Peer and selection effects on youth smoking in California

Brian Krauth
Simon Fraser University*

March 24, 2004

Abstract

A number of studies have indicated that peer smoking is a highly influential factor in a young person's decision to smoke. However, these results are suspect because the studies often fail to account for selection and simultaneity bias. This paper develops an econometric model of youth smoking which incorporates both peer effects and selection effects, and estimates its parameters using data on California youth. Identification is achieved by using the degree of selection on observables as a proxy for the degree of selection on unobservables. The results indicate that the influence of peers on a young person's decision to smoke is much weaker than is suggested by reduced form models.

JEL Codes: I12, J13, C35
Keywords: Social interactions, peer effects, youth smoking

1 Introduction

Youth smoking is a major public health concern in much of the world. The World Health Organization (Mackay and Eriksen 2002, p. 36) estimates that 4.2 million individuals die every year from smoking-related conditions. As a result, governments often spend large sums on programs to reduce tobacco use. Because tobacco is highly addictive and most smokers begin when they are teenagers (Mackay and Eriksen 2002, p.28), tobacco control efforts often focus on discouraging young people from starting to smoke. Many of these efforts aim to alter the social context of youth smoking, due to a consensus among policymakers and researchers (Mackay and Eriksen 2002, Tyas and Pederson 1998) that a young person's decision to smoke is highly influenced by the behavior and attitudes of his or her peers. From an economic standpoint, the case for extensive government intervention in tobacco markets may be strengthened by the existence of significant peer effects, as they imply an externality which may in principle lead to suboptimal outcomes even among rational individuals. For example, Kenkel, Reed, and Wang (2002) calibrate a "rational addiction"

*8888 University Dr. Burnaby BC V5A 1S6 Canada. Telephone: (604)291-4438. Fax: (604)291-5944. Email: bkrauth@sfu.ca.

model of lifetime tobacco use and find that the optimal tobacco tax increases by over 50% when peer effects are included. In spite of the apparent consensus among researchers and policymakers, the influence of peers in youth smoking is far from well-established. Most empirical studies in the literature fail to account for both endogenous peer selection and the simultaneity of choice among peers, both of which may lead a researcher to dramatically overestimate the strength of peer influence (Manski 1993). More recent studies which attempt to account for selection and/or simulteneity bias (Engels, Knibbe, Drop and de Haan 1997, Gaviria and Raphael 2001, Norton, Lindrooth and Ennett 1998, Wang, Eddy and Fitzhugh 2000) have found mixed results.

This paper uses data from the 1994-2002 California Youth Tobacco Survey (CYTS) (California Department of Health Services 2003) to estimate the strength of peer influence in smoking among young people. I use a simulation-based structural estimation method developed in Krauth (2002) which explicitly allows for both endogenous peer selection and simultaneity of choice. This approach is complementary to recent attempts to estimate peer effects in youth smoking using instrumental variables (Gaviria and Raphael 2001, Norton et al. 1998) or panel data methods (Engels et al. 1997, Wang et al. 2000), and has some significant advantages over these other approaches. The structural estimator addresses simultaneity by treating the group outcome as an endogenous variable, and addresses selection by explicitly allowing correlation in unobservables between peers. In order to achieve point identification of the model parameters, this correlation is assumed to be equal to the correlation in observables between peers. While this is a strong assumption, it is no stronger than those assumptions used in IV and panel studies, and can be weakened substantially while still generating informative interval estimates of the peer effect.

The empirical results reported in this paper include point estimates under the assumption of equal correlation in observables and unobservables as well as interval estimates consistent with much weaker restrictions. The results indicate that observable characteristics are highly correlated among peer group members, with an estimated correlation coefficient of more than 0.5. If we assume that peer group members have the same degree of correlation in unobservables, the structural estimate of the peer effect is less than one-tenth as large as the reduced form estimate and fails to be statistically significant. The results from interval estimation reduce the starkness of this finding somewhat. A correlation in unobservables which is somewhere between zero and 0.5 produces estimates that imply peers have some influence. However, even in these results the estimated peer effect is well below the reduced form estimates. Taken as a whole, the results in this paper suggest that peer are less influential in youth smoking than one might conclude from

previous studies.

1.1 Related literature

There is a vast literature in public health and a smaller related literature in economics on the determinants of tobacco use. Among adults, the strongest predictor of current smoking is smoking as a young person. For example, Gruber and Zinman (2001) find that 75% of adult current or former smokers began before their 19th birthday. The literature on youth smoking has found a number of variables that are closely associated with a young person's smoking behavior in multiple surveys. These variables include parental or sibling smoking, performance in school, race and ethnicity, and prices[1] (Gruber and Zinman 2001). Peer smoking is also, in the words of a recent review article, "consistently found to be related to smoking initiation, maintenance, and intentions" (Tyas and Pederson 1998).

However, the finding of a close statistical association between peer smoking and a respondent's own smoking does not necessarily imply a causal relationship. Indeed, selection and simultaneity imply that the statistical association may dramatically overstate the strength of any causal relationship. While this issue has been known for some time, it is only recently that empirical studies of youth smoking that attempt to account for selection and simultaneity have appeared. These studies have used either instrumental variables or panel data methods. In the economics literature, the dominant approach uses group characteristics as instrumental variables for group behavior. When these group characteristics are simply the group averages of the individual-level regressors, as in Gaviria and Raphael (2001), this approach can lead to consistent estimates under some strong but plausible assumptions. However, other researchers (Norton et al. 1998, for example) have used other group characteristics (for example, population density of the neighborhood the group lives in) which cannot in principle be valid instruments - if they have a direct impact on the group's average behavior, they must have a corresponding direct impact on the behavior of individual group members. This difficulty in selecting truly exogenous instruments provides some motivation for the use of alternative approaches to identification.

