Worked Exercise: Case/Control Sampling

A case/control sample design draws a sample from a stratum corresponding to an outcome that is the target of study. A second sample is drawn from the stratum with other outcomes. For example, in biostatistics, one may draw a sample of individuals with a rare disease, the *cases*. One then draws a sample of individuals who do not have this disease, perhaps by some convenient protocol such as sampling from individuals hospitalized for reasons other than the rare disease. In drawing these *controls*, it is common to screen subjects to match the distribution of age, sex, and race found among the cases. The idea behind this is that once these demographic variables are controlled, differences in the distribution of remaining explanatory variables will reflect their influence on the probability of developing the rare disease. Then, this sampling protocol provides a relatively fast and inexpensive method for screening populations to identify potential risk factors for the rare disease.

To put case/control sampling in an economic context, suppose one wants to know what variables determine whether a person will receive an appointment as a Federal judge. The sample of *cases* is drawn from the population of Federal judges. The *controls* are drawn from the general population, excluding federal judges, with screening so that the sample of controls have the same distribution of age, sex, and race as the sample of cases.

Let (v,x,y) denote the variables collected on sampled individuals, where y = +1 for Federal judges and y = -1 otherwise, v is the vector of explanatory variables (e.g., age, sex, race) on which the distribution in the controls will be matched to the distribution in the cases, and x is a vector of additional explanatory variables, which may include parents' occupation and income, test scores for verbal ability and for reasoning, and personality scales for bellicosity and self-confidence. The stratum of cases is indexed s = 1 and the stratum of controls is indexed s = 2.

Let P(y|v,x) denote the *structural model,* or conditional probability for the target outcome, given the covariates. Let p(x|v)h(v) denote the density of the covariates in the general population. Let

$$(1) \qquad q(y) = \sum_v \sum_x P(y|v,x)p(x|v)h(v)$$

denote the shares of Federal judges and others in the general population. Let

$$(2) \qquad Q(v,x|y) = Q_2(x|v,y)Q_1(v|y) = P(y|v,x)p(x|v)h(v)/q(y)$$

denote the conditional density of (v,x) given y, obtained using Bayes law. This conditional density can be written as the product of the conditional density $Q_1(v|y)$ of v given y times the conditional density $Q_2(x|v,y)$ of x given v and y. Then,

(3) $$Q_1(v|y) = \sum_x P(y|v,x)p(x|v)h(v)/q(y),$$

(4) $$Q_2(x|v,y) = P(y|v,x)p(x|v)/ \sum_x P(y|v,x)p(x|v).$$

The cases, stratum s = 1, have the simple qualification rate R(v,x,y,1) = **1**(y==1), and the qualification factor q(1). The conditional distribution of v among the population of Federal judges is $Q_1(v|1)$, and among the remaining population is $Q_1(v|2)$.

The controls, stratum s = 2, are drawn with the qualification rate R(v,x,y,2) = **1**(y==-1)$Q_1(v|1)/Q_1(v|-1)$, so that the controls contain no Federal judges and their distribution of v will match that of the cases. Their qualification factor is

(5) $$r(2) = \sum_v \sum_x P(-1|v,x)p(x|v)h(v)Q_1(v|1)/Q_1(v|-1) = \sum_v q(-1)Q_1(v|1)h(v),$$

where the second form comes from substituting (3) into (5).

A typical structural model is the parametric form

(6) $$P(y|v,x) = F(y(\alpha + \gamma_v + x\beta)),$$

where F is a CDF, such as logit or probit, for a density that is symmetric about zero, and α, γ_v, β are parameters. The interpretation of the parameter γ_v is that it takes on some value for each possible configuration of v, including interactions. For example, if v contains a dummy for sex, a dummy for race, and a series of dummies for age by decade, then γ_v would take a separate value for each sex-race-decade configuration. Some normalizations are normally imposed for identification, such as the requirement that the γ_v's sum to zero over each component of v.

