Tom Hertz
Monday, February 12, 2007

Thoughts on Restricted Samples, with reference to Gregory Clark's Data Memo

The problems raised by non-representative, restricted samples are tricky. Gregory Clark may well be right in his conclusion about the sign of the bias that is created by not observing many sons of low-wealth parents. But I think the argument he makes rests not just on restriction of X, but on an implicit restriction of Y as well. I begin with a discussion of the problem of X-based selection only, as this seems a general concern in Sam's memo about the memos. I then talk about Clark's argument, and end with some thoughts on the primogeniture problem.

1) In the simplest linear model, in which both slopes and intercepts are the same for all people, even fairly radical sample selection (or non-random restriction, or flat out truncation) that is based on an X variable should not create bias in the estimation of a *regression coefficient*. For example, if the elasticity of father-son wealth is the same for rich and poor fathers, it does not matter what mix of the two you happen to observe, from a bias point of view. However, you will, as always, get more *precise* estimates with a wider range of X's in the sample, because the sample variance of X appears in the denominator of the expression for the standard error of betahat.

Here is an example I simulated. Both X and the error term (e) are standard normals, and Y=X+e. The data are:

```
. sum y x e

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |      5000   -.0011983    1.418157  -5.275992   5.459172
           x |      5000    .0071502    .9925645   -3.81654   3.641588
           e |      5000   -.0083485    1.008727  -3.918931   3.188653
```

And the regression in the full sample is:

```
. regress y x

      Source |       SS       df       MS              Number of obs =    5000
-------------+------------------------------           F(  1,  4998) = 4880.82
       Model |  4967.28662     1  4967.28662           Prob > F      =  0.0000
    Residual |   5086.5465  4998  1.01771639           R-squared     =  0.4941
-------------+------------------------------           Adj R-squared =  0.4940
       Total |  10053.8331  4999  2.01116886           Root MSE      =  1.0088
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.00429   .0143752    69.86   0.000     .9761088    1.032472
       _cons |  -.0083792   .0142672    -0.59   0.557    -.0363492    .0195908
------------------------------------------------------------------------------
```

To prove the point about regression coefficients not being affected by selection on X, try running it on only the X>0's. (Here the loss in precision is due more to cutting N in half than restricting X's range, but whatever.)

```
. regress y x if x>0

      Source |       SS       df       MS              Number of obs =    2542
-------------+------------------------------           F(  1,  2540) =  903.98
       Model |   943.09137     1   943.09137           Prob > F      =  0.0000
    Residual |  2649.90462  2540  1.04326954           R-squared     =  0.2625
-------------+------------------------------           Adj R-squared =  0.2622
       Total |  3592.99599  2541  1.41400866           Root MSE      =  1.0214

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.023626   .0340457    30.07   0.000     .9568654    1.090386
       _cons |  -.0200287   .0336039    -0.60   0.551    -.0859226    .0458652
------------------------------------------------------------------------------
```

2) For the correlation coefficient, the story is trickier. What counts is whether your X-based restrictions, or under-representations, have a proportionately larger effect on sd(X) than on sd(Y). In the case of actual truncation at a certain value of X, Sam is right that the correlation between Y and X will fall. This occurs because you restrict sd(yhat) in proportion to the restriction of sd(X), but the sd of the error terms is not affected, so observed sd(Y) falls less than proportionately. Thus for a given beta, corr(Y,X)=beta*sd(X)/sd(Y) falls, and so does R^2. Here is the simulation, again restricting X>0: sd(X) fell from 1 to .6, but sd(Y) only fell from 1.4 to 1.2, as sd(e) is unchanged.

```
. corr y x
(obs=5000)

             |        y        x
-------------+------------------
           y |   1.0000
           x |   0.7029   1.0000


. corr y x if x>0
(obs=2542)

             |        y        x
-------------+------------------
           y |   1.0000
           x |   0.5123   1.0000


. sum y x e if x>0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |      2542    .7860662    1.189121  -3.252486   5.459172
           x |      2542    .7874899    .5951595    .000986   3.641588
           e |      2542   -.0014237    1.021301  -3.918931   3.188653
```

In less dramatic cases, however, where the low values are present but under-represented, the effects on sd(X) and sd(Y) might not be so different. Remember that what affects sd most are extreme values – so if the *range* of X is still intact, its sd may not be all that different. As an example, I tried randomly dropping half of the bottom quartile of X, so 1/8th the sample, which I thought would have been a pretty dramatic move, but it made very little difference to the correlation (reduced from 0.70 to 0.68):

```
. corr y x if insamp2
(obs=4323)
             |        y        x
-------------+------------------
           y |   1.0000
           x |   0.6784   1.0000

. sum y x e if insamp2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |      4323    .1855935    1.360765  -5.275992   5.459172
           x |      4323    .1998524    .9100829   -3.81654   3.641588
           e |      4323   -.0142589    .9998747  -3.918931   3.188653
```

3) Now, there are other cases in which the *regression estimates* themselves are affected by selection on X. One is the emphasized by Solon (AER 1992) in his critique of the white twins from Minnesota type of studies. There the problem with restricting X relates to measurement error: in restricting the variance of true X, but not restricting the variance of the measurement error term (or annual volatility component), the reliability ratio Var(Xtrue)/Var(Xobs) falls, leading to low intergenerational elasticities.