In the public health literature, concern with selection issues has been addressed mostly using panel data. These papers (Engels et al. 1997, Wang et al. 2000, for example) do not use regression-based econometric analysis, but rather ask whether smoking tends to precede or follow membership

[1] DeCicca, Kenkel, and Mathios (2002) argue that the common finding of a high price elasticity of youth smoking is an artifact of correlation between prices and unobserved local factors.

in a peer group of smokers. If initial membership in a peer group of smokers predicts subsequent transition into smoking, this is interpreted as evidence for a peer effect. If initial smoking predicts subsequent transition into a higher-smoking peer group, this is interpreted as evidence for a selection effect. They find evidence for the transition from smoking to membership in a peer group of smokers, but no evidence for transition from membership in a peer group of smokers to smoking. This result is generally taken to mean that the correlation in smoking among peers is caused primarily by selection effects and not peer effects. However, these results are somewhat weakened by the lack of a formal model of youth decision making. For example, if peer effects operate without a significant lag, current behavior of group members may not be strongly associated with future changes in group behavior even if peers are highly influential. While the results from these longitudinal studies provide some evidence for selection, they may actually have little to say about the strength of peer influence.

As discussed earlier, this paper pursues a structural approach. Although this approach is fairly novel in the analysis of youth smoking, there is a recently-growing literature in econometrics on empirical models of binary games. Kooreman (1994) and Tamer (2002) develop econometric models of interdependent binary choice for the special case with two group members and no selection effect. Krauth (2002) and Kooreman and Soetevent (2002) extend this approach to an arbitrary group size and incorporate inter-peer correlation in characteristics (i.e., peer selection). The structural approach has several advantages over the IV and panel methods. First, it allows for more flexibility in the choice of identifying assumptions. While point identification requires assumptions as strong as those required by the IV methodology, it is only under the structural method used here that informative interval estimates can be obtained with weaker assumptions. Second, the results can be much more easily used to generate quantitative behavioral predictions than is possible with the panel data approach in public health.

2 Data

The data set used for this study is the California Youth Tobacco Survey (CYTS), an annual cross-sectional household-based survey of California youth aged 12 to 17 (California Department of Health Services 2003). This study uses results from 1994 through 2002, and restricts attention to respondents from age 14 through 17. The original data set contains 13,878 observations in this age range, of which 13,037 have sufficient information for use in estimating the structural

model. Summary statistics, after the data cleaning and imputations outlined below, are reported in Table 1.

For each observation, there are two endogenous variables: a binary variable indicating the current smoking status of the respondent, and a discrete variable indicating the fraction of the respondent's closest same-sex friends that smoke. The respondent is defined as a "current smoker" if he or she reports having smoked at least one cigarette in the previous 30 days. This measure of smoking prevalence is standard in the literature. Because current smoking is self reported in the CYTS, underreporting may be a concern. Section 3.4 describes how this issue is addressed within the structural model. The peer smoking rate is defined as the respondent's estimate of the fraction of his or her same sex "best friends" that smoke. The construction of this variable varies somewhat in different years of the CYTS. In the 1994-1999 surveys, the respondent is asked how many best friends of each sex he or she has, and then how many of these friends smoke. The number of best friends of each sex can vary from zero to seventy. In the 2000-2002 surveys, the respondent is not asked the number of best friends, but is simply asked how many of his or her four best friends of each sex smoke. In order to keep the computational time for the structural estimator reasonable, it is necessary to place a cap on the number of friends. Any respondent who reports more than six close friends is recoded as having exactly six friends, and the fraction of friends who smoke is rounded to the nearest sixth. For example, if a respondent reports ten close friends, seven of whom smoke, he is coded as having six close friends, four ($6 * 7/10 = 4.2 \approx 4$) of whom smoke. Those respondents who did not report their own smoking behavior (76 obs.) or that of their peers (92 obs.), or who reported having zero best friends (673 obs.) were dropped from the sample. For comparison, summary statistics for the observations that were dropped from the sample are reported in Table 1. As the table shows, those respondents who were dropped from the sample are similar in observable characteristics to those who are included.

In addition to these variables, the CYTS includes information on demographics, family, workplace and school environment, attitudes, and other risky behavior. The explanatory variables included in the model include year, age, ethnicity (African-American, Hispanic, and other), whether parents, teachers, or older siblings smoke, immigrant status, weekly disposable income, employment, performance in school, and classroom exposure to information about the risks of smoking. Disposable income is top-coded at $200 per week, and missing values for disposable income are replaced with the sample median of $15. Missing values for the other explanatory variables are replaced by the sample mean for that variable.

2.1 Trends in youth smoking and tobacco control in California

The sample period saw two major changes in California's tobacco control policy. The first was the passage of legislation which required almost all workplaces, including restaurants, to be smoke-free beginning in January 1995, and bars to be smoke free beginning in January 1998. The second was the January 1999 increase in state excise taxes from 37 cents to 87 cents per pack, as a result of the passage of the voter initiative known as Proposition 10. This was the only change in state excise taxes during the sample period. California's excise tax rates – tied for 15th in 1994, 18th in 2002 (Orzechowski and Walker 2002) – and level of per-capita expenditure on tobacco control – 19th in 2002 (Centers for Disease Control and Prevention 2002a) – are above the median among US states, though far from the highest.

Figure 1 shows the rate of self-reported current smoking among age 14-17 CYTS respondents by year, as well as the rate of self-reported current smoking among US high school students as calculated by the Centers for Disease Control and Prevention (CDC) (2002b, 2003). Throughout the sample period, California youth are approximately half as likely to smoke as youth in the United States as a whole. California also shows a large decline in youth smoking from 1997 to 1999, possibly related to the policy changes during that period. Figure 2 shows the smoking rates of California youth by year and birth cohort. As the figure shows, the smoking rate of each cohort increases with age. In addition, the figure shows that the overall downward movement in smoking rates over the sample period appears among all age groups.

3 Model

The econometric model is similar in spirit to the standard model of discrete choice with social interaction effects due to Brock and Durlauf (2001). Both the model and estimation method are described in greater detail in Krauth (2002).

3.1 Preferences and peer groups

In the model, each individual is a member of a peer group. Peer groups are indexed by $g \in Z+$, and group g has $n_g \geq 2$ members, where n_g is exogenous and may vary across groups. Within each peer group, individuals are indexed by i, so that the pair i, g identifies an individual. Each member of a peer group is influenced symmetrically by each other member, and there are no cross-group influences.

