On Misclassification Errors in Self-Reported Drug Use

Ryan Bush[†], Yingyao Hu[†], and Yajing Jiang[†]
Johns Hopkins University

Abstract

The misreporting problem of drug use in self-reported surveys can severely affect the validity of estimation results in empirical work. In this paper we use an eigen-decomposition method to nonparametrically estimate the misclassification errors under various assumptions and settings. We use the longitudinal data of NYSL97 and focus on the years from 2005 to 2009, when the cohort is aged in their mid-20s. We find that the overall proportion of participants who actually use marijuana is higher than the reported proportion. Moreover, participants' inclination to misreport their drug use status is related to their current and previous actual drug use as well as their "habit" for misreporting in surveys. In general, males are more likely to underreport their drug use than females.

1 Introduction

Misclassification of drug use is pervasive in self-reported surveys and has important implications to both research and policy. It is well known that estimators ignoring measurement error in the independent variable can be biased and inconsistent, complicating how to interpret results and analysis. Many factors lead to reporting error in population surveys, and the stigma associated with drugs only adds to the difficulty of finding an accurate measure of self-reported drug use. We estimate a nonlinear model with nonclassical measurement error in covariates using data from the National Longitudinal Survey of Youth 1997 in order to find the extent to which measurement error is present in self-reported drug use. Our identification and estimation strategy allow us to consider misclassification in multiple variables, giving us a more realistic framework for analyzing the problem.

Our approach to correcting the misclassification error has important implications to many areas of research. There exists a wide body of literature in labor economics and health economics focusing on the effects of drug use as they relate to youth behavior, wage rates, labor supply, and employment. Kaestner (1994) uses the NLSY to measure the effect of drug use (both marijuana and cocaine) on labor supply. He finds that when looking at a cross section there is a significant negative effect of drug use on labor supply, but when using a longitudinal sample there is no significant effect. MacDonald and

[†]Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA. The authors can be reached at rbush11@jhu.edu, yhu@jhu.edu, and yjiang20@jhu.edu. Keywords: misclassification, measurement error, drug use. JEL classification: C40, I1

Pudney (2000) use two cross sections from the British Crime Survey data to study the relationship between drug use and labor market outcomes. Other studies by DeSimone (2002), French et al. (2001), and French et al. (1998) all rely on self-reported data when estimating their results related to drug use and employment. Some of the labor economics literature addresses issues of endogeneity of drug use, but the literature has not given the same attention to measurement error that could be significantly affecting the results.

Some studies from the health economics literature are concerned with how drug use is related to educational attainment and abuse of other substances. Bray et al. (2000) use longitudinal data on students in the US public schools to estimate the relationship between marijuana use and dropping out of school. They find that students using marijuana are roughly 2.3 times more likely to drop out than students who abstain. The authors also investigate how these odds of dropping out vary across the age of the students. A study by Beenstock and Rahav (2002) uses a self-reported measure of drug use to test the Gateway Theory of drug use, and it concludes that the use of cigarettes leads to marijuana use. Norton et al. (1998) attempt to account for endogenously selected peer groups when analyzing peer effects on substance abuse. Using longitudinal data from a drug use prevention program, they find a significant role for peer effects even after controlling for peer selection. It would be interesting to investigate how the existing results change after correcting the misreporting error in self-reported drug use.

Our identification strategy is also important to the econometrics literature. Nonlinear modeling with measurement error has been an active area of research recently (see Chen et al. (2011) for a survey). The literature focusing on classical error, where the measurement error is independent of the latent variable, has built off the work by Hausman et al. (1991). Recent studies include Schennach (2004) and Schennach (2007), which use an instrumental variable (IV) approach also featured in this paper. The work by Hui and Walter (1980), Mahajan (2006), Hu (2008), and Hu and Schennach (2008) all relax the assumption of independence between the latent true variable and measurement error. The results of Hu (2008) have recently been used by Balat (2011), Sasaki (2011), and An et al. (2010). This paper also uses the results of Hu (2008) in identifying and estimating our model with misclassification. This paper is the first to estimate such a complicated misclassification error. We allow for two latent variables when estimating the joint distribution of reported and actual drug use, and our method is well suited to be adopted by researchers who find it realistic to have both present and lagged latent variables affect misclassification.

Misreporting errors in self-reported drug use surveys have been extensively studied. Mensch and Kandel (1988) compare the self-reported drugs use in the 1984 NLSY survey to other surveys and conclude that the NLSY data was subject to underreporting. Some studies attempt to quantify the measurement error in drug use responses using advanced statistical techniques, but none use the methodology presented in this paper. Biemer and Wiesen (2002) employ latent class analysis in order to characterize the classification error in the US National Household Survey on Drug Abuse. Biemer and Witt (1996) use techniques that require repeated measures of the variable of interest, largely focusing on methods first developed by Hui and Walter (1980). Our paper adds to this literature by offering a way to measure this misclassification while flexibly allowing for multiple variables to have measurement error.

In our model, we assume a nonparametric distribution of misclassification errors which are allowed to be correlated with explanatory variables. With relatively few restrictions we are able to identify the misclassification probabilities and directly estimate latent variables whenever instrumental variables are available by using an eigen-decomposition method. We show that in general, when people indeed use drugs during the current or previous periods, the probability for him or her to misreport is significantly larger than zero. Also, this misreporting probability is contingent on the past misreporting behaviors as well as drug use history.

The rest of the paper is organized as follows. Section 2 describes our model and identification strategy that allows for two latent variable. Section 3 discusses simulation results employing our identification strategy. Section 4 describes data we use in our estimation and some assumptions we make regarding key variables. This section also presents our estimation results. Section 5 concludes.

2 Model and identification

In this section, we use the nonparametric identification method from Hu (2008) to estimate the misclassification error of reported marijuana use. Specifically, we are interested in the conditional distribution of reported marijuana use, D_t, on the unobserved latent variables indicating the true marijuana use status, D^*, namely $\Pr(D_t|D^*)$.[1] We restrict our attention to discrete measures of all variables. We present results using two approaches: the basic approach using data from four consecutive periods and assuming misclassification of only one variable, and a more general approach using data from five consecutive periods that allows for misclassification of multiple variables. Specific assumptions for each case will be illustrated and discussed in this section. Lastly, two major theorems will be presented which directly address the identification and estimation of our model.

2.1 A basic approach

In this part, we consider five types of discrete variables: the reported health condition variable at time t, H_t, the self-reported drug use at time t, D_t, the latent true level of drug use at time t, D_t^*, the self-reported drug use at time $t-2$, D_{t-2}, and lastly the two other independent variables, (D_{t-1}, H_{t-1}). They are displayed in Table 1.

Table 1: Summary of variables - basic case

Variable types	Variable names	Description
Dependent variable	H_t	self-reported health condition at time t
Proxy variable	D_t	self-reported drug use status at time t
Latent variable	D_t^*	true drug use status st time t
Instrument variable	D_{t-2}	self-reported drug use status at time $t-2$
Other independently observed variables	D_{t-1}	self-reported drug use status at time $t-1$
	H_{t-1}	self-reported health condition at time $t-1$

[1] D^* can be a scalar or a vector. In our basic model, D^* simply refers to the true drug use status D_t^*. In our general case, it is a vector indicating the true drug use status both at time t and $t-1$.

We are interested in the misreporting error distribution $\Pr(D_t|D_t^*, W_t)$ where W_t refers to all the other covariates. In order to derive the major theorem for identification and estimation, let us make the following assumptions:

Assumption 1.1

$$\Pr\left(H_t|D_t^*, D_t, D_{t-1}, H_{t-1}, D_{t-2}\right) = \Pr\left(H_t|D_t^*, D_{t-1}, H_{t-1}\right) \tag{1}$$

Assumption 1.1 states that conditional on the last-period report about drug use and health condition, D_t and D_{t-2} provide no relevant information beyond D_t^* to predict the current period health condition. We are implicitly assuming that after considering the report from the previous period, the misclassification error is completely independent of H_t. This immediately indicates that the bias of misreporting current-period marijuana use is the same for healthy and unhealthy people alike if they report the same health and drug use status in the previous period.

Next we impose the conditional independence restrictions on the misclassification error.

Assumption 1.2

$$\Pr\left(D_t|D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) = \Pr\left(D_t|D_t^*, D_{t-1}\right) \tag{2}$$

This assumption states the misclassification error is independent of D_{t-2} and H_{t-1}, conditional on the true latent drug use status, D_t^*, as well as last-period's reported drug use status, D_{t-1}. This assumption is reasonable because misreporting of marijuana use should only be dependent on whether in fact they used drugs this period and how they reported drug use last period as captured by D_{t-1}.

Note that we only assume conditional independence between the misclassification error and (H_t,D_{t-2},H_{t-1}); other than this, we do not restrict the independence between the error and any other covariates such as gender, education level, and marital status. Furthermore, these assumptions are weak in the sense that we do not impose any specific functional forms on the error term.

As mentioned earlier, we focus on discrete cases where:

$$D_\tau = \begin{cases} 1 & \text{if report using marijuana at least once in the period or "no response"} \\ 0 & \text{if report no marijuana use during the period} \end{cases}$$

and the latent variables are:

$$D_\tau^* = \begin{cases} 1 & \text{if using marijuana for at least once during the period} \\ 0 & \text{otherwise} \end{cases}$$

And for the dependent variable:

$$H_\tau = \begin{cases} 1 & \text{if in good health} \\ 0 & \text{if in bad health} \end{cases}$$

for any time period τ. Under Assumptions 1.1 and 1.2, one can show that for any function $\omega(\cdot)$

$$\begin{aligned}&\sum_{H_t}\omega(H_t)\Pr(D_t,H_t,D_{t-1},H_{t-1},D_{t-2})\\&=\sum_{D_t^*}\Pr(D_t|D_t^*,D_{t-1})\times E[\omega(H_t)|D_t^*,D_{t-1},H_{t-1}]\times\Pr(D_t^*,D_{t-1},H_{t-1},D_{t-2})\end{aligned}\tag{3}$$

These equations relate observed distributions to the underlying latent distributions, and they will be used to prove identification of our model.

2.1.1 Identification

We define for any given d_{t-1},h_{t-1},

$$\begin{aligned}&L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})}\\&=\left[\sum_{H_t}\omega(H_t)\Pr(D_t=i,H_t,d_{t-1},h_{t-1},D_{t-2}=j)\right]_{i,j\in\{1,0\}},\end{aligned}$$

$$L_{D_t|D_t^*,d_{t-1}}=\left[\Pr(D_t=i|D_t^*=j,d_{t-1})\right]_{i,j\in\{1,0\}},$$

$$D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}=\begin{bmatrix}E[\omega(H_t)|D_t^*=1,d_{t-1},h_{t-1}] & 0\\0 & E[\omega(H_t)|D_t^*=0,d_{t-1},h_{t-1}]\end{bmatrix},$$

$$L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}=\left[\Pr(D_t^*=i,d_{t-1},h_{t-1},D_{t-2}=j)\right]_{i,j\in\{1,0\}},$$

and

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}}=[\Pr(D_t=i,d_{t-1},h_{t-1},D_{t-2}=j]_{i,j\in\{1,0\}}.$$

We may then show that equation (3) is equivalent to

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})}=L_{D_t|D_t^*,d_{t-1}}\times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}\times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}\tag{4}$$

and corresponding to a degenerated $\omega(\cdot)=1$

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}}=L_{D_t|D_t^*,d_{t-1}}\times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}.\tag{5}$$

We need to make an assumption on an observable matrix in order to allow us to proceed with our eigen-decomposition technique.

