Institute for Research on Poverty
Discussion Paper no. 1092-96

The Impact of Alcohol and Drug Use on Employment: A Labor Market Study Using the National Longitudinal Survey of Youth

Richard R. Bryant
Department of Economics
University of Missouri–Rolla

Ananda Jayawardhana
Department of Mathematics and Statistics
University of Missouri–Rolla

V. A. Samaranayake
Department of Mathematics and Statistics
University of Missouri–Rolla

Allen Wilhite
Department of Economics and Finance
University of Alabama in Huntsville

June 1996

We are grateful for financial support from the Institute for Research on Poverty through the U.S. Department of Labor. Any opinions expressed in this paper are those of the authors alone.

Abstract

The purpose of this study was, first, to estimate of the impact of alcohol and drug use on the employment status of men and women, and second, to examine whether a history of past use, as opposed to current use, adversely affects the propensity to be employed.

Using data from the National Longitudinal Survey of Youth we conducted a cross-sectional and a longitudinal analysis with logistic regression estimation to model the probability that a person was employed in 1992. In addition to usual regressors, interactions between substance use measures, between substance use measures and human capital variables, and between substance use measures and race dummies were included in the equation. The longitudinal analysis utilized a conditional likelihood method based on employment data in 1992 and 1988 and included the difference between 1992 regressors and their 1988 counterparts. A comparison was made between the prediction accuracy of the logit choice model, linear discriminant analysis, k-nearest neighbor analysis, and three modern classification methods that are used extensively in the area of machine learning. Results showed that the logit model performs relatively well in classifying individuals into employed and unemployed categories based on individual attributes.

Results of the cross-sectional and longitudinal analysis were mixed, but not inconsistent with our prior expectations that use of alcohol or drug has a negative impact on a person's propensity to be employed. Cross-sectional results show a clear negative impact of past substance use on a person's employment probability among all demographic groups examined (by gender: all persons, blacks, Hispanics, families with income below the poverty line, and high users of alcohol or drugs). However, when current and past use are considered together, only women seem to experience negative impacts. The results of the longitudinal analysis are less clear, although they do indicate that negative impacts are associated with the interaction between substance use measures and human capital variables. Limitations of the study are pointed out and suggestions are made for future research.

The Impact of Alcohol and Drug Use on Employment:
A Labor Market Study Using the National Longitudinal Survey of Youth

In the last few years, several studies have investigated the impacts of drinking or drug use on labor market success, using micro data. They have reached a variety of conclusions. Berger and Leigh (1988) found the counterintuitive result that drinking increases wages, although they offered no convincing explanation. Recent studies of drug use also have reported surprising results. In Kaestner's initial study of drug use he concluded, "the increased frequency of drug use leads to higher wages" (Kaestner, 1991). Two papers published in the same 1992 issue of Industrial and Labor Relations Review found "that a one unit increase in marijuana use . . . was associated with about a 3 to 5 percent increase in wages" (Register and Williams, 1992), and "drug users actually received higher wages than non-drug users" (Gill and Michaels, 1992). Kaestner, in two 1994 papers, used data from National Longitudinal Survey of Youth over the period 1984 to 1988 to model the effect of illicit drug use on total hours worked and wages. He found a positive impact in a cross-sectional analysis of wages, and some negative impacts on total hours of work, but in either case failed to confirm these results in a more detailed longitudinal study (Kaestner, 1994a, 1994b).

Bryant, Samaranayake, and Wilhite (1992, 1993) also found wage premiums for alcohol use with a methodology similar to Berger and Leigh, but they went on to show that different approaches negated this result: their 1992 paper controlled for an income effect on alcohol use, and wage premiums for alcohol use disappeared; their 1993 study investigated the importance of drinking patterns over time and found that after including an individual's drinking history, wage premiums disappeared and persons who had been heavy drinkers over an extended period received wage penalties. These results are consistent with their hypothesis that the negative consequences of drinking accumulate over time. The most recent paper by Bryant, Samaranayake, and Wilhite (1995) examined the effect of drug use over an extended period on wages and found that a history of drug use reduced the expected wage, and persons with longer histories had larger wage penalties.

To date, the most comprehensive set of alcohol studies come from Mullahy and Sindelar (1989, 1991, 1993). Their life-cycle approach suggests that alcohol use affects several dimensions of behavior and these can, in turn, affect earnings. For example, they find a statistically significant negative impact of alcohol use on education: individuals who drink receive less education and enter the labor market earlier. This premature entry may lead to higher wages in the earlier segment of an individual's life (as job experience is accumulated more rapidly), but eventually the lower level of education limits their wage growth, occupational choice, and earnings. In sum, a consensus seems to be emerging that alcohol use reduces earnings, but the extent of the reduction and the mechanism by which that occurs is less clear. In the area of drug use, however, more research is needed on how use, past or current, affects labor market success.

The relationship between alcohol and/or drug use and labor market success is not a trivial issue. At any point in time (depending on how alcohol use is measured), as much as 20 percent of the U.S. population can be classified as heavy drinkers (Fingarette, 1988). Further, although drug use may have fallen in recent years, it remains at a high level. In addition to the direct costs of heavy drinking or drug use (accidents, health, etc.) there may be large effects on productivity, income, and our standard of living. The work of Mullahy and Sindelar calls to attention the importance of including the impact of alcohol use on labor supply as well as on labor earnings.

This paper concentrates on the issue of employability, and differs from Mullahy and Sindelar's work in two ways. First, Mullahy and Sindelar (1993) use Epidemiological Catchment Area data, which provide medically sophisticated alcohol use measures, but more restrictive economic measures. Our study uses data from the National Longitudinal Survey of Youth (NLSY), which contains less desirable alcohol and drug use measures, but richer socioeconomic information. The availability of this additional data leads to a more comprehensive set of explanatory variables, which in turn affects the

structural model. Second, this study concerns the impact of drug use as well as the impact of alcohol use on employment.

Traditionally, economists have modeled labor force participation and similar decisions using models such as probit and logit. These are classification tools that build a discriminant rule based on a set of training data. This rule can then be used to classify individuals into two or more categories based on observed attributes. Recent advances in artificial intelligence and machine learning techniques have resulted in several modern classification tools that seem to outperform the traditional discriminant rules, at least in some situations. In addition, there are nonparametric statistical methods that can also be used for such classification. Apart from the major goals of this research, a comparison of some of these classification methods is also carried out, using the NLSY data.

Apart from the above comparison, the paper has two ultimate goals. The first is to construct estimates of the impact of alcohol and drug use on the employment status of men and women. Second, our alcohol and earnings studies suggest that while current drinking levels may not have large impacts on wages, a history of drinking, especially heavy drinking over many years, reduces earnings substantially. We want to see if a similar effect occurs with respect to employment status—that is, does a history of past use adversely affect the propensity to be employed?

EMPLOYMENT AND ALCOHOL (DRUG) USE

The supply of labor is typically derived from a utility-maximizing model in which individuals allocate their scarce time among consumption activities and labor activities. For a given wage, workers can raise the level of consumption goods by working more hours (thus earning more), but this necessarily takes away from their time to consume these items.

While intuitively simple, this framework gives structure to a variety of substantive issues in labor supply. Three main factors determine an individual’s choice between labor and leisure: expected

wages, nonlabor income, and the individual's tastes and preferences for work. Wages are a measure of the opportunity cost of leisure, and as they increase people are inclined to substitute from leisure to labor.[1] Other sources of income, income not dependent on working, generally have the opposite impact from wages. As income from nonlabor sources increases, the net marginal utility of additional income declines, earnings from work become less attractive, and labor supply declines. Finally, an individual's tastes and preferences for work, and consequently tastes and preferences for leisure, explain different decisions by individuals facing identical prices and incomes. Faced with these options, each person maximizes utility by allocating a particular amount of time to work or consumption activities. Their decisions yield a labor supply.

However, an individual's decision to work does not directly translate into employment unless that individual is selected for employment. Such selection is based on an employer's decision regarding the benefits of employing that individual as well as the individual's personal decision on acceptance of employment if offered. The former decision is assumed to be based primarily on an individual's accumulated human capital and personal characteristics as well as the economic conditions prevailing at the time. The latter decision is assumed to be based on the reservation wage (the minimum wage necessary to induce a person to work) and the market wage. In short, the employment or unemployment status of an individual is hypothesized to be a function of individual characteristics, including accumulated human capital, the prevailing economic conditions, demographic variables, and other factors that determine his or her choice between labor and leisure.