Individuals choose either to smoke ($s_{i,g} = 1$) or not ($s_{i,g} = 0$). An individual's utility function $u_{i,g}(s_{i,g})$ satisfies the following:

$$u_{i,g}(1) - u_{i,g}(0) = \alpha + \beta x_{i,g} + \lambda z_g + \gamma \bar{s}_{i,g} + \epsilon_{i,g} \tag{1}$$

where $x_{i,g}$ is a vector of individual-level exogenous variables (e.g., ethnicity, sex, parental and sibling behavior, disposable income), z_g is a vector of group-level exogenous variables (e.g., prices, school characteristics, or aggregate fixed effects), $\bar{s}_{i,g}$ is the fraction of the other group members that smoke:

$$\bar{s}_{i,g} = \frac{1}{n_g - 1} \sum_{j \neq i} s_{j,g} \tag{2}$$

and $\epsilon_{i,g}$ is an unobserved individual-level term. The model thus incorporates an endogenous[2] peer effect ($\gamma \bar{s}_{i,g}$) into an otherwise standard discrete choice model of smoking. An individual will prefer to smoke ($s_{i,g} = 1$) if and only if his or her incremental utility from doing so is positive. However, because this incremental utility depends in part on the average choice of peers ($\bar{s}_{i,g}$), predictions on the behavior of individuals cannot be derived in isolation. Instead, it is necessary to analyze behavior in equilibrium.

3.2 Equilibrium

Given the exogenous variables, the observed choices of individuals are assumed to be a Nash equilibrium in pure strategies of a complete information simultaneous move game, where player i, g's strategy is given by $s_{i,g}$ and his or her payoff function is described by equation (1). Because there may be multiple Nash equilibria for some draws of the exogenous variables, the model is supplemented with an equilibrium selection rule: the endogenous variables are assumed to take on the values associated with the lowest-activity Nash equilibrium for the given exogenous variables. This selection rule is chosen because it corresponds to the steady state of the myopic best-reply dynamics reached when agents start as nonsmokers. Given that young people are born as nonsmokers and that teen smokers rarely quit while young, such a dynamic process may be a reasonable model of behavior. Monte Carlo results reported in Krauth (2002) suggest that for low

[2]The model does not include what are known as "contextual effects", in which the characteristics (as opposed to smoking behavior) of one's peers have a direct effect on the relative utility of smoking. Since Manski (1993) pointed out the difficulties in doing so, empirical researchers rarely attempt to simultaneously estimate endogenous effects and contextual effects. It is more common (Gaviria and Raphael 2001, Hoxby 2000, Sacerdote 2001, for example) to simply ignore one or the other and note that the estimated endogenous (contextual) effect could be reinterpreted as contextual (endogenous). That caveat applies to this article as well.

to moderate peer effects, misspecification in the equilibrium selection rule has minimal impact on the resulting estimates.

3.3 Peer selection and correlation in characteristics

As indicated in the introduction, the structural model estimated in this paper allows for the likelihood that a young person's choice of peer group is not entirely random. This is done by incorporating a reduced-form correlation in both observable and unobservable characteristics across members of the same peer group. That is, the exogenous variables are identically distributed across all individuals in the population, and are independent across members of different peer groups but are not independent across members of the same peer group. Although it might be desirable to formally model the process of selecting one's peer group, estimating such a model would require substantially more detail on group composition than is available.

The details are as follows. The joint distribution of the individual-level exogenous variables is assumed to be multivariate normal across the group members. For a group of size $n_g = 3$:

$$\begin{bmatrix} \beta x_{1,g} \\ \beta x_{2,g} \\ \beta x_{3,g} \\ \epsilon_{1,g} \\ \epsilon_{2,g} \\ \epsilon_{3,g} \end{bmatrix} \sim N \left(\begin{bmatrix} \mu \\ \mu \\ \mu \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho_x\sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \rho_x\sigma^2 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_\epsilon & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & 1 & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & \rho_\epsilon & 1 \end{bmatrix} \right) \tag{3}$$

with the distribution being defined analogously for other values of n_g.

Equation (3) defines a probability distribution with several useful characteristics. First, with the exception of the dependence across peers, the distribution corresponds to that in a standard probit model. In particular, for all individuals i, g the conditional distribution of unobservables has $\epsilon_{i,g}|x_{i,g} \sim N(0,1)$. Unlike a standard probit model, this model allows for correlation in both observable characteristics (ρ_x) and unobservable characteristics (ρ_ϵ) across group members, as would occur when young people are prone to selecting friends who are similar to themselves. The "naive" analysis pursued in much of the literature on peer effects implicitly assumes that peer grouop selection is random ($\rho_\epsilon = \rho_x = 0$).

In order to obtain point estimates of model parameters, it is necessary to impose an additional restriction on ρ_ϵ, the between-peer correlation in unobservables. The primary restriction used

in this paper is that the correlation is the same as the correlation in observables, i.e. $\rho_\epsilon = \rho_x$. The idea of using the degree of selection on observables as a proxy for the degree of selection on unobservables was first proposed by Altonji, Elder, and Taber (2000) to correct for selection effects in measuring the effect of attending a Catholic school. These authors demonstrate that equality in these two correlations will hold (in expectation) if the observables are a random subset of a large set of relevant variables. Alternatively, if the observed variables are more highly correlated between peers than the unobserved variables, the equal-correlation point estimate of the peer effect will be biased downwards. This is a distinct possibility, as personal information that is particularly easily gathered in surveys (race, sex, age) may also be more easily observed by potential friends. In any case, the model can also be estimated under alternative restrictions on ρ_ϵ, including interval restrictions. As a result, the results reported in this paper allow readers with different prior beliefs about the selection process to construct their own range of estimates consistent with both the data and their prior beliefs.

3.4 Actual smoking and reported smoking

One serious issue with this and similar data sets is that adolescent self-reports of their smoking and that of their peers are known to be biased. Norton, Lindrooth, and Ennett (2003) provide a good recent discussion of these issues. First, despite efforts by survey collectors to emphasize and ensure confidentiality, young people tend to underreport their own smoking. Audit studies (Wagenknecht, Burke, Perkins, Haley and Friedman 1992, for example) that compare self-reported smoking with breath or saliva tests often find substantial underreporting. Second, young people also tend to overestimate the extent to which their peers smoke (Norton, Lindrooth and Ennett 2003). Smokers have a particularly high propensity to overestimate the fraction of their peers that smoke, an example of what psychologists call the 'false consensus' effect. The CYTS data exhibit clear signs of some combination of underreported own smoking and overreported peer smoking: 13.1% of respondents are current smokers, whereas 21.3% of their close friends are smokers. As the respondents and their friends are drawn from the same population, and should thus have similar smoking rates, biased reporting is a clear issue. Biased reporting has particular implications for the estimation method developed in this paper. Closing the model in order to have a well-defined likelihood function requires modeling the relationship between a person's actual smoking behavior, his or her self-report of that behavior to the survey interviewer, and the perception of that behavior by his or her peers.

