Fitting models to stratified samples

Alastair Scott
(joint work with Chris Wild)

University of Auckland

Problem

We want to fit a model, $f(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{\theta})$ say, for the distribution of a response,$\boldsymbol{y}$, as a function of explanatory variables, $\boldsymbol{X}$, to data from stratified samples.

- Particularly interested in situations where the probability that a unit is in a particular stratum may depend on the value of $\boldsymbol{y}$

Example 1: Case-control family study

Study of glioma, the most common type of malignant brain tumour (see Wrensch et al (1997)):

- all newly diagnosed cases in a specified time interval in the San Francisco Bay area
- a population-based sample of controls through random digit dialling.
- brain tumour status and covariate information from the original case-control sample and, in addition, from members of their families.

Example 1

Want to fit a mixed logistic model

$$\text{logit } P(Y_{ij} = 1 \mid b_i) = \beta_0 + \beta_1 ep_{ij} + \beta_2 ca_{ij} + b_i$$

where b_i is a random family effect with $b_i \sim N(0, \sigma^2)$. Here

- Y is a binary indicator of brain cancer status,
- ep is a binary indicator of epilepsy history,
- ca is a binary indicator of other cancers.

Example 1

- If we only had information from the original case-control probands, we would just fit a standard logistic ignoring the differential sampling rates (with an adjustment to the constant if wanted).
- Natural to try to mimic this and fit a mixed logistic model ignoring the case-control sampling
  - unfortunately, this doesn't work!

General set-up

- Population of N units, split into strata $\mathcal{S}_1, \ldots, \mathcal{S}_L$ with N_h units in $\mathcal{S}_h$.
- We draw independent random samples, D_h, of n_h units from the N_h units in $\mathcal{S}_h$ $(h = 1, \ldots, L)$ and observe the corresponding $(\boldsymbol{y}, \boldsymbol{X})$ values.
- Data are of the form:
  $\{(\boldsymbol{y}_{hj}, \boldsymbol{X}_{hj}, j \in D_h), N_h; h = 1, \ldots, L\}$.
  - N_hs provide information about stratum probabilities

General set-up

- No problem if the probability of being in a stratum depends on $\boldsymbol{X}$ alone.
- However, if stratum probabilities depend on $\boldsymbol{y}$, then the sampling is informative and cannot be ignored in the analysis.
  - Sometimes (in many retrospective medical studies, for example) the dependence is explicit and clear.

Example 2

Stratified cluster sample of individuals enrolled in the Federal Employees Health Benefit Plan (Neuhaus & Jewell (1990)).

- Response variable indicates whether someone used outpatient mental health services in the previous year for each of 1979-81.
- Clusters consist of the three values for a single person
- Four strata defined by the four possible values of the cluster totals with sampling fractions

$$f_0 = .03, \quad f_1 = \frac{1}{3}, \quad f_2 = \frac{2}{3}, \quad f_3 = 1$$

General case

In most surveys the dependence is less clearcut with stratification determined by such things as administrative convenience or the availability of a suitable list. Often, the survey will be designed by someone else and it is not always clear why a particular stratification was chosen.

- "Even stratification could be informative to someone who knows less about the population than the survey designer" (Scott (1975))

General case

In general, if we are not sure about the stratifying variables (or even if we are sure, but don't want to include them all in the model) we may have to fit a parametric model, $p_h(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\gamma})$ say, for the conditional probability of a unit being in the hth stratum given values of $\boldsymbol{y}$ and $\boldsymbol{X}$.

- In our glioma example, only about one case in 70 falls in the specified sampling period.
  - natural model $p_1(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\gamma}) = \gamma \sum_j y_j$

General set-up

One posssible strategy: include stratum indicator as a covariate in the model.

- obviously not sensible if the indicator is the response itself
- can't if we don't know the stratifying variable
- may distort the relationship of interest even if possible (eg SIDS study)

Weighted Estimators

The standard alternative is to use weighted estimating equations.