In the simplest models of labor supply, no distinction is made between labor force participation and hours of work. That is, if an individual decides to work more than zero hours, the individual participates; otherwise not. However, various researchers have suggested that the labor force participation decision is unique and should be considered separate from the decision on the hours of work. Cogan (1981) addresses the fixed costs of work (commuting time, day care, etc.) and shows that

these costs lead to a minimum number of hours necessary to recover those fixed costs. To internalize those costs he introduces a "reservation hours" equation. If available work exceeds those hours the individual participates in the labor market. Moffitt (1982) recognizes that it may be inefficient for firms to offer very low amounts of work (say two hours per week). Consequently, firms typically require some minimum level of work hours. If an individual's desired hours are less than the available minimum, then once again the decision to work differs from the supply of hours.

Zabel (1993) offers the most general model that characterizes the labor supply decision as having three components; wages, desired hours, and the decision to participate. In this study, we concentrate on employment rather than labor force participation. The rationale is that not only does nonparticipation in the labor force add to the social cost of substance abuse, but so does unemployment (of labor force participants). Thus, we add an employment equation in our model in place of a labor force participation equation. Both the labor force participation decision and subsequent employment are modeled as a compound event whose probability of occurrence is hypothesized to be a function of demographic, economic, human capital, personal, and substance use attributes of an individual. In the spirit of Zabel, our generalized employment model can be expressed by the following equations:

$$H_{i,t} = \alpha_1 \ln W_{i,t} + \alpha_2 OI_{i,t} + \mathbf{X}_{i,t}\,\alpha_3 + e_{1i,t} \quad (1a)$$

$$\ln W_{i,t} = \mathbf{Y}_{i,t}\alpha_4 + e_{2i,t} \quad (1b)$$

$$P_{i,t} \sim \text{Bernoulli}\,(P(\mathbf{Z}_{i,t}, w_{i,t}, W_{i,t}, OI_{i,t})) \text{ where} \quad (1c)$$

$$P(\mathbf{Z}_{i,t}, w_{i,t}, W_{i,t}, OI_{i,t}) = 1/[1 + \exp(-g(\mathbf{Z}_{i,t}, w_{i,t}, W_{i,t}, OI_{i,t}))]$$

$$\text{and } g(\mathbf{Z}_{i,t}, w_{i,t}, W_{i,t}, OI_{i,t}) = \mathbf{Z}_{i,t}\alpha_5 + \alpha_6 \ln w_{i,t} + \alpha_7 \ln W_{i,t} + \alpha_8 OI_{i,t} \quad (1d)$$

where: H is hours worked, lnW is the log of the market wage, OI is other (nonlabor) income, P is the employment variable set equal to one for individuals who are working, zero otherwise, and **X**, **Y**, and **Z** are vectors of demographic, economic, human capital, and personal characteristic variables expected to affect hours, wages, and employment. Note that some attributes may appear in all three vectors: **X**, **Y**,

and **Z**. Vectors **X**, **Y**, and **Z** may also include squares of quantitative variables to account for possible nonlinear relations. The subscript I refers to the i^{th} respondent, and t represents the time. All variables considered are taken from the 1979 through 1992 NLSY.

One of the explanatory variables in the employment equation (1c), is the reservation wage, w_i. Unfortunately, the reservation wage is not observed and market wage data are only observed for individuals who are working, that is, those who have decided to join the labor force and are currently employed. We do not, or cannot, observe a market wage for individuals who are presently not working. This complicates estimation.

If one is interested in estimating the structural equations, the initial step is to estimate a wage equation using only the group of individuals currently working, with necessary corrections for self-selection. Using those estimated coefficients, a predicted wage can be obtained for nonworking individuals. This predicted wage may then be used as a proxy for the reservation wage in model (1c) and estimation can proceed using logistic regression.

If estimates of structural coefficients are less important, a reduced-form equation can be used instead. In that method, the explanatory variables in the wage equation, vector **Y**, are substituted into the employment equation (and in the general case also in the hours equation) in place of the market wage. It can also be assumed that these same factors, together with other personal attributes, determine the reservation wage and hence should take the place of the reservation wage in a reduced-form equation. The resulting equation incorporates the direct effects of variables that affect employment as well as those indirect effects that affect the wage and hours worked, and thus employment. Estimated coefficients from a reduced-form equation are often called impact multipliers, because they measure the response of the endogenous variable to a change in the predetermined variables. In this study we are interested in the effect of alcohol or drug consumption on the propensity to be employed; hence, reduced-form estimation is suited for our problem.

Because the focus of our paper is the impact of alcohol and drug use on employment, alcohol and drug use must to be integrated into the model. These substances can affect employment directly, as individuals choose to consume alcohol or use drugs instead of working or, if they are heavy drinkers, or drug users, lose their interest in work. To capture the effects of alcohol or drug use, the reduced-form employment equation includes vectors, $\mathbf{A}_t$, $\mathbf{A}_{t-4}$, of alcohol and drug use measures for the current year and a past year taken from the NLSY interviews in 1984, 1988, and 1992 for drug use, and from the NLSY interviews in 1984, 1988, and 1992 for alcohol use.[2] Table 1 provides definitions for all variables used in this study. The resulting logit model incorporating the substance use terms is:

$$\text{Prob}\,(P_{i,t} = 1) = E[P_{i,t}] = [1+ \exp(-g(\mathbf{Y}_{i,t}, \mathbf{Z}_{i,t}, \mathbf{X}_{i,t}, OI_{i,t}, \mathbf{A}_{i,t}, \mathbf{A}_{i,t-4}))]^{-1} \quad (2)$$

where $g(\mathbf{Y}_{i,t}, \mathbf{Z}_{i,t}, \mathbf{X}_{i,t}, OI_{i,t}, \mathbf{A}_{i,t}) = \mathbf{Y}_{i,t}\boldsymbol{\beta}_1 + \mathbf{Z}_{i,t}\boldsymbol{\beta}_2 + \mathbf{X}_{i,t}\boldsymbol{\beta}_3 + \beta_4 OI_{i,t} + \mathbf{A}_{i,t}\boldsymbol{\beta}_5 + \mathbf{A}_{i,t-4}\boldsymbol{\beta}_6$. A major drawback of the above model is that it ignores any heterogeneity among the individuals in the sample. We address this problem by rewriting $g(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, OI_i, \mathbf{A}_i)$ as:

$$g(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, OI_i, \mathbf{A}_i) = \alpha_i + \mathbf{Y}_{i,t}\boldsymbol{\beta}_1 + \mathbf{Z}_{i,t}\boldsymbol{\beta}_2 + \mathbf{X}_{i,t}\boldsymbol{\beta}_3 + \beta_4 OI_{i,t} + \mathbf{A}_{i,t}\boldsymbol{\beta}_5 + \mathbf{A}_{i,t-4}\boldsymbol{\beta}_6 \quad (3)$$

where the α_i are individual specific constants that are assumed to be time invariant. The addition of these heterogeneity terms poses a problem in that the maximum likelihood estimates of the slope parameters $\boldsymbol{\beta}_j$, j = 1, 2, 3, 4, 5, 6, are no longer consistent when the number of time periods considered are finite. One commonly used technique for obtaining consistent estimates of the above parameters is the use of a conditional likelihood function. This is the likelihood function conditioned on a minimum sufficient statistic for α_i. Another option is to use one or more observed attributes that would act as a proxy for the heterogeneity term α_i.

TABLE 1
Variable Definitions

Variable Names	*Definitions*
Continuous Variables	
AFQT, AFQT2	Armed Forces Qualification Test score calculated from the Armed Services Vocational Aptitude Battery administered to all respondents in 1980.
AGE92	Respondent's age in 1992.
EDUC92, EDUC922	Education, the highest grade completed as of 1992.
HRSWRK84, 87, 91	Total hours worked by respondent through the 1984 interview, through the 1987 interview, and through the 1991 interview.
KIDS	KIDS092: The number of children the respondent has from 0 to 1 years of age in 1992. KIDS2392: The number of children the respondent has from 1 to 5 years of age in 1992. KIDS592: The number of children the respondent has over 5 years of age in 1992.
NLY92	1992 nonlabor income; net family income minus wage and salary income.
ROS7988	A measure of self-esteem, as measured in 1979 or 1988 in the National Longitudinal Survey of Youth. In response to 10 yes/no questions, a scale of one to ten, 1 being high and 10 low self-esteem. This practice originated with Rosenberg (1965).
ROTTER	A measure of an individual's "external/internal" view of life's events, as measured in 1988 in the National Longitudinal Survey of Youth. An external-view person thinks his life is determined by forces beyond his control; an internal view reflects the ability to alter one's environment. A scale from 0 to 4 measures an increasingly external view of life's events. The questions used for this construction come from Rotter (1966).
SHYNESS	Average measure of shyness at age 6 and as adult. Inversely related to the degree of shyness, range is 1 to 4.
UNEMP92	Local unemployment rate for the region in 1992.