Because actual behavior is not observed in our survey, the model of reporting behavior used here is quite simple. First, it is assumed that peer smoking is reported truthfully. Second, it is assumed that smokers falsely report as nonsmokers with a fixed probability p_r[3]. Let $r_{i,g}$ indicate whether a person self-reports as a smoker. Then:

$$r_{i,g} = \begin{cases} s_{i,g} & \text{with probability } p_r \\ 0 & \text{with probability } 1 - p_r \end{cases} \tag{4}$$

where p_r is a parameter. Note that conditional on $s_{i,g}$, $r_{i,g}$ is independent of all other variables, and that

$$p_r = \frac{E(r_{i,g})}{E(\bar{s}_{i,g})} \tag{5}$$

The parameter p_r can thus be estimated using the sample analogue to this expression.

4 Estimation

Given the model outlined in Section 3, this section describes how the model is estimated using the CYTS data. Let N be the number of observations in the data . Because the CYTS is a household-based random sample from a large population, it can be assumed that each observation represents a member of a different peer group. Without loss of generality, let the gth observation in the data set be identified as describing person 1 of group g.

4.1 Naive estimator

In order to facilitate comparison with previous work, I first apply a naive probit estimator to the data. The naive estimator is simply the standard maximum likelihood probit estimator which treats average peer behavior $\bar{s}$ as an exogenous variable. In other words, I use maximum likelihood to estimate a model in which

$$\Pr(r_{1,g} = 1 | x_{1,g}, z_g, \bar{s}_{1,g}) = \Phi\left(\tilde{\alpha} + \tilde{\beta} x_{1,g} + \tilde{\lambda} z_g + \tilde{\gamma} \bar{s}_{1,g}\right) \tag{6}$$

[3] I do allow p_r to vary across years.

where Φ is the CDF for the standard normal distribution. In addition I estimate a second naive model which corrects for reporting bias as described in Section 3.4. In this case:

$$\Pr(r_{1,g} = 1|x_{1,g}, z_g, \bar{s}_{1,g}) = p_r \Phi\left(\tilde{\tilde{\alpha}} + \tilde{\tilde{\beta}} x_{1,g} + \tilde{\tilde{\lambda}} z_g + \tilde{\tilde{\gamma}} \bar{s}_{1,g}\right) \tag{7}$$

The parameter p_r is estimated by the formula:

$$\hat{p}_r = \frac{\frac{1}{n}\sum_{g=1}^{N} r_{1,g}}{\frac{1}{n}\sum_{g=1}^{n} \bar{s}_{1,g}} \tag{8}$$

and the remaining parameters are estimated by maximum likelihood. Because both these naive estimators treat peer behavior as exogenous, the (pseudo) maximum likelihood estimators of $\tilde{\gamma}$ and $\tilde{\tilde{\gamma}}$ are not consistent estimators of the true peer effect γ. These estimates are reported for comparison with the structural estimates, and to demonstrate the cost of ignoring selection and simultaneity.

4.2 Structural estimator: Point estimates

Next, the model is estimated by the simulated maximum likelihood (SML) method developed in Krauth (2002). The vector of parameters to be estimated is $\theta \equiv (p_r, \alpha, \beta, \lambda, \gamma, \mu, \sigma, \rho_x, \rho_\epsilon)$. Observations are indexed by g; observation g in the data set is treated as describing person 1 in group g. For each observation, the GHK simulator (Hajivassiliou, McFadden and Ruud 1996) is used to estimate the conditional likelihood $\Pr(r_{1,g}, \bar{s}_{1,g}|x_{1,g}, z_g; \theta)$. This approximate conditional likelihood is then logged and added up across all observations to give

$$\hat{L}(\theta) \equiv \sum_{g=1}^{n} \ln \hat{\Pr}(r_{1,g}, \bar{s}_{1,g}|x_{1,g}, z_g; \theta) \tag{9}$$

where $\hat{\Pr}$ is the simulation-based estimate of the true probability Pr, and $\hat{L}(\theta)$ is thus a simulation-based estimate of the true log-likelihood function $L(\theta)$. The parameter vector θ is then chosen to maximize the conditional log-likelihood subject to the restrictions:

$$\rho_\epsilon, \rho_x \in \left(\frac{-1}{\max_g(n_g)}, 1\right) \tag{10}$$

$$\gamma \geq 0 \tag{11}$$

$$\hat{p}_r = \frac{\frac{1}{n}\sum_{g=1}^{N} r_{1,g}}{\frac{1}{n}\sum_{g=1}^{n} \bar{s}_{1,g}} \tag{12}$$

$$\rho_\epsilon \quad = \quad \rho_x \equiv \rho \tag{13}$$

Restriction (10) implies that the peer group covariance matrix is positive definite. The nonnegativity constraint on the peer effect in restriction (11) is necessary for computational purposes; this constraint does not bind for this particular data set. However, it should be noted that if γ is exactly zero, the bootstrap method used to estimate the covariance matrix of estimates is inconsistent (Andrews 2000), and the t-statistic for γ has a nonstandard asymptotic distribution. As a result, significance tests for γ are not performed in this paper. Finally equation (13) gives the equal-correlation restriction discussed in Section 3.3.

Krauth (2002) describes the estimation method and its properties in detail, and reports the outcome of various Monte Carlo experiments. Results in that paper include:

1. If the model is correctly specified, the structural estimator of the peer effect is consistent and has much lower bias than the naive probit estimator.
2. Deviations from normality in the explanatory variables do not have a substantial effect on parameter estimates.
3. The equilibrium selection rule has no effect on parameter estimates when the peer effect is zero; in that case equilibrium is always unique. The potential bias due to an incorrectly specified equilibrium selection rule increases gradually in the strength of the peer effect.
4. The variance of the estimator depends strongly and negatively on the explanatory power of the individual-level observables. When these variables have strong explanatory power, the peer effect will be more precisely estimated.