Assumption 1.3 The matrix $L_{D_t,d_{t-1},h_{t-1},D_{t-2}}$ is invertible.

This assumption is testable, and our results for this test are presented later in the paper. Given our assumptions, we invert both sides of (5) and multiply these by the corresponding sides of (4) to get:

$$\begin{aligned}&L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})}\times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}}\\&=L_{D_t|D_t^*,d_{t-1}}\times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}\times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}\times\\&L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}\times L^{-1}_{D_t|D_t^*,d_{t-1}}\\&=L_{D_t|D_t^*,d_{t-1}}\times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}\times L^{-1}_{D_t|D_t^*,d_{t-1}}\end{aligned}\tag{6}$$

for any given (d_{t-1}, h_{t-1}). For the remainder of this subsection we will refer to $L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} \times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}}$ as the left-hand side (LHS) matrix. Looking at equation 6, we see that the right-hand side is in the form of an eigen-decomposition of the LHS matrix. Thus, each column of the $L_{D_t|D_t^*,D_{t-1}}$ matrix is a eigenvector of the LHS matrix. The diagonal elements of the $D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}$ are the corresponding eigenvalues. Therefore we can directly identify the distribution of misclassification errors, or $L_{D_t|D_t^*,D_{t-1}}$, from this matrix diagonalization process.

We can see that all the LHS variables are observable from the data, which means the LHS matrix can be directly calculated. In order to complete our identification, we need to place an ordering on the eigenvectors and guarantee uniqueness of the eigenvalues.

Assumption 1.4 For any given d_{t-1},

$$\Pr\left(D_t = 1 | D_t^* = 0, D_{t-1} = d_{t-1}\right) < \Pr\left(D_t = 1 | D_t^* = 1, D_{t-1} = d_{t-1}\right) \tag{7}$$

This means the element of the upper-right corner of the misclassification matrix $L_{D_t|D_t^*,d_{t-1}}$ should be smaller than that of the upper-left corner. This assumption is reasonable because the probability of those who do not use drugs but report using drugs during the period should be very small. This assumption has the flavor or truth-telling that will also be present in our general model. Hence, by checking this criteria we can determine the correct order of the eigenvectors along with corresponding eigenvalues.

Lastly, we need to impose one additional assumption on the eigenvalues:

Assumption 1.5 For any given d_{t-1} there exists an h_{t-1} and a function $\omega(\cdot)$ such that,

$$E\left[\omega\left(H_t\right) | D_t^* = i, d_{t-1}, h_{t-1}\right] \neq E\left[\omega\left(H_t\right) | D_t^* = j, d_{t-1}, h_{t-1}\right], \text{for any } i \neq j. \tag{8}$$

This assumption ensures that the two eigenvalues are not identical to each other. Without this assumption we cannot successfully identify the misclassification matrix because it could be singular, and there would be no variation in the conditional distribution of $D_t|D_t^*$. Note that here we do not require this inequality to hold for each pair of (d_{t-1}, h_{t-1}). Instead, we only need one subgroup of people sharing the same h_{t-1} for any given d_{t-1}, such that their health-related function ω is contingent on true latent drug use at t.

So far we have made a series of assumptions, and the following theorem justifies our identification and estimation:

Theorem 1.1 *Suppose Assumptions 1.1-1.5 hold, then the misclassifiction probability* $\Pr\left(D_t|D_t^*, D_{t-1}\right)$ *is nonparametrically identifiable and directly estimable.*

This basic approach uses an eigen-decomposition technique and identifies 2×2 misclassification matrices. The next section will cover a more complicated model, where we identify and estimate 4×4 misclassification matrices.

2.2 A general approach

Now we use a more general approach, where we impose assumptions that are empirically more reasonable than those in the first model. Namely, we generalize our conditional independence assumptions by also incorporating a latent variable from a previous period. We still have in total five types of discrete variables, yet in each type, we include more variables. The detailed description is illustrated in Table 2.

Here, we use conditional probability $\Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, W_t)$ to describe the misreporting behavior, where W_t refers to all the other covariates. Similar to the basic case, we make the following assumptions for identification and estimation:

Assumption 2.1

$$\Pr\left(H_t | D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) = \Pr\left(H_t | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}\right) \quad (9)$$

This assumption indicates that the proxy drug use variables, (D_{t+1}, D_t), and the instrumental variables, (D_{t-2}, D_{t-3}), do not tell us anything more useful about the person's current health condition as long as we know the actual two-period drug use status and their previous-period reported health and drug use. In other words, the misclassification error, conditional on $\left(D_t^*, D_{t-1}^*\right)$ and (D_{t-1}, H_{t-1}), is independent of H_t. Compared with our previous assumption in the basic approach, now we allow the current-period health condition to rely on both current-period true drug use and true drug use from the previous period. This assumption is more realistic in the sense that people's health condition is usually related to his or her drug use history.

Table 2: Summary of variables - general case

Variable types	Variable names	Description
Dependent variable	H_t	self-reported health condition at time t
Proxy variables	D_{t+1}	self-reported drug use status at time $t+1$
	D_t	self- reported drug use status at time t
Latent variables	D_t^*	true drug use status at time t
	D_{t-1}^*	true drug use status at time $t-1$
Instrument variables	D_{t-2}	self-reported drug use status at time $t-2$
	D_{t-3}	self-reported drug use status at time $t-3$
Other independently observed variables	D_{t-1}	self-reported drug use status at time $t-1$
	H_{t-1}	self-reported health condition at time $t-1$

We now impose restrictions on the conditional independence between the misclassification error and the instrument variable.

Assumption 2.2

$$\Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) = \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right) \quad (10)$$

Assumption 2.2 implies that the misclassification error is contingent on actual drug use over two periods. We assume people's misreporting decisions only depend on whether they used drugs and how they reported previously. This assumption is more general when compared to that in the basic model, and this can be illustrated in a simple example. Consider a person who used marijuana last year but did not report using it, which is represented by $D^*_{t-1} = 1, D_{t-1} = 0$. Given that he or she is using marijuana this year, $D^*_t = 1$, his or her probability of misreporting could be higher than those who did not use drug and did not report last year, but indeed use drug this year, $D^*_{t-1} = 0, D_{t-1} = 0, D^*_t = 1$. This can be justified by assuming the former person has a "habit" for misreporting his or her drug use, whereas the latter one is more likely to tell the truth for both periods. On the other hand, there is another driving force in the opposite direction. The former person who has been using drugs for two years might be more likely to report truthfully since he identifies himself as a "frequent" drug user; the latter person who just switched to marijuana this year might be less likely to report this drug use out of fear of getting into trouble. These two forces makes the conditional probabilities $\Pr\left(D_t|D^*_t = 1, D^*_{t-1} = 1, D_{t-1} = 0, h_{t-1}\right)$ and $\Pr\left(D_t|D^*_t = 1, D^*_{t-1} = 0, D_{t-1} = 0, h_{t-1}\right)$ hard to compare, but we have to separate them. In our previous model, however, we assume these two probabilities are the same regardless of an individual's behavior the previous year.

Furthermore, Assumption 2.2 is weaker than Assumption 1.2 since in this model we assume the misclassification error is independent of (D_{t-2}, D_{t-3}), given true latent variables $\left(D^*_t, D^*_{t-1}\right)$ and two extra observable variables. This is implied by Assumption 1.2 where we assume the error is independent of $D_{\tau-2}$ given D^*_τ and other observable variables, for all τ.

Under Assumptions 2.1 and 2.2, we may have

$$\begin{aligned}
&\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1}, D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right) \\
&= \sum_{D^*_t, D^*_{t-1}} \Pr\left(D_{t+1}, D_t | D^*_t, D^*_{t-1}, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D^*_t, D^*_{t-1}, D_{t-1}, H_{t-1}\right] \times \\
&\quad \Pr\left(D^*_t, D^*_{t-1}, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right). \qquad (11)
\end{aligned}$$

Equation (11) relates observed distributions to the underlying latent distributions, and they will be used to prove identification of our model.

2.2.1 Identification

Again, we define matrices which are analogous to the components of equations

$$\begin{aligned}
&L_{\omega(H_t)(D_{t+1}, D_t, H_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3})} \\
&= \begin{bmatrix}
g_1(1,1,1,1) & g_1(1,1,1,0) & g_1(1,1,0,1) & g_1(1,1,0,0) \\
g_1(1,0,1,1) & g_1(1,0,1,0) & g_1(1,0,0,1) & g_1(1,0,0,0) \\
g_1(0,1,1,1) & g_1(0,1,1,0) & g_1(0,1,0,1) & g_1(0,1,0,0) \\
g_1(0,0,1,1) & g_1(0,0,1,0) & g_1(0,0,0,1) & g_1(0,0,0,0)
\end{bmatrix}
\end{aligned}$$

where $g_1(i,j,r,k) = \sum_{H_t} \omega(H_t) \Pr(D_{t+1}=i, D_t=j, H_t, D_{t-1}, H_{t-1}, D_{t-2}=r, D_{t-3}=k)$, for any i,j,r,k;

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = \begin{bmatrix} g_2(1,1,1,1) & g_2(1,1,1,0) & g_2(1,1,0,1) & g_2(1,1,0,0) \\ g_2(1,0,1,1) & g_2(1,0,1,0) & g_2(1,0,0,1) & g_2(1,0,0,0) \\ g_2(0,1,1,1) & g_2(0,1,1,0) & g_2(0,1,0,1) & g_2(0,1,0,0) \\ g_2(0,0,1,1) & g_2(0,0,1,0) & g_2(0,0,0,1) & g_2(0,0,0,0) \end{bmatrix}$$

where $g_2(i,j,r,k) = \Pr(D_{t+1}=i, D_t=j, d_{t-1}, h_{t-1}, D_{t-2}=r, D_{t-3}=k)$,for any i,j,r,k;

$$L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} = \begin{bmatrix} g_3(1,1|1,1) & g_3(1,1|1,0) & g_3(1,1|0,1) & g_3(1,1|0,0) \\ g_3(1,0|1,1) & g_3(1,0|1,0) & g_3(1,0|0,1) & g_3(1,0|0,0) \\ g_3(0,1|1,1) & g_3(0,1|1,0) & g_3(0,1|0,1) & g_3(0,1|0,0) \\ g_3(0,0|1,1) & g_3(0,0|1,0) & g_3(0,0|0,1) & g_3(0,0|0,0) \end{bmatrix}$$

where $g_3(i,j|r,k) = \Pr(D_{t+1}=i, D_t=j|D_t^*=r, D_{t-1}^*=k, d_{t-1})$, for any i,j,r,k;

$$D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} = \begin{bmatrix} E[1,1] & & & \\ & E[1,0] & & \\ & & E[0,1] & \\ & & & E[0,0] \end{bmatrix}$$

where, $E[i,j] = E_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}[i,j] = E\left[\omega(H_t)|D_t^*=i, D_{t-1}^*=j, d_{t-1}, h_{t-1}\right]$, for any i,j;

$$L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = \begin{bmatrix} g_4(1,1,1,1) & g_4(1,1,1,0) & g_4(1,1,0,1) & g_4(1,1,0,0) \\ g_4(1,0,1,1) & g_4(1,0,1,0) & g_4(1,0,0,1) & g_4(1,0,0,0) \\ g_4(0,1,1,1) & g_4(0,1,1,0) & g_4(0,1,0,1) & g_4(0,1,0,0) \\ g_4(0,0,1,1) & g_4(0,0,1,0) & g_4(0,0,0,1) & g_4(0,0,0,0) \end{bmatrix}$$

where $g_4(i,j,r,k) = \Pr(D_t^*=i, D_{t-1}^*=j, d_{t-1}, h_{t-1}, D_{t-2}=r, D_{t-3}=k)$, for any i,j,r,k. Thus, the matrix notation would be written as:

$$\begin{aligned} L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} =& L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \\ & \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \end{aligned} \tag{12}$$

and corresponding to a degenerated $\omega(\cdot) = 1$

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \tag{13}$$

We need to make an assumption on an observable matrix in order to allow us to proceed with an eigen-decomposition technique.