(table continues)

TABLE 1, continued

Variable Names	*Definitions*
Dummy Variables	
HLIMIT92	Respondent had health limitations in 1992.
MARRY92	= 1 for married workers whose spouse is present, measured in 1992.
RACE	BLACK = 1 if respondent is black. HISPANIC = 1 if respondent is Hispanic.
REGION	A vector of regional dummy variables measured in 1992: northeast (NERD92), northcentral (NCRD92), south (SRD92), and west (WRD92).
SCHOOL	School attendance during survey period: ATTSCL84 survey period was 1984. ATTSCL92 survey period was 1992.
SMSA92	= 1 for residence in a Standard Metropolitan Statistical Area in 1992.
URBAN92	= 1 for residence in an urban area in 1992
Substance Use Variables	
Drug Use Variables:	DRUG84: Measured as the sum of the midpoints of categorical responses to questions as to the number of times the respondent reported use of marijuana or hashish, use of cocaine, use of psychedelics, and use of heroin, in the month preceding the 1984 interview. DRUG88: Measured as the sum of the midpoints of categorical responses to questions as to the number of times the respondent reported use of marijuana or hashish, and use of cocaine, in the month preceding the 1988 interview. DRUG92: Measured as the sum of the midpoints of categorical responses to questions as to the number of times the respondent reported use of marijuana or hashish, use of cocaine, and use of crack cocaine, in the month preceding the 1992 interview.
Alcohol Use Variables:	DRKLMT84, 88, 92: Total drinks reported in the month preceding the 1984, 1988, or 1992 interview. Calculated as the product of the reported number of days respondent drank in the preceding month and the typical number of drinks on those days.

(table continues)

TABLE 1, continued

Variable Names	*Definitions*
Interaction Terms:	XDRBL92 = DRUG92 * BLACK
	XDRHI92 = DRUG92 * HISPANIC
	XALBL92 = DRKLMT92 * BLACK
	XALHI92 = DRKLMT92 * HISPANIC
	XDRAL84 = DRUG84 * DRKLMT84
	XDRAL88 = DRUG88 * DRKLMT88
	XDRAL92 = DRUG92 * DRKLMT92
	XDRAF92 = DRUG92 * AFQT
	XDRED92 = DRUG92 * EDUC92
	XALAF92 = DRKLMT92 * AFQT
	XALED92 = DRKLMT92 * EDUC92
	XDRED88 = DRUG88 * EDUC92
	XDRED84 = DRUG84 * EDUC92
	XALED88 = DRKLMT88 * EDUC92
	XALED84 = DRKLMT84 * EDUC92
	XDRAF88 = DRUG88 * AFQT
	XDRAF84 = DRUG84 * AFQT
	XALAF88 = DRKLMT88 * AFQT
	XALAF84 = DRKLMT84 * AFQT
	XDR88H91 = DRUG88 * HRSWRK91
	XDR92H91 = DRUG92 * HRSWRK91
	XAL88H91 = DRKLMT88 * HRSWRK91
	XAL92H91 = DRKLMT92 * HRSWRK91
	XDRBL84 = DRUG84 * BLACK
	XDRBL88 = DRUG88 * BLACK
	XDRHI84 = DRUG84 * HISPANIC
	XDRHI88 = DRUG88 * HISPANIC
	XALBL84 = DRKLMT84 * BLACK
	XALBL88 = DRKLMT88 * BLACK
	XALHI84 = DRKLMT84 * HISPANIC
	XALHI88 = DRKLMT88 * HISPANIC

Note: Continuous variables that are squared are denoted by a "2" following the variable name.

CROSS-SECTIONAL ANALYSIS

A cross-sectional analysis was carried out using model (3), with total hours worked through the 1984 interview used as a proxy for α_i. The year 1984 was chosen as the benchmark because the cumulative hours worked through a later year would be correlated with the substance use variables, from 1984 through 1992, used in the model. The assumption that the heterogeneity factor α_i is time invariant and that total hours worked through 1984 reflects each individual's attitude toward employment is essential for this approach to be valid. Since individuals attending school in 1984 will not have had a chance to accumulate too many hours in the work force, a dummy variable indicating school attendance in 1984 was also included. The response variable used is employment status in 1992. In addition to the substance use vectors representing 1992 and 1988, another vector representing substance use in 1984 was also included. Further, interactions between alcohol and drug use variables in a given year were also introduced into the model, as were terms that represent interaction of substance use and human capital variables. Last, terms representing interaction between race dummies and substance abuse variables were included in the model.

Estimation of this expanded version of equation (3) yields the probability that a person I will be employed in 1992. The equation was fitted using logistic regression. By setting all or some measures of substance use to zero, the employment status of an individual under the assumption of nonuse of one or more substances can be determined. For instance, let P_i and P_i^* denote the predicted decision of an individual when that person's substance use measures are unchanged and are changed to zero, respectively. The difference between P_i and P_i^* can be associated with substance use. This was the approach taken in the cross-sectional analysis, the results of which are given in tables 4a and 4b.

LONGITUDINAL ANALYSIS

The above cross-sectional analysis has several potential drawbacks. First, it does not fully utilize the longitudinal information available in the NLSY data. Second, the use of a proxy for α_i may not fully eliminate the problem of heterogeneity. We therefore employed a longitudinal approach utilizing the conditional likelihood method mentioned above.

Before proceeding with the details of the conditional likelihood method used in this study, we point out certain limitations inherent in the current NLSY data set. Since equation (3) contains $\mathbf{A}_{i,t}$, the current substance use vector, and $\mathbf{A}_{i,t-4}$, the vector of substance use measures available for the most recent past, we can only let t = 1988 and 1992. Setting t = 1984 would necessitate the use of drug use measures for 1980 which are categorized into frequency classes incompatible with those for 1984, 1988, and 1992. Thus, the conditional likelihood is based on employment data for t = 1992 and t = 1988 only. Further, the minimum sufficient statistic we used is $P_{i,1988} + P_{i,1992}$. This reduces the conditional logistic model to:

$$P(b_i = 1 \mid P_{i,1988} + P_{i,1992} = 1) = 1/[1 + \exp[\boldsymbol{\beta}'w]] \qquad (4)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \boldsymbol{\beta}_4', \boldsymbol{\beta}_5', \boldsymbol{\beta}_6')'$ and $w = (\mathbf{Y}'_{i,1992} - \mathbf{Y}'_{i,1988}, \mathbf{Z}'_{i,1992} - \mathbf{Z}'_{i,1988}, \mathbf{X}'_{i,1992} - \mathbf{X}'_{i,1988}, OI'_{i,1992} - OI'_{i,1988}, \mathbf{A}'_{i,1992} - \mathbf{A}'_{i,1988}, \mathbf{A}'_{i,1988} - \mathbf{A}'_{i,1984})'$, with $\boldsymbol{X}'$ denoting the transpose of the vector $\boldsymbol{X}$. It should be noted that the Bernoulli variable b_i is defined such that $b_i = 1$ if $P_{i,1988} = 0$ with $P_{i,1992} = 1$ and $b_i = 0$ if $P_{i,1988} = 1$ with $P_{i,1992} = 0$. The individuals with $P_{i,1992} = P_{i,1988} = 0$ and $P_{i,1992} = P_{i,1992} = 1$ do not contribute to the conditional likelihood function and hence can be dropped (see Hsiao, 1986, p. 162). In addition to the above, we also included the differences between the interactions of the 1992 human capital variables with $\mathbf{A}_{i,1992}$ and $\mathbf{A}_{i,1988}$ and the interactions of 1988 human capital variables with $\mathbf{A}_{i,1988}$ and $\mathbf{A}_{i,1984}$.

An issue that brings into question the validity of the results obtained by the above longitudinal approach is the possible endogeneity of the substance use variables. A solution to this problem is the use of predicted substance use values, based on reduced-form drug and alcohol use equations for each year, in place of the actual values. One can obtain such predictions by fitting substance use equations to model the actual drug/alcohol use frequencies as a function of demographic, human capital, other personal characteristics, and socioeconomic variables.

