8 October 2007

Eric Rasmusen, visitor, Nuffield College, erasmuse@indiana.edu

Notes on Standard Errors in the Case Control Method and the Meaning of Repeated Sampling

A standard counterintuitive result in statistics is that if the true model is logit, then it is okay to use a sample selected on the Y's, which is what the "case-control method" amounts to. You may select 1000 observations with Y=1 and 1000 observations with Y=0 and do estimation of the effects of every variable but the constant in the usual way, without any sort of weighting. This was shown in Prentice & Pyke (1979). They also purport to show that the standard errors may be computed in the usual way— that is, using the curvature (2nd derivative) of the likelihood function. This, I was skeptical of. If the constant is misestimated, how can you deduce the variance of the disturbance term, and if you can't deduce that, how can you deduce the standard error of any of the coefficients? Nowhere have I seen a clear demonstration or an intuition for the result, so I thought there might be a crucial unnoticed mistake in the math somewhere, as is not unknown in famous papers (e. g. Hotelling on location, Tullock on overdissipation, Viner on average cost curves, and Rothschild-Stiglitz on risk).

Since I did not follow all the steps of the Prentice-Pyke proof and so did not know of any error in what they did, I tried doing a Monte Carlo study which seemed to confirm my intuition.

Since then, however, I have seen where my Monte Carlo study went wrong, and now I believe Prentice and Pyke. I am writing these notes up because they will help me remember where I went wrong and perhaps will be helpful to other people too. I'll distribute these to people with whom I've discussed the idea too, so they can see why I've abandoned it.

1.Here is some intuition– a bit shaky, I fear. Suppose that a coefficient is estimated correctly by some estimator. We want to estimate the estimator's standard error, to know how variable the estimate would be if we repeated the estimation with different disturbances. For this, we need to know how noisy the data is. We do not need to know how noisy the data in the whole population is, however, just how noisy in the kind of sample we draw. If our procedure is to draw a biased sample, then we need to know what will happen in other biased samples, not in the population. It is okay to use the sample for this purpose. In using a standard error, we are not generalizing anything to the population (not estimating goodness of fit, for example), we are just generalizing to repeated samples.

2. Here is how to think about repeated sampling and how to do a Monte Carlow study. What I did was to construct a population of 60,000 data points, drawing X from a uniform distribution on [0,1] and a disturbance epsilon from a logit density with an alpha "constant" coefficient of -4 and a beta X coefficient of 0. If alpha; + epsilon < 0 then Y=0; if alpha + epsilon ≥ 0 then Y = 1. That yields 1,039 points with Y= 1, about 1.7% of them.

Our estimation procedure is to combine two random samples of 1,000 observations with Y=0 and 1,000 observations with Y=1 and do a logit estimate of alpha and beta. We would expect the estimate of alpha to be wrong– not close to 0.017– and the estimate of beta to be right– close to 0.000– since we have a large enough sample that consistent estimates ought to be close to the true parameters.

The maximum likelihood estimate would give us standard errors based on the second derivative of the likelihood function or on bootstrapping. In repeated sampling, we would expect the standard deviation of the alpha estimates not to be close to the average of the estimates of its standard error. The question to be investigated is whether the the standard deviation of the beta estimates is close to the average of the estimates of its standard error.

So far, so good. Where I made my mistake, I think, is in the definition of "repeated sampling". Ordinarily in frequentist thinking, in repeated sampling we keep the X values the same in each sample, and we draw new disturbances, which combine with the fixed X's to give new Y's. That also amounts to conditioning on the X's, though we wouldn't have had to condition the X's, since our estimator should work fine even if we changed the X's in each sample too. (If we did change the X's, though, that change the information content in each sample— a sample in which X only varied between .3 and .4 would have less information and yield worse estimates than one with X varying widely between .02 and .94. So in small samples, especially, we'd have to make some allowance for that.)

Here, though, we can't keep the X's fixed. If we did, then although our first sample would have 1,000 observations with Y=1, our succeeding samples would have about 34. We wouldn't be using the case-control method.

So what we have to do is to think about repeated samples with 1,000 Y=0 observations and 1,000 Y=1's. Turning our usual thinking upside down, we need to keep the Y's fixed, draw new disturbances, and let the X's vary. This is especially hard to think about here, because knowing Y and epsilon does not tell us X– remember, Y is coarse and contains less information than alpha + beta*X + epsilon,and beta is zero here too, making things even worse.

The best way to proceed is to think about repeating the entire scientific procedure, including the sampling as well as the estimation. The way I did this was to take 100 n=2000 samples from the 60,000-point population, each time combining equal-sized subsamples with Y=0 and with Y=1.

Recall, however, that there are only 1,037 Y=1 values in the entire population. Thus, my repeated sampling had to be with replacement, and was using the same Y=1 observations over and over. It is OK to use the same X values repeatedly, but these observations also had the same epsilon values each time, so the samples are not independent in the way needed for the law of large numbers to work. The standard errors computed by maximum likelihood came out wrong— not equal to the standard deviation of the estimates, but that is to be expected when the draws are not independent.

Realizing this, I also tried doing the procedure with 100 n=200 samples instead of 100 n=2000 samples. I still used sampling with replacement, but now there was less overlap between replacements, less dependence between samples. And now the estimated standard errors were close to the standard deviations.

This, I expect is what would happen if I did the kind of repeated sampling that is our thought experiment for the kind of real studies that use the case-control method. That thought experiment is to take repeated draws of 60,000-point populations, with the same X's each time but with different epsilons and hence Y's. Each of the 100 Monte Carlo samples would be from a different population draw.

Prentice, R. L. & R. Pyke (1979) "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66(3): 403-411 (December 1979).