Course EPIB-681: Data Analysis II [Winter 2004] Assignment 8

[See SAS/Stat programs provided under Resources in www.epi.mcgill.ca/hanley/c681/cox

1. Sharper and Fairer Comparisons: effect of sexual activity on the longevity of male fruitflies

[For all but part h, limit analysis to fruitflies with 1 partner .. the effect is obvious in those with 8]

When we first analyzed this dataset, student PE, now on McGill faculty , argued that thorax size cannot be used as a predictor or explanatory variable since fruitflies who die young may not be fully grown, i.e., it is also an "intermediate" variable. Later, student NK (now on faculty elsewhere) had studied entomology & assured us that fruitflies do not grow longer after birth; i.e., thorax length is not time- (age)-dependent!

a *Use a c621-type regression model (and datafile 'fruitfly') to estimate the (absolute) difference in mean longevity of sexually active flies (index category) relative to sexually inactive flies (reference category), ignoring other covariates. Is this difference (i) substantial? (ii) statistically significant at the conventional alpha=0.05 level?*

b *How different are the mean thorax lengths of the active and inactive flies? Is this difference "statistically" significant? Is statistical significance a non-issue here anyway? Explain.*

c *If -- other things being equal -- flies 0.01 mm larger live on average 1 day longer, how much of a longevity "advantage" would the active flies have as a result of their larger average thorax size? On this basis, how much lower is the mean longevity of active than inactive flies if "adjusted" for the difference in thorax size?*

d *Instead of using the "out of the air" value of 1 day/ 0.01 mm, use multiple regression to simultaneously estimate the additional mean days/mm and the decrease in days associated with (due to) activity i.e., fit the model: average[longevity | thorax type] =* $B_0 + B_{thorax} \times thorax + B_{active} \times active$

e *Verify that if you correct/adjust the comparison in (a) using the fitted* b_{thorax} *from (d) and the thorax difference in (b), you arrive at the* b_{active} *obtained in (d). Hint: cf JH notes on confounding.*

f The p-value for the activity contrast in (d) is smaller (& the associated CI narrower) than the corresponding one in (a). One reason is that the larger adjusted estimate of the effect (the numerator of t-test on *adjusted* difference); another is the *smaller* SE of the estimated effect (denominator of t-test). *Why is the SE of the estimated longevity difference from analysis (d) smaller?*

The longevity of the 2 groups can also be compared by survival analysis methods.

g *Use p h models to obtain "crude" & adjusted hazard ratios (again treat thorax as continuous, & use datafile 'fruitfly'). Did the adjusted coefficient move in the direction, and by the amount, you expected? Explain.*

h *Repeat question g, but for flies with 8 partners: Do the adjustments go in the direction you expected? Explain.*

i *Use logs to write the p h models so that the right hand side has the same additive forms as in (a) and (d)*

j Some flies began adult life on Mondays, some on Tuesdays etc. The research assistant entered data for each fly each Saturday, making a separate (partial) record for each week, or part thereof: for example, a fly who began as an adult on Tuesday, and ultimately lived 62 days, has 10 records -- each of the first 9, namely those where (t_0,t_1) = (0,4), (4,11) to.(53,60) is accompanied by a "censored" indicator, and a 10th record, with (t_0,t_1) = (60,62) is accompanied by a "complete" indicator *(Run SAS/Stata program and inspect the "byweek " datafile created). Repeat the crude & survival comparisons with these split observations. Explain why the likelihoods, beta_hats, SE's etc. are identical to those in (g). A diagram, with timelines & risksets, may help.*

k On the weekend, the RA's boss, not knowing how to get SAS/Stata to put the 1154 split records back together, decided that one should only use the 50 split records that terminated in a death (see near the end of the program file). *Run the boss's analysis, and explain why it gives a very different (wrong?) answer.*

l Rather than use thorax size as a term in (and a coefficient to be estimated from) the regression, *use the variable thorax_Q (quintiles 1 to 5) as a stratum variable and re-estimate the HR for the active relative. to the inactive group. Why is the likelihood much larger (the logL less negative) in the stratified analysis? Hint: examine sizes of the risksets, and likelihood contributions, in Figures 1 and 3 of JH's draft article on Survival analysis; risk sets; case control studies: part II.*

m *Use the Schoenfeld residuals, & log[-Log[S]] plots, to visually assess if the p h assumption is reasonable in this dataset. (consult onlinedoc or other documentation)*

n *What role should the variable SLEEP have in this analysis? Based on the data, does it seem to be influential/relevant?*

o *Compare beta_hats in (g) with those from "ran_out" dataset. Explain why the larger SE's, and by how much.*

2 Clayton &Hills, Ch 30; Exercises 30.1 and 30.2 (attached, below)

Sexual activity and the lifespan of male fruitflies

ABSTRACT: A cost of increased reproduction in terms of reduced longevity has been shown for female fruitflies, but not for males. The flies used were an outbred stock. [In a randomized trial] sexual activity was manipulated by supplying individual males with one or eight receptive virgin females per day. The longevity of these males was compared with that of two control types. The first control consisted of two sets of individual males kept with one or eight newly inseminated females. Newly inseminated females will not usually remate for at least two days, and thus served as a control for any effect of competition with the male for food or space. The second control was a set of individual males kept with no females. There were 25 males in each of the five groups, which were treated identically in number of anaesthetizations (using CO2) and provision of fresh food medium.

SOURCE:
Figure 2 in the article "Sexual Activity and the Lifespan of Male Fruitflies" by Linda Partridge and Marion Farquhar. Nature, 294, 580-581, 1981.