Each of the substance use models for a given year was fitted using only those individuals who use the substance under question for that given year. This was done to avoid fitting an equation to data that contain zero response values for a large number of observations. Appropriate corrections were made to account for the bias due to selection into the user group. Specifically, what was modeled is S_i^*, an individual's "potential" substance use frequency, given his or her personal and other related attributes. It is assumed that an individual's expected substance use frequency, $S_{i,}$ is zero if S_i^* is zero or negative and equals S^*_I otherwise. Mathematically, one can express

$$S_{i,t}^* = \beta_0 + U_{i,t}\boldsymbol{\beta}_1 + \epsilon_{i,t} \tag{5}$$

where U_i is a vector of personal characteristics, demographic and socioeconomic variables, and contains several attributes not contained in the regressor vectors $\mathbf{X}_i$, $\mathbf{Y}_i$, and $\mathbf{Z}_i$. These include family poverty status, several religion dummies, and a shyness index. For a given substance, the same model was used for each year. However, the values taken by the variables correspond to the level of each variable for that year. Only data from those who used the substance in year t were used to fit equation (5), however, the inverse Mills ratio associated with selection into the user group was included in the equation to correct for possible bias.

The estimated equation was used to predict $S_{i,t}$ for all individuals, including nonusers. The expected substance use, $S_{i,t}$, was then predicted as follows:

$$PS_{i,t} = PS_{i,t}^* \text{ if } PS_{i,t}^* > 0 \text{ and } PS_{i,t} = 0 \text{ if } PS_{i,t}^* \leq 0. \tag{6}$$

Note that $PS_{i,t}^{*}$ denotes the predicted value of $S_{i,t}^{*}$ and $PS_{i,t}$ denotes the predicted value of $S_{i,t}$. The longitudinal analysis was carried out with the actual substance use values in vectors $\mathbf{A}_{i,t}$ and $\mathbf{A}_{i,t-1}$ replaced by the predicted values obtained by using (5) and (6).

Recently, concern has been raised as to the utility of using predicted values of endogenous variables in eliminating bias. For instance, Bound et al. (1995) demonstrate that problems can arise when instrumental variables are utilized to predict endogenous variables as a means of removing bias. Such problems arise when there is a weak correlation between the endogenous variable and the instrument variables used to predict it. This is the situation in this study, where there is a very weak association between the substance use frequencies and the regressors in the drug and alcohol use equations.[3] Thus, the longitudinal analysis was carried out using both predicted and actual substance use data.

DATA DESCRIPTION

The National Longitudinal Survey of Youth Labor Market Experience (NLSY), 1979-1992, prepared by the Center for Human Resources Research at Ohio State University, provides the data base for this investigation. The NLSY is a multistage random sample of 12,686 individuals representative of persons born in the years 1957–1964, surveyed annually beginning in 1979. The initial 1979 survey consisted of 6,283 females (1,002 Hispanic and 1,561 black), and 6,403 males (832 Hispanic and 1,488 black). In 1992, 71.1 percent of the original cohort was interviewed, including 4,481 males and 4,535 females. All of the respondents were between the ages of 27 and 35, and whites outnumbered blacks by a little more than two to one. Observations that were unusable owing to missing values further reduced the sample size to 5,934 respondents, 2,886 males (737 blacks and 446 Hispanics) and 3,048 females (846 blacks and 472 Hispanics). Table 2a provides summary data on the variables by gender for the

TABLE 2A
Cross-Sectional Model, Summary Statistics

	Males: N = 2886		*Females: N = 3048*	
Variable	Mean	Standard Deviation	Mean	Standard Deviation
Continuous Variables				
AFQT	65.40	22.43	65.58	20.38
AGE92	30.81	2.23	30.99	2.22
EDUC92	12.85	2.42	12.99	2.26
HRSWRK84	6,502.67	4,321.16	5,070.03	3,810.12
KIDS092	0.13	0.35	0.11	0.32
KIDS592	0.78	1.09	1.12	1.19
KIDS2392	0.44	0.65	0.40	0.60
NLY92	10,779.57	13,433.06	19,441.81	18,302.00
ROS7988	1.63	0.41	1.67	0.41
ROTTER	1.46	1.03	1.50	1.05
S84	0.19	0.39	0.17	0.37
S92	0.05	0.21	0.07	0.26
UNEMP92	7.97	2.62	7.99	2.62
Drug Variables				
COKE84MT	0.41	2.71	0.20	1.75
COKE88MT	0.28	2.15	0.17	1.63
COKE92MT	0.41	3.36	0.15	2.06
CRK92MT	0.17	2.17	0.06	1.18
HER84MT	0.01	0.76	0.01	0.59
MRJH84MT	3.54	8.48	1.65	6.00
MRJH88MT	1.97	6.43	1.00	4.70
MRJH92MT	1.98	6.77	1.18	5.46
PSY84MT	0.04	0.94	0.02	0.81
Alcohol Variables				
DRKLMT84	23.49	29.83	8.94	16.92
DRKLMT88	25.25	57.90	9.13	23.86
DRKLMT92	53.75	81.03	33.35	68.79
Dummy Variables (% 1's)				
BLACK	25.53	43.61	27.75	44.78
HISPANIC	15.45	36.15	15.48	36.18
HLIMIT92	5.95	23.67	8.72	28.22
MARRY92	5.58	49.66	5.64	49.59
NCRD92	26.16	43.95	24.67	43.11
NERD92	14.51	35.23	13.28	33.94
SMSA92	76.05	42.68	77.88	41.50
WRD92	21.20	40.88	20.53	40.40

cross-sectional analysis. Table 2b provides this information for the data used in the longitudinal analysis.

ESTIMATES OF THE IMPACT OF ALCOHOL AND DRUG USE ON EMPLOYMENT STATUS: CROSS-SECTIONAL RESULTS

Table 1 gives the list of variables used in the employment model described in (1c) and (3). Since the impact of the independent variables on employment status could differ across gender, separate models were fit for males and females. The estimated logistic regression coefficients are given for males and females in Table 3.

Observe that in addition to fitting different models for males and females, the models also include additional variables that give the number of children between the ages 0 to 1, 1 to 5, and over 5. As expected, the number-of-children variables show a negative and significant impact on the employment of women, but in the equation for men only the variable for number of children over age 5 is negative, showing a marginally significant impact on their employment.

Other variables that have a negative impact on the probability of employment of women and are statistically significant at standard levels include having a higher level of family nonlabor income, having a health limitation, living in a region with a higher unemployment rate, and attending school in 1992. Variables that positively influence a typical women's employment probability include previous work experience (through 1984), the stock of human capital as measured by the armed forces qualification test (AFQT) score (dampened somewhat by the negative impact of AFQT squared), being married in 1992, having attended school in 1984, and being black or Hispanic.

In the male equation, being black or Hispanic, having a health limitation, a higher family nonlabor income ($p = 0.06$), attending school in 1992, or having lower self-esteem as measured by the Rosenberg index reduces the chances of employment. A typical male is more likely to be employed if

TABLE 2B
Longitudinal Model, Summary Statistics

Variable	*Males: N = 233* Mean Difference (92-88)	*Males: N = 233* Standard Deviation of Difference	*Females: N = 576* Mean Difference (92-88)	*Females: N = 576* Standard Deviation of Difference
Continuous Variables				
EDUC	0.29	0.78	0.27	0.78
EXPCLF	33.96	42.63	42.74	42.48
KIDS0	-0.12	0.51	-0.14	0.56
KIDS5	0.41	0.65	0.59	0.68
KIDS23	-0.04	0.85	-0.12	0.98
NLY	-141.32	16,651.90	3,289.77	17,852.12
UNEMP	1.57	2.72	1.68	2.79
Drug Variables				
DRUG (past use)	-3.32	15.61	-1.37	7.59
DRUG (current use)	0.24	12.91	0.57	7.51
Alcohol Variables				
DRKLMT (past use)	6.15	49.91	-0.69	21.36
DRKLMT (current use)	16.68	72.53	26.94	76.02
Dummy Variables (% 1's)				
ATTSCL	-2.14	40.94	-13.71	39.20
HLIMIT	8.58	41.65	4.69	37.70
MARRY	6.43	47.36	0.87	45.10
URBAN	-0.43	30.08	1.04	27.01

TABLE 3
Estimated Logistic Regression Equations for Cross-sectional Models for Males and Females

