Biostatistics 140.654
Fourth Term, 2021
Problem Set 3

Instructions: follow the general directions for the class that permit group discussion and implementation of the analyses but requires each student to write-up their own problem set.

Due date: Friday May 14 by 5pm

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:

A. 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?
B. 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.

Summarize the methods of analysis and results. Give a one or two sentence conclusion that is numerate and avoids unnecessary statistical jargon. Pay attention to the quality of scientific writing.

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

Upon successful completion of this problem, a student should be able to:

- Formulate a log-linear Poisson regression model that estimates the relative incidence/risk/hazard of an event as a function of covariates
- Define the "baseline hazard function" and compare multiple models for the "baseline" hazard function
- Test the assumption of proportional hazards

Below find a set of times (in weeks) to hospitalization for persons with a diagnosis of schizophrenia who have been randomized to standard therapy (Trt=0) or a new drug treatment (Trt=1). A plus sign indicates censoring, i.e. the patient dropped out of the study or the patient was still enrolled without a hospitalization at the end of the study period (administrative censoring). We will assume the censoring to be independent of hospitalization/disease state (strong assumption and it is likely that patient drop-out was related to disease state for schizophrenics).

Trt=0: 6 8 11+ 13 16 16 19 21+ 22+ 28 28+ 29 31 35 40+ 41+ 41+ 59+ 86+ 132+

Trt=1: 6 9+ 9 10 11+ 12+ 13+ 17+ 18 19+ 19 20+ 22 24 28+ 31 43+ 48 51+ 57+

For example, in the standard therapy group (Trt=0), Patient1 was followed for 6 weeks and then was hospitalized, Patient2 was followed for 8 weeks and then hospitalized, Patient3 was followed for 11 weeks and then dropped out of the study.

For the standard therapy group (Trt=0), 10 patients were hospitalized and 10 patients were censored. For the new drug treatment group (Trt=1), 9 patients were hospitalized and 11 patients were censored.

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

	Standard Therapy (Trt=0)		New Drug Therapy (Trt=1)	
Interval	Person-time	Events	Person-Time	Events
0-10	6+8+ 18*10=194	2		
11-20	1+3+6+6+9+ 13*10=155	4		
21-30				
31-40				
41-50				
51+				

2. Before we do some modeling, we will establish some definitions:

We will be building several models for the incidence rate or "hazard" of a hospitalization in each interval of time, i.e. we want to describe patterns of how the incidence rate or "hazard" of a hospitalization changes over time and/or as a function of other exposures (e.g. treatment). The incidence is the risk of hospitalization per unit time among those that enter the time interval. The term hazard is usually reserved for the limit of the incidence rate as the interval width goes to zero. We can get a crude estimate by computing the number of events in the interval divided by the person-time experienced in the interval. For example, we estimate the incidence rate to be 2/194 = 0.01 events per week during 0 to 10 weeks among patients receiving the standard therapy.

But this is a crude estimate based upon few events. We want to smooth these rates using a log-linear model.

We assume the incidence rate $\lambda_i(X_i)$ satisfies a log-linear regression

$$\lambda_i(X_i) = \exp(X_i'\beta)$$

In this example, we will consider two X variables: time represented by the mid-point of each interval (t) and Trt.

The number of events in an interval is assumed to be a Poisson variable since this count is a sum of independent random events assuming each person is independent of the others. The expected number of events in an interval is the incidence rate, $\lambda_i\,(X_i)$, multiplied by the person-time for which this rate is experienced.

Hence, we have

$$E(Y_i) = \lambda_i\,(X_i) \times PT_i = \exp\,(\log(PT_i) + X_i'\beta)$$

Here, the term log(PT_i) is called an "offset" because it is added to the linear predictor without needing a regression coefficient. You can think of an offset as a predictor variable whose coeficient is known to be 1.

In survival analysis, the incidence rate or hazard is indexed by time and other exposure variables, so that the model may look like:

$$\lambda_i(t_i, Trt_i) = \exp\,(f(t_i, \beta_t) + \beta \times Trt_i)$$

NOTE: $f(t_i, \beta_t)$ indicates some function of time with corresponding coefficients.

When we set all the exposure variables to 0, e.g. $Trt_i = 0$, we refer to this incidence rate or hazard function as the "baseline hazard"

$$\text{i.e. } \lambda_i(t_i, Trt_i = 0) = \exp\,(f(t_i, \beta_t))$$

In addition, the model we specified above is a "proportional hazards model" in that the relative effect of the exposure variable on the incidence rate is the same regardless of time.

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

	Control Group (Trt=0)				Treatment Group (Trt=1)			
Interval	Person-time	Events	Incidence: risk of hosp per unit time	Prob survive past interval	Person-Time	Events	Incidence: risk of hosp per unit time	Prob survive past interval
0-10	6+8+ 18*10=194	2	2/194= 0.01	1- 0.01*10=0.90				
11-20	1+3+6+6+9+ 13*10=155	4	4/155= 0.026	0.90*(1- 0.026*10) = 0.67				
21-30								
31-40								
41-50								
51+								

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

Model	Log Rel Risk	Std Error	95% CI	Model df	Deviance	AIC
A						
B						
C						

5. For Models A and B, extend the models by including the appropriate interaction terms and use a likelihood ratio test of the null hypothesis that the treatment hazards are proportional.

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.