EPIB-609: Assignment based on parts I and II of "Survival analysis; risk sets; matched case control studies: a unified view of some epidemiologic data-analyses" (v2012.11.11)

The purposes are to (i) stress the principle of efficiency (and informative strata) in the estimation of rate ratios (ii) familiarize you with statistical techniques for 'matched-set' and ‘not-fully-matched-set’ data, and (iii) understand how risksets and likelihoods are set up in the Cox Model and in conditional logistic regression more generally.

1. For the vasectomy study (see website, and part I), write out the full formula for the summary M-H rate ratio (aliases: summary incidence ratio [Rothman terminology], summary incidence density ratio [Miettinen terminology]). Show that the resulting 1.25 is a ratio of the weighted sum of 4830 products to the (same) weighed sum of 4830 other products. Note the ratio-estimator’s 'architecture': a ratio of a single, and thus more stable, number, to another single, more stable, number; the 4830 'weights' used to combine the a*PT_0 products into a single M-H numerator and the 4830 b*PT_1 products into a single M-H denominator are w = 1/PT = 1/(PT_1+PT_0). (Mantel spoke to Miettinen of ‘slapping down’ the unruly a*d and b*c products in his original summary OR formula).

2. What would happen if (as some authors teach) the formula were written as a weighed average of 4830 individual (i.e. pair-specific) ratios? In other words, you could take the M-H formula

SummaryRateRatio = Sum{a*PT_0/PT} / Sum{b*PT_1/PT},

divide and multiply each a*PT_0/PT by b*PT_1, and re-arrange to get

SummaryRateRatio = Sum{ [a*PT_0/ b*PT_1] * [b*PT_1/PT] } / Sum{b*PT_1/PT}.

Written this way, it ‘looks like’ a weighed average

SummaryRateRatio = Sum{ rateratio * W} / Sum{W},

where rateratio = a*PT_0/ b*PT_1 = (a/PT_1) / (b/PT_0), W = b*PT_1/PT, and Sum{ } has 4830 items.

If you used Excel or R, or your hand calculator, to compute this weighted average of 4830 pair-specific rate ratios, what would be the result?

3. For the 4-pairs example in Figure 3, compute the likelihood (and thus the log Likelihood) at HR=2.5, HR=3 and HR=3.5. *Hint*: the likelihood is the product of the (4) probabilities of observing what you observed in the 4 risksets, i.e., the MI in riskset #1 occurred in the vasectomized man, in #2 in the vasectomized man, in #3 in the non-vasectomized man, and in #4 in the vasectomized man. So Lik = P(‘Mr. V’) x P(‘Mr. V’) x P(‘Mr. NV’) x P(‘Mr. V’), so LogLik = sum of the logs of these.

4. Compute the M-H summary rate ratio using (i) only the matched-on-smoking (ii) only the matched-on-smoking-and-obesity pairs. (R hardly necessary, but code is available on website)

5. For the 36-pairs example in Figure 4, and using a calculator, or a spreadsheet or your favourite statistical package, compute the likelihood (and thus the log Likelihood) at HR=1.25 and at HR=1.5, where HR refers to the ratio of hazard rate in vasectomized to that in non-vasectomized men. As in the calculations in Figure 4, assume that the hazard rate in smokers is <u>2</u> times that in nonsmokers. *Hint*: the likelihood is the product of the (36) probabilities of observing what you observed in the 36 risksets, i.e., 4 risksets where the MI occurred in the non-smoking non-vasectomized man, 6 where the MI occurred in the non-smoking vasectomized man, etc. So, taking S, V, and NV as shorthand for ‘smoker’,

‘vasectomized’ and ‘not-vasectomized’,
Lik = $P(NV)^4$ x $P(V)^6$ x $P(S\ NV)^4$ x $P(S\ V)^5$ x $P(NV)^1$ x $P(S\ NV)^6$ x $P(S\ NV)^7$ x $P(V)^3$
so, LogLik = sum of the logs of these.

Figure 4 *fixes* the HR for smoking at 2, and using this fixed (assumed) value, finds the HR for vasectomy that maximizes the LogLik. Describe how you could extend your calculations to find the ML estimates of both the HR for smoking and the HR for vasectomy (hint: see Fig 5).

6. Via software (see website), obtain an adjusted-for-smoking-obesity hazard ratio using conditional logistic regression, or (equivalently) a stratified version of Cox regression analysis. Also, explain in plain language how *conditional* differs from *unconditional* logistic regression.

7. The self-paired data obtained from Ayas (percutaneous injuries in the ICU, 3rd row of Table 3, see website), the Walker study, and the mmr-autism study (déjá vu), are all examples of the role of statistical efficiency in the estimation of rate ratios. JH’s Discussion section in part I refers to the ‘matched sets’ data analyzed in “Short and long term mortality associated with foodborne bacterial gastrointestinal infections” by Helms and Evans commentary (“Matched cohorts can be useful” -- cf. website). Interestingly, Evans doesn’t address statistical efficiency. Prepare a suitable paragraph on this topic that could be inserted into Evans’ commentary.

8. Refer to the *fruitfly* study (see website, and part II of the article). For the 10-flies example in Figure 1, compute the likelihood (and thus the log Likelihood) at HR = 3, where HR is the hazard ratio for the Active versus non-active contrast, and thorax size is ignored.

9. For the 10-flies example in Figure 2, compute the likelihood (and log Likelihood) at HR_A=3 and HR_S = 2, where HR_A = 3 is the hazard ratio for the Active versus non-active contrast, and HR_S =2 is the hazard ratio for the Shorter thorax versus longer thorax contrast.

10. For the ‘stratified Cox’ example in Figure 3, compute the log likelihood at HR_A=3.

11. From the (8 informative) stratified risksets, compute a summary M-H ratio. *[Aside, in relation to Mantel’s comment to OSM: what if you `rolled your own’ ratio formula, with no `slapping down’ of products.. ie what if you used the formula Ratio = Sum(ad)/Sum(bc)?]*

12. Briefly describe how ‘case-controlling’ as a *data-analysis outlook* is central to how risksets are used to fit the parameters in Cox regression.

13. Which of the above examples takes (a) a ‘matching/stratification only’ (b) ‘regression only’ (c) ‘a mix of these’ approach to confounding-control? What are some upside/downsides of each?

14. How does the model used to predict Oscar winners (see Conditional Logistic Regression link in bios602 website) relate to conditional logistic regression in epidemiology? How close is the latter to the McFadden model in choice-based sampling in marketing and transportation-choices research, and mentioned often in Stata?
(See http://en.wikipedia.org/wiki/Logit and http://en.wikipedia.org/wiki/Daniel_McFadden)

E-mail preliminary answers before the beginning of class on November 19, 2012.