknown vs. no) plus interaction terms were included. Note that the main term for age was excluded from the interaction model given age was one of the central matching variables. Likelihood ratio tests were used to determine if age and obesity were moderators. Lastly, we considered a model where the effect of estrogen use and history of gallbladder disease could act synergistically (i.e. interaction of estrogen use and history of gallbladder disease).

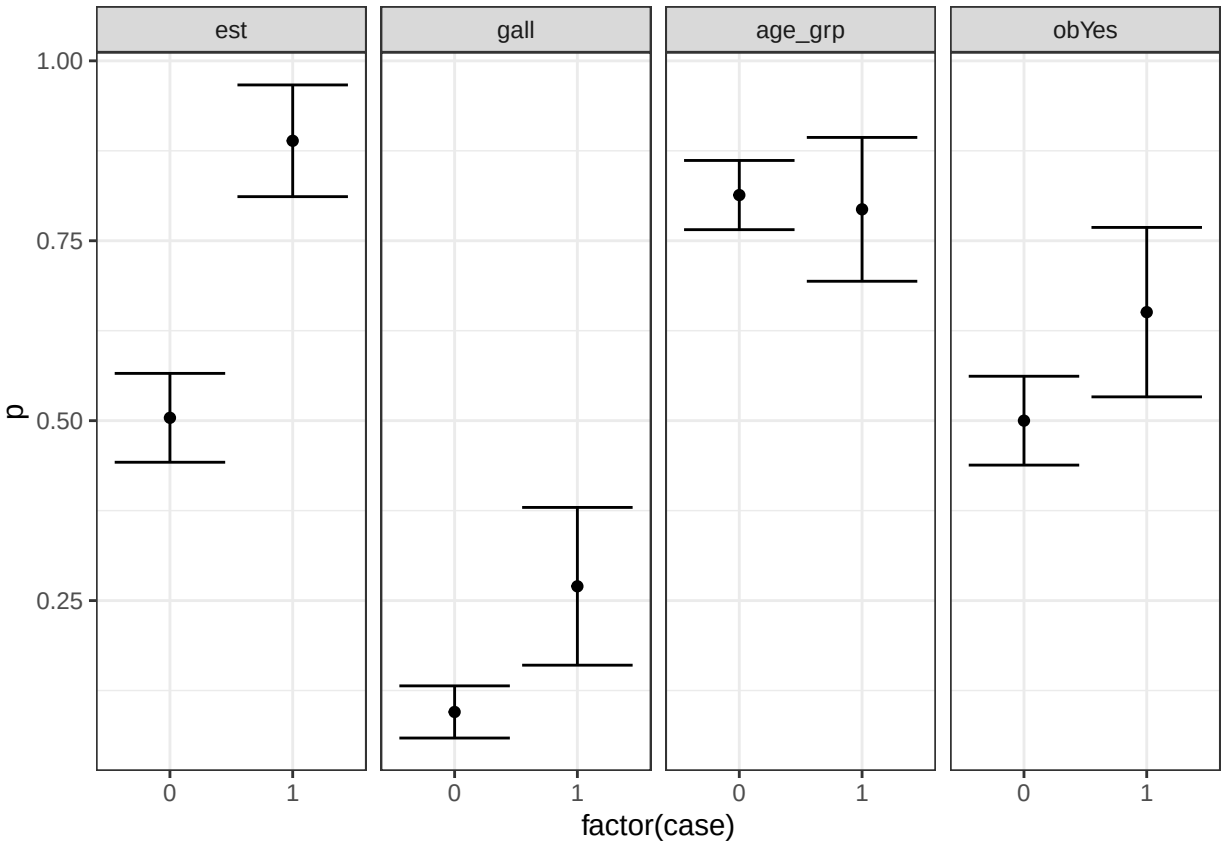
Results: The figure below displays the proportion of cases and controls with estrogen use, history of gallbladder disease, age ≥ 65 and whom are obese. There are clinically relevant differences between proportion of cases and controls reporting estrogen use, history of gallbladder disease and obesity. Based on the design we didn't anticipate differences by age groups.