4.3 Structural estimator: Interval estimates

As the equal-correlation assumption (13) reflects a strong and possibly incorrect restriction on the model, one might also want to explore the effect of alternative or weaker restrictions. While some sort of restriction on ρ_ϵ is needed to achieve identification, the structural model allows one to obtain interval estimates of the peer effect under interval restrictions on ρ_ϵ.

Interval estimates are composed as follows. Choose some arbitrary $P \in \left(\frac{-1}{\max_g(n_g)}, 1\right)$. Estimate the model, replacing the equal-correlation restriction (13) with the restriction:

$$\rho_\epsilon = P \tag{14}$$

Define the function $\hat{\gamma}(P)$ as the point estimate of the peer effect γ under the restriction $\rho_\epsilon = P$. This function can then be used to construct interval estimates. Let $[P_0, P_1]$ be an arbitrary closed subinterval of $\left(\frac{-1}{\max_g(n_g)}, 1\right)$. Then:

$$\rho_\epsilon \in [P_0, P_1] \Rightarrow \hat{\gamma} \in \left[\min_{P\in[P_0,P_1]} \hat{\gamma}(P), \max_{P\in[P_0,P_1]} \hat{\gamma}(P)\right] \tag{15}$$

Because the selection of a plausible interval restriction on ρ_ϵ is a subjective matter, in reporting results I report values[4] for the function $\hat{\gamma}(\rho_\epsilon)$ and leave the reader to apply the formula (15).

Another alternative to interval estimation is to estimate both ρ_x and ρ_ϵ under the restriction:

$$\gamma = 0 \tag{16}$$

in place of (13). The resulting estimates answer the question "How high does the within-group correlation in unobservables need to be to explain the observed within-group correlation in behavior without peer effects?"

5 Results

5.1 Point estimates

Table 2 reports point estimates using both naive and structural estimators. Estimated standard errors are calculated for the naive estimators using the inverse Hessian of the log-likelihood function, and calculated for the structural model using the simple bootstrap with 100 replications. In addition to the explanatory variables reported in the table, year effects are included to control for any aggregate shocks such as changes in cigarette taxes. The year effects are treated as aggregate variables (z_g in the model) and all other variables are treated as individual-level characteristics ($x_{i,g}$ in the model).

The first column shows the results from a naive probit estimator using a baseline specification that includes the same set of variables used in the structural model. In particular, the peer effect is calculated using the measure of friend smoking that has been rounded to the nearest

[4]It should be noted that it is only possible to calculate a value of $\hat{\gamma}(P)$ at a finite number of points in the interval, implying that the true minimum or maximum may not be found. However, it can be proved that $\hat{\gamma}(P)$ is continuous, implying that checking a sufficiently fine grid of points in the interval will yield a reasonable approximation to the true minimum and maximum. In addition, experience with the estimator suggests that $\hat{\gamma}(P)$ is strictly decreasing, though that has yet to be proved. If true, this would imply one only needs to calculate $\hat{\gamma}(P_0)$ and $\hat{\gamma}(P_1)$.

sixth for all respondents who report more than six friends, and the number of friends is not included as an explanatory variable. According to the baseline estimates, a young person's smoking has a strong association with the fraction of his or her friends that smoke, and a positive and statistically significant association with the smoking of teachers, parents, and siblings. Youth smoking exhibits significant variation by ethnicity, with African-American youth much less likely to smoke than others. Smoking increases with age and income, and is higher among young people who work (possibly through increased income, increased independence, or exposure to older smokers). Finally, students who report they are doing well in school are less likely to smoke than those who report they are doing poorly. These results are all consistent with previous findings in the literature.

The second column in Table 2 shows the results from the naive probit with the underreporting correction. Note that the sign of all significant coefficients is the same as in the first column, but the magnitude of each significant coefficient is between 10% and 30% higher. This does not imply a great difference in the predictions of the two models: calculated marginal effects on the probability of being a self-reported current smoker are similar with and without the underreporting correction. However, because the structural estimator includes the underreporting correction, parameter estimates from that model should be compared with estimates from the underreporting-corrected naive probit estimator in the second column rather than the uncorrected estimator in the first column.

The third and fourth columns in Table 2 provide a simple robustness check on model specification. They are estimated using the original unadjusted data for the fraction of friends smoking, and include the reported number of friends as an explanatory variable. As the table shows, these two changes have almost no effect on the results. This result provides some confidence that the rounding of the friends smoking variable for computational ease does not materially affect the results.

The fifth column in Table 2 displays parameter estimates from the structural model. The non-peer explanatory variables generally have signs and magnitudes which are consistent with the results from the baseline probit models. However, the measured peer effect has fallen dramatically from 2.789 to 0.235, a decline of over 90%. The reason for this is that the measured selection effect (between-peer correlation in characteristics) is quite high at 0.544.

It is potentially informative at this point to consider how the selection effect and peer effect are simultaneously identified in this model. The estimated selection effect is identified essentially

from the correlation between the respondent's observed characteristics and the average choice of his or her peers. If the respondent's characteristics (as opposed to his or her choice) are strongly associated with the choice of the peers, then this is evidence supporting a high correlation between the respondent's observed characteristics and the characteristics of his or her peers. Because of the assumption of equal correlation in observables and unobservables, this high correlation in observable characteristics implies high correlation in unobservables as well. Once the high correlation in characteristics is considered, the observed correlation in behavior between peers can be explained without a strong peer effect. As a result, the estimated peer effect is much lower than is implied by a naive model without correlation in unobservables.