	Males		*Females*	
Variable	Parameter Estimate	p-value	Parameter Estimate	p-value
INTERCEPT	3.4344	0.0318	1.1524	0.3780
HRSWRK84	0.000123	0.0001	0.000179	0.0001
AGE92	-0.0802	0.0188	-0.0673	0.0088
AFQT	0.00787	0.6208	0.0634	0.0001
EDUC92	0.0399	0.8251	-0.0449	0.7704
BLACK	-0.6306	0.0066	0.4922	0.0028
HISPANIC	-0.7653	0.0059	0.4226	0.0180
HLIMIT92	-2.1677	0.0001	-1.3627	0.0001
MARRY92	0.9226	0.0001	0.5525	0.0001
NERD92	-0.3451	0.0748	-0.1584	0.3033
NCRD92	0.0810	0.6414	-0.1581	0.2075
WRD92	0.0617	0.7429	-0.1914	0.1640
NLY92	-0.00001	0.0611	-0.00003	0.0001
KIDS092	0.1178	0.5633	-0.9560	0.0001
KIDS2392	0.0720	0.4983	-0.8297	0.0001
KIDS592	-0.1034	0.0825	-0.1617	0.0006
SMSA92	0.0862	0.5958	-0.0592	0.6243
ROTTER	0.0712	0.2758	-0.0379	0.4259
ROS7988	-0.4978	0.0030	-0.1643	0.1889
UNEMP92	-0.0366	0.1690	-0.0454	0.0190
ATTSCL84	0.3133	0.1623	0.6374	0.0002
ATTSCL92	-1.2692	0.0001	-0.6441	0.0003
AFQT2	6.257E-7	0.9963	-0.0004	0.0002
EDUC922	0.00265	0.7186	0.00653	0.2836
DRUG84	0.0144	0.7408	0.0801	0.2070
DRUG88	0.0681	0.1719	-0.0302	0.6828
DRUG92	0.0129	0.7790	-0.0882	0.1656
DRKLMT84	0.00149	0.9098	-0.0504	0.0338
DRKLMT88	-0.00259	0.8170	-0.0191	0.2560
DRKLMT92	-0.00393	0.5048	0.0120	0.0228
XDRBL92	0.0168	0.3619	-0.0302	0.1419
XDRHI92	0.0554	0.1301	-0.0141	0.5487
XALBL92	0.0044	0.0656	-0.00678	0.0052
XALHI92	0.00592	0.0567	-0.00340	0.1788
XDRAL84	0.000047	0.7126	-0.00008	0.7642
XDRAL88	-0.00018	0.1957	0.000143	0.4392
XDRAL92	-0.00013	0.0499	-0.00014	0.1960
XDRAF92	-0.00013	0.7944	-0.00056	0.2846
XDRED92	0.000556	0.8963	0.00974	0.0884
XALAF92	0.000053	0.3162	0.000047	0.2889
XALED92	0.000247	0.6683	-0.00102	0.0392
XDRED88	-0.00555	0.2484	-0.00272	0.6498
XDRED84	-0.00336	0.4046	-0.00758	0.1625
XALED88	0.00124	0.2915	0.00108	0.5271
XALED84	-0.00133	0.2885	0.00388	0.0921

(table continues)

TABLE 3, continued

	Males		*Females*	
Variable	Parameter Estimate	p-value	Parameter Estimate	p-value
XDRAF88	-0.00007	0.8769	0.000811	0.3161
XDRAF84	0.00014	0.7235	-0.00011	0.8524
XALAF88	-0.00015	0.2200	0.000026	0.8980
XALAF84	0.000174	0.2427	0.000076	0.7230
XDRBL84	-0.0194	0.1969	0.000689	0.9731
XDRBL88	0.0103	0.6210	-0.00173	0.9515
XDRHI84	0.00554	0.8093	-0.0244	0.3168
XDRHI88	-0.0240	0.4548	0.0191	0.6220
XALBL84	0.00823	0.1667	-0.0006	0.9467
XALBL88	-0.00594	0.1852	0.00946	0.2223
XALHI84	0.0143	0.0821	-0.00537	0.6243
XALHI88	-0.00686	0.2316	0.000779	0.9242
	n = 2,886		n = 3,048	

he has worked many hours in the past (up through 1984), and perhaps if he lived in the northeast region of the United States as compared to the south. Significant positive association is also found between employment and being married in 1992.

The models for both men and women show a negative relationship between employment and age. This somewhat contradictory result may be explained by the presence of human capital variables such as AFQT and education in the model. If two individuals have the same amount of accumulated human capital, but one is older than the other, then it is reasonable to conclude that the additional age need not give an employment edge to the older individual. In fact, the younger person is likely to be more aggressive in seeking not only accumulation of human capital, but also employment. On a related topic, the nonsignificance of AFQT in the model for men, and education variables in both models, may be due to the high collinearity that exists between the human capital variables as well as between a variable and its square term. These being standard employment equations, the direction and significance of most variables are not unusual, although the lack of significance of the regional unemployment rate in the male equation is surprising. Our main concern, however, is the impact of substance use on the propensity to be employed.

In the model for men, none of the alcohol and drug use variables are significant (at 0.05 level), except for the 1992 drug and alcohol use interaction term, which shows a negative impact associated with concurrent use of alcohol and drugs. A marginally significant positive association is found between employment and 1992 alcohol consumption by blacks and between employment and 1992 and 1984 alcohol consumption by Hispanics.

In the women's model, significant negative impacts on the probability of employment are associated with drinking in 1984. Interaction between the black dummy and drinking in 1992 also shows a significant negative association, indicating that the employment probability of blacks is reduced by current drinking. Interaction between 1992 alcohol use and education is also associated with

a significant negative impact. Only 1992 drinking shows a significant positive impact at the 0.05 level. Marginally significant positive associations are shown between employment and the interaction terms, drug use in 1992 and education, as well as 1984 alcohol use and education.

A number of drug use and drinking variables are not significant, but the nonsignificance of these variables does not necessarily suggest that they have no effect on employment. Drinking and drug use tend to be correlated with one another and, when fit simultaneously into a model, they may appear nonsignificant. Chi-square tests do not directly test the overall impact of each variable, but are a measure of how much additional information a given variable can provide about the response variable after adjusting for the rest of the variables in the model. Two or more correlated variables that directly affect employment can appear nonsignificant if they are fit together in the model. The hypothesis that drinking and drug use variables taken together have no impact on the propensity to be employed was tested (for both the male and the female model) and was rejected.

Since in all models the impact of substance use variables is not consistently negative or positive, inspection of the regression coefficients does not immediately reveal the net impact of substance use. The matter is further complicated by the presence of interaction terms. As such, the net impact of substance use on employment was studied by estimating $E[P_{i,t}]$ in equation (2), for each individual, under their existing substance use pattern, a so-called status quo scenario, and two variations of nonuse scenarios. The quantity $E[P_{i,1992}]$ is the probability that an individual having the attributes, $\mathbf{X}_i$, OI_i, $\mathbf{A}_i$, $\mathbf{Z}_i$, would have been employed in 1992. By setting all or some variables in the substance use vector, $\mathbf{A}_i$, to zero, in equations (1c), (2), and (3), one can determine the probability of employment under any nonuse scenario. The difference between $E[P_{i,1992}]$ under an individual's existing substance use profile and under a nonuse scenario can be attributed to the impact of the substance use variables that were set equal to zero.

In this study, two nonuse scenarios were used. First, all substance use variables, alcohol and drug at the different years, were set equal to zero, and second, all substance use variables except the 1992 ones were set equal to zero. The latter scenario allows the estimation of the effect of past substance use. The net effect of both past and current use (1992) are investigated by the first scenario. The values $E[P_{i,1992}]$ were averaged by gender for several demographic groups: all persons, blacks, Hispanics, families with income below the poverty line, and high users of alcohol or drugs. High users were defined as those who consumed alcohol or drugs at or above the median level for users in two of the three years considered. The results of this estimation process are reported in Tables 4a and 4b.

For every demographic group, except for Hispanic men, setting past substance use variables equal to zero leads to an increase in the expected probability of employment. Furthermore, for women, similar increases in the expected probability of employment is observed when past and current substance use variables are set equal to zero. For men, setting both current and past substance use variables equal to zero leads to a decrease in the expected probability of employment. This result, coupled with the results of setting past use equal to zero, suggests that for men current substance use is associated with an increase in employment probability, whereas past use is associated with a decrease. On the other hand, for women both past and current use are associated with deceases in the probability of employment. The positive association between current substance use and employment of men is perhaps due to an income effect. That is, current employment leads to more discretionary income, which in turn leads to higher substance use. Other demographic categories have results similar to those associated with results for men and women: when past use is set to zero, the result is that expected employment probability increases for that group by one to three percentage points (again the sole exception is for Hispanic men).