The table below displays the results from the conditional logistic regression models. Both estrogen use and history of gallbladder disease were found to increase the risk of endometrial cancer. Specifically, the odds of endometrial cancer are 8.29 times greater for women reporting estrogen use compared to otherwise similar women with no estrogen use (95% CI: 3.5 to 19.62) and are 3.58 times greater for women with a history of gallbladder disease compared to otherwise similar women without (95% CI: 1.6 to 8). Although being an older woman (Age ≥ 65) increased the odds of endometrial cancer for estrogen users (Odds ratio: 7.91 vs. 8.48 for older and younger women, respectively) and those with a history of gallbladder disease (Odds ratio: 1.5 vs. 4.41 for older and younger women, respectively); these differences did not reach statistical

significance (p-value for effect modification: 0.584). Unlike for age, when assessing whether obesity status moderates the relationship between endometrial cancer with estrogen use and history of gallbladder disease we found no clear pattern for the direction of the effect modification (p-value: 0.603). Lastly, we found that estrogen use and history of gallbladder act synergistically on the risk of endometrial cancer (p-value for effect modification 0.039). The estimated odds of endometrial cancer among women with only estrogren use are 14.88 (95% CI:[4.49,49.36]) times the odds for women with neither estrogen use or history of gallbladder disease. The estimated odds of endometrial cancer among women with only a history of gall bladder disease are 18.07(95% CI:[3.20,102.01]) times the odds for women with neither estrogen use or history of gallbladder disease. Finally, the estimated odds of endometrial cancer for women with both estrogen use and gall bladder disease are 34.53(95% CI:[8.33,142.59]) times the odds for women with neither estogren use or history of gallbladder disease. This is approximately double the odds ratio from either risk factor alone. We note that our analysis with one case and four controls often yielded more narrow confidence intervals compared to when we considered the case matched to a single control (data not shown).

Discussion: In this small matched case-control study, we estimated that older age may modify the relationship between risk of endometrial cancer as a function of estrogen use and history of gallbladder disease; however we could not make a strong conclusion given the size and uncertainty in the data. The strongest evidence in the data of moderation comes from estrogen use and gallbladder disease acting synergistically on the risk of endometrial cancer.

```
## `summarise()` regrouping output by 'case' (override with `.groups` argument)
```




```
out = xtable(out,digits=2,align="rccc")
print(out,hline.after = c(-1,0,2,6,12,nrow(out)))
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Wed May 12 16:21:05 2021

	Odds Ratio	Lower limit	Upper limit
Estrogen use	8.29	3.50	19.62
HX gallbladder	3.58	1.60	8.00
Estrogen use: Age<65	7.91	1.75	35.79
Estrogen use: Age>=65	8.48	3.42	21.03
HX gallbladder: Age<65	1.50	0.23	9.63
HX gallbladder: Age>=65	4.41	1.77	10.97
Estrogen use: Not obese	12.89	1.52	109.36
Estrogen use: Obese	6.25	2.05	19.02
Estrogren use: Unknown	20.78	2.34	184.36
HX gallbladder: Not obese	3.63	0.83	15.84
HX gallbladder: Obese	3.94	1.47	10.55
HX gallbladder: Unknown	2.77	0.14	53.57
Estrogen use ONLY	14.88	4.49	49.36
HX gallbladder ONLY	18.07	3.20	102.01
Both	34.53	8.37	142.48

II. Log-linear Poisson regression with application to a survival outcome

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.6.3
```

```
dat = data.frame(trt=c(rep(0,20),rep(1,20)),
                 event_time=c(6,8,11,13,16,16,19,21,22,
                              28,28,29 ,31,35,40,41,41,
                              59,86,132,6,9,9,10,11,12,
                              13,17,18,19,19,20,22,24,
                              28,31,43,48,51,57),
                 event=c( 1, 1 ,0, 1, 1, 1, 1, 0, 0, 1,
                          0, 1, 1, 1, 0, 0, 0, 0, 0, 0,
                          1, 0, 1, 1, 0, 0, 0, 0, 1, 0,
                          1, 0, 1, 1, 0, 1 ,0 ,1 ,0 ,0))
```

1. Create discrete time, grouped data by completing the table below.

```
py = pyears(Surv(event_time, event) ~
              tcut(rep(0, length(dat$event_time)),
                   breaks = c(-1, 10, 20, 30, 40, 50, 140),
                   labels = c("0-10", "11-20", "21-30", "31-40",
                              "41-50", "51+")) + trt,
            data = dat,
            scale = 1,
            data.frame = TRUE)
pydat = py$data
colnames(pydat) = c("label", "treat", "ptime", "n", "event")
pydat = cbind(pydat, midpoints = rep(c(5, 15, 25, 35, 45, 55), times = 2),
    ncensor = c(0, 1, 3, 1, 2, 3, 1, 6, 1, 0, 1, 2), incidence = pydat$event/pydat$ptime)
pydat[,c("label","treat","ptime","n","event")]
```

```
##    label treat ptime  n event
## 1   0-10     0   194 20     2
## 2  11-20     0   155 18     4
## 3  21-30     0   108 13     2
```

```
## 4  31-40     0    66  8     2
## 5  41-50     0    32  5     0
## 6    51+     0   127  3     0
## 7   0-10     1   194 20     3
## 8  11-20     1   129 16     2
## 9  21-30     1    64  8     2
## 10 31-40     1    41  5     1
## 11 41-50     1    31  4     1
## 12   51+     1     8  2     0
```

3. Using the definitions provided and your binned survival data, compute the incidence rate and probability of surviving past each interval of time

NOTE: The incidence rate per bin of time should be computed as $event/ptime$.