Assumption 2.3 The matrix $L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$ is invertible.

As was the case in the basic model, this assumption is testable and our results for this test are presented later in the paper. Given these assumptions, we invert both sides of (13) and multiply these by the corresponding sides of (12) to obtain:

$$\begin{aligned}
&L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \\
=&L_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}} \times D_{\omega(H_t)|D^*_t,D^*_{t-1},d_{t-1},h_{t-1}} \times L_{D^*_t,D^*_{t-1},d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times \\
&L^{-1}_{D^*_t,D^*_{t-1},d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L^{-1}_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}} \\
=&L_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}} \times D_{\omega(H_t)|D^*_t,D^*_{t-1},d_{t-1},h_{t-1}} \times L^{-1}_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}} \qquad (14)
\end{aligned}$$

As was done with the basic model, we will refer to $L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$ as the LHS matrix. Thus, the $L_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}}$ matrix is the eigenvector matrix of the LHS matrix, and the diagonal elements of the $D_{\omega(H_t)|D^*_t,D^*_{t-1},d_{t-1},h_{t-1}}$ are the corresponding eigenvalues. We can directly identify the distribution of misclassification errors, or $L_{D_{t+1},D_t|D^*_t,D^*_{t-1},d_{t-1}}$, from this matrix diagonalization process.

As discussed in previous section, after recovering the eigenvectors and eigenvalues, we need to determine the correct ordering that is consistent with the LHS matrix. Therefore we make the following assumption:

Assumption 2.4 For $D_{t-1} = 1$,

$$\begin{aligned}
Pr(D_t = 1|D^*_t = 1, D^*_{t-1} = 1, D_{t-1} = 1) &> Pr(D_t = 1|D^*_t = 1, D^*_{t-1} = 0, D_{t-1} = 1) \\
&> Pr(D_t = 1|d^*_t, d^*_{t-1}, D_{t-1} = 1) \text{ for other } (d^*_t, d^*_{t-1}),
\end{aligned}$$

and

$$E\left[H_t|D^*_t = 0, D^*_{t-1} = 1, D_{t-1} = 1\right] > E\left[H_t|D^*_t = 0, D^*_{t-1} = 0, D_{t-1} = 1\right].$$

For $D_{t-1} = 0$,

$$\begin{aligned}
Pr(D_t = 0|D^*_t = 0, D^*_{t-1} = 0, D_{t-1} = 0) &> Pr(D_t = 0|D^*_t = 0, D^*_{t-1} = 1, D_{t-1} = 0) \\
&> Pr(D_t = 0|d^*_t, d^*_{t-1}, D_{t-1} = 0) \text{ for other } (d^*_t, d^*_{t-1}),
\end{aligned}$$

and

$$E\left[H_t|D^*_t = 1, D^*_{t-1} = 1, D_{t-1} = 0\right] > E\left[H_t|D^*_t = 1, D^*_{t-1} = 0, D_{t-1} = 0\right].$$

This assumption directly gives us the ordering of the eigenvalues and eigenvectors needed to establish identification. For $D_{t-1} = 1$ case, we assume people primarily consider their current-period true drug-use status as a reference of how they report it. Specifically, we assume that people who are using marijuana in current period would be more likely to report using it than would people who are not currently using marijuana. Another factor that influences people's current reporting behavior is their last-period misreporting behavior. We assume people who have been using marijuana for both periods and reported using it last period have the highest probability of reporting this period again; and people who use marijuana this period but not last period but report using it last

period have the second highest probability of reporting drug use this period. Thus, we are able to identify two columns of our misclassification matrix, and for the other two we use eigenvalues to distinguish them, as is illustrated in the second equation for $D_{t-1} = 1$ case. We assume that the expected reported health condition is better for those who do not use marijuana either period, compared with those who use marijuana last period but not this period. Similarly, for $D_{t-1} = 0$ case, we assume that people who do not use marijuana in either period have the highest probability of not reporting marijuana use this period. People who do not use marijuana this period but used it last period would have the second highest probability of not reporting it now. For the other two cases, we assume the expected health condition for those who use marijuana both periods is worse than those who only used it in the current period. These assumptions have the flavor of "truth telling" as a eigenvalue and eigenvector ordering mechanism.

Lastly, to make the identification effective, we need to impose restrictions on the eigenvalues:

Assumption 2.5 For all values of D_{t-1} there existing some h_{t-1} such that

$$\begin{aligned} E\left[\omega\left(H_t\right)|D_t^*, D_{t-1}^*, d_{t-1}, h_{t-1}\right] &\neq E\left[\omega\left(H_t\right)|D_t^*, d_{t-1}, h_{t-1}\right] \\ &\neq E\left[\omega\left(H_t\right)|D_{t-1}^*, d_{t-1}, h_{t-1}\right]. \end{aligned} \tag{15}$$

It states that for those who reported using drugs last period, there exists at least one subgroup of people who share (d_{t-1}, h_{t-1}) for whom their current health condition depends both on current drug use status and previous drug use. Otherwise, we cannot effectively tell apart the difference of the misclassification errors between $\left(D_t^* = i, D_{t-1}^* = j\right)$ type people and $\left(D_t^* = i, D_{t-1}^* = k\right)$ type people when $j \neq k$. In that case we would have duplicates in the eigenvalues, and the identification fails. The following theorem justifies our identification and estimation:

Theorem 2.1 *Suppose Assumptions 2.1-2.5 hold, then the conditional probability* $\Pr\left(D_{t+1}, D_t|D_t^*, D_{t-1}^*, D_{t-1}\right)$ *is nonparametrically identifiable and directly estimable.*

In the next section we run simulation for both the 2×2 and the 4×4 cases to justify the validity of our models, and then we use NYSL97 data to estimate the misclassification errors in marijuana use.

3 Simulation

3.1 The basic approach

In this section we generate a set of data using underlying parametric values, and then we use our nonparametric identification method to estimate the misclassification matrices. Information regarding the data generation process can be found in the appendix. We compare these estimates with the true underlying matrices in order to validate our method. We first generate all the data that are related to our estimation, namely $(D_t^*, H_t, H_{t-1}, D_t, D_{t-1}, D_{t-2})$. Using both observable data, $(H_t, H_{t-1}, D_t, D_{t-1}, D_{t-2})$, and unobservable latent variable D_t^*, we are able to calculate the sample average of the

misclassification matrices of interest, $\Pr(D_t|D_t^*, D_{t-1}, H_{t-1})$. Lastly, we use only observable data to estimate $\widehat{\Pr}(D_t|D_t^*, D_{t-1}, H_{t-1})$ according to Theorem 1.1. If the difference of these two matrices converges as the sample size increases, then our basic model is valid.

3.1.1 Simulation results

The mean, median, and standard errors for simulation results are displayed in Table 7. There are several things to point out here. Firstly, the means and medians are converging to the true value as the sample size increases. Secondly, the standard errors are decreasing as the sample size increases. This convergence indicates that our method of identification is at least correct asymptotically. In fact when the sample size is 7000, which is very close to that of our real data, the simulation results are already significant. This further validates our estimation using the basic model.

The results indicate that some estimated probabilities are much more accurate than others. This is due to the loss in accuracy associated with inverting the LHS matrix if it is near singular. This loss in accuracy can be remedied, however, by increasing sample size and therefore canceling out inaccurate estimates from a single trial.

3.2 The more general approach

In the general model, we need more data regarding previous drug use and health conditions. We require data on $(H_t, H_{t-1}, D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, D_{t-2}, D_{t-3})$. We use the data generation process in the next subsection to generate all these data, then we use all information, observable and unobservable, to calculate the sample mean of the true values of the misclassification matrices. Lastly, as before, we only use observable variables to estimate this matrix according to Theorem 2.1 and compare this estimation result with the underlying true values.

This simulation process, compared with that in the basic model, is more complex and needs careful consideration. Instead of estimating two 2×2 matrices, we are now estimating two 4×4 matrices. As was the case in the basic model, any inaccuracy can be resolved as we increase the sample size. These features will be discussed in the simulation results section.

3.2.1 Simulation results

The mean, median, and standard errors for simulation results are displayed in Table 11. We can see that when we have a sample size of 7000 and 10000, the estimation results are farther from the true values when compared to the basic model. The standard errors are larger than those found in the basic model as well. However, as the sample size increases to 100000, the results are statistically significant and accurate. This shows that our general model is also correct at least asymptotically, or when we have relatively large sample size.

4 Estimation

4.1 Data

The data we use in our analysis come from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 is a panel survey conducted by the Bureau of Labor Statistics that captures the transition from youth to adulthood. The data consist of detailed information on a cohort of approximately 9000 youths who were between the ages of 12 and 16 when the survey was first administered in 1997. These individuals were asked questions covering areas such as background, employment, drug and alcohol use, health, family, and education. Previous literature that has investigated self-reported drug use sometimes focused on the National Household Survey on Drug Abuse data, but we take advantage of the panel nature of the NLSY97 to obtain our estimates of misclassification.

The variables central to our study include marijuana use and health condition. Additionally, we can account for gender, age, marital status, ethnicity, and education level, but these are not crucial to our estimation and identification of the model. We select 2008 as the basis year for our analysis, because this survey year places the respondents between the ages of 23 and 27. Since our particular focus is on drug use, we want the respondents to be at an age where peer effects were considerably weaker than they would be in high school and early college years. Additionally, we want the respondents to be young enough that they were likely to be making important labor market decisions. We make use of 5 years of data in our estimation, and 2008 also happens to be the latest year we can choose as our basis year because data are only available through 2009 at this time.