Suppose n_1 Federal judges and n_2 others are drawn from the respective strata, with $N = n_1 + n_2$. The CML probability of y given v,x in the pooled sample of cases and controls has the generic form

(6) $$\Pr(1|v,x) = \frac{\sum_s R(v,x,1,s)P(1|v,x)n_s/Nr(s)}{\sum_y \sum_s R(v,x,y,s)P(1|v,x)n_s/Nr(s)} .$$

In this application, the term in the denominator of (6) for y = 1 is

(7) $\sum_s$ R(v,x,1,s)P(1|v,x)n_s/Nr(s) = P(1|v,x)n_1/Nq(1),

and the term in the denominator of (6) for y = -1 is

(8) $\sum_s$ R(v,x,-1,s)P(-1|v,x)n_s/Nr(s) = [Q_1(v|1)/Q_1(v|-1)]P(-1|v,x)n_2/Nr(2).

Putting these together,

$$(9) \qquad \Pr(1|v,x) = \frac{P(1|v,x)n_1/Nq(1)}{P(1|v,x)n_1/Nq(1) + Q_1(v|1)P(-1|v,x)n_2/Nr(2)Q_1(v|-1)} \quad .$$

$$= \frac{P(1|v,x)}{P(1|v,x) + P(-1|v,x)n_2q(1)Q_1(v|1)/n_1r(2)Q_1(v|-1)} \quad .$$

This probability model for the sample weights the structural model, with the probability of being a Federal judge in the pooled sample, given v,x, lower for v configurations that are positively associated with judgeship. The CML criterion maximizes the sum, over the pooled sample of cases and controls, of the log of (9) with respect to unknown parameters in the structural model P(1|v,x). It requires that one know, or alternatively be able to parameterize and identify, the terms that appear in the weight in the denominator of (9). Often, it will be possible to estimate these terms using some combination of population statistics (e.g., for q(1)), sampling information (e.g., for n_1, n_2, r(2), and possibly Q_1(v|1) and Q_1(v|-1).), and parameterization (e.g., for Q_1(v|1)/Q_1(v|-1)).

Consider the special case that the structural model is logistic, P(1|v,x) = 1/(1+exp(-α-γ_v-xβ)). Substituting the logistic formula in (9) yields

$$(10) \qquad \Pr(1|v,x) = \frac{1}{1 + \exp(-\alpha-\gamma_v-x\beta)n_2q(1)Q_1(v|1)/n_1r(2)Q_1(v|-1)}$$

$$= \frac{1}{1 + \exp(-\alpha*-\gamma_v*-x\beta)} \quad .$$

with

(11) $\alpha^* = \alpha - \log(n_2q(1)/n_1r(2))$ and $\gamma_v^* = \gamma_v - \log(Q_1(v|1)/Q_1(v|-1))$.

Thus, the effects of the sample design are confined to the intercept and to the v-interactions, and the coefficients β on x remain unchanged. This implies in particular that fitting a random sample logistic to the data generated by the case/control sampling protocol will yield consistent estimates of β, provided that the model contains an intercept term and a full set of v interactions which can absorb the sampling effects as well as any corresponding population effects. If estimates of population α and γ_v are needed, they can be obtained from (11) if the remaining terms in these equations can be estimated.

Thought exercises on case/control sampling:

1. Show in the case of a logistic structural model that if the model excludes some v-interactions that enter the protocol for matching controls, then the resulting estimator of β is in general inconsistent.

2. Show in the case of a logistic structural model that if the structural model includes all v-interactions, and in addition includes some v-x interactions, then the coefficients of v-x interactions are affected by the sample design.

3. Show that it is possible to estimate structural model parameters using the WESML method, with WESML weights $w(y,v,x)$ satisfying $w(1,v,x) = Nq(1)/n_1$ and $w(-1,v,x) = Nr(2)Q_1(v|-1)/n_1Q_1(v|1)$.

4. Carry out a Monte Carlo study of case/control designs when v and x are both univariate indicators. Is there a gain in precision in estimating the coefficient on x when matching is done on v?

5. Suppose x is a single explanatory variable that is itself discrete (e.g., x is an indicator for whether an individual's father was a lawyer), and the objective is to test whether x is a risk factor for the outcome y. A statistician proposes putting x and y in a 2×2 contingency table and testing whether they are independent, arguing that the matching of cases and controls on v eliminates the possibility that they could confound the test. What are the advantages and disadvantages of this method compared with a LR test for $\beta = 0$ in the random sample logit model applied to the data? Are the methods asymptotically equivalent?

6. Suppose that instead of drawing controls with probabilities that will give the same population distribution of v, one first gets the empirical distribution of v among the cases, and then accepts controls from a target population until one has exactly matched this empirical distribution. How would the analysis above change? Would the asymptotic analysis change?