NOTE: To compute the probability of surviving past each interval of time, you would compute $1 - event/ptime \times 10$, i.e. the incidence rate per week -> calculate the incidence rate per 10 weeks -> calculate 1 - the incidence rate per 10 weeks. This will give you the probability of surviving the given interval of time given that you had survived to the beginning of the interval of time.

NOTE: I didn't ask you to compute the survival function, i.e. probability of surviving past the interval of time. Here you would compute the cumulative product of the probabilities described above.

```
pydat = cbind(pydat, ProbInt = round((1-pydat$event/pydat$ptime*10),2),survProb = round(c(cumprod((1-py
    cumprod((1-pydat$event[7:12]/pydat$ptime[7:12]*10))),2))
pydat
```

```
##    label treat ptime  n event midpoints ncensor incidence ProbInt
## 1   0-10     0   194 20     2         5       0   0.01031    0.90
## 2  11-20     0   155 18     4        15       1   0.02581    0.74
## 3  21-30     0   108 13     2        25       3   0.01852    0.81
## 4  31-40     0    66  8     2        35       1   0.03030    0.70
## 5  41-50     0    32  5     0        45       2   0.00000    1.00
## 6    51+     0   127  3     0        55       3   0.00000    1.00
## 7   0-10     1   194 20     3         5       1   0.01546    0.85
## 8  11-20     1   129 16     2        15       6   0.01550    0.84
## 9  21-30     1    64  8     2        25       1   0.03125    0.69
## 10 31-40     1    41  5     1        35       0   0.02439    0.76
## 11 41-50     1    31  4     1        45       1   0.03226    0.68
## 12   51+     1     8  2     0        55       2   0.00000    1.00
##    survProb
## 1      0.90
## 2      0.67
## 3      0.54
## 4      0.38
## 5      0.38
## 6      0.38
## 7      0.85
## 8      0.71
## 9      0.49
## 10     0.37
## 11     0.25
## 12     0.25
```

4. Use Poisson regression with the grouped data above to estimate the relative hazard of hospitalization for treatment as compared to control assuming that the hazards are proportional and that the baseline log incidence rate is a:

A. linear function of weeks B. linear spline function of weeks with breaks at 20 and 40 weeks C. Step function with a separate rate in each interval

Complete the table below using the results for the 3 models

```
pydat$time_sp1 = ifelse(pydat$midpoints > 20, pydat$midpoints - 20, 0)
pydat$time_sp2 = ifelse(pydat$midpoints > 40, pydat$midpoints - 40, 0)
fit1 = glm(event ~ treat + midpoints, family = "poisson",
           offset = log(ptime), data = pydat)
fit1p = glm(event ~ treat + midpoints, family = "quasipoisson", offset = log(ptime), data = pydat)
fit2 = glm(event ~ treat + midpoints + time_sp1 + time_sp2, family = "poisson",
           offset = log(ptime), data = pydat)
fit2p = glm(event ~ treat + midpoints + time_sp1 + time_sp2, family = "quasipoisson", offset = log(ptim
fit3 = glm(event ~ treat + factor(midpoints), family = "poisson", offset = log(ptime), data = pydat)
fit3p = glm(event ~ treat + factor(midpoints), family = "quasipoisson", offset = log(ptime), data = pyda
RR = exp(c(coef(fit1)[2], coef(fit2)[2], coef(fit3)[2]))
SE = c(coef(summary(fit1))[2, 2], coef(summary(fit2))[2, 2], coef(summary(fit3))[2, 2])
CI_LB = exp(log(RR) - 1.96 * SE)
CI_UB = exp(log(RR) + 1.96 * SE)
df = c(length(coef(fit1))-1, length(coef(fit2))-1, length(coef(fit3))-1)
deviance = c(summary(fit1)$deviance, summary(fit2)$deviance, summary(fit3)$deviance)
AIC = c(AIC(fit1), AIC(fit2), AIC(fit3))
table_fit = cbind(RR, CI_LB, CI_UB, df, deviance, AIC)
rownames(table_fit) = c("model A", "model B", "model C")
table_fit
```

```
##            RR  CI_LB CI_UB df deviance   AIC
## model A 1.254 0.4982 3.154  2    8.004 37.33
## model B 1.174 0.4747 2.901  4    2.699 36.03
## model C 1.148 0.4643 2.840  6    2.196 39.52
```