Variable	Description
ID	Serial No. (1-25) within each group of 25 (the order in which data points were abstracted)
PARTNERS	Number of companions (0, 1 or 8)
ACTIVE	Type of companion 0: newly pregnant female 1: virgin female 9: not applicable (when PARTNERS=0)
LONGEVITY	Lifespan, in days
THORAX	Length of thorax, in mm (0.xx)
SLEEP	Percentage of each day spent sleeping

NOTES: "Compliance" of the males in the two experimental groups was documented as follows: On two days per week throughout the life of each experimental male, the females that had been supplied as virgins to that male were kept and examined for fertile eggs. The insemination rate declined from approximately 7 females/day at age one week to just under 2/day at age eight weeks in the males supplied with eight virgin females per day, and from just under 1/day at age one week to approximately 0.6/day at age eight weeks in the males supplied with one virgin female per day. These `compliance' data were not supplied for individual males, but the authors say that "There were no significant differences between the individual males within each experimental group."

If interested, see also: Hanley, J. A. (1983), "Appropriate Uses of Multivariate Analysis," *Annual Review of Public Health*, 4, 155-180. (on c697 website under other Resources) and *Nature* article -- and especially the original Figure from the Nature article (on c622 website under Datasets)

Life cycle of Drosophila: The drosophila egg is about half a millimeter long. It takes about one day after fertilisation for the embryo to develop and hatch into a worm-like larva. The larva eats and grows continuously, moulting one day, two days, and four days after hatching (first, second and third instars). After two days as a third instar larva, it moults one more time to form an immobile pupa. *Over the next four days, the body is completely remodelled to give the adult winged form, which then hatches from the pupal case and is fertile within about 12 hours*. (timing is for 25°C; at 18°, development takes twice as long.) [http://www.ceolas.org/fly/intro.html]

Table 30.1. A cohort of 10 subjects

		Entry to Study		End of Study	
Subject	Sex	Date	Age	Date	Age
A	F	13/ 6/65	29.3	31/12/89	53.8
B	M	23/10/72	25.2	31/12/89	42.4
C	M	3/ 3/59	22.1	31/12/89	52.8
D	F	10/10/67	32.2	31/12/89	54.4
E	M	2/ 1/60	33.1	4/ 7/79	52.6
F	M	9/ 1/75	42.1	31/12/89	57.1
G	F	5/ 8/53	35.2	3/10/68	50.4
H	M	10/10/69	27.0	31/12/89	47.2
I	M	2/ 3/72	44.8	31/12/89	62.7
J	F	1/11/70	51.5	31/12/89	70.6

likelihood as the amount of data increases.

The composition of risk sets (and hence the results of the analysis) depend upon the choice of time scale for the analysis, as is demonstrated by the following exercise.

Exercise 30.1. The data set out in Table 30.1 refer to 10 subjects from a cohort study. Subjects *E* and *G* died at the second date while the remaining eight subjects survived until the date of analysis (31/12/89). List the members of the risk sets for both deaths when the appropriate time scale is (a) calendar date (b) age (c) time since entry into the study.

The difference between these analyses is that they represent three different models. In each case the λ_C^t parameters represent variation of baseline rates along different time scales.

30.4 Choice of time scale

Our derivation of Cox's method allows for time to be interpreted in the most appropriate manner for a particular analysis. Usually this will mean the time scale with the strongest relationship to failure rate. Regrettably it is still the case that some major software packages do not allow such flexibility. This reflects the fact that the method was motivated by problems of survival following medical treatment. In such studies the appropriate time scale is time since start of follow-up so that all observation of all subjects starts at time zero. In such studies, risk sets always become smaller (as a result of failure and censoring) as time advances.

On other time scales there will be *late entry* of subjects (observation starting at time > 0) and risk sets may be supplemented by new entrants as time advances. In order to be able to select the most appropriate time scale for an analysis, the software must be capable of allowing for late entry.

30.5 Confounders other than time

The confounding effect of time is allowed for by including time in the first part of the model. For example, taking age as the time variable, the multiplicative model

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{A} \times \text{B}},$$

includes the effect of age in the baseline rate parameters. The most obvious way to deal with another confounder, such as sex, is to include it in the second part of the model, as in

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age}} \times \boxed{\text{Sex} \times \text{A} \times \text{B}}.$$

This model assumes that the effect of sex is constant with age so that the baseline rates for males are a constant multiple of those for females. To extend the model to allow for different patterns of baseline rates for each sex, the interaction between age and sex must be included in the model. When the age scale is divided into clicks this interaction term involves a very large number of parameters, so it is best to absorb these parameters in the baseline rate part of the model, giving

$$\text{Rate} = \boxed{\text{Corner} \times \text{Age} \times \text{Sex} \times \text{Age}\cdot\text{Sex}} \times \boxed{\text{A} \times \text{B}}.$$

This model has the effect of allowing different sets of baseline rate parameters for males and females. If we estimate these algebraically as before, we find that the profile likelihood for the rate ratio part of the model still has the form of a partial likelihood:

$$\sum_{\text{Failures}} \log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right)$$

but the risk set is now restricted to contain only those subjects who (a) were under study at the time of failure of the case, and (b) belonged to the same sex as the case. Thus the analysis simulates a matched case-control study in which controls are matched to cases with respect to sex.

This extension of Cox's method is usually referred to as a stratified analysis, although more properly it should be referred to as *doubly* stratified — Cox's method stratifies by time alone, while the extended method stratifies by both time and a further variable. In our example stratification is by age and sex.

Exercise 30.2. Repeat Exercise 30.1 for an analysis which is to be stratified by sex.

It can be seen from the last exercise that when an analysis is doubly strat-