Health is measured by the respondents on a 1-5 scale, representing "excellent", "very good", "good", "fair", and "poor", respectively. We group the respondents who answered "very good" and "good" into the same category, which we call responding 2.5, because these groups are very similar and the increase in sample size allows for more accurate estimates. We would use a more objective measure of health if one was available in the NLSY97, but since the responses are all self-reported it is unclear how to create a more accurate measure of health using responses from other questions (such as height and weight). For marijuana use, we use a variable in the data that indicates whether the respondents claim to have used marijuana at all since the date of the previous interview. To make this variable binary, we group those who admitted to using marijuana with those who did not respond to the question or said they could not remember. Participants who said they did not use marijuana in the last year make up the other group. For background information, we use the responses available from the initial 1997 survey and the 2008 survey when appropriate.

In order to run our estimation we take some steps in dropping observations. We disregard all observations (individuals) who were not interviewed for any of the 5 years our variables span. Respondents were not interviewed because they could not be tracked down or were otherwise unavailable, and since we require data on all 5 years we need to drop these people. Additionally, we drop respondents who did not respond to the health condition question. The percentage of participants who did not respond to this question in 2008 and 2007 was less than 1%. We do not believe that dropping these observations bias our results. Our sample thus reduces to 6298 individuals. Table 12 displays summary statistics for our sample. When conditioning on covariates, we drop

observations that did not have a response for any of the variables on which we were conditioning. While an omission of a response about drug use may be informative about whether that person has actually used drugs, we do not believe the same is true for omissions on questions regarding background. Assuming that agreeing to be interviewed but not answering questions about background information is independent of true drug use, the results presented here are not biased. Tables 13 and 14 provide further details about our sample including additional covariates. In those tables, all of the variables are binary except for health (H) which takes on the values discussed above. For the remaining variables, $D = 1$ if the participant used marijuana in the previous year, Education = 1 if awarded any degree beyond a high school diploma by 2008, Marital Status = 1 if married, Ethnicity = 1 if white, and Gender = 1 if male.

4.2 The basic approach

We use 2008 as the basis year to do the estimation. In total there are 6298 effective observations. As is depicted in Assumption 1.5, when we do estimation for each subgroup of d_{t-1}, we would like to find the best subsample of h_{t-1} such that the eigenvalues are most distinct from each other and therefore the estimation is most valid. After trying different strategies, we decide to combine the samples where H_τ equals 2 or 3.[2] They account for nearly 70% of the total population, and this increase in the sample size for a particular subgroup facilitates more accurate estimation. Thus, the dependent variable takes the following possible values:

$$H_\tau = \begin{cases} 1 & \text{if in "excellent health condition"} \\ 2.5 & \text{if in "very good or good health condition"} \\ 4 & \text{if in "fair health condition"} \\ 5 & \text{if in "poor health condition"} \end{cases},$$

In addition, we assume,

$$E\left[H_t | D_t^* = i, d_{t-1}, H_{t-1} = 2.5\right] \neq E\left[H_t | d_{t-1}, H_{t-1} = 2.5\right], \text{for any } i.$$

which is a specification based on Assumption 1.5.

The main estimation results are shown in Table 15, while Table 16 and Table 17 are the estimation results conditional on gender covariate. There are several interesting results from this table. Firstly, for any given D_{t-1}, the probability of reporting drug use for actual drug users is higher than those non-drug users. For the first two rows of each table, by assumption the truth-telling dominant rule is true; but for the third and fourth row, it still holds. For those who did not report marijuana use last year and did not use marijuana this year, the estimated probability of reporting drug use is zero. This is consistent with any intuition that leads one to believe the misreporting problem is mostly related to individuals trying to hide true drug use. The misreporting problem is most prevalent with people who did not report drug use last year and did use drug this year ($D_t^* = 1, D_{t-1} = 0$). The estimated probability of telling the truth is only 30.89%, much lower than the 82.23% where participants reported using marijuana last

[2]For robustness check, we also provide estimation results by combining $H_\tau = 1, 2$ and 3 in the appendix.

year. Our estimation for the true marijuana use proportion in the cohort is 32.13%, significantly higher than the reported proportion of 17.61%. These results show the significant role of misclassification error in the data.When we condition on gender and re-estimate the misclassification errors, the only major difference between females and males arises when the individuals do not report using marijuana last year but use it this year(i.e. $D_t^* = 1, D_{t-1} = 0$). In that situation, males are more likely than females to hide the truth and misreport this year.

One confusing result in Table 15 is the $\Pr(D_t = 1|D_t^* = 0, D_{t-1} = 1) = 0.5317$. This could be partially explained by people having a proclivity for reporting a particular type of behavior even when they do not necessarily undertake that behavior anymore. Intuitively this probability should be fairly small, though. Therefore, we use a different way of modifying the data and conduct the estimation process again.

4.3 The more general approach

In this section, we adopt the general model and estimate the misclassification matrices according to Theorem 2.2. Similar as in the basic case, we combine $H_\tau = 2$ and 3.[3] The results are presented in Tables 18, and results when conditioning on gender covariates are presented in Table 19 and Table 20.

Firstly, the general results are mostly consistent with those from the basic model. For instance, the people who have been using marijuana for both periods and reported using it last period have a probability of reporting usage as high as 75.02%. By contrast, people who have not used marijuana for either period and did not report usage last period have a very low probability (1.78%) of reporting usage this period. Secondly, conditional on true drug use in the previous period, people who reported using marijuana last period are more likely to report using it again this period. These results seem to display certain habits in behavior.

Another exciting result lies in the fifth row of Table 18, where people used drug last year but misreported this drug use. When continuing to use drugs this year, the probability for them to report the truth is less than 40%, much less than those who already told the truth last period (75.02%). This helps show the severity of misreporting problem in the survey.

Recall the example we used to motivate our more general model where we stressed that two individuals who were likely to have different reporting behavior would be captured by the same probability when only one latent variable was present. The comparison of the seventh and eighth row shows us the importance of incorporating last-period marijuana usage status. The rows tell us that, when people used marijuana last year but hid the truth, he or she is still much more likely to report using marijuana(24.72%) compared with those who do not use drugs for two years and do not report using it last year(1.78%).

The estimates for expected (reported) health conditions are more dispersed than in the basic case, as is shown in the ninth through sixteenth row of the tables. Also the

[3]In the appendix we also re-estimate the measurement error matrices by combining $H_\tau = 1, 2$ and 3, as in the basic case.

standard errors are much higher compared with those in the basic case. At a price of losing the estimation accuracy of the eigenvalues, we are able to gain more accuracy of estimated eigenvectors which represent the misclassification errors of our primary interest. Here the estimated eigenvalues only serve a role to distinguish different columns of eigenvectors, therefore it is acceptable in this application to have some eigenvalues that are out of proportion.

When we condition on gender covariates, the distinction between males' and females' misreporting behavior become clearer than in the basic case. Firstly consistent with the basic case, when males reported no marijuana use last period, they are more likely to follow their "habit" of saying "no" to the same question this period even though they in fact are using marijuana, as is shown by the fifth and sixth rows of Table 19 and Table 20. Moreover, when males used marijuana last year and told the truth, they are more likely to report "yes" to the same question this year, even though they actually are not using marijuana this year, than are females(62.34% versus 39.18%). In a word, males appear to be more stuck with their last-period reporting behavior than are females, at least in some cases.

One thing to note is the probability of $\Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$ for males (10.11%). It is not very different from $\Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0) =$ 11.16% for males, which indicates that males who are non-drug users this year and did not report using it last year has a probability as high as 10% of reporting usage this year. Somewhat counterintuitive as it seems, when we bootstrap for the mean and median, we found that they are much more distinctive from each other. The median for $\Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$ reduces to 6.21%, which looks more reasonable. We suspect that the high point estimation is resulted from some deeper unobserved patterns within the male subgroup, which calls for further study into the misreporting problem, conditional on more covariates.

One last observation is the marginal probability of true marijuana usage. It can be seen from the last rows of Table 18, Table 19 and Table 20. Males have higher probability of using marijuana than do females(17.02% versus 10.43%); but all of the marginal probabilities are not quite different from the probability of reporting marijuana use. Again, if we look at the bootstrap means and medians we see results that are higher than our point estimates. This tells us that people are not always underreporting their marijuana usage, instead, there exists noisy reporting phenomenon in both directions.

5 Conclusion

In this paper we use an eigen-decomposition method from Hu (2008) to estimate the misclassification errors in self-reported drug use. Using reasonable and relatively weak assumptions in our models, we find that the self-reported responses from the NLSY97 sometimes underreport the true level of marijuana use. Both our basic model and general model give us insights into severity and pattern of this misclassification problem. We conclude that if this misclassification problem is simply ignored in empirical work, results could be severely biased and difficult to interpret. In future work, if we can obtain more complete data detailing the different levels of marijuana use, we can attempt a continuous

version of this nonparametric identification which could give us more insight in this topic.

References

An, Y., Y. Hu, and M. Shum (2010). Estimating first-price auctions with an unknown number of bidders: A misclassification approach. *Journal of Econometrics 157*(2), 328–341.

Balat, J. (2011). Highway procurement and the stimulus package: Identification and estimation of dynamic auctions with unobserved heterogeneity.

Beenstock, M. and G. Rahav (2002). Testing gateway theory: do cigarette prices affect illicit drug use? *Journal of Health Economics 21*(4), 679–698.

Biemer, P. and C. Wiesen (2002). Measurement error evaluation of self-reported drug use: a latent class analysis of the us national household survey on drug abuse. *Journal of the Royal Statistical Society: Series A 165*(1), 97–119.

Biemer, P. and M. Witt (1996). Estimation of measurement bias in self-reports of drug use with applications to the national household survey on drug abuse. *Journal of Official Statistics 12*(3), 275–300.

Bray, J., G. Zarkin, C. Ringwalt, and J. Qi (2000). The relationship between marijuana initiation and dropping out of high school. *Health Economics 9*(1), 9–18.

Chen, X., H. Hong, and D. Nekipelov (2011). Nonlinear models of measurement errors. *Journal of Economic Literature 49*(4), 901–37.

DeSimone, J. (2002). Illegal drug use and employment. *Journal of Labor Economics 20*(4), 952–977.

Duncan, G., B. Wilkerson, and P. England (2006). Cleaning up their act: the effects of marriage and cohabitation on licit and illicit drug use. *Demography 43*(4), 691–710.

French, M., M. Roebuck, and P. Alexandre (2001). Illicit drug use, employment, and labor force participation. *Southern Economic Journal 68*(2), 349–368.

French, M., G. Zarkin, T. Mroz, and J. Bray (1998). The relationship between drug use and labor supply for young men. *Labour Economics 5*(4), 385–409.

Harrison, L. and A. Hughes (1997). Validity of self-reported drug use: Improving the accuracy of survey estimates. *NIDA research monograph 167*, 1.

Hausman, J., W. Newey, H. Ichimura, and J. Powell (1991). Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics 50*(3), 273–295.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

Hu, Y. and S. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Hui, S. and S. Walter (1980). Estimating the error rates of diagnostic tests. *Biometrics 36*(1), 167–171.

Kaestner, R. (1994). The effect of illicit drug use on the labor supply of young adults. *Journal of Human Resources 29*(1), 126–155.

Kandel, D. and J. Logan (1984). Patterns of drug use from adolescence to young adulthood: I. periods of risk for initiation, continued use, and discontinuation. *American Journal of Public Health 74*(7), 660–666.

