5. LONGITUDINAL SCREENING

Introduction

The final logistic regression model for heavy drug use above can be used as a screening tool to predict future heavy cocaine use. The variables in this final logistic regression model point out questions that could be used in future surveys or general population applications. The answers can then be input into this model to predict the likelihood of each respondent becoming a heavy drug user. It is also possible to use this likelihood to oversample the "most" likely future heavy drug users. As long as this is done on a probability basis, we can use weights to get a nationally representative and unbiased sample no matter how strong the oversampling is. Of course, the more oversampling there is, the more variability we would have in our nationally representative estimates using the weights (precision of the estimates is mathematically related to the variation in the weights). More oversampling, of course, also should lead to a larger sample of heavy cocaine users.

Screening the Sample

One goal was to develop a logistic regression model to help screen for a longitudinal survey that includes appropriate numbers of likely heavy drug users. As above, the screener questions would be designed with the model in mind, with the responses input to predict the likelihood of becoming a heavy drug user. The sample could be selected with a probability proportional to this estimated likelihood of becoming a heavy drug user. For example, since the odds ratio for males indicates a (roughly) doubled risk of becoming heavy drug users versus females, then males will be selected at twice the rate of females.

It is important to realize that age is a critical factor in any longitudinal survey that would study heavy drug use. The age of the youth need to be determined to balance the benefits and costs of starting with younger children who could develop into heavy drug users versus older children (e.g., teenagers and college-age) who could already be using heavy drugs. According to our analyses of age of first use, 14.2 percent of the NLSY sample had smoked a cigarette by age 10, and 10.7 percent had used cocaine by age 18.

While it will be difficult to ask about marijuana and other sensitive topics during a screening interview, there are several possible approaches. One straightforward approach is to have a screening interview only to determine age-eligible children and performing a short interview to ask the more sensitive questions as part of a further screener.

This would involve a lot of screening, so another suggestion is to try to "piggy-back" on another large screening sample. In this way, screening money could be saved if permission could be granted to simply target age-eligible children found by another project's large screening operation. The 1997 Profiles of American Youth (PAY97) sponsored by the Department of Defense did just this in partnership with the 1997 National Longitudinal Survey of Youth (NLSY97) sponsored by the Bureau of Labor Statistics. NLSY97 (the newer cohort to NLSY79) screened a national sample of 90,000 housing units for 12- to 16-year-old youth. At the same time, PAY97 screened for 18- to 23-year-old youth and 10^{th}- to 12^{th}-grade students.

Selecting the Sample

Let us assume that screener data are available for all the variables in Exhibit 4.7: Number of times smoked marijuana/hashish in the past year, gender, race/ethnicity, smoking status (most recent cigarette), any school suspension history, amount of illegal income in last year, selling of hard drugs, and religious attendance. Taking the idea expressed above, a sample proportional to their predicted probability of becoming a heavy cocaine user can be selected.

In order to simplify this discussion, we use unweighted NLSY79 data. The rate of future heavy drug users among the NLSY79 respondents was 551/8,033 = 6.86 percent. If we used an equal probability sample of 8,033 screener respondents (including the variables in Exhibit 4.7), we would then expect to have 551 future heavy cocaine users in our sample. However, by selecting our sample based on the screener variables shown to be related to future heavy cocaine use, we can increase our yield in a sample of 8,033. For example, the odds ratio for gender is 1.87, which suggests that the rate of future heavy cocaine users is 4.78 percent for females and 1.87 * 4.78 = 8.94 percent for males (please note that the average of 4.78 percent and 8.94 percent is 6.86 percent). If instead of a 50/50 sample of males and females, we selected 87 percent more males (34.84 percent female), we would obtain 2,799 females (134 future heavy cocaine users) and 5,234 males (468 future heavy cocaine users). This would result in 468 + 134 = 602 future heavy cocaine users (7.49 percent), an increase of over 9 percent over a gender-balanced sample. This assumes, of course, that the screening data contains at least 5,234 males.

