On the robustness of weighted methods for fitting models to case-control data

Alastair Scott

Chris Wild

University of Auckland

University of Auckland, Department of Statistics

Private Bag 92019, Auckland, New Zealand

scott@stat.auckland.ac.nz, wild@stat.auckland.ac.nz

The question of whether or not to use survey weights when analysing data from complex surveys is a contentious one that has generated considerable discussion. This paper is a contribution to that discussion in the context of a very special (albeit very important) type of survey, viz population-based case-control studies, where maximum likelihood methods are well developed and easy to implement.

If the logistic model is valid, then the maximum likelihood (ML) method has full semi-parametric efficiency and, in particular, is more efficient than the survey-weighted (SW) method. Moreover, the loss of efficiency from using the SW method can be quite large in some circumstances. The appeal of the SW method comes from its robustness to departures from the underlying model. When the model is not true, then the estimates produced by the SW method can still be interpreted as estimating the best fitting logistic model for the whole population. Since the form of the regression surface is never known exactly, we are always working with a model that is not quite right. In this paper we look at what happens when the model is 'not quite right'.

The estimating equations underlying both methods are special cases of the general class

$$(1-\lambda)\sum_{\text{cases}}\frac{\boldsymbol{x}_t p_0(\boldsymbol{x}_t;\boldsymbol{b})}{n_1}-\lambda\sum_{\text{controls}}\frac{\boldsymbol{x}_t p_1(\boldsymbol{x}_t;\boldsymbol{b})}{n_0}=\boldsymbol{0}, \tag{1}$$

for any $\lambda > 0$. Setting $\lambda = n_0/n$ gives the ML estimator, while setting $\lambda = W_0$ gives the SW estimator.

If the true model is logistic , then only the intercept is affected by the choice of weighting used in the estimating equations. It is the remaining regression coefficients that are of primary interest, however, since they determine the relative risk of becoming a case associated with a change in the values of explanatory variables. These coefficients are unaffected by the choice of weightings. By contrast, if the model is misspecified, every single element of $\widehat{\boldsymbol{\beta}}_\lambda$, the solution to (1), depends upon the weightings used. Choosing survey weights ($\lambda = W_0$) leads to a consistent estimator of $\boldsymbol{B}$ ($= \boldsymbol{B}_{W_0}$), the quantity we would be estimating if we fitted the model to the whole population. The interpretation for other values of λ is less clearcut. It is this interpretation that we want to explore.

For simplicity , we consider the case of a single explanatory variable. We assume that the primary focus of interest is on the slope of the logit curve, since this tells us about the change in the odds of becoming a case associated with a change in x. We show that any value of λ will lead to an estimate of the true slope at some point, x_λ say, along the curve. The 'best' choice will depend on where we want to estimate effects or make predictions. If x has very little effect on the response, then $\widehat{\beta}_{1\lambda}$ and x_λ will be essentially the same for any λ and, in particular, for the SW and ML methods. The differences get larger as the effect of x increases and the case and control populations move further apart. In most applications, cases will be rare and the SW method will be estimating the correct slope for values of x out in the upper tail of the X- distribution. This will be a good strategy if we are particularly interested to what happens to the people who are at high risk when we change the value of the risk factor. The ML method corresponds to a value of $\lambda = n_0/n$ which is typically close to $\frac{1}{2}$, and tends to give the correct slope for people at moderate risk. In our numerical study, for example, the ML method gave a better approximation to the true slope for up to 95% of the population while survey weighting was better just for the 5% or so of people in the highest risk category.

Our conclusion is that a prescriptive approach that says that we should always use one or other approach is wrong and that the method should be tailored to the particular application. If we are particularly interested in high risk individuals, then the survey-weighted method may be appropriate. If we are interested in more typical members of the population, then the likelihood-based method will have smaller bias as well as being more efficient. A compromise between the two methods might also be appropriate.