MacDonald, Z. and S. Pudney (2000). Illicit drug use, unemployment, and occupational attainment. *Journal of Health Economics 19*(6), 1089–1115.

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica 74*(3), 631–665.

Mensch, B. and D. Kandel (1988). Underreporting of substance use in a national longitudinal youth cohort. *Public Opinion Quarterly 52*(1), 100–124.

Morral, A., D. McCaffrey, and S. Chien (2003). Measurement of adolescent drug use. *Journal of Psychoactive Drugs 35*(3), 301–309.

Norton, E., R. Lindrooth, and S. Ennett (1998). Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use. *Health Economics 7*(5), 439–453.

Sasaki, Y. (2011). Heterogeneity and selection in dynamic panel data.

Schennach, S. (2004). Estimation of nonlinear models with measurement error. *Econometrica 72*(1), 33–75.

Schennach, S. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica 75*(1), 201–239.

Schennach, S., Y. Hu, and A. Lewbel (2007). Nonparametric identification of the classical errors-in-variables model without side information. *Boston College Working Papers in Economics*.

Weatherby, N., R. Needle, H. Cesari, R. Booth, C. McCoy, J. Watters, M. Williams, and D. Chitwood (1994). Validity of self-reported drug use among injection drug users and crack cocaine users recruited through street outreach. *Evaluation and Program Planning 17*(4), 347–355.

6 Appendix

6.1 Proofs of theorems

In the appendix we give brief proofs of our two theorems in the model section.

6.1.1 Proof of Theorem 1.1

The first main equation we have is,

$$\begin{aligned}
&\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}\right) \\
&= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right] \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \\
&= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}, H_{t-1}\right] \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)
\end{aligned}$$

according to Assumption 1.1 and 1.2. Similarly,

$$\begin{aligned}
&\Pr(D_t, D_{t-1}, H_{t-1}, D_{t-2}) \\
&= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right) \\
&= \sum_{D_t^*} \Pr\left(D_t | D_t^*, D_{t-1}\right) \times \Pr\left(D_t^*, D_{t-1}, H_{t-1}, D_{t-2}\right)
\end{aligned}$$

Then we put the probabilities on both sides into matrix forms. For the first equation, the left-hand side probabilities could be written as,

$$\begin{aligned}
& L_{\omega(H_t)(D_t, H_t, d_{t-1}, h_{t-1}, D_{t-2})} \\
= & \left[\begin{array}{cc} g_5(1,1) & g_5(1,0) \\ g_5(0,1) & g_5(0,0) \end{array}\right]
\end{aligned}$$

where $g_5(i,,j) = \sum_{H_t} \omega\left(H_t\right) \Pr\left(D_t = i, H_t, d_{t-1}, h_{t-1}, D_{t-2} = j\right)$.

And for the right-hand side probabilities,

$$\begin{aligned}
& L_{D_t^*, d_{t-1}, h_{t-1}, D_{t-2}} \\
= & \left[\begin{array}{cc} g_6(1,1) & g_6(1,0) \\ g_6(0,1) & g_6(0,0) \end{array}\right]
\end{aligned}$$

where, $g_6(i,j) = \Pr(D_t^* = i, d_{t-1}, h_{t-1}, D_{t-2} = j)$.

$$\begin{aligned}
& L_{D_t, d_{t-1}, h_{t-1}, D_{t-2}} \\
= & \left[\begin{array}{cc} g_7(1,1) & g_7(1,0) \\ g_7(0,1) & g_7(0,0) \end{array}\right]
\end{aligned}$$

where, $g_7(i,j) = \Pr(D_t = i, d_{t-1}, h_{t-1}, D_{t-2} = j)$.

And,

$$D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} = \begin{bmatrix} E[1] & \\ & E[0] \end{bmatrix}$$

where, $E[i] = E_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}}[i] = E[\omega(H_t)|D_t^* = i, d_{t-1}, h_{t-1}]$,

Thus, equation (3) is equivalent to

$$L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} = L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}$$

And for the second equation,

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}} = \begin{bmatrix} g_8(1,1) & g_8(1,0) \\ g_8(0,1) & g_8(0,0) \end{bmatrix}$$

where,$g_8(i,j) = \Pr(D_t = i|D_t^* = j, d_{t-1})$.

Thus, we are able to rewrite the second main equations as,

$$L_{D_t,d_{t-1},h_{t-1},D_{t-2}} = L_{D_t|D_t^*,d_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}}.$$

Given Assumption 1.3 which ensures invertibility of the left-hand side matrix in equation (5), we take the inverse of both sides in this equation and the equation above therefore becomes,

$$L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}} = L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times L^{-1}_{D_t|D_t^*,d_{t-1}}.$$

Finally, we right-multiply each side of the equation with the corresponding side in equation(4), to get equation (6).

$$\begin{aligned} &L_{\omega(H_t)(D_t,H_t,d_{t-1},h_{t-1},D_{t-2})} \times L^{-1}_{D_t,d_{t-1},h_{t-1},D_{t-2}} \\ &= L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times \\ &L^{-1}_{D_t^*,d_{t-1},h_{t-1},D_{t-2}} \times L^{-1}_{D_t|D_t^*,d_{t-1}} \\ &= L_{D_t|D_t^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_t|D_t^*,d_{t-1}} \end{aligned}$$

In the last line of the equation, the RHS matrcies include the misclassification error which is of our central interest, and they could be recovered via a eigenvector-eigenvalue decomposition of the LHS matrix. But as we know from basic matrix arithmatics, the exact position of each eigenvector in the matrix is not determinant. In order to reconcile this problem, we introduce Assumption 1.4 and 1.5 which ensure the most reasonable ordering of eigenvectors from an economic point of view. Once the order of the columns of eigenvectors is determined, the misclassification error is uniquely identified and could be estimated. Therefore, Theorem 1.1 is proved.

6.1.2 Proof of Theorem 2.1

Following the similar logic as in the previous proof, we firstly write down the two main equations for the general approach,

$$\begin{aligned}
&\sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1}, D_t, H_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right) \\
&= \sum_{D_t^*, D_{t-1}^*} \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, D_{t-2}, D_{t-3}, H_{t-1}\right) \times \\
&\quad E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}^*, D_{t+1}, D_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right] \times \\
&\quad \Pr\left(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right) \\
&= \sum_{D_t^*, D_{t-1}^*} \Pr\left(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}\right) \times E\left[\omega\left(H_t\right) | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}\right] \times \\
&\quad \Pr\left(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}\right).
\end{aligned}$$

and the second equality follows from Assumption 2.1 and 2.2. Similarly, we write down the second main equation as,

$$\begin{aligned}
&\Pr(D_{t+1}, D_t, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}) \\
&= \sum_{D_t^*, D_{t-1}^*} \Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}, H_{t-1}, D_{t-2}, D_{t-3}) \times \Pr(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-2}, D_{t-2}, D_{t-3}) \\
&= \sum_{D_t^*, D_{t-1}^*} \Pr(D_{t+1}, D_t | D_t^*, D_{t-1}^*, D_{t-1}) \times \Pr(D_t^*, D_{t-1}^*, D_{t-1}, H_{t-2}, D_{t-2}, D_{t-3})
\end{aligned}$$

Now we put all the probabilities in the equations into matrix forms. For the first one,

$$\begin{aligned}
&L_{\omega(H_t)(D_{t+1}, D_t, H_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3})} \\
&= \begin{bmatrix}
g_1(1,1,1,1) & g_1(1,1,1,0) & g_1(1,1,0,1) & g_1(1,1,0,0) \\
g_1(1,0,1,1) & g_1(1,0,1,0) & g_1(1,0,0,1) & g_1(1,0,0,0) \\
g_1(0,1,1,1) & g_1(0,1,1,0) & g_1(0,1,0,1) & g_1(0,1,0,0) \\
g_1(0,0,1,1) & g_1(0,0,1,0) & g_1(0,0,0,1) & g_1(0,0,0,0)
\end{bmatrix}
\end{aligned}$$

where $g_1(i,j,r,k) = \sum_{H_t} \omega\left(H_t\right) \Pr\left(D_{t+1} = i, D_t = j, H_t, D_{t-1}, H_{t-1}, D_{t-2} = r, D_{t-3} = k\right)$, for any i, j, r, k;

$$\begin{aligned}
&L_{D_{t+1}, D_t, d_{t-1}, h_{t-1}, D_{t-2}, D_{t-3}} \\
&= \begin{bmatrix}
g_2(1,1,1,1) & g_2(1,1,1,0) & g_2(1,1,0,1) & g_2(1,1,0,0) \\
g_2(1,0,1,1) & g_2(1,0,1,0) & g_2(1,0,0,1) & g_2(1,0,0,0) \\
g_2(0,1,1,1) & g_2(0,1,1,0) & g_2(0,1,0,1) & g_2(0,1,0,0) \\
g_2(0,0,1,1) & g_2(0,0,1,0) & g_2(0,0,0,1) & g_2(0,0,0,0)
\end{bmatrix}
\end{aligned}$$

where $g_2(i,j,r,k) = \Pr(D_{t+1} = i, D_t = j, d_{t-1}, h_{t-1}, D_{t-2} = r, D_{t-3} = k)$,for any i, j, r, k;

$$\begin{aligned}
&L_{D_{t+1}, D_t | D_t^*, D_{t-1}^*, d_{t-1}} \\
&= \begin{bmatrix}
g_3(1,1|1,1) & g_3(1,1|1,0) & g_3(1,1|0,1) & g_3(1,1|0,0) \\
g_3(1,0|1,1) & g_3(1,0|1,0) & g_3(1,0|0,1) & g_3(1,0|0,0) \\
g_3(0,1|1,1) & g_3(0,1|1,0) & g_3(0,1|0,1) & g_3(0,1|0,0) \\
g_3(0,0|1,1) & g_3(0,0|1,0) & g_3(0,0|0,1) & g_3(0,0|0,0)
\end{bmatrix}
\end{aligned}$$

where $g_3(i,j|r,k) = \Pr(D_{t+1}=i, D_t=j|D_t^*=r, D_{t-1}^*=k, d_{t-1})$, for any i,j,r,k;

$$D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} = \begin{bmatrix} E[1,1] & & & \\ & E[1,0] & & \\ & & E[0,1] & \\ & & & E[0,0] \end{bmatrix}$$

where, $E[i,j] = E_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}}[i,j] = E\left[\omega(H_t)|D_t^*=i, D_{t-1}^*=j, d_{t-1}, h_{t-1}\right]$, for any i,j;

$$L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = \begin{bmatrix} g_4(1,1,1,1) & g_4(1,1,1,0) & g_4(1,1,0,1) & g_4(1,1,0,0) \\ g_4(1,0,1,1) & g_4(1,0,1,0) & g_4(1,0,0,1) & g_4(1,0,0,0) \\ g_4(0,1,1,1) & g_4(0,1,1,0) & g_4(0,1,0,1) & g_4(0,1,0,0) \\ g_4(0,0,1,1) & g_4(0,0,1,0) & g_4(0,0,0,1) & g_4(0,0,0,0) \end{bmatrix}$$

where $g_4(i,j,r,k) = \Pr(D_t^*=i, D_{t-1}^*=j, d_{t-1}, h_{t-1}, D_{t-2}=r, D_{t-3}=k)$, for any i,j,r,k. Thus, the matrix notation would be written as:

$$\begin{aligned} L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} =& L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \\ & \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \end{aligned}$$

and corresponding to a degenerated $\omega(\cdot) = 1$

$$L_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}}$$

Given Assumption 2.3 which ensures invertibility of the left-hand side matrix in equation (13), we take the inverse of both sides and obtain the new equation,

$$L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} = L^{-1}_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}}$$

and then we right-multiply each side with the corresponding side in equation(12). Finally, we get equation (14).

$$\begin{aligned} & L_{\omega(H_t)(D_{t+1},D_t,H_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3})} \times L^{-1}_{D_{t+1},D_t,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \\ =& L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times \\ & L^{-1}_{D_t^*,D_{t-1}^*,d_{t-1},h_{t-1},D_{t-2},D_{t-3}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \\ =& L_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \times D_{\omega(H_t)|D_t^*,D_{t-1}^*,d_{t-1},h_{t-1}} \times L^{-1}_{D_{t+1},D_t|D_t^*,D_{t-1}^*,d_{t-1}} \end{aligned}$$

Similarly as in the basic case, this equation above could help us to recover the eigenvectors that represents the misclassification errors of our central interest on the right-hand side. in addition, Assumption 2.4 and 2.5 enable us to determine the correct ordering of each eigenvector, therefore completes the proof for Theorem 2.1.

