WLS MEMO 176

Estimating the cognitive test score distribution of high school drop-outs in the WLS cohort using the freshman centile ranks of WLS graduates

Jeremy Freese

February 22, 2006

All WLS respondents are high school graduates. WLS does not have direct measures of cohort members who dropped out of high school, nor does it even have direct information on how many such dropouts there were.

Of course there is ample reason to expect those who dropped out of college to be systematically different from those who did not. Generally, information on the consequences of truncated samples relies on the use of side information about the population distribution of either the excluded population or the overall population. For example, if one assumed the percentage of males in the WLS cohort was no less than 50% (or , we might conclude that men are underrepresented in the WLS sample because 51.6% of WLS sample members are females ($p < .001$ for a one-sided exact test of proportions). This implies that men are slightly overrepresented among unobserved members of the cohorts, possibly due to being disproportionately likely to drop out.

We might expect that those with lower cognitive ability (as measured from test scores) would be more likely to drop out of high school. (For one thing, we expect ability *per se* to be related to retention, but anticipations about dropping out would seem like it might have obvious effects on test performance.) Since dropouts are not observed, it might seem like there is no way to use the WLS to gain information on the cognitive test score distribution of drop-outs from the WLS cohort.

However, the WLS has archival information on freshman test scores, and these freshman scores record information by Wisconsin centile rank. If one assumes that the centile ranks were uniformly distributed over the population that took the test, one can use the ranks of observed cases to infer what the distribution would need to be of unobserved cases for the resulting distribution to be uniform.

Extremely helpful for this purpose is that, although freshman scores were not observed for the entire WLS sample, both nonparametric (rank-sum) and parametric (t) tests suggest no differences in junior-year scores depending on whether or not freshman year scores are observed. For this reason, the sample with freshman year scores can be treated as representative of the WLS sample (that is, of Wisconsin 1957 high school graduates) as a whole.

The first step is to transform the existing centile ranks to a continuous distribution from 0 to 100. Because the H-N is less than 100 questions and so certainly had less than 100

unique scores, not all centile ranks are used. Imagine if centile ranks 19 22 and 24 were observed, and there were 215 cases with the rank 22. We would want the reconstructed variable to have 215 evenly spaced values between 20.5 (the mean of 19 and 22) and 23 (the mean of 22 and 24). Which observations are assigned which particular values in this interval is of no consequence for this analysis.

From the resulting data, we can construct a simple histogram dividing the WLS graduates into ten groups based on their centile ranks (i.e, groups from 0-10, 10-20, … 90-100). If cognition was not related to dropping out or leaving Wisconsin between one's freshman and junior years for other reasons, then each centile should include about 10% of the distribution. (In a simpler version of this exercise using five groups and the junior scores, the resulting distribution is indeed basically uniform.) [1]

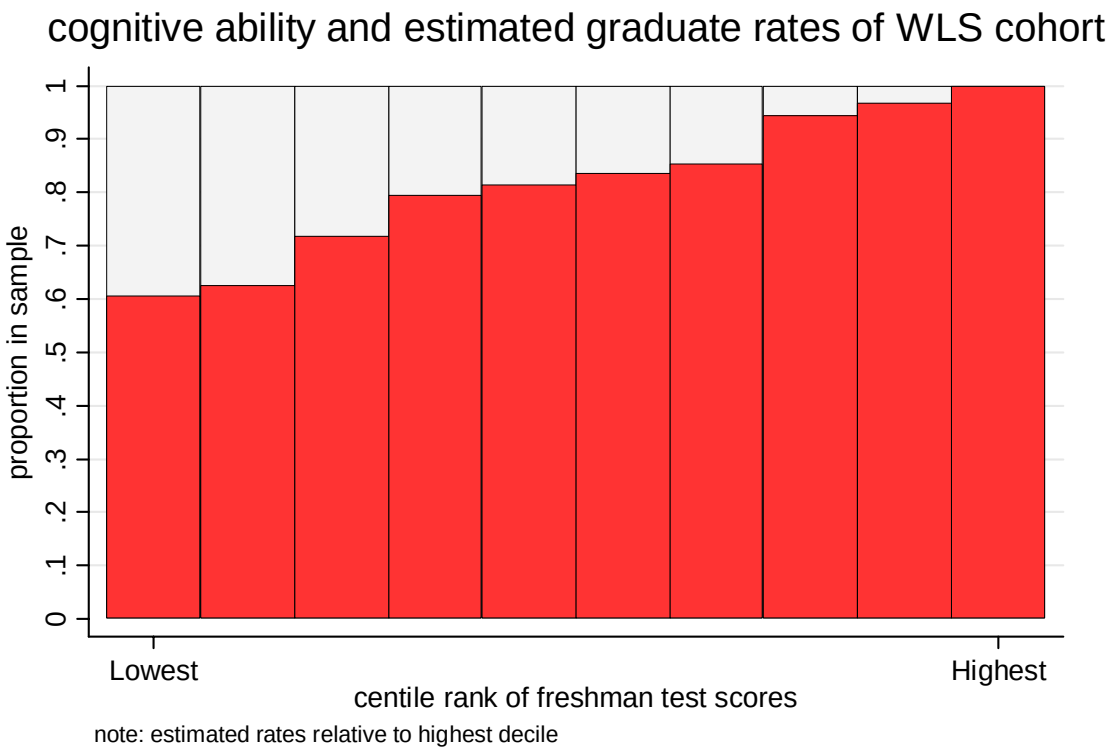
The resulting distribution is not uniform. One can use this information to figure out the rate of exclusion relative to those in the highest cognitive decile. Further, if one assumes that the distribution of those leaving the sample for other reasons (i.e., moving) between the freshman and junior years are uniformly distributed in terms of cognition, one can consider the rate of exclusion to be an estimate of relative differences in dropout rates.

The accompanying do file generates this graph and table reprising the results. These are also appended to the end of this memo.

The mean of the deciles provides a lower-bound estimate of the dropout rate itself. In this case, that estimate if 18.5%. That estimate assumes both that the dropout rate in the highest decile was 0% and that the dropout rate prior to the freshman test administration was 0%. Both are obviously implausible, and so it is not surprising that this rate is lower than estimates of the WLS dropout rate derived from side information (about 30%).

If one was willing to make assumptions about (1) the overall dropout rate, (2) the dropout rate for the highest decile, (3) and that the relative cognitive distribution of those who dropped out prior to the freshman test was like those who dropped out after the freshman test, then one could generate estimates of the cognitive distribution of dropouts that were absolute rather than relative to the upper decile. That exercise is easily conducted but outside the scope of this memo.

[1] An exactly parallel form of this analysis to the freshman score analysis should be conducted before drawing the conclusion that there was not a cognitive gradient to any attrition that occurred between the junior test administration and the senior year.



decile	graduate	dropout
lowest	.6048951	.3951049
2	.6258741	.3741259
3	.7179487	.2820513
4	.7948718	.2051282
5	.8135198	.1864802
6	.8368298	.1631702
7	.8531469	.1468531
8	.9440559	.0559441
9	.9685315	.0314685
highest	1	0