STAT 536B - Biostatistics (Second-half of Term 2, 2014-15)

ASSIGNMENT

NOTE: More questions will be added as we cover material.

NOTE: Some of the problems are deliberately open-ended, giving you the opportunity to investigate as you best see fit. One tradeoff is as follows. I will not assign a large number of problems, but in return I will be looking for well-documented solutions. At a minimum, you should be reporting your findings in complete sentences/paragraphs (not computer-code fragments!), and using tables and/or figures to summarize empirical work as appropriate. One guideline that cuts across all scientific work is that you should provide enough detail so that an interested reader could replicate what you have done. In terms of using mathematical notation versus words, I'm not fussed. That is, some of you will be more comfortable using less/more mathematical notation than others, and that's fine. Clarity can be achieved either way!

1. [added Feb. 27] A medical researcher comes to you for advice. She has enough funding to carry out an unmatched case-control study with 300 study participants in total, and she has unlimited flexibility as to how many of these should be cases versus controls. Upon discussion with her, you learn that:

- She would like to estimate the exposure-disease log odds-ratio as accurately as possible.
- Her best *a priori* guess is that 10% of healthy people are exposed.
- Her best *a priori* guess is that there is a very strong exposure-disease association, with an odds ratio of four.

Give a recommendation to the researcher on how many cases to sample.

2. [added Mar. 4] Say the population distribution of binary variables (Y, X, C) is completely characterized according to:

- X and C are independent of one another, each having a Bernoulli(0.5) distribution.
- $Pr(Y = 1|X = 0, C = c)$ is 0.2 when $c = 0$ and 0.5 when $c = 1$.
- The stratified odds-ratios are $OR(Y, X|C = c) = 4$, for both $c = 0$ and $c = 1$.

Is C a confounder in this population? What is the value of the crude odds-ratio, $OR(Y, X)$? And what have you thus learned about about the behaviour of odds-ratios compared to the behaviour of risk-ratios?

3. [added Mar. 8] At `www.clinicalpredictionmodels.org` you can find the 'West Region' subset of the Gusto-I data. (The datafile is in SPSS format, but I found that the 'read.spss' function in the 'foreign' library easily imports this into an R data frame.) This portion of the data constitutes 2188 subjects. Table 24.3 in Steyerberg gives variable labels.

For this question, we will *pretend* that these 2188 subjects comprise our whole **population** of interest, and will see how well we can estimate population quantities from **samples** from this population. Particularly, say we are interested in the association between 30-day mortality as the outcome and HIG (the 'high risk' indicator) as the exposure, conditioned on confounders AGE and $ST4$ (a binary indicator of a more abnormal electrocardiogram). Consider the logistic regression of $DAY30$ on (1,HIG,AGE,AGE^2,$ST4$), fit to all the data, to be 'the' population relationship of interest. In particular, the association between HIG and $DAY30$ given AGE and $ST4$ is the target of inference.

(a) Note that your population only contains 135 cases. For each case, sample a matching control, so that your data are comprised of 135 pairs. (For AGE, consider a difference of 1 year or less to qualify as a match.) Carry out an appropriate analysis of these paired data, and comment briefly on what you find.

(b) Now consider a different study design that would also result in data on 270 subjects. In particular, simply sample 270 population members at random, as per a cross-sectional study. Compare an estimate from these data to what you got in (a). To keep things simple, let's say that you know the correct terms to put in the model you fit.

(c) Of course we are wary about inferring differences between study designs or methods of analysis based on just one realization of data. So replicate what you did in (a) and (b) 200 times, and comment briefly on what you find.

4. [added Mar. 8] A worry is that an unscrupulous researcher could use dichotomization of exposure in a sneaky way, to overstate the evidence in favour of an exposure-disease relationship. In particular, a researcher could try many analyses with different cutpoints, but only report the "most favourable" analysis, without stating that the others were carried out. Let's construct a small demonstration concerning how misleading this might be.

Say that in reality X and Y are independent, with $X \sim N(0,1)$ and $Y \sim \text{Bern}(0.3)$. The study will involve a cross-sectional sample of $n = 500$ subjects. The investigator's hidden plan is to conduct seven analyses, based on dichotomizing X at cutpoints ranging from -1.5 to 1.5 in increments of 0.5. But he/she will present the analysis with the smallest P-value for testing the null of no (X, Y) association, and, in particular, he/she will claim evidence for an association if this smallest P-value is below 0.05.