TABLE 4A
Probability of Employment for Males, Based on Cross-Sectional Model

	Status Quo[a]	*Past Use*[b]	*All Use*[c]
Substance Use Scenarios			
Mean	0.8669	0.8775	0.8537
Median	0.9284	0.9341	0.9149
Std. of mean	0.0030	0.0028	0.0030
Sample size	2,886	2,886	2,886
Blacks (BLACK=1)			
Mean	0.7720	0.7950	0.7575
Median	0.8324	0.8509	0.8056
Std. of mean	0.0071	0.0066	0.0068
Sample mean	737	737	737
Hispanics (HISPANIC=1)			
Mean	0.8542	0.8513	0.7927
Median	0.9118	0.9065	0.8527
Std. of mean	0.0077	0.0075	0.0083
Sample size	446	446	446
Below Poverty Level (FAMILY POVERTY STATUS=1)			
Mean	0.6961	0.7263	0.6903
Median	0.7573	0.7865	0.7452
Std. of mean	0.0118	0.0109	0.0112
Sample size	400	400	400
High Substance Use (HIALDRIN=1)			
Mean	0.8852	0.8986	0.8640
Median	0.9387	0.9432	0.9220
Std. of mean	0.0039	0.0035	0.0042
Sample size	1,360	1,360	1,360

[a]Current use and past substance use unchanged, in the years 1984, 1988, and 1992.
[b]Past substance use set equal to zero, in the years 1984 and 1988.
[c]All substance use set equal to zero, in the years 1984, 1988, and 1992.

TABLE 4B
Probability of Employment for Females, Based on Cross-Sectional Model

	Status Quo[a]	*Past Use*[b]	*All Use*[c]
Substance Use Scenarios			
Mean	0.7106	0.7177	0.7229
Median	0.7791	0.7833	0.7876
Std. of mean	0.0042	0.0040	0.0039
Sample size	3,048	3,048	3,048
Blacks (BLACK=1)			
Mean	0.6867	0.6938	0.7217
Median	0.7607	0.7629	0.7971
Std. of mean	0.0087	0.0083	0.0079
Sample mean	846	846	846
Hispanics (HISPANIC=1)			
Mean	0.6673	0.6851	0.6938
Median	0.7285	0.7413	0.7542
Std. of mean	0.0114	0.0105	0.0110
Sample size	472	472	472
Below Poverty Level (FAMILY POVERTY STATUS=1)			
Mean	0.5156	0.5402	0.5584
Median	0.5267	0.5713	0.5924
Std. of mean	0.0105	0.0102	0.0100
Sample size	590	590	590
High Substance Use (HIALDRIN=1)			
Mean	0.7271	0.7542	0.7636
Median	0.8065	0.8268	0.8293
Std. of mean	0.0103	0.0090	0.0084
Sample size	560	560	560

[a]Current use and past substance use unchanged. In the years 1984, 1988, and 1992.
[b]Past substance use set equal to zero. In the years 1984 and 1988.
[c]All substance use set equal to zero. In the years 1984, 1988, and 1992.

ESTIMATES OF THE IMPACT OF ALCOHOL AND DRUG USE ON EMPLOYMENT STATUS : LONGITUDINAL ANALYSIS

The results of the longitudinal analysis for males are reported in Table 5a; for females, in Table 5b. Recall that model (4) utilizes the difference between regressors in a 1992 version of model (3) and the 1988 version. These differences are used to obtain consistent estimates of the coefficients in model (3). Thus, the estimated coefficients are reported alongside the corresponding variable in model (3). For instance, the coefficient associated with the education variable in equation (3) is estimated by using a term giving the difference between 1992 and 1988 education levels. This estimate is reported in Tables 5a and 5b as the coefficient associated with the EDUC92 (education) variable. Further, each table gives estimates based on models employing actual substance use frequencies as well as models that used their predicted counterparts.

The only statistically significant substance use terms in the male models appear in the version that utilizes predicted substance use variables. These terms include the 1988 drinking variable and the interaction between 1988 drug use and labor force experience as measured by total hours worked through the 1991 NLSY interview.[4] The latter show a negative association between substance use and employment probability (although it is only marginally significant, with a p-value equal to 0.0625).

The model for women utilizing actual substance use values shows that the interaction between 1992 drug use and labor force experience (through 1991) has a marginally negative association with employment (the p-value is 0.0859). No other substance use variable shows any significant association in this model. The model that employs predicted substance use values indicates that 1988 drinking is positively associated with employment, with a p-value of 0.0714. It also shows a significant positive association between 1992 drug use and employment probability. In the same model, the interaction between alcohol use and education shows a marginally significant negative (p-value 0.0949) association with 1992 employment. The female model utilizing actual substance use values shows negative

TABLE 5A
Longitudinal Model, Males

	Actual		*Predicted*	
	Parameter Estimate	P	Parameter Estimate	P
INTERCPT	-0.1490	0.6157	-0.6521	0.3171
HLIMIT92	-1.2064	0.0077	-1.3195	0.0152
MARRY92	0.2019	0.5865	0.1947	0.6540
KIDS092	0.3052	0.4816	0.8495	0.0778
KIDS2392	-0.0674	0.8386	0.0864	0.8092
KIDS592	0.2287	0.5712	0.5494	0.2186
UNEMP92	-0.2199	0.0013	-0.2280	0.0027
ATTSCL92	-1.3080	0.0037	-1.2922	0.0071
URBAN92	0.8633	0.1439	0.8617	0.2080
NLY92	-0.00003	0.0046	-0.00003	0.0090
HRSWRK91	0.00127	0.7969	0.0346	0.0289
EDUC92	0.5035	0.0613	0.8870	0.0476
DRKLMT88	-0.0241	0.2806	0.1486	0.0455
DRKLMT92	-0.0129	0.4203	0.0634	0.1922
DRUG88	-0.0962	0.3601	0.1900	0.1632
DRUG92	-0.1358	0.2376	0.0045	0.8008
XDRED88	0.0113	0.1463	-0.00913	0.3839
XDRED92	0.0134	0.1487	0.000886	0.4435
XDR88H91	-0.00025	0.5786	-0.00121	0.0625
XDR92H91	0.000022	0.9609	-0.00012	0.2569
XALED88	0.00140	0.4267	-0.00804	0.1475
XALED92	0.000688	0.5546	-0.00232	0.4882
XAL88H91	0.000038	0.6213	-0.0003	0.1901
XAL92H91	0.000037	0.6591	-0.00024	0.3166
P > chi-square		N = 233		N = 202

TABLE 5B
Longitudinal Model, Females

	Actual		*Predicted*	
	Parameter Estimate	P	Parameter Estimate	P
INTERCPT	0.0659	0.7314	-1.4311	0.0063
HLIMIT92	-0.8104	0.0011	-0.8884	0.0007
MARRY92	0.00411	0.9863	-0.1851	0.4637
KIDS092	-1.3201	0.0001	-1.2781	0.0001
KIDS2392	-0.9188	0.0001	-0.8331	0.0001
KIDS592	-0.5058	0.0474	-0.3983	0.1369
UNEMP92	-0.00643	0.8481	-0.0149	0.6723
ATTSCL92	-0.7066	0.0043	-0.5479	0.0341
URBAN92	-0.1190	0.7380	-0.2222	0.5400
NLY92	-5.12E-6	0.3924	-2.7E-6	0.6582
HRSWRK91	0.00508	0.0406	0.0102	0.2452
EDUC92	0.0669	0.6073	0.2560	0.2117
DRKLMT88	-0.00674	0.7877	0.1173	0.0714
DRKLMT92	-0.00722	0.6319	-0.0128	0.3355
DRUG88	0.00428	0.9673	0.0677	0.1905
DRUG92	0.0318	0.7837	0.0336	0.0442
XDRED88	-0.00219	0.7876	-0.00440	0.2866
XDRED92	-0.00209	0.7993	-0.00074	0.5531
XDR88H91	0.000084	0.8224	0.00005	0.8206
XDR92H91	-0.00077	0.0859	-0.0001	0.1935
XALED88	0.000791	0.7087	-0.00864	0.0949
XALED92	0.00103	0.3914	0.000821	0.3817
XAL88H91	-0.00009	0.3428	-0.00011	0.6808
XAL92H91	-0.00007	0.1821	9.381E-6	0.8775
P> chi-square	N = 576		N = 534	

coefficients associated with 1988 and 1992 drinking variables; neither, however, is statistically significant.

