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

Assignment 5

material in www.epi.mcgill.ca/hanley/c681/alr_1 unless otherwise specified

Exercise 4. The Prostate Cancer Study (from Hosmer and Lemeshow text; p 30 in 2nd edition)

Do all parts except (d) and (j). For part (h) use a PSA level of 10.

Homework 1 revisited ...

Q1 You observe 3 "positives" in a simple random sample of size n = 20 from a certain 'source' and wish to obtain a CI for the proportion positive (π) in the source.

(a) Set the data up as 20 separate observations and use a logistic regression software program to obtain/calculate a 'logit-based' 95% (frequentist) CI for π. [see p6 of JH Notes for Chapter 8.1 in course 607]

(b) Set the data up as 1 (or in Stata at least 2) aggregated observation(s) and again estimate the logit of π. Comment

(c) Repeat either (a) or (b) using a generalized regression software program (one where you have to specify which distribution is used (here binomial) and which link (here logit)

The ultimate purposes of (a) and (b) are

- *to become used to aggregated data, i.e., where the units of analysis are "cells" i.e. groups of individuals having the same covariate pattern. This saves time/effort if a dataset is very large; gets us mentally prepared for (Poisson) regression with person-time data, where the units are 'collections' of person time, often aggregated over (parts) of many persons' experiences; is the preferred way to judge the 'goodness-of-fit' for models where responses are binary -- Hosmer and Lemeshow's test groups together into the same "cell" subjects who may have unique covariate patterns, but have close to the same predicted probabilities.*
- *appreciate that the 'intercept-only' model is the 'beginning of all regressions' and that even the single sample estimation/testing problem for means or proportions in 607 can be set up as a regression model.*

the ultimate purposes of (c) are

- *to become comfortable with generalized linear models, which we will need in any case when we come to Poisson (and even other binary) regressions, and to start seeing the unification of approaches and principles that comes from viewing most regressions as special cases of a general structure [the Cox model for survival data takes a bit more manipulation to force it into a common mould]*

Q3 Patients undergoing surgery were randomly assigned to routine intraoperative thermal care or additional warming. Surgical-wound infections were found in 18/96 patients assigned to hypothermia (19 percent) but in only 6/104 patients assigned to normothermia (6 percent, P=0.009).

(a) Use a logistic regression or generalized linear model software program to obtain/calculate a point- and a 95% interval- estimate for the odds ratio (use the routine intraoperative thermal care as the reference category)

(b) Use a generalized linear model software program to obtain/calculate a point- and a 95% interval-estimate for the risk difference, and of the number required to treat.

There are several purposes to Q3:

find ways to avoid having to type in 200 individual observations(!); formulate a comparison of two comparisons as a regression problem and avoid having to calculate CI's for odds ratios and risk ratios by hand (moving up from the beginning of all regressions); provide a counterpoint to H&L who think that one should fit regressions with a continuous X before one with a two-point X; to have the alternative tools and be able to say to others that just because we have the software for logistic regression doesn't mean it is right for every situation. In this study, how much it costs to prevent an infection depends on the risk difference, not the risk or odds ratio; we often have to compute risk differences that are adjusted for other covariates, so saying "I can do it all by hand, who needs regression?" is not a realistic option.