- Stratum sizes (and hence weights) are random in most cases
  - strictly need to account for this in standard errors
- can use results from double sampling for stratification to get
  $\widehat{V}\left\{\widehat{\boldsymbol{\theta}}\right\} = \left\{\boldsymbol{\mathcal{I}}_W^{-1} + (\frac{1}{N}\boldsymbol{\mathcal{I}}_W)^{-1}\left(\sum_h W_h^2 \frac{(1-f_h)}{n_h}\widehat{\Sigma}_h\right)(\frac{1}{N}\boldsymbol{\mathcal{I}}_W)^{-1}\right\}$.

Weighted Estimators

- Works well if weights don't vary too much in size.
- Result always has a natural population interpretation independent of the particular sampling scheme even if the model is wrong.
- However, it can be extremely inefficient with very unequal weights (as in our example and most retrospective studies)

Likelihood Approaches

We explore more efficient approaches based on the likelihood:

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}, g) = \prod_{h=1}^{L} \{ \prod_{j \in D_h} \text{pr}(\boldsymbol{y}_{hj}, \boldsymbol{X}_{hj} \mid \text{unit in } \mathcal{S}_h) \} \text{pr}(\text{unit in } \mathcal{S}_h)^{N_h}$$

$$= \prod_{h} \left(\prod_{D_h} \{ p_h(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\gamma}) f(\boldsymbol{y}_{hj} | \boldsymbol{X}_{hj}; \boldsymbol{\theta}) g(\boldsymbol{X}_{hj}) \} Q_h^{N_h - n_h} \right)$$

- $g(\boldsymbol{X})$ denotes the marginal (population) density of X
- $Q_h = Q_h(\boldsymbol{\theta}, \boldsymbol{\gamma}, g)$ denotes the marginal probability that a unit is in $\mathcal{S}_h$.

Likelihood Approaches

- With simple random sampling we can make inferences about $\boldsymbol{\theta}$ conditional on the observed values of $\boldsymbol{X}$ and ignore terms involving $g(\boldsymbol{X})$.
- Unfortunately, we can't ignore $g(\boldsymbol{X})$ here: since $Q_h = Q_h(\boldsymbol{\theta}, \boldsymbol{\gamma}, g) = \int \int p_h(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\gamma}) f(\boldsymbol{y}|\boldsymbol{X}; \boldsymbol{\theta}) g(\boldsymbol{X}) d\boldsymbol{y} d\boldsymbol{X}$
  - $g(\boldsymbol{X})$ becomes a (potentially infinite dimensional) nuisance parameter in the likelihood.

Likelihood Approaches

- A full likelihood approach requires us to model the covariate distribution, $g(\boldsymbol{X})$
  - possible in simple cases (e.g. see Pearson (1903), DeMets & Halperin (1977))
  - problems involved in modelling $g(\boldsymbol{X})$ are overwhelming as soon as we get past two or three covariates.
- We want to develop semi-parametric methods in which $g(\boldsymbol{X})$ is left completely unspecified
  - sounds formidible - turns out to be relatively simple in practice.

Semi-parametric approach

Let

$$\ell^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) = \sum_h \sum_{D_h} log(f^*(\boldsymbol{y}_{hj}|\boldsymbol{X}_{hj}; \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\pi}))$$

$$- \sum_h \left[(N_n - n_h) log(1 - \pi_h) + n_h log \pi_h\right],$$

with $f^*(\boldsymbol{y}_{hj}|\boldsymbol{X}_{hj}; \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\pi})$ defined by

$$f^*(\boldsymbol{y}_{hj}|\boldsymbol{X}_{hj}; \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\pi}) \propto \pi_h p_h(\boldsymbol{y}_{hj}, \boldsymbol{X}_{hj}; \boldsymbol{\gamma}) f(\boldsymbol{y}_{hj}|\boldsymbol{X}_{hj}; \boldsymbol{\theta}).$$

The quantity $\ell^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \widehat{\boldsymbol{\pi}}_n(\boldsymbol{\theta}, \boldsymbol{\gamma}))$ is the profile likelihood obtained by maximizing the likelihood over g($\boldsymbol{X}$).

Semi-parametric approach

Although $\ell^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\pi})$ is not itself a likelihood, we obtain efficient semi-parametric estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ by acting as if it were.

More generally, for making inferences about $\begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \end{pmatrix}$, we can act just as if the pseudo-likelihood, $\ell^*(\boldsymbol{\phi})$, where $\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \\ \boldsymbol{\pi} \end{pmatrix}$, is an ordinary likelihood.