What is the investigator's Type I error rate here?

This is the batch #1 cutoff. Problems 1 through 4 are due Tuesday March 24th. Other problems will be posted before then, but they will be part of the next batch.

5. [added Mar. 21] Take a quick look at "Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials" (Dias et al, *Med. Decis. Making* 2013;33:607617). In particular, consider the data in Table 1, and the analysis in the right-half of Table 3. This analysis is based on a "full-blown" Bayesian hierarchical model, in which each study contributes two binomial counts.

In class we studied a simpler approach to random-effect meta-analysis whereby each study simply contributes an estimate and a standard error. Carry out this simpler analysis for these data, and see if you obtain a similar answer to the full-blown analysis.

(To save you some typing, these data area available various places on line, for instance at `mathstat.helsinki.fi/openbugs/Examples/Blockersdata.html` for an easy to cut-and-paste into R format.)

6. [added Mar. 27] We briefly mentioned in class a situation where a **nested case-control** study might be particularly cost-effective. The purpose of this question is to construct a demonstration to back that up.

Consider a binary exposure X for which 10% of the study population are exposed. A study will recruit a random sample of n population members and follow them for six months. For the disease being studied, the time from study entry to disease outcome (in years) is exponentially distributed with hazard rate 0.25 for unexposed subjects and 0.37 for exposed subjects. But of course you only observe the disease outcome if it occurs within 6 months of study entry.

For this question you will need to be able to simulate data under these conditions, and carry out two analyses to assess the exposure-disease relationship. One analysis is a simple survival proportional hazards analysis for the outcome time given X, taking into account the censored data. The other analysis is the nested case-control study as described in class. For every observed outcome (case), a matched control is picked at random from all those subjects who were still "at risk" at the time the case reached the outcome. Going back some weeks in the course, we know that the pertinent data will be the X values for each matching pair. And we know how to obtain an estimate and SE for the exposure-disease association from these data.

To make things interesting, say that in fact the X is both expensive to measure and can be measured after the fact (think X is a genotype, can be measured on a frozen blood sample...). In fact, say that the cost of acquiring a subject without X measured is only one-third the cost of acquiring a subject with X is measured.

Generate a "telescoping" sequence of datasets, i.e., the first dataset has $n = n_1$, the next one has $n = n_2 > n_1$ and is comprised of the original n_1 subjects plus $n_2 - n_1$ additional subjects, and so on. For each dataset, compute an estimate and standard error for both methods. And display these *as a function of the total cost of acquiring the data* in each case. What do you conclude?

7. [added Mar. 28] Locate a dataset which has a continuous outcome variable Y and a bunch of explanatory variables, at least one of which is binary, and can be treated as exposure X (while the others are treated as C).

[a] Provide four different estimates (along with standard errors) of $\Delta = E\{E(Y|X = 1, C) - E(Y|X = 0, C)\}$, using what we referred to in class as the regression estimator, the inverse-probability weighted estimator, and the double-robust estimator, as well as the estimator based on grouping the data according to quintiles of propensity.

[b] Also determine a fifth estimate (and standard error) by using a tweaked version of the IPW estimator:

$$\hat{\Delta}_{IPW2} \frac{\sum_{i=1}^n y_i\{x_i/\hat{\pi}(c_i)\}}{\sum_{i=1}^n\{x_i/\hat{\pi}(c_i)\}} - \frac{\sum_{i=1}^n y_i\{(1-x_i)/(1-\hat{\pi}(c_i))\}}{\sum_{i=1}^n\{(1-x_i)/(1-\hat{\pi}(c_i))\}}$$

Can you give an argument for why $\hat{\Delta}_{IPW2}$ is estimating the same target as $\hat{\Delta}_{IPW}$?

8. [added Mar. 31] Give a brief demonstration (either empirical or theoretical) showing that in the instrumental variable set-up we can estimate the target parameter better when the association between the instrumental variable and the exposure variable is stronger.

This is the second/final batch cutoff. Problems 5 through 8 are due Thursday April 9th (the last day of classes).