6.2 Hypothesis Testings of the Conditional Independence Assumptions

In this section we would like to test the validity of the conditional independence assumptions we have made in deriving our identification models. Specifically we test the following hypotheses:

6.2.1 The basic approach

Firstly we would like to test whether D_{t-1} always plays a role in determining the distribution of D_t given D_t^*. Namely, the null hypothesis is:

$$H_0 : \Pr(D_t|D_t^*, D_{t-1}) = \Pr(D_t|D_t^*)$$

versus the alternative hypothesis,

$$H_1 : \Pr(D_t|D_t^*, D_{t-1}) \neq \Pr(D_t|D_t^*)$$

Here we define

$$\Lambda := \begin{bmatrix} \Pr(D_t = 1|D_t^* = 1, D_{t-1} = 1) - \Pr(D_t = 1|D_t^* = 1, D_{t-1} = 0) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1} = 1) - \Pr(D_t = 1|D_t^* = 0, D_{t-1} = 0) \end{bmatrix},$$

and test $H_0 : \Lambda(i) = 0$ versus $H_1 : \Lambda(i) \neq 0$ separately for $i = 1, 2$. Table 3 below displays the 95% bootstrap confidence intervals for $\Lambda(i)$. It can be indicated from the table that the Λ is significantly different from zero, therefore we could effectively reject the null hypothesis.

Table 3: Testing of Validation of Conditional Independence – $H_t = 2, 3$ case

Null hypothesis H_0	95% confidence interval of $L(i)$
$Pr(D_t = 1\|D_t^* = 1, D_{t-1} = 1) = Pr(D_t = 1\|D_t^* = 1, D_{t-1} = 0)$	[0.0000,0.9208]
$Pr(D_t = 1\|D_t^* = 0, D_{t-1} = 1) = Pr(D_t = 1\|D_t^* = 0, D_{t-1} = 0)$	[0.0645,0.8078]

6.2.2 The general approach

Now we test the validity of the conditional independence of our general model. Similar as in the basic model, we want to test whether latent variables and the independently observed variables always play a role in determining the distribution of the proxy. Here, however, we need to do three classes of testings in total. Firstly we would like to test whether the following equality holds,

$$H_0 : \Pr(D_t|D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t|D_t^*, D_{t-1}^*),$$

for each given values of (D_t^*, D_{t-1}^*).

Specifically, we define Λ_1 as the L^2 norm of the difference of two probability vectors, and test whether Λ_1 is statistically significantly larger than zero. In other words,

$$\Lambda_1 := ||\Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = 1) - \Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = 1)||_{L^2},$$

where for $d = 1, 0$

$$\Pr(D_t = 1|D_t^*, D_{t-1}^*, D_{t-1} = d) := \begin{bmatrix} \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = d) \\ \Pr(D_t = 1|D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = d) \end{bmatrix}.$$

And we test

$$H_0 : \Lambda_1 = 0$$

versus

$$H_1 : \Lambda_1 > 0.$$

The second hypothesis we would like to test is

$$H_0 : \Pr(D_t | D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t | D_t^*, D_{t-1}),$$

for each given values of (D_t^*, D_{t-1}). Following the same logic, we define Λ_2 as the L^2 norm of the difference of two probability vectors, and test whether Λ_2 is significantly larger than zero. In other words,

$$\Lambda_2 := ||\Pr(D_t = 1 | D_t^*, D_{t-1}^* = 1, D_{t-1}) - \Pr(D_t = 1 | D_t^*, D_{t-1}^* = 0, D_{t-1})||_{L^2},$$

where for $d = 1, 0$

$$\Pr(D_t = 1 | D_t^*, D_{t-1}^* = d, D_{t-1}) := \begin{bmatrix} \Pr(D_t = 1 | D_t^* = 1, D_{t-1}^* = d, D_{t-1} = 1) \\ \Pr(D_t = 1 | D_t^* = 1, D_{t-1}^* = d, D_{t-1} = 0) \\ \Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = d, D_{t-1} = 1) \\ \Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = d, D_{t-1} = 0) \end{bmatrix}.$$

We then test

$$H_0 : \Lambda_2 = 0$$

versus

$$H_1 : \Lambda_2 > 0.$$

Lastly we want to test whether the following equality holds

$$H_0 : \Pr(D_t | D_t^*, D_{t-1}^*, D_{t-1}) = \Pr(D_t | D_t^*),$$

for any given values of $(D_{t-1}^*,\ D_{t-1})$. To test this, we first calculate the conditional probability given any particular values of the pair $(D_{t-1}^* = d_1, D_{t-1} = d_2)$, where $d_1, d_2 \in \{0, 1\}$, and then define $\Lambda_3(d_1, d_2)$ as the L^2 norm of the difference of these two probability vectors. In other words,

$$\Lambda_3(d_1, d_2) := ||\Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2) - \Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2)||_{L^2},$$

where,

$\Pr(D_t = 1 | D_t^*, D_{t-1}^* = d_1, D_{t-1} = d_2) := \begin{bmatrix} \Pr(D_t = 1 | D_t^* = 1, D_{t-1}^* = d_1, D_{t-1} = d_2) \\ \Pr(D_t = 1 | D_t^* = 0, D_{t-1}^* = d_1, D_{t-1} = d_2) \end{bmatrix}$, for $d_1, d_2 \in \{0, 1\}$. We test whether Λ_3 is significantly larger than zero [4], i.e.

$$H_0 : \Lambda_3(d_1, d_2) = 0$$

versus

$$H_1 : \Lambda_3(d_1, d_2) > 0.$$

Table 4 displays the bootstrap confidence intervals for each of the statistics we use in our testings. It can be seen from the table that, none of the bootstrap confidence intervals is significantly close to zero, which means we can reject all the null hypotheses, therefore our conditional independence assumptions are valid for the general case.

[4] Note that Λ_3 is a L^2-norm which is always nonnegative, so we do one-sided test here

Table 4: Testing of Validation of Conditional Independence – $H_t = 2, 3$ case

Null hypothesis H_0	95% confidence interval of $L_3(d_1, d_2)$
$Pr(D_t = 1\|D_t^*, D_{t-1}^*, D_{t-1} = 1) = Pr(D_t = 1\|D_t^*, D_{t-1}^*, D_{t-1} = 0)$	[0.0570,1.1796]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1}) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1})$	[0.0000,0.7177]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 0)$	[0.0142, 0.8735]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0000, 0.5571]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0105, 0.9371]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0070, 0.7779]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0000, 0.5661]
$Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 1) = Pr(D_t = 1\|D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0128, 0.8590]

6.3 Simulation: the basic approach

6.3.1 Data generation

Data of size N is generated according to the following procedure for 1000 repetition:

Step 1 Generate D_{t-1} using the marginal distribution $\Pr(D_{t-1})$ from the true data.

Step 2 Conditional on $D_{t-1} \in \{0, 1\}$ simultaneously generate (D_t^*, D_{t-2}) using the joint distribution of $\Pr(D_t^*, D_{t-2}|D_{t-1})$.

Step 3 Conditional on D_{t-1} and D_t^*, generate D_t using $\Pr(D_t|D_t^*, D_{t-1})$.

Step 4 Generate the health condition data, H_t, using $\Pr(H_t = 1|D_t^*, D_{t-1})$.

In order to get the distribution of H_t, it is better to assume it takes on binary values. This is because only $E[H_t|D_t^*, D_{t-1}, H_{t-1}]$ can be identified using our current method. Therefore a binary H_t could give us accurate estimation of $\Pr(H_t|D_t^*, D_{t-1}, H_{t-1})$. In our real estimation, H_t takes four possible values, $(1, 2.5, 4, 5)$. To get the parameters for simulation, we assume $H_\tau = 0$ if $H_\tau \leq 3$ originally and 1 otherwise.

Thus we have a data set containing information about $(H_t, H_{t-1}, D_t, D_t^*, D_{t-1,}D_{t-2})$. We use our identification method to estimate $\Pr(D_t|D_t^*, D_{t-1})$ and compare the results with the underlying true values.

6.3.2 Parametrization

The underlying parameter values now are given in the following equations and tables:

$$\Pr(D_t^* = 1) = 0.5000, \Pr(D_{t-1} = 1) = 0.5000$$
$$\Pr(D_t^* = 1|D_{t-1}^* = 1) = 0.5000, \Pr(D_t^* = 1|D_{t-1}^* = 0) = 0.5000$$

Table 5: Joint distribution of $f(D_t^*, D_{t-2}|D_{t-1})$

$f(D_t^*, D_{t-2}\|D_{t-1})$	$D_{t-1}=1$	$D_{t-1}=0$
$D_t^*=1, D_{t-2}=1$	0.2000	0.3000
$D_t^*=1, D_{t-2}=0$	0.1000	0.2000
$D_t^*=0, D_{t-2}=1$	0.1000	0.1000
$D_t^*=0, D_{t-2}=0$	0.6000	0.4000

Table 6: Conditional distribution of D_t and H_t

	$D_t^*=1$	$D_t^*=0$
$Pr(D_t=1\|D_t^*, D_{t-1}=1)$	0.2000	0.2500
$Pr(D_t=1\|D_t^*, D_{t-1}=0)$	0.0500	0.1000
$E(H_t\|D_t^*, H_{t-1}=1)$	0.6000	0.5000
$E(H_t\|D_t^*, H_{t-1}=0)$	0.9000	0.8000

6.3.3 Results

Table 7: Simulation results: means, medians and standard errors

		Sample size				
Estimated probabilities		7,000	10,000	50,000	100,000	True value
$Pr(D_t=1\|D_t^*=1, D_{t-1}=1)$	mean	0.2073	0.2036	0.2000	0.2009	0.2000
	median	0.2040	0.2012	0.2000	0.2002	
	(std.err.)	(0.0578)	(0.0481)	(0.0212)	(0.0150)	
$Pr(D_t=1\|D_t^*=0, D_{t-1}=1)$	mean	0.0492	0.0482	0.0495	0.0500	0.0500
	median	0.0522	0.0501	0.0503	0.0500	
	(std.err.)	(0.0198)	(0.0169)	(0.0075)	(0.0053)	
$Pr(D_t=1\|D_t^*=1, D_{t-1}=0)$	mean	0.2527	0.2506	0.2499	0.2501	0.2500
	median	0.2500	0.2499	0.2498	0.2500	
	(std.err.)	(0.0289)	(0.0242)	(0.0100)	(0.0073)	
$Pr(D_t=1\|D_t^*=0, D_{t-1}=0)$	mean	0.0982	0.0984	0.0994	0.0999	0.1000
	median	0.0999	0.0995	0.0997	0.0999	
	(std.err.)	(0.0182)	(0.0147)	(0.0062)	(0.0042)	

6.4 Simulation: the general approach

6.4.1 Data generation

Data of size N is generated according to the following procedure for 1000 repetition:

Step 1 Generate D_{t-1} using the marginal distribution $\Pr(D_{t-1})$ from the true data.