5.2 Interval estimates

The point estimates from the structural model reported in the previous section were derived by employing the assumption of equal correlation in observable and unobservable characteristics. While this assumption is both plausible and useful for generating point estimates, it may of course be false. Table 3 displays estimates of $(\gamma, \rho_x, \rho_\epsilon)$ from the structural model with the equal correlation restriction (13) replaced by alternative restrictions, as described in Section 4.3. Standard errors are calculated using the simple bootstrap with 30 replications. Table 3 can be used, in combination with the formula in equation (15), to construct interval estimates of the peer effect under user-specified restrictions on the selection effect. For example, suppose that $\rho_\epsilon \in [0.0, 0.544]$, i.e., the correlation in unobservables is positive, and no more than the estimated correlation in observables (from the point estimates). In that case the interval estimate of the peer effect (interpolating linearly between 0.5 and 0.6) would be approximately $[1.548, 0.185]$. Note that even with the assumption of no correlation in unobservables, the estimated peer effect of 1.548 is nearly 45% lower than the estimated peer effect of 2.789 from the underreporting-corrected naive probit. This is because the structural model eliminates one source of bias in the naive probit, the simultaneity of choice among peers. In other words almost half of the apparent peer effect is the result of simple simultaneity bias. At the same time, a peer behavior coefficient of 1.548 still indicates a nontrivial peer effect.

Figure 3 summarizes the results of the paper graphically by plotting the estimated peer effect (γ) against the assumed or estimated selection effect (ρ_ϵ). The two points on the left side of the graph represent the naive estimates (which implicitly assume no selection effect). The point in the center represents the point estimate from the structural model, and the dotted ellipse surrounding

it is a 95% joint asymptotic confidence region. The solid line represents the estimated $\hat{\gamma}(\rho_\epsilon)$ function, and the dotted lines represent a pointwise 95% asymptotic confidence band.

5.3 Interpretation and comparison to previous findings

To give the estimated coefficients a more intuitive interpretation which is comparable across models, Table 4 reports the results from a simple thought experiment. Consider a representative individual who has four close friends (the median in the CYTS data) who are all nonsmokers, and has a set of observed characteristics such that the model predicts his or her probability of self-reporting as a smoker to be exactly 13.177%, the mean in the CYTS data. Suppose that 25% of the relevant peer group (e.g., one of the four friends) suddenly switches from a non-smoker to a smoker. By how many percentage points does the model predict the representative individual's probability of smoking[5] would increase? That quantity ($\Delta \Pr(y = 1)$), for both the models estimated here and for several other recent studies, is reported in Table 4.

The estimates in Table 4 come from researchers in economics, public health, and psychology, and employ a variety of methods and definitions of peer group. However, a few common patterns appear across the studies. First, peer effects are generally found to be substantial, in some cases implausibly so. Second, close friends appear to be more influential than larger peer groups such as schoolmates or neighbors. However, it should be noted that the simultaneity bias in naive estimators is likely to be larger with smaller peer groups. Third, those researchers who have employed instrumental variables methods (Gaviria and Raphael, Norton et al., and Powell et al.) have tended to find that these methods have very little impact on the estimated peer effect. Given the similarity in choice of instruments among the three studies, the consistency of this finding strongly suggests either that there is no correlation in unobservable characteristics among

[5] For standard probit or logit models with coefficient on peer smoking β, I calculate this as:

$$\Delta \Pr(y = 1) = F(F^{-1}(0.13177) + \beta * 0.25) - 0.13177$$

where F is the logistic CDF, standard normal CDF, or identity function, depending on whether the researcher is estimating a logit, probit, or linear probability model respectively. In two papers (Lloyd-Richardson and Wang) the number of friends smoking is treated as a categorical variable, in which case β is the coefficient on the binary variable "one friend smokes", with "no friends smoke" being the excluded category, and:

$$\Delta \Pr(y = 1) = F(F^{-1}(0.13177) + \beta) - 0.13177$$

Finally, for the models with underreporting corrections in this paper and in Krauth (2004), it is also necessary to include the estimated reporting rate (r):

$$\Delta \Pr(y = 1) = r\Phi(\Phi^{-1}(0.13177/r) + \beta * 0.25) - 0.13177$$

schoolmates and neighbors, or that the instruments are invalid.

In contrast, the two reported sets of estimates using the structural method show a substantial decline in the measured peer effect. In addition to the results for this study, I report the results from a related study of youth smoking in Canada (Krauth 2004). In both cases, the structural model reports much lower peer effects. In the Canada data, the structural model estimate of the peer effect remains economically significant: the effect of one of the representative individual's friends becoming a smoker is to increase the representative individual's probability of smoking by 5.8 percentage points. In the California data, the decline in the estimated peer effect is much larger: the representative individual's probability of smoking increases by only 1.1 percentage points.

6 Conclusion

A variety of different approaches to youth tobacco control efforts compete for public and philanthropic resources. Some, including tax increases or access and use restrictions aim at affecting incentives. Others attempt to change social norms about smoking through advertising, celebrity endorsement of nonsmoking, in-school antismoking programs, etc. Interventions to change social norms are based in part on the belief that social influences have a great deal to do with a young person's decision to smoke. If more careful analysis of the data overturn this belief, the implications for tobacco control policy are substantial.

The structural estimation approach applied in this paper provides a complementary methodology to that employed in previous studies. I model members of a peer group as having correlated observed and unobserved characteristics, as would be implied by nonrandom peer selection, and I model the interdependent choices of respondents and their peers as equilibrium phenomena. Applied to data on youth smoking in California, I find that there is a high degree of correlation in characteristics between groups of close friends, and that this correlation explains over 90% of the apparent peer effect.

This result is consistent with those from longitudinal studies of young people, which generally find that smokers tend to acquire smoking friends over time to a much greater extent than nonsmokers tend to become smokers after being friends with smokers. However, it is inconsistent with those studies using instrumental variables, which generally find no evidence for endogeneity bias in naive estimators of peer effects. The apparent consensus of studies using a given method and disagreement between studies using different methods is a clear puzzle in this literature. It

might be of interest in the future to apply all three methods to a single data set, for example the National Longitudinal Study of Adolescent Health, that features both detailed school-based samples (necessary for the IV approach) and longitudinal data (necessary for the panel approach), to see more clearly why the different methods yield such different conclusions. Alternatively, it may be useful to investigate follow-up data from experimental program evaluations of successful antismoking programs, using the methodology of Boozer and Cacciola (2001): if the program is given to randomly selected classes in a given grade, and students from the treatment and control groups are exogenously mixed in subsequent grades, the fraction of classmates in the treatment group can be a valid instrumental variable for peer smoking. In any case, the conflict in findings remains an open puzzle.

