

Online Appendix

A Note on Choice-Based Sampling and WESML

In samples where the fraction of $y=1$ observations (the “ones”) is very small, the information content is much greater in the ones rather than the zeroes. To see this, recall that the asymptotic covariance matrix for the MLE for logit is given by (see Greene, 2003, p. 672)

$$\left[\sum_{i=1}^n \Lambda_i (1 - \Lambda_i) x_i x_i' \right]^{-1}$$

If the logit model has some explanatory power, Λ_i is larger (i.e. closer to 0.5 for rare events) when $y_i=1$. Thus $\Lambda_i(1-\Lambda_i)$ is larger, implying that having a higher fraction of 1’s in the sample would reduce variance. Choice-based sampling tries to achieve this by over-sampling on the “ones” from the population. The sample is formed by taking a fraction α of the population’s dyads with $y = 0$, and a fraction γ of the dyads with $y = 1$, where α is much smaller than γ . The probability of a citation *conditional on the dyad being in the sample* flows from Bayes’ rule:

$$\Lambda_i' = \frac{\gamma \Lambda_i}{\gamma \Lambda_i + \alpha (1 - \Lambda_i)} = \frac{\gamma}{\gamma + \alpha e^{-\beta X_i}} = \frac{1}{1 + e^{-\left(\ln\left(\frac{\gamma}{\alpha}\right) + \beta X_i\right)}}$$

The extra term $\ln(\gamma/\alpha)$ in the exponent leads to a bias. However, since the functional form is still logistic, a simple estimation strategy is to simply subtract $\ln(\gamma/\alpha)$ from the estimate for the constant term of the usual logit. The efficiency of the correction, however, depends crucially on the logit functional form not being misspecified (Manski and Lerman, 1977; Cosslet, 1981). An alternate method, which is not as sensitive to model misspecification, is the *weighted exogenous sampling maximum likelihood* (WESML) estimator suggested by Manski and Lerman (1977). The WESML estimator is obtained by maximizing the following weighted “pseudo-likelihood” function:

$$\ln L_w = \frac{1}{\gamma} \sum_{\{y_i=1\}} \ln(\Lambda_i) + \frac{1}{\alpha} \sum_{\{y_i=0\}} \ln(1 - \Lambda_i) = - \sum_{i=1}^n w_i \ln(1 + e^{(1-2y_i)x_i\beta})$$

where $w_i = (1/\gamma) y_i + (1/\alpha)(1 - y_i)$.

In other words, each sample observation is weighted by the number of elements it represents from the overall population in order to make the choice-based sample “simulate” a random exogenous sample. Here is some intuition on why WESML works: Let the joint probability density be $g(x,y)$ for the sample, and $g^*(x,y)$ for the population. Let the fraction of elements with $y = j$ be $f(j)$ in the sample, and $f^*(j)$ in the population ($j = 0,1$). Let n and N be sample size and population size respectively, and n_j and N_j be the number with $y = j$. Using conditional probability rules,

$$g(x, j) = \Pr(x | y = j) f(j) = \frac{g^*(x, j) f(j)}{f^*(j)} = \frac{g^*(x, j) (n_j / n)}{N_j / N} = \frac{N / n}{w(j)} g^*(x, j)$$

where $w(j) = N/n_j$ is the reciprocal of the sampling rate for observations with $y = j$. Let $P(y_i)$ be the probability of $y = y_i$ conditional on $x = x_i$ in the population. Then, the expected value of the weighted likelihood function is

$$E \ln L_w = \int \left(\sum_{i=1}^n w(y_i) [\ln P(y_i)] \right) g(x, y_i) dx = \sum_{i=1}^n \left(\int w(y_i) [\ln P(y_i)] \frac{N/n}{w(y_i)} g^*(x, y_i) dx \right) = \frac{N}{n} \int \left(\sum_{i=1}^n [\ln P(y_i)] \right) g^*(x, y_i) dx$$

Thus, ignoring the constant scaling factor N/n , the expected value of the weighted log likelihood equals the expected log likelihood for the same sample resulting through random exogenous sampling from the population. As shown formally in Amemiya (1985, section 9.5.2), this ensures consistency of WESML estimation.

The choice-based WESML procedure described above can be extended to allow “matched samples”. This involves taking all actual citations ($y=1$) and matching each of these with k “control citations” ($y=0$) along a dimension z (e.g., the “cells” indexed by the vector combination of the citing technological class and cited technological class). Without loss generality, denote the values z can take as $1, 2, \dots, T$. For a matching-based sampling design, it is easier to think of not just y but (z, y) as the dependent variable. In forming the likelihood function, I will use the result that

$$\begin{aligned} \Pr(z = z_i \text{ and } y = j | x = x_i) &= \Pr(z = z_i | x_i) \Pr(y = j | z = z_i \text{ and } x = x_i) \\ &= \Pr(z = z_i | x_i) \Pr(y = j | x = x_i) \end{aligned}$$

The second equality assumes that the vector x includes all information about z that affects citation outcome y , i.e., x is a sufficient statistic for z . The log-likelihood function for estimation using an exogenous random sample of size n would therefore be

$$\begin{aligned}\ln L &= \sum_{i=1}^n \ln[\Pr(z = z_i \text{ and } y = y_i | x_i)] \\ &= \sum_{i=1}^n \{y_i \ln[\Pr(z = z_i | x_i)\Lambda(x_i\beta)] + (1 - y_i) \ln[\Pr(z = z_i | x_i)(1 - \Lambda(x_i\beta))]\}\end{aligned}$$

This forms the basis for deriving the pseudo-likelihood function for choice-based sampling. Each log likelihood function term has to be weighted by the inverse of the probability that the corresponding population element will be included in the sample. To derive these weights, denote the number of elements with $z = t$ and $y=j$ as n_{ij} for the sample and N_{ij} for the population. Matching ensures that, from each cell, I pick all elements with $y=1$ and k times as many elements with $y=0$. In other words, $n_{t1} = N_{t1}$ and $n_{t0} = kN_{t1}$. Also, since N_{ij} is known, the probability p_{ij} of a population element with $z = t$ and $y = j$ getting selected in our sample is easily calculated as $p_{t1} = n_{t1}/N_{t1} = 1$ and $p_{t0} = n_{t0}/N_{t0} = kN_{t1}/N_{t0}$ for all values of t . Denoting $w_{ij} = 1/p_{ij}$, the weighted likelihood function for choice-based sampling is the given by

$$\begin{aligned}\ln L_w &= \sum_{i=1}^n \{y_i w_{z_i 1} \ln[\Pr(z = z_i | x_i)\Lambda(x_i\beta)] + (1 - y_i) w_{z_i 0} \ln[\Pr(z = z_i | x_i)(1 - \Lambda(x_i\beta))]\} \\ &= C - \sum_{i=1}^n w_i \ln(1 + e^{(1-2y_i)X_i\beta})\end{aligned}$$

$$\text{where } w_i = y_i w_{z_i 1} + (1 - y_i) w_{z_i 0} \quad \text{and} \quad C = \sum_{i=1}^n w_i \ln[\Pr(z = z_i | x_i)]$$

Since C is independent of β , it can be ignored in the maximum likelihood procedure. Thus, a weighted logit estimation can be used, where the weights of the observations are now given by w_i . Unlike the simple WESML with random sampling from the $y=0$ observations, the weights now depend not just on the value of y but also on the cell that the observations falls into.