Because it is possible that the nonsignificance of some of the substance use variables is due to the presence of other substance use variables in the model, a stepwise procedure was carried out to eliminate variables that are not significant at the 0.15 level. The elimination was limited to the interaction terms only, with the main effects forced into the model. The results of this variable selection process (implemented by the SAS procedure LOGIT) are reported in Tables 6a and 6b.

The model for males utilizing predicted substance use values show a positive association between 1988 drinking and 1992 employment. However, the interaction between 1988 drug use and labor force experience through 1991 indicates a negative impact that is marginally significant. A similar result is seen for the term representing the interaction between the above labor force experience term and 1988 alcohol use.

The model for females (Table 6b) using actual substance use values shows a negative impact resulting from interaction of 1988 drug use and labor force experience through 1991. A marginally significant negative association between employment in 1992 and the interaction between 1988 alcohol use and labor force experience through 1991 is also present in this model. The model using predicted values shows two statistically significant substance use variables. The 1992 drug use variable is positively associated with 1992 employment, yet the interaction of 1988 drug use and labor force experience through 1991 is significantly negative. In addition, the 1988 drug use variable shows a positive impact that is marginally significant.

Clearly, the longitudinal study yields mixed signals as to the impact of current and past substance use on employment. In spite of these mixed signals, the estimates for the other variables are to a large extent what one would expect. For instance, having children reduces the probability of women's employment, and the unemployment rate has a negative impact on the probability of

TABLE 6A
Longitudinal Model, Males: Stepwise Procedure

	Actual		*Predicted*	
	Parameter Estimate	P	Parameter Estimate	P
INTERCPT	-0.1501	0.5940	-0.4102	0.4916
HLIMIT92	-1.1983	0.0061	-1.3975	0.0086
MARRY92	0.1955	0.5856	0.4255	0.2929
KIDS092	0.4288	0.3098	0.5338	0.2398
KIDS2392	-0.0933	0.7721	0.0328	0.9244
KIDS592	0.1799	0.6437	0.4782	0.2657
UNEMP92	-0.2088	0.0013	-0.2397	0.0013
ATTSCL92	-1.2156	0.0043	-1.2315	0.0073
URBAN92	0.8259	0.1540	0.8847	0.1970
NLY92	-0.00003	0.0040	-0.00003	0.0090
HRSWRK91	0.00309	0.4082	0.0177	0.0661
EDUC92	0.6260	0.0106	0.5430	0.0387
DRKLMT88	-0.00409	0.2965	0.0470	0.0202
DRKLMT92	-0.00226	0.3769	0.0134	0.1429
DRUG88	0.0173	0.2576	0.0746	0.1171
DRUG92	0.0226	0.1133	-0.00261	0.6739
XDR88H91	—	—	-0.00115	0.0662
XAL88H91	—	—	-0.00035	0.0944
P > chi-square		N = 233		N = 202

TABLE 6B
Longitudinal Model, Females: Stepwise Procedure

	Actual		*Predicted*	
	Parameter Estimate	P	Parameter Estimate	P
INTERCPT	0.0666	0.7272	-1.5054	0.0007
HLIMIT92	-0.8072	0.0011	-0.9017	0.0006
MARRY92	0.000664	0.9978	-0.1814	0.4650
KIDS092	-1.3223	0.0001	-1.2469	0.0001
KIDS2392	-0.9003	0.0001	-0.8104	0.0001
KIDS592	-0.4826	0.0560	-0.3944	0.1318
UNEMP92	-0.00576	0.8630	-0.00985	0.7739
ATTSCL92	-0.7079	0.0040	-0.6302	0.0124
URBAN92	-0.1014	0.7736	-0.1577	0.6604
NLY92	-5.28E-6	0.3742	-2.9E-6	0.6306
HRSWRK91	0.00451	0.0636	0.0104	0.0034
EDUC92	0.0955	0.4414	0.0837	0.5222
DRKLMT88	-0.00185	0.6900	-0.00054	0.9547
DRKLMT92	0.00636	0.1461	-0.00120	0.5593
DRUG88	-0.0154	0.2454	0.0150	0.0724
DRUG92	0.00507	0.8697	0.0252	0.0001
XDR88H91	-0.00076	0.0389	-0.00011	0.0032
XAL88H91	-0.00008	0.0980	.------	.-----
P > chi-square		N = 576		N = 534

employment. This indicates that the models are, to an extent, a reflection of the real nature of things and not an artifact created by the idiosyncrasies of the data. Moreover, there is one result that appears consistently across most of the longitudinal models, namely the significant negative impact of the interaction between the past labor force experience variable and one or more of the substance use variables. It seems that the positive impact of past experience is tempered to some extent by substance use.

There are drawbacks to the longitudinal analysis carried out in this study. First, the limitation of two time periods (1988 and 1992) restricts the scope of the analysis by reducing the effective data base to those who were employed in exactly one of the two years. We are interested in determining the effect of not only current substance use, but also such use in the past. However, the use of more than two time periods is infeasible, owing to the nonstandard scales used for 1980 drug use measurements. Further, no alcohol use variables are available for 1980. Thus, we need to wait for later data to be available.

COMPARISON OF CLASSIFIERS

Three modern classifiers, with roots in machine learning, and three traditional classifiers were selected for comparison. The selected machine learning classifiers are *Classification and Regression Tree* (CART), CN2, and C4.5. These are a representative group of the pattern recognition and classification algorithms that are currently the most prominent. CART and C4.5 are top-down, decision-tree based classifiers, whereas CN2 uses a set of “if-then” rules to carry out the classification. C4.5 is an improved version of the well-known machine learning algorithm ID3. The CART classifier is described in McLachlan (1992), and details of the others are given in Mitchie et al. (1994).

Three traditional methods, logistic regression, linear discriminant analysis, and k-nearest neighbor analysis, were also selected for comparison. These methods are described in detail in

McLachlan (1992). Since the machine-learning algorithms and the k-nearest neighbor method are nonparametric in nature, it is very difficult to incorporate the complex system of structural equations, discussed earlier, in the analysis. Further, to keep the comparative study simple, 1992 employment status was modeled in a cross-sectional sense. The emphasis here is not to obtain estimates of the impact of substance use, but to determine if one or more of the modern classifiers do a better job of modeling the employment status of individuals as a function of socioeconomic, demographic, and personal attributes.

The data set was first divided into male and female groups. The substance use variables were predicted for each individual, as was done in the longitudinal analysis. The male data set was further randomly subdivided into two roughly equal subsets, preserving the employed-to-unemployed ratio found in the full data set. Each classifier was trained (estimated) on one subset and its prediction accuracy was determined using the other subset. The role of the subsets were then reversed and prediction accuracy was measured again. This process was repeated for the female data set. For the k-nearest neighbor methods, the number of neighborhood points, k, was set to 8 for males and 7 for females. This was based on a preliminary study that revealed the above values as optimal among a range of such values. The logistic regression method requires a cut-off probability to determine the individuals for classification into the employed category. This probability was selected so as to obtain near-equal classification accuracies for both employed and unemployed categories. It should be noted that this selection does not in any way influence the parameter estimates of the logit function. The average prediction accuracies obtained for each method are given in Tables 7a and 7b. Observe that the CN2 classifier was run using unordered rules as well as ordered rules. The CART algorithm was first used without a utility matrix, then using such a matrix. The use of a utility matrix allows the weighting

TABLE 7A

Accuracy of Discriminant Methods for Males:
Percentage Correctly Classified

Method	*Accuracy for Employed*	*Accuracy for Unemployed*
CART without utility matrix	100.00%	0.00%
CART with utility matrix	89.48	59.75
C4.5	96.52	31.87
CN2 with unordered rules	99.25	9.46
CN2 with ordered rules	97.54	13.26
Logitistic regression	75.00	75.43
k-nearest neighbor (k=8)	68.52	70.59
Linear discriminant analysis	86.99	61.94

TABLE 7B

Accuracy of Discriminant Methods for Females:
Percentage Correctly Classified

Method	*Accuracy for Employed*	*Accuracy for Unemployed*
CART without utility matrix	93.21%	35.95%
CART with utility matrix	79.40	60.45
C4.5	90.76	39.30
CN2 with unordered rules	94.80	26.12
CN2 with ordered rules	92.69	35.07
Logitistic regression	74.26	74.50
k-nearest neighbor (k=7)	65.26	64.43
Linear discriminant analysis	78.56	66.17

of individual observations differently, so that misclassifying an unemployed individual is given a higher penalty than misclassifying an employed person. Otherwise, the algorithm tends to favor employed individuals and ignore the less frequent unemployed individuals.