A key idea is that the number of future heavy cocaine users can be increased further by increasing the oversampling of males. However, using only gender, we could not expect to exceed a percentage of future heavy cocaine users of 8.94 percent (which would consist of a 100 percent male sample).

Similar to gender, we could repeat the above analysis for all other screener variables. The most productive of these would surely be the number of times a respondent smoked marijuana or hashish in the last year, as shown by the odds ratio of 6.67 in Exhibit 4.7 for those who smoked marijuana or hashish at least 50 times in the last year. Examining this variable shows that 1,815 reported more than 50 uses in the last year, 817 reported 11–50 uses, 1,359 reported 3–10 uses, and 6,229 reported no use. The parameters in Exhibit 4.7 suggest that the rates of future heavy cocaine use are 20.22 percent (more than 50 uses), 9.79 percent (11–50 uses), 7.21 percent (3–10 uses), 4.67 percent (1–2 uses), and 3.03 percent (no use). Selecting the sample proportional to these predicted probabilities would result in a sample of (total sample of 8,033) 3,667 sample members with more than 50 uses (742 future heavy cocaine users), 799 with 11–50 uses (78 future heavy cocaine users), 980 with 3–10 uses (71 future heavy cocaine users), 701 with 1–2 uses (33 future heavy cocaine users), and 1,887 with 0 uses (57 future heavy cocaine users). This sample results in 980 future heavy cocaine users (an increase of 78 percent)—using only the one variable.

Using selection probabilities based on using all of the variables in Exhibit 4.7 will obviously lead to the possibility of even more future heavy cocaine users in the sample. Calculations, however, are highly dependent on the distribution of the screening sample across all cells for these variables, and are therefore not shown here. However, the above example for marijuana use shows that the rate of future heavy cocaine users can easily be increased from the NLSY97 rate of 6.86 percent to more than 10 percent using only this one variable in the selection of the sample.

Weighting the Sample

It should be noted that the survey need not be restricted to "likely" heavy drug users; "likely" heavy drug users would simply be overrepresented. For example, the above examples still include females and non-marijuana smokers in the sample. Therefore, this sample would still be nationally representative with the proper weights. Weights are commonly used to adjust for differential probabilities of selection. In fact, a Horvitz-Thompson[15] estimator is simply a weighted mean where the weight is the reciprocal of the selection probability:

$$\hat{y} = \frac{1}{N}\sum \frac{y_i}{p_i},$$

where N is the population size, y_i is the observation, and p_i is the selection probability. Horvitz-Thompson estimators have been studied extensively, but their main benefit is that they are unbiased. In this case, this unbiased property results in nationally representative estimates.

Taking our example samples selected above, if males were selected with a probability (p_i) 87 percent greater than females, the base weight ($1/p_i$) for females would be 87 percent greater. Taking a very simple example, let's assume that the sample of 2,799 females and 5,234 males was taken from a population of 25,000 females and 25,000 males. The selection probabilities in this case would be 2,799/25,000 = 11.20 percent for the females and 5,234/25,000 = 20.94 percent for the males. The base weight would be 1/0.1120 = 8.93 for the females and 1/0.2094 = 4.78 for the males. Since we would then expect 134 female and 468 male future heavy drug users, the above formula would imply our estimate of the future heavy drug user percentage (where 1= future heavy drug user, 0 = not) in the entire population of 50,000 males and females is:

$$\hat{y} = \frac{1}{50{,}000}\left(\frac{134*1}{.1120} + \frac{468*1}{.2094}\right) = .0686$$

This estimate of 6.86 percent matches our NLSY97 (unweighted) data, and implies 3,431 future heavy drug users in the theoretical population of 50,000 youth.

Conclusion

A longitudinal study would fill a currently large gap in our knowledge base of drug use. Currently, most data on drug use comes from cross-sectional surveys, from which it is very difficult to learn the temporal order of the factors that lead to drug use and abuse.

[15] W. G. Cochran, *Sampling Techniques*, 3rd edition, New York: Wiley, 259-261, 1977.