4) Either effect 1) or 2), above, may be analogous to the *Psychol. Bull.* (unfortunate journal name) article that Sam linked to in an email, but I'd have to dig deeper to understand the nature of the analogy.

5) Next there is the issue of group-based heterogeneity in slopes or intercepts, or of non-linearity in the slope, across members of the sample. I discuss this at some length in my Rags, Riches, and Race chapter. If the parent-child elasticity is *different* for rich and poor parents, or if the intercept is different for blacks versus whites, then non-representative samples will indeed yield different results than representative ones. In such cases, weights that bring the sample means of X and Y back into alignment with the population, if such can be had, are indeed useful: they should allow us to recreate the descriptive statistics we would obtain from the population.

(Aside: You will hear much debate about the value of weights in regressions, but Deaton (1997 book) makes it clear that they have value for descriptive exercises such as these. Their shortcoming, which many cite but do not understand, is the following. Suppose there are two betas, say betablack and betawhite, and what you want to estimate is a population-group-size weighted average of the two. But your sample has too few blacks in it. You might think that upweighting the blacks would allow you to obtain the right population-group-size weighted average of the two betas, just as it allows us to get group mean incomes right...but it doesn't. It does, however, allow you to retrieve the descriptive stat you would have gotten if your sample were representative to begin with, which is what we are after.)

6) Similar problems arise if the true (in a descriptive sense) relationship is non-linear in X. Suppose it is flat for X's below the mean, then rises with slope one thereafter. With all the data, the regression line would split the difference – it is "wrong" for both branches, but right as a linear summary of the population. Then if you leave out a lot of cases from the flat part, the regression line will get steeper, and no longer be correct as a linear summary of the population. Weights would again correct the problem.

7) Assuming Clark's calculations about the joint distributions of Y and X are correct, I think his argument rests on the assumption that the true descriptive pattern is linear, and implies that sample selection has occurred not just on the basis of X, but also of Y. It is not just that there are too few low-wealth dads in the sample, but also that the sons-of-low-wealth-dads that *do* appear are too wealthy (or the sons that don't appear are the poorest) which pulls the regression line up.

But can we be sure of this? The fact that the box in the lower left of his Figure 6 would have two cases if the correlation were zero and 14 cases if it were unity, would seem to suggest that it *should* have between two and 14, and thus that observing zero is sign of sample selection on Y. But it could also be that the relationship is non-linear, and that box *should* be empty (within statistical limits of certainty), in which case the regression line is not biased per se, it is merely a linear approximation of a non-linear phenomenon. Put still otherwise, his statement "...since the missing observations are concentrated below the regression line on the left hand side" can only be made with confidence if we are sure that the regression line is indeed a line! If it is a curve, all bets are off.

The bottom line is that weights that restore the means of X and Y in the father-son sample to numbers closer to their population averages should help, if such can be constructed. The problem then is: what are the variables that we should use to reweight? The answer is: those that led to cases being omitted, which we don't usually know. In any event, until such weights are in hand, I would hedge my bets about what the outcome will look like.

8) A final thought about the primogeniture problem: that is clearly selection on Y, if you observe a disproportionate numnber of first-born, and hence wealthier, sons. However, this is one case in which a Heckman model could really make sense: we KNOW the variable that determines selection: it is being the eldest son. And we are also justified in excluding that variable from the intergenerational equation, not because it has no other effect on wealth (it might, via education, etc) but because we don't care about its other effects, and just want to correct our descriptive statistics. It might work, but you have to have data on SOME non-first-born-sons in order to get the selection probit off the ground. Also, I seem to recall that you cannot base a Heckman selection story on a single dummy variable (first born son) so birth-order (continuous) would be better.

The result would be an estimate of the intergen elasticity that we would observe if we could observe *all* kids, not a sample that is biased towards first-born sons. This again would represent a kind of average of two potentially very different elasticities: the connection between paternal and filial wealth is presumably very strong for first born sons and much weaker for the unfortunate seconds, and daughters.