As the results show, all the modern discriminant models tend to build rules that favor classification into the more prevalent group, which happens to be the employed category. This problem is somewhat corrected in CART by the use of the utility matrix. A seeming strength of these modern classifiers is that they are very flexible and can thus model very complex phenomena. However, this can become a weakness in situations where the classification data are very noisy. They tend to model not only the signal in the data but the noise as well. In addition, in the absence of a clear-cut set of attributes that gives accurate classification information, these methods will develop spurious rules that would bias the classification towards the more dominant category. This is apparently what occurred in this study.[5] There is no simple method, such as selecting a cutoff probability in logistic regression, to balance the accuracy for each employment category. The actual implementation of the machine-learning methods requires judgmental and heuristic setting of parameters. We have keep such ad hoc fine tuning to a minimum, although it would have been possible to improve the performance of these classifiers by employing some of the ad hoc fine-tuning techniques. Since these are more indirect than, say, selecting a simple cutoff probability, and can affect the very structure of the estimated model, there is no assurance that the resulting models are a good approximation of the true underlying causal relationship between employment and the individual attributes. Overdependence on such heuristic methods can result in models that are massaged into fitting the data. Thus, no extensive efforts were made to further improve the performance of the machine-learning classifiers by employing data-driven fine-tuning techniques.

Among the traditional methods, logistic regression seem to perform well, classifying both employed and unemployed with equal accuracy. The latter result is no surprise, since a cutoff

probability was selected to achieve equal accuracy for both employment groups. What is important, however, is that the overall accuracy of the logistic method is very respectable compared to that of other methods. Although both the k-nearest neighbor and linear discriminant methods showed respectable accuracy overall, they also showed a tendency to prefer the dominant category over the other. Clearly, the above results give some justification to the use of a logit function to model the probability of employment.

SUMMARY

Our results concerning the impact of substance use, although mixed, are not inconsistent with our prior expectations that use of alcohol or of drugs will have a negative impact on a person's propensity to be employed. At the least, the cross-sectional results show a pattern of negative association between past substance use and employment. Even though the results of the longitudinal study provide no clear answer, they do indicate that negative impacts are associated with interaction terms, suggesting that substance use may negatively affect employment probability by affecting how human capital variables influence employment.

As in any study, the investigation has shortcomings, and other areas remain to be explored. First, the equations used to predict alcohol and drug use frequencies for each of the three years produced models that explained very little of the variation found in the actual frequencies, despite the fact that NLSY data are rich in attributes associated with personal characteristics of the respondents. As indicated earlier, a weak correlation between the substance use variables and their postulated determinants can give rise to serious problems. Note that the choice between use and nonuse and the frequency of such use is a personal decision that was not well modeled by the limited number of psychological and personal characteristics variables available. Second, the time series observed in the longitudinal study is too short to account for the possible negative impact of long-term substance use.

The two-period study employed in this research limited the number of observations that could be effectively used in the analysis.

One need for further research results from the fact that the fitted models are simplistic in that they do not consider all of the possibilities for interactions between substance use variables and the other independent variables. Moreover, the model assumes that, with a slight exception, the effect of being black or Hispanic is purely additive. In ongoing research we are looking at models that assume complete interaction between the black and Hispanic dummies and other variables. Second, our research has concentrated on the direct impact of substance use. Indirect impacts may also be as important, or as important as direct impacts, as found by Mullahy and Sindelar (1989). Hypotheses that substance use has effects through other variables, such as education levels, remain to be tested. Third, this paper has examined the impact of substance use on productivity as measured by the employment status. Another important dimension of labor supply is hours worked if employed. Research is currently being undertaken on the relationship between substance use and hours worked. Last, there is the question of measurement error that may be especially important for the substance use variables.

Notes

[1]There are confounding income effects, but they are not pertinent to the current discussion.

[2]Drug use measures are only available at four-year intervals. Moreover, even though drug use measures for 1980 are available, the range of drug use frequencies used allowed in these variables is not consistent with those used in later years. Hence the 1980 measures were not used.

[3]We have not included the estimated drug and alcohol use equations that were employed to predict the substance use frequencies. They are available from the authors on request.

[4]Total hours worked through 1991, and not through 1992, was used to avoid simultaneity bias. Hours worked through 1992 would include labor force experience in 1992, which is correlated with the Bernoulli variable defining employment in 1992. Use of hours worked through 1991, but not in 1992, avoid this problem.

[5]Training the classifiers with data sets having equal numbers of employed and unemployed improved the results to some extent. However, the results still showed a lower prediction accuracy for the unemployed group. These results are available from the authors on request.

References

Berger, M. C., and J. P. Leigh. 1988. "The Effect of Alcohol Use on Wages." Applied Economics, 20:143–1351.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variable Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." Journal of the American Statistical Association, 90(430): 443–450.

Bryant, R., V. Samaranayake, and A. Wilhite. 1992. "Alcohol Use and Wages of Young Men: Whites and Nonwhites." International Review of Applied Economics, 6(2): 184–202.

Bryant, R., V. Samaranayake, and A. Wilhite. 1993. "The Influence of Current and Past Alcohol Use on Earnings: Three Approaches to Estimation." Journal of Applied Behavioral Science, 29: 9–31.

Bryant, R., V. Samaranayake, and A. Wilhite. 1995. "Effect of Drug Use on Wages: A Human Capital Approach." Unpublished paper, Department of Economics, University of Missouri–Rolla.

Cogan, J. F. 1981. "Labor Supply with Costs of Labor Market Entry." In Female Labor Supply, ed. J.P. Smith. Princeton, N.J.: Princeton University Press.

Cronbach, L. J., and Furby, L. 1970. "How Should We Measure 'Change'—or Should We?" Psychological Bulletin, 74: 68–80.

Fingarette, H. 1988. Heavy Drinking: The Myth of Alcoholism as a Disease. Berkeley: University of California Press.

Gill, A. M., and Michaels, R. J. 1992. "Does Drug Use Lower Wages?" Industrial and Labor Relations Review, 45(3): 419–434.

Hsiao, Chang 1986. Analysis of Panel Data. New York: Cambridge University Press.

Kaestner, Robert. 1994a. "New Estimates of the Effect of Marijuana and Cocaine Use on Wages." Industrial and Labor Relations Review, 47(3): 454–470.

Kaestner, Robert. 1994b. “The Effect of Illicit Drug Use on the Labor Supply of Young Adults.” Journal of Human Resources, 29(1): 126–152.

Kaestner, Robert. 1991. “The Effect of Illicit Drug Use on the Wages of Young Adults.” Journal of Labor Economics, 9(4): 381–412.

McLachlan, G. J. 1992. Discriminant Analysis and Statistical Pattern Recognition. New York: John Wiley.

Mitchie, D., D. J. Spiegelhalter, and C. C. Taylor eds. 1994. Machine Learning, Neural and Statistical Classification. London: Ellis Horwood Series in Artificial Intelligence.

Moffitt, R. 1982. “The Tobit Model, Hours of Work and Institutional Constraints.” Review of Economics and Statistics, 64(August): 510–515.

Mullahy J., and J. Sindelar. 1989. “Life-Cycle Effects of Alcoholism on Education, Earnings and Occupation.” Inquiry, 26: 272–282.

Mullahy J., and J. Sindelar (1991). “Gender Differences in Labor Market Effects of Alcoholism.” American Economic Review, 81(2): 161–165.

Mullahy J., and J. Sindelar. 1993. “Alcoholism, Work and Income.” Journal of Labor Economics, 11: 494–520.

Register, Charles A., and Donald R. Williams. 1992. “Labor Market Effects of Marijuana and Cocaine Use among Young Men.” Industrial and Labor Relations Review, 45(3): 435–448.

Rosenberg, M. 1965. Society and the Adolescent Self-Image. Princeton, N.J.: Princeton University Press.

Rotter, J. B. 1966. “Generalized Expectancies for Internal versus External Control of Reinforcement.” Psychological Monographs, 80(whole No. 609).

Zabel, J. E. 1993. “The Relationship between Hours of Work and Labor Force Participation in Four Models of Labor Supply Behavior.” Journal of Labor Economics, 11(2): 387–416.