Step 2 Conditional on $D_{t-1} \in \{0, 1\}$, we use the joint distribution of $\Pr(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|d_{t-1})$ to generate simultaneously $(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3})$ given d_{t-1}.

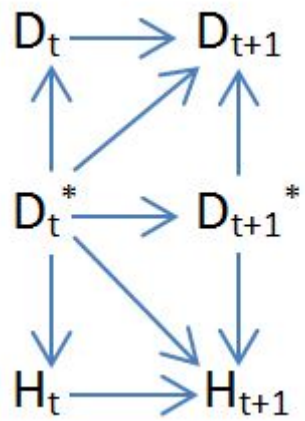
Step 3 We use the estimated conditional probability $\Pr(D_\tau|D_\tau^*, D_{\tau-1}^*, D_{\tau-1})$ to generate D_t .

Step 4 Using the conditional distribution of $\Pr(D_{t+1}^*|D_t^*)$ we can generate D_{t+1}^*.

Step 5 Similar to that in step 3, we use the estimated conditional probability $\Pr(D_\tau|D_\tau^*, D_{\tau-1}^*, D_{\tau-1})$ to generate D_{t+1} .

Step 6 The next step is to generate H_τ given $H_{\tau-1}$, D_τ^* and $D_{\tau-1}^*$ for $\tau \in \{t, t+1\}$. This can be achieved by applying the estimated result of $\Pr(H_\tau|D_\tau^*, D_{\tau-1}^*, H_{\tau-1})$.

Ideally the data should be generated as an evolution from past periods to future ones, as is shown in the following flow chart:



where first a joint distribution of $(H_{t-4}, D_{t-4}, D_{t-4}^*)$ is assumed, then D_τ^* is generated from $\Pr(D_\tau^*|D_{\tau-1}^*)$, D_τ is generated from $\Pr(D_\tau|D_\tau^*, D_{\tau-1}^*, D_{\tau-1})$, and H_τ is generated from $\Pr(H_\tau|D_\tau^*, D_{\tau-1}^*, H_{\tau-1})$. Yet by using this process, we cannot effectively control the invertibility of the LHS matrix within a limited sample size. Therefore we adopt a more direct approach described above, without violating assumptions we made to get nonsingular matrices.

6.4.2 Parametrization

The underlying parameter values are given in the following equations and in tables:

$$\Pr(D_t^* = 1) = 0.5000, \Pr(D_{t-1} = 1) = 0.5000$$
$$\Pr(D_t^* = 1|D_{t-1}^* = 1) = 0.5000, \Pr(D_t^* = 1|D_{t-1}^* = 0) = 0.5000$$

Table 8: Joint distribution of $f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|D_{t-1}=1)$

$f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}\|D_{t-1}=1)$	$D_{t-2}=1$, $D_{t-3}=1$	$D_{t-2}=1$, $D_{t-3}=0$	$D_{t-2}=0$, $D_{t-3}=1$	$D_{t-2}=0$, $D_{t-3}=0$
$D_t^*=1, D_{t-1}^*=1$	0.1500	0.0500	0.0250	0.0250
$D_t^*=1, D_{t-1}^*=0$	0.0250	0.1750	0.0250	0.0250
$D_t^*=0, D_{t-1}^*=1$	0.0250	0.0250	0.1500	0.0500
$D_t^*=0, D_{t-1}^*=0$	0.0250	0.0250	0.0250	0.1750

Table 9: Joint distribution of $f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}|D_{t-1}=0)$

$f(D_t^*, D_{t-1}^*, D_{t-2}, D_{t-3}\|D_{t-1}=0)$	$D_{t-2}=1$, $D_{t-3}=1$	$D_{t-2}=1$, $D_{t-3}=0$	$D_{t-2}=0$, $D_{t-3}=1$	$D_{t-2}=0$, $D_{t-3}=0$
$D_t^*=1, D_{t-1}^*=1$	0.1500	0.0500	0.0250	0.0250
$D_t^*=1, D_{t-1}^*=0$	0.0250	0.1750	0.0250	0.0250
$D_t^*=0, D_{t-1}^*=1$	0.0250	0.0250	0.1500	0.0500
$D_t^*=0, D_{t-1}^*=0$	0.0250	0.0250	0.0250	0.1750

Table 10: Conditional distribution of D_t and H_t

	$D_t^*=1$, $D_{t-1}^*=1$	$D_t^*=1$, $D_{t-1}^*=0$	$D_t^*=0$, $D_{t-1}^*=1$	$D_t^*=0$, $D_{t-1}^*=0$
$Pr(D_t=1\|D_t^*, D_{t-1}^*, D_{t-1}=1)$	0.8000	0.1000	0.7000	0.1000
$Pr(D_t=1\|D_t^*, D_{t-1}^*, D_{t-1}=0)$	0.6000	0.1000	0.5000	0.0500
$E(H_t\|D_t^*, D_{t-1}^*, H_{t-1}=1)$	0.8000	0.5000	0.3000	0.1000
$E(H_t\|D_t^*, D_{t-1}^*, H_{t-1}=0)$	0.6000	0.4000	0.2000	0.0500

6.4.3 Results

Table 11: Simulation results: means, medians and standard errors

		Sample size				
Estimated probabilities		7,000	10,000	50,000	100,000	True value
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	mean	0.7660	0.7779	0.8018	0.8012	0.8000
	median	0.7813	0.7859	0.8008	0.8002	
	(std.err.)	(0.1541)	(0.1399)	(0.0401)	(0.0284)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	mean	0.1463	0.1267	0.1006	0.0981	0.1000
	median	0.1302	0.1234	0.1006	0.0986	
	(std.err.)	(0.1288)	(0.0997)	(0.0551)	(0.0407)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	mean	0.5490	0.5740	0.6490	0.6790	0.7000
	median	0.5867	0.6157	0.6763	0.6932	
	(std.err.)	(0.2630)	(0.2603)	(0.1966)	(0.1397)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	mean	0.1271	0.1151	0.0972	0.0984	0.1000
	median	0.1130	0.1032	0.1007	0.1027	
	(std.err.)	(0.1355)	(0.1082)	(0.0575)	(0.0438)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	mean	0.6068	0.6076	0.6027	0.6016	0.6000
	median	0.6026	0.5987	0.5993	0.5995	
	(std.err.)	(0.1415)	(0.1136)	(0.0446)	(0.0306)	
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	mean	0.1403	0.1173	0.0988	0.0981	0.1000
	median	0.1240	0.1132	0.1023	0.1000	
	(std.err.)	(0.1293)	(0.0869)	(0.0488)	(0.0383)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	mean	0.4002	0.4120	0.4809	0.4936	0.5000
	median	0.4126	0.4248	0.4850	0.4994	
	(std.err.)	(0.2407)	(0.2254)	(0.1351)	(0.0947)	
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	mean	0.0891	0.0724	0.0519	0.0500	0.0500
	median	0.0544	0.0520	0.0496	0.0496	
	(std.err.)	(0.1234)	(0.0820)	(0.0410)	(0.0325)	

6.5 Estimation

Table 12: Means and Standard Deviations of Key Variables From 2008 Survey

Variable	Population	Females	Males
D_{t+1}	0.1750	0.1415	0.2108
	(0.3800)	(0.3485)	(0.4079)
D_t	0.1904	0.1645	0.2180
	(0.3926)	(0.3708)	(0.4129)
D_{t-1}	0.1851	0.1504	0.2223
	(0.3884)	(0.3575)	(0.4158)
D_{t-2}	0.1980	0.1581	0.2406
	(0.3985)	(0.3649)	(0.4275)
D_{t-3}	0.2190	0.1814	0.2590
	(0.4136)	(0.3854)	(0.4382)
H_t	2.2484	2.3264	2.1652
	(0.9428)	(0.9454)	(0.9330)
H_{t-1}	2.2188	2.3090	2.1225
	(0.9490)	(0.9497)	(0.9389)
Education	0.3069	0.3373	0.2744
	(0.4612)	(0.4728)	(0.4463)
Marital Status	0.2885	0.3236	0.2510
	(0.4531)	(0.4679)	(0.4337)
Ethnicity	0.5087	0.4871	0.5318
	(0.5000)	(0.4999)	(0.4991)

Table 13: Regression results with covariates

	D_{t+1}	
Variable	Coefficient	Standard Error
Intercept	0.0081	0.0069
D_t	0.3466	0.0120
D_{t-1}	0.1708	0.0129
D_{t-2}	0.1582	0.0125
D_{t-3}	0.1018	0.0112
Education	0.0051	0.0078
Marital Status	-0.0028	0.0079
Ethnicity	0.0131	0.0073
Gender	0.0173	0.0071

Table 14: Regression results with covariates

	H_t	
Variable	Coefficient	Standard Error
Intercept	1.1588	0.0323
D_{t+1}	0.0278	0.0353
D_t	0.0415	0.0358
D_{t-1}	-0.0342	0.0364
D_{t-2}	0.0224	0.0354
D_{t-3}	0.0404	0.0316
H_{t-1}	0.5326	0.0106
Education	-0.2235	0.0221
Marital Status	-0.0073	0.0221
Ethnicity	-0.0021	0.0204
Gender	-0.0842	0.0199

Table 15: Estimation results: basic approach

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8223	0.8382	0.8325	0.1200
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5317	0.5157	0.5259	0.1793
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.3089	0.3848	0.3008	0.3067
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0328	0.0138	0.0472
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.4896	2.4976	2.4932	0.0563
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.3985	2.3798	2.3888	0.0761
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.4419	2.4481	2.4393	0.0657
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.3973	2.3936	2.3972	0.0378
$Pr(D_t^* = 1)$	0.3213	0.3409	0.2529	0.2344