An additional avenue for future research using this method is to estimate peer effects for alternative peer groups, such as classrooms, schools, and neighborhoods, as well as extending the model to include the simultaneous influence of more than one type of group. Although it is valuable to know whether peers are influential in the smoking decisions of young people, design of effective peer-based tobacco control programs will tend to depend critically on identifying exactly which peers are influential.

References

Alexander, Cheryl, Marina Piazza, Debra Mekos, and Thomas Valente, "Peers, schools, and adolescent cigarette smoking," *Journal of Adolescent Health*, 2001, *29*, 22–30.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," Working Paper 7831, NBER 2000.

Andrews, Donald W. K., "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space," *Econometrica*, 2000, *68* (2), 399–406.

Boozer, Michael A. and Stephen E. Cacciola, "Inside the black box of Project STAR: Estimation of peer effects using experimental data," Discussion Paper 832, Yale University Economic Growth Center 2001.

Brock, William A. and Steven N. Durlauf, "Discrete choice with social interactions," *Review of Economic Studies*, 2001, *68* (2), 235–260.

California Department of Health Services, "California Youth Tobacco Survey: SAS Dataset Documentation and Technical Report, 1994-2002," Technical Report, Cancer Surveillance Section 2003.

Centers for Disease Control and Prevention, *Tobacco Control State Highlights 2002: Impact and Opportunity*, US Department of Health and Human Services, 2002.

___ , "Trends in Cigarette Smoking Among High School Students — United States, 1991–2001," *Morbidity and Mortality Weekly Report*, 2002, *51* (19), 409–412.

___ , "Tobacco Use Among Middle and High School Students — United States, 2002," *Morbidity and Mortality Weekly Report*, 2003, *52* (45), 1096–1098.

DeCicca, Philip, Donald Kenkel, and Alan Mathios, "Putting out the fires: Will higher taxes reduce the onset of youth smoking?," *Journal of Political Economy*, 2002, *110* (1), 144–169.

Engels, Rutger C.M.E., Ronald A. Knibbe, Maria J. Drop, and Ypie T. de Haan, "Homogeneity of cigarette smoking within peer groups: Influence or selection?," *Health Education and Behavior*, 1997, *24*, 81–811.

Gaviria, Alejandro and Steven Raphael, "School-based peer effects and juvenile behavior," *Review of Economics and Statistics*, 2001, *83* (2), 257–268.

Gruber, Jonathan and Jonathan Zinman, "Youth smoking in the U.S.: Evidence and implications," in Jonathan Gruber, ed., *Risky Behavior Among Youths: An Economic Analysis*, University of Chicago Press, 2001.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud, "Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results," *Journal of Econometrics*, 1996, *72* (1), 85–134.

Hoxby, Caroline M., "Peer effects in the classroom: Learning from gender and race variation," Working Paper 7867, NBER 2000.

Kenkel, Donald S., Robert R. Reed III, and Ping Wang, "Rational addiction, peer externalities, and long run effects of public policy," Working Paper 9249, NBER 2002.

Kooreman, Peter, "Estimation of econometric models of some discrete games," *Journal of Applied Econometrics*, 1994, *9*, 255–268.

__ **and Adriaan Soetevent**, "A discrete choice model with social interactions: An analysis of high school teen behavior," Working Paper, University of Groningen 2002.

Krauth, Brian V., "Simulation-based estimation of peer effects," Working Paper, Simon Fraser University 2002.

__ , "Peer and selection effects on smoking among Canadian youth," Working Paper, Simon Fraser University 2004.

Lloyd-Richardson, Elizabeth E., George Papandonatos, Alessandra Kazura, Cassandra Stanton, and Raymond Niaura, "Differentiating stages of smoking intensity among adolescents: Stage-specific psychological and social influences," *Journal of Consulting and Clinical Psychology*, 2002, *70* (4), 998–1009.

Mackay, Judith and Michael Eriksen, *The Tobacco Atlas*, World Health Organization, 2002.

Manski, Charles F., "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 1993, *60* (3), 531–542.

Norton, Edward C., Richard C. Lindrooth, and Susan T. Ennett, "Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use," *Health Economics*, 1998, *7*, 439–453.

__ **,** __ **, and** __ , "How measures of perception from survey data lead to inconsistent regression results: Evidence form adolescent and peer substance use," *Health Economics*, 2003, *12*, 139–148.

Orzechowski and Walker, *The Tax Burden on Tobacco: 2002*, Washington: Orzechowski and Walker, 2002.

Powell, Lisa M., John A. Tauras, and Hana Ross, "Peer Effects, Tobacco Control Policies, and Youth Smoking Behavior," Working Paper, University of Illinois - Chicago 2003.

Sacerdote, Bruce, "Peer effects with random assignment: Results for Dartmouth roommates," *Quarterly Journal of Economics*, May 2001, *116* (2), 681–704.

Tamer, Elie T., "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 2002. Forthcoming.

Tyas, Suzanne L. and Linda L. Pederson, "Psychosocial factors related to adolescent smoking: A critical review of the literature," *Tobacco Control*, 1998, *7*, 409–420.

Wagenknecht, K.E., G.L. Burke, L.L. Perkins, N.J. Haley, and G.D. Friedman, "Misclassification of smoking status in the CARDIA study: a comparison of self-report with serum cotinine levels," *American Journal of Public Health*, 1992, *82*, 33–36.

Wang, Min Qi, Eugene C. Fitzhugh, R. Carl Westerfield, and James M. Eddy, "Family and peer influences on smoking behavior among American adolescents: An age trend," *Journal of Adolescent Health*, March 1995, *16*, 200–203.

__ **, James M. Eddy, and Eugene C. Fitzhugh**, "Smoking acquisition: Peer influence and self-selection," *Psychological Reports*, 2000, *86*, 1241–1246.