5. Starting with Model B, extend the model by including the appropriate interaction terms and use a likelihood ratio test of the null hypothesis that the treatment hazards are proportional.

```
fit1pint = glm(event ~ treat * midpoints, family = "poisson", offset = log(ptime), data = pydat)
summary(fit1pint)
```

```
##
## Call:
## glm(formula = event ~ treat * midpoints, family = "poisson",
##     data = pydat, offset = log(ptime))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4115  -0.7393  -0.0329   0.4265   1.1884
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.8166     0.4918   -7.76  8.4e-15 ***
## treat            -0.3939     0.7436   -0.53     0.60
## midpoints        -0.0188     0.0198   -0.95     0.34
## treat:midpoints   0.0333     0.0304    1.10     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 8.5459  on 11  degrees of freedom
## Residual deviance: 6.8271  on  8  degrees of freedom
## AIC: 38.15
##
## Number of Fisher Scoring iterations: 6
```

```
anova(fit1p,fit1pint,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: event ~ treat + midpoints
## Model 2: event ~ treat * midpoints
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         9       8.00
## 2         8       6.83  1     1.18     0.28
```

```
fit2int = glm(event ~ treat * (midpoints + time_sp1 + time_sp2), family = "poisson",
    offset = log(ptime), data = pydat)
```

```
## Warning: glm.fit: fitted rates numerically 0 occurred
```

```
summary(fit2int)
```

```
##
## Call:
## glm(formula = event ~ treat * (midpoints + time_sp1 + time_sp2),
##     family = "poisson", data = pydat, offset = log(ptime))
##
## Deviance Residuals:
##       1        2        3        4        5        6        7
## -0.1925   0.4618  -0.5345   0.2117  -0.0001   0.0000   0.0737
##       8        9       10       11       12
## -0.2517   0.3510  -0.3257   0.2579  -0.3968
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.7138     0.9390   -5.02  5.2e-07 ***
## treat              0.3926     1.2802    0.31     0.76
## midpoints          0.0544     0.0693    0.79     0.43
## time_sp1          -0.0561     0.1296   -0.43     0.67
## time_sp2          -4.0099  3028.6280    0.00     1.00
## treat:midpoints   -0.0326     0.0997   -0.33     0.74
## treat:time_sp1     0.0663     0.1878    0.35     0.72
## treat:time_sp2     3.8862  3028.6280    0.00     1.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8.5459  on 11  degrees of freedom
## Residual deviance: 1.1029  on  4  degrees of freedom
## AIC: 40.43
##
## Number of Fisher Scoring iterations: 20
```

```r
anova(fit2,fit2int,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: event ~ treat + midpoints + time_sp1 + time_sp2
## Model 2: event ~ treat * (midpoints + time_sp1 + time_sp2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         7        2.7
## 2         4        1.1  3      1.6     0.66
```

```r
#fit3int = glm(event ~ treat * factor(midpoints), family ="poisson", offset = log(ptime),data = pydat)
#summary(fit3int)
#anova(fit3,fit3int,test="LRT")
```

Based on the results of the likelihood ratio tests for the linear model (p = 0.28) and linear spline model (p = 0.660), we do not find evidence against the proportional hazards assumption.

6. Write a one page summary of your analysis of these data to address the question (QQQ): Is the distribution of time to hospitalization similar for persons randomized to receive treatment 0 as compared to treatment 1. Use the class format for a brief report: question, data display, methods, findings/discussion.

Include in your report a paragraph that addresses two questions: (1) are your main findings sensitive to assumptions about the baseline hazard; and (2) is there strong evidence in these data that the proportional hazards assumption is incorrect.

In the report, be quantitative and remember that absence of evidence is not the same as evidence of absence.