Table 16: Estimation results: basic approach for males

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7349	0.8076	0.8122	0.1652
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5492	0.4179	0.4399	0.2912
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1901	0.3801	0.2498	0.3502
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0376	0.0116	0.0583
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.4528	2.4726	2.4650	0.0670
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.4154	2.3919	2.4019	0.1064
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.3770	2.3781	2.3772	0.0853
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.3559	2.3508	2.3566	0.0538
$Pr(D_t^* = 1)$	0.4419	0.3945	0.3058	0.2461

Table 17: Estimation results: basic approach for females

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8632	0.8591	0.8631	0.1270
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.5346	0.5143	0.5336	0.1798
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.3661	0.4273	0.3718	0.3165
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0276	0.0357	0.0300	0.0509
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.5404	2.5543	2.5459	0.1015
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.3595	2.3506	2.3647	0.1211
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5271	2.5358	2.5179	0.1134
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.4283	2.4178	2.4260	0.0554
$Pr(D_t^* = 1)$	0.2032	0.2993	0.1846	0.2462

Table 18: Estimation results: general approach

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.7502	0.7820	0.8055	0.2778
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.6340	0.6504	0.6415	0.2253
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3022	0.4238	0.4148	0.2264
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2412	0.4341	0.4326	0.2383
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.3956	0.4599	0.4606	0.1998
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.4025	0.4631	0.4543	0.1980
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2472	0.2674	0.2542	0.1629
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0178	0.1233	0.0754	0.1653
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	-21.6839	2.6398	2.4672	4.3786
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.1625	2.3381	2.4308	2.0964
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4537	3.3446	2.5025	7.7230
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4217	0.8459	2.1924	15.1518
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.6832	3.4901	2.6696	6.1663
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.1974	1.2858	2.1426	6.9415
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5035	2.5830	2.4285	5.0442
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.3729	1.4602	2.3888	19.7184
$Pr(D_t^* = 1)$	0.1380	0.2410	0.1931	0.1656

Table 19: Estimation results: general approach of males

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	1.0000	0.7614	0.7836	0.1767
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.7099	0.6190	0.6066	0.1996
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.6234	0.4208	0.4154	0.2210
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.3573	0.3723	0.3681	0.2344
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2909	0.4588	0.4528	0.1964
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.5278	0.5189	0.5004	0.2262
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1116	0.2243	0.1919	0.1454
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1011	0.0951	0.0621	0.1025
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5559	2.1450	2.4369	8.9161
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	3.0644	2.5098	2.4468	7.1481
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4689	3.9783	2.5317	23.2692
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3456	1.7889	2.3204	3.4740
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5426	3.1131	2.5466	10.6401
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.5439	0.4875	1.7965	31.7647
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.3243	2.7043	2.4747	7.2401
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	3.2677	2.8256	2.3581	14.2710
$Pr(D_t^* = 1)$	0.1702	0.2687	0.2370	0.1465

Table 20: Estimation results: general approach of females

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.7461	0.7957	0.8258	0.1617
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.6377	0.6592	0.6641	0.1918
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3918	0.4573	0.4644	0.2131
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.4031	0.4253	0.4255	0.1992
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.6477	0.5352	0.5361	0.2185
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.6527	0.4651	0.4739	0.2369
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1265	0.2459	0.2186	0.1802
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0000	0.0963	0.0461	0.1200
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	3.7445	2.7840	2.6129	12.1329
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	1.7189	2.3308	2.5285	3.4989
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6786	3.8359	2.5554	7.9331
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3810	1.0889	2.1884	7.8487
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	5.0667	6.0379	2.7972	82.1887
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5889	1.4805	2.4274	8.9413
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.4333	-8.6810	2.4171	423.8590
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.3383	2.1029	2.4237	11.8090
$Pr(D_t^* = 1)$	0.1043	0.2495	0.1737	0.2102

6.6 An alternative way of dealing with H_τ: Combining $H_\tau = 1, 2, 3$

6.6.1 The basic approach

In this section we combine all the samples with at least "good" health conditions in their surveys answers. Specifically, we redefine $H_\tau = 2.5$ if it originally is $1, 2$ or 3. The results from this second estimation are shown in Table 21. In addition, we re-estimate everything conditional on gender covariates, and the results are presented in Table 22 and Table 23. Compared with those in Section 4, the results here are not very different, except for the estimation for male subgroup. It can be seen that the point estimates for misclassification errors are the same for $D_t^* = 1, D_{t-1} = 1$ case and for $D_t^* = 0, D_{t-1} = 1$ case, which does not make much sense. Nonetheless, the bootstrap means and medians provide more useful information about this subgroup, using the newly defined subsample.

Table 21: Estimation results: basic approach, combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7992	0.8475	0.8394	0.1152
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.7100	0.6801	0.6790	0.0930
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1022	0.4159	0.2763	0.3664
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0425	0.0218	0.0538
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6325	2.6442	2.6401	0.0364
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6102	2.5971	2.5978	0.0255
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5860	2.5756	2.5828	0.0322
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5631	2.5736	2.5802	0.0218
$Pr(D_t^* = 1)$	0.5555	0.3315	0.2140	0.2591

Table 22: Estimation results: basic approach for males,combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.7443	0.8371	0.8421	0.1360
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.7443	0.6583	0.6585	0.1539
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	1.0000	0.4772	0.4102	0.3670
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0817	0.0525	0.0431	0.0642
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6098	2.6316	2.6255	0.0461
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6098	2.5871	2.5892	0.0331
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.5371	2.5454	2.5486	0.0353
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5695	2.5615	2.5671	0.0228
$Pr(D_t^* = 1)$	0.1309	0.3049	0.1944	0.2372

Table 23: Estimation results: basic approach for females,combining $H_\tau = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1)$	0.8666	0.8656	0.8770	0.1219
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1)$	0.6689	0.6587	0.6554	0.1106
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	0.1719	0.3879	0.2553	0.3472
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	0.0000	0.0341	0.0151	0.0457
$E(H_t \mid D_t^* = 1, D_{t-1} = 1)$	2.6724	2.6830	2.6775	0.0679
$E(H_t \mid D_t^* = 0, D_{t-1} = 1)$	2.6066	2.5981	2.5983	0.0414
$E(H_t \mid D_t^* = 1, D_{t-1} = 0)$	2.6098	2.6158	2.6065	0.0603
$E(H_t \mid D_t^* = 0, D_{t-1} = 0)$	2.5890	2.5810	2.5912	0.0358
$Pr(D_t^* = 1)$	0.3589	0.3353	0.2118	0.2769

6.6.2 The general approach

We combine $H_\tau = 1, 2$ and 3 and get the estimation results in Table 24. The results are qualitatively similar to those in Table 18. One advantage for this setup over the previous one can be found from the ninth to the sixteenth row. $E\left[H_t \mid D_t^*, D_{t-1}^*, d_{t-1}, h_{t-1} = 2.5\right]$ differ more in this table than in Table 18. This tells us that the estimation results are more reliable than those in Table 18, because the eigenvalues have more variation and thus can be better identified.

Table 24: Estimation results: general approach, combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.8225	0.7621	0.7993	0.1658
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.4260	0.5900	0.5797	0.2153
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.2273	0.3488	0.3258	0.2302
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.1891	0.3071	0.2650	0.2502
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.6449	0.5098	0.5134	0.2218
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.7814	0.5337	0.5367	0.2471
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.2164	0.2630	0.2301	0.1908
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1256	0.0996	0.0471	0.1293
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6096	2.5165	2.5897	3.6577
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	1.8385	3.2092	2.5267	23.3473
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5168	2.7619	2.5317	1.0847
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.2861	1.7174	2.4180	6.7044
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	3.1835	3.1515	2.6905	6.7574
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.2231	2.1712	2.4967	2.6271
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.6110	2.6396	2.5865	1.2866
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5846	2.1804	2.5810	13.3059
$Pr(D_t^* = 1)$	0.1926	0.2478	0.12093	0.1499

Table 25: Estimation results: general approach of males, combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.5158	0.7392	0.7575	0.1794
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2811	0.5781	0.5803	0.2390
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.0000	0.3388	0.3299	0.2296
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.0000	0.3164	0.2501	0.2640
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1608	0.3865	0.3419	0.2437
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.4849	0.4891	0.4920	0.2221
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1608	0.2037	0.1507	0.1610
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0402	0.0853	0.0628	0.0922
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5799	2.0940	2.5413	8.6562
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5240	2.3785	2.5000	2.6628
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3276	2.9407	2.5444	8.8653
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.3276	2.1845	2.4522	1.7809
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7009	2.7169	2.5732	0.7657
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.4693	2.4031	2.4759	0.7829
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7009	3.4648	2.5697	27.5230
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5623	2.6390	2.5734	1.9286
$Pr(D_t^* = 1)$	0.2869	0.2931	0.2395	0.1915

Table 26: Estimation results: general approach of females, combining $H_t = 1, 2$ and 3

Estimated probabilities	Point estimation	Bootstrap mean	Bootstrap median	Standard error
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 1)$	0.9337	0.8286	0.8713	0.1513
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 1)$	0.7523	0.6528	0.6596	0.1934
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 1)$	0.3020	0.3853	0.4030	0.2285
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 1)$	0.2378	0.4072	0.4259	0.2153
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 1, D_{t-1} = 0)$	1.0000	0.5619	0.5724	0.2502
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1}^* = 0, D_{t-1} = 0)$	0.1354	0.5443	0.5684	0.2747
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 1, D_{t-1} = 0)$	0.1162	0.2627	0.1992	0.2151
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1}^* = 0, D_{t-1} = 0)$	0.0000	0.0774	0.0282	0.1167
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.6982	2.6954	2.6696	2.9087
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	-0.0859	2.5871	2.5063	0.3876
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 1)$	2.5000	4.6111	2.5132	34.1491
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 1)$	2.4168	2.0740	2.5000	3.4499
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	2.7177	3.5561	2.7918	6.5064
$E(H_t \mid D_t^* = 1, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	2.5971	1.7223	2.5308	4.7850
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 1, H_{t-1} = 2.5, D_{t-1} = 0)$	3.5107	2.8225	2.5991	5.4624
$E(H_t \mid D_t^* = 0, D_{t-1}^* = 0, H_{t-1} = 2.5, D_{t-1} = 0)$	1.2059	2.5537	2.5947	6.8411
$Pr(D_t^* = 1)$	0.9377	0.2428	0.1729	0.2007

Table 27: Testing of Validation of Conditional Independence – $H_t = 1, 2, 3$ case

Null hypothesis H_0	95% confidence interval of $L(i)$
$Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 1, D_{t-1} = 0)$	[0.0000,0.9659]
$Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^* = 0, D_{t-1} = 0)$	[0.4287,0.8532]

Table 28: Testing of Validation of Conditional Independence – $H_t = 1, 2, 3$ case

Null hypothesis H_0	95% confidence interval of $L_3(d_1, d_2)$
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^*, D_{t-1} = 0)$	[0.0626,1.1417]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1}) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1})$	[0.0000,0.7888]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0)$	[0.0091, 0.8332]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0000, 0.6127]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0107, 0.8966]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1)$	[0.0186, 0.8169]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 1, D_{t-1} = 0) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0000, 0.6427]
$Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 1) = Pr(D_t = 1 \mid D_t^*, D_{t-1}^* = 0, D_{t-1} = 0)$	[0.0073, 0.8706]