Semi-parametric approach

Specifically:

- setting the pseudo score, $S^*(\boldsymbol{\phi}) = \partial \ell^*(\boldsymbol{\phi})/\partial \boldsymbol{\phi}$, equal to zero gives the semi-parametric MLE;
- the resulting estimators are asymptotically normal;
- an estimate of $\mathrm{Cov}\{\widehat{\boldsymbol{\theta}}\}$ can be obtained directly from the appropriate block of the inverse of the observed pseudo information matrix, $\boldsymbol{\mathcal{J}}^*(\widehat{\boldsymbol{\phi}})^{-1}$.

In effect, we have replaced an infinite dimensional nuisance parameter, $g(\boldsymbol{X})$, by an (L$-$1)-dimensional parameter,π. (See Alan Lee's talk tomorrow for more details.)

Example

We have two strata here: one containing all 463 families with someone who has been diagnosed in the specified time interval, and the other containing the remaining $N_2 = 1,942,490$ families without such a member.

- We need to fit a model for the probability of a family being in $\mathcal{S}_1$ given $\boldsymbol{y}$ and $\boldsymbol{X}$.
- Natural to model the probability of being in $\mathcal{S}_1$ (i.e. of containing a recently diagnosed case) given $\boldsymbol{y}$ and $\boldsymbol{X}$ as proportional to the total number of members with brain cancer, say $p_h(\boldsymbol{y}_i, \boldsymbol{X}_i; \gamma) = \gamma \sum_j y_{ij}$

Results for the glioma data

	$\widehat{\beta}_0$(se)	$\widehat{\beta}_1$(se)	$\widehat{\beta}_2$(se)	$\widehat{\sigma}$(se)
Semipar:	$-5.29(.18)$	$2.12(.25)$	$0.32(.16)$	$.88(.19)$
Conditional:	–	$2.09(.36)$	$0.31(.21)$	–

Results for the glioma data

	$\widehat{\beta}_0$(se)	$\widehat{\beta}_1$(se)	$\widehat{\beta}_2$(se)	$\widehat{\sigma}$(se)
Semipar:	$-5.29(.18)$	$2.12(.25)$	$0.32(.16)$	$.88(.19)$
Conditional:	–	$2.09(.36)$	$0.31(.21)$	–
Probands only:	–	$1.78(.39)$	$-.02(.21)$	–

Results for the glioma data

	$\widehat{\beta}_0$(se)	$\widehat{\beta}_1$(se)	$\widehat{\beta}_2$(se)	$\widehat{\sigma}$(se)
Semipar:	$-5.60(.18)$	$2.12(.25)$	$0.32(.16)$	$.88(.19)$
Conditional:	–	$2.09(.36)$	$0.31(.21)$	–
Semip. (no N):	$-5.64(.54)$	$2.12(.26)$	$0.32(.16)$	$.00(.00)$

Results for the glioma data

	$\widehat{\beta}_0$(se)	$\widehat{\beta}_1$(se)	$\widehat{\beta}_2$(se)	$\widehat{\sigma}$(se)
Semipar:	$-5.60(.18)$	$2.12(.25)$	$0.32(.16)$	$.88(.19)$
Semip. (no N):	$-5.64(.54)$	$2.12(.26)$	$0.32(.16)$	$.00(.00)$
Semip.(γ, no N):	$-5.64(.54)$	$2.12(.26)$	$0.32(.16)$	$.89(.27)$

Results for the glioma data

	$\widehat{\beta}_0$(se)	$\widehat{\beta}_1$(se)	$\widehat{\beta}_2$(se)	$\widehat{\sigma}$(se)
Semipar:	$-5.29(.18)$	$2.12(.25)$	$0.32(.16)$	$.88(.19)$
Conditional:	–	$2.09(.36)$	$0.31(.21)$	–
Weighted:	$-5.61(.44)$	$-1.78(.55)$	$2.01(.68)$	$0(0)$

The semi-parametric approach . . .

- can be implemented simply with a reasonably general likelihood program;
- requires absolutely no modelling of the covariate distribution;
- more efficient (sometimes much more efficient) than weighted methods;
- more robust than (and almost as efficient as?) full maximum likelihood.