Report

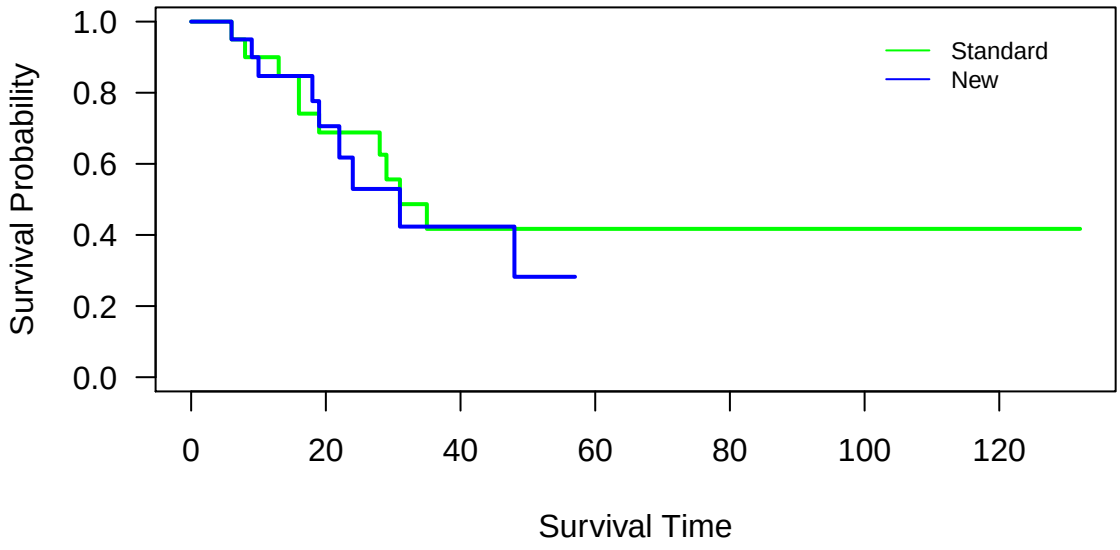
Research question

We are interested in examining whether the distributions of time to hospitalization are similar for persons randomized to receive the standard treatment as compared to the new treatment.

Data display

The available dataset contained data on 40 patients diagonised with schizophrenia, of whom 20 were randomized to the new drug treatment and the remaining 20 were randomzed to receive standard therapy. Each patient was followed from randomization / treatment initiation until hospitalization or censoring. Length of follow up was recorded in weeks. Whether a patient was hospitalized or censored was also captured in the data.

The figure below displays Kaplan-Meier survival function estimates for each of the two treatment group. It is hard to tell visually if there is a difference between the two groups in terms of survival.



Methods

We grouped follow-up time into six discrete intervals, i.e. 0-10, 11-20, 21-30, 31-40, 41-50, and 50+ weeks, and recorded the number of events and the total length of follow up within each time interval (i.e. person-weeks).Then we applied Poisson regression models to estimate the risk of hospitalization for subjects with schizophrenia on a standard versus new drug treatment. We modeled the baseline log incidence rate as a linear function, a linear spline with knots at 20 and 40 weeks, as well as a step function for time. We selected a model based on the Akaike information criterion (AIC). To test whether the relative risks are proportional, we extended the model by including interaction terms between time and treatment status, allowing the treatment hazards to differ across time. Following that, we conducted a likelihood ratio test to test the null hypothesis that the treatment hazards are proportional.

```
kable(table_fit)
```

	RR	CI_LB	CI_UB	df	deviance	AIC
model A	1.254	0.4982	3.155	2	8.004	37.33
model B	1.174	0.4747	2.901	4	2.699	36.02
model C	1.148	0.4643	2.840	6	2.196	39.52

The linear spline model was selected as it minimized AIC. From this model, the risk of hospitalization among the persons receiving the new treatment is 1.174 (95% CI [0.475, 2.901]) times the risk among the persons receiving standard treatment. The estimate is not statistically significant at alpha = 0.05, as the 95% confidence interval overlaps one. Therefore, we don't have enough evidence to show that the distribution of time to hospitalization for patient on a standard versus new drug treatment are different.

Our findings where similar when we allowed the baseline log incidence rate to be linear or a step function of the binned survival times. Further, we found no evidence that the relative risk of hospitalization varied as a function of time from randomization (likelihood ratio test comparing the linear spline model to an extended model that included treatment x linear spline interaction terms, p = 0.6603).

Conclusion and Discussion

We find that the risk of hospitalization for subjects with schizophrenia on a standard versus new drug treatment is best represented with a proportional hazards model that allows for a nonlinear baseline log incidence rate. We estimated an increased risk of hospitalization among subjects receiving the new drug treatment; but this increased risk was not statistically significant which may be driven by the relative small size of the trial. The choice to model the baseline log incidence rate as a nonlinear function did not have a substantial effect on the estimated relative risk of hospitalization for the new versus the standard treatment group. Within this small trial, there was no evidence that the proportional hazards assumption was violated.

Additional note: Many of you may have used a *quasipoisson* model. If you did, your results would have been very similar to the results above. Small differences in the confidence intervals but these differences were qualitatively similar, i.e. the overall findings were the same.