A Tables and Figures

	Included		Dropped	
Variable name	Mean	Std.Dev	Mean	Std.Dev.
Current smoker	0.132	0.338	–	–
Fraction of friends smoking	0.213	0.308	–	–
Number of friends	4.043	1.370	–	–
Age	15.429	1.104	15.360	1.108
Teacher's smoking (1-3 scale)	2.162	0.700	2.092	0.719
Discussed health risks in class	0.772	0.416	0.764	0.421
At least one parent smokes	0.327	0.469	0.321	0.467
At least one sibling smokes	0.173	0.378	0.158	0.365
Male	0.503	0.500	0.537	0.499
African-American	0.065	0.247	0.082	0.275
Hispanic	0.364	0.481	0.382	0.486
Asian/other	0.092	0.289	0.090	0.286
US born	0.868	0.338	0.853	0.355
Weekly disposable income, $	29.982	40.204	30.038	41.220
Paid job	0.263	0.440	0.231	0.421
Above average in school	0.560	0.493	0.563	0.492
Below average in school	0.031	0.173	0.039	0.191
# observations	13,037		841	

Table 1: Summary statistics for 1994-2002 CYTS data. Columns labeled "Included" describe data included in estimation, columns labeled "Dropped" describe data dropped because of missing data on endogenous variables (current smoking and/or friend smoking).

Variable Name	Naive probit				Structural Estimator
	Baseline Specification		Alternate Specification		
Selection effect (ρ)	–	–	–	–	0.544
					(0.058)
Peer effect (γ)	1.993	2.789	1.991	2.789	0.235
	(0.049)	(0.086)	(0.049)	(0.087)	(0.180)
Age	1.897	2.187	1.879	2.173	0.757
	(0.539)	(0.723)	(0.539)	(0.724)	(0.407)
Age^2	-0.056	-0.064	-0.055	-0.063	-0.016
	(0.017)	(0.023)	(0.017)	(0.023)	(0.013)
Teacher smoking	0.102	0.126	0.101	0.126	0.241
	(0.026)	(0.035)	(0.026)	(0.035)	(0.023)
Discuss health risks in class	-0.027	-0.017	-0.028	-0.018	-0.079
	(0.042)	(0.057)	(0.042)	(0.057)	(0.034)
Parent smokes	0.079	0.093	0.079	0.093	0.272
	(0.035)	(0.049)	(0.035)	(0.049)	(0.036)
Sibling smokes	0.393	0.535	0.392	0.535	0.589
	(0.039)	(0.056)	(0.039)	(0.056)	(0.046)
Male	0.005	-0.038	0.006	-0.031	0.062
	(0.034)	(0.046)	(0.034)	(0.046)	(0.033)
African-American	-0.649	-0.806	-0.649	-0.807	-0.550
	(0.091)	(0.118)	(0.091)	(0.118)	(0.072)
Hispanic	-0.132	-0.142	-0.132	-0.142	-0.094
	(0.039)	(0.052)	(0.039)	(0.052)	(0.038)
Asian/Other	-0.155	-0.166	-0.155	-0.170	-0.123
	(0.066)	(0.088)	(0.066)	(0.088)	(0.053)
U.S. born	0.151	0.177	0.149	0.171	0.238
	(0.059)	(0.077)	(0.059)	(0.077)	(0.055)
Disposable income (100$)	0.151	0.193	0.151	0.196	0.340
	(0.041)	(0.059)	(0.041)	(0.059)	(0.040)
Paid job	0.173	0.225	0.172	0.227	0.118
	(0.040)	(0.055)	(0.040)	(0.055)	(0.036)
Above avg in school	-0.249	-0.322	-0.248	-0.323	-0.511
	(0.036)	(0.048)	(0.036)	(0.048)	(0.035)
Below avg in school	0.448	0.719	0.444	0.709	0.743
	(0.078)	(0.128)	(0.078)	(0.128)	(0.104)
# of friends	–	–	0.000	-0.005	–
			(0.003)	(0.005)	
Year effects?	Yes	Yes	Yes	Yes	Yes
Underreporting correction?	No	Yes	No	Yes	Yes

Table 2: Regression results, with estimated standard errors in parentheses.

Identifying Restriction	$\hat{\gamma}$	$\hat{\rho}_x$	$\hat{\rho}_\epsilon$
$\rho_\epsilon = \rho_x$	0.235	0.544	0.544
	(0.180)	(0.058)	(0.058)
$\rho_\epsilon = 0.0$	1.548	0.455	–
	(0.026)	(0.023)	
$\rho_\epsilon = 0.1$	1.245	0.499	–
	(0.027)	(0.022)	
$\rho_\epsilon = 0.2$	0.978	0.529	–
	(0.026)	(0.022)	
$\rho_\epsilon = 0.3$	0.736	0.552	–
	(0.028)	(0.023)	
$\rho_\epsilon = 0.4$	0.503	0.575	–
	(0.029)	(0.021)	
$\rho_\epsilon = 0.5$	0.281	0.596	–
	(0.030)	(0.023)	
$\rho_\epsilon = 0.6$	0.063	0.614	–
	(0.035)	(0.027)	
$\gamma = 0.0$	–	0.578	0.669
		(0.099)	(0.092)

Table 3: Estimated peer/selection effect under alternative identifying restrictions. Estimated standard errors in parentheses.

Source (peer group)	Estimation Method	Implied $\Delta \Pr(smoke)$
Alexander et al., 2001		
(school)	Naive (logit)	24.2%
Gaviria and Raphael, 2001		
(schools)	Naive (LPM)	4.0%
	IV (2SLS LPM)	3.9%
Lloyd-Richardson et al., 2002		
(3 best friends)	Naive (logit)	28.4%
Norton et al.,1998		
(neighborhood)	Naive (probit)	25.4% - 54.5%
	IV (2-stage probit)	26.4% - 86.0%
Norton et al., 2003		
(3 best friends)	Naive (LPM)	13.2%
(school)	Naive (LPM)	9.9%
Powell et al., 2003		
(school)	Naive (probit)	13.9%
	IV (AGLS probit)	13.5%
Wang et al., 1995		
(4 best friends)	Naive (logit)	31.4% - 48.7%
Krauth (2004)		
(close friends)	Naive (probit)	15.7%
	Naive w/underrep.	18.5%
	Structural, no corr.	8.4%
	Structural, equal corr.	5.8%
This paper		
(close friends)	Naive (probit)	13.6%
	Naive w/underrep.	15.3%
	Structural, no corr.	7.9%
	Structural, equal corr.	1.1%

Table 4: Comparison of estimated peer effects in this paper and in previous studies. To facilitate comparison across different models, results are stated in terms of the increase in probability that a representative individual will be a (self-reported) smoker in response to a 25% increase in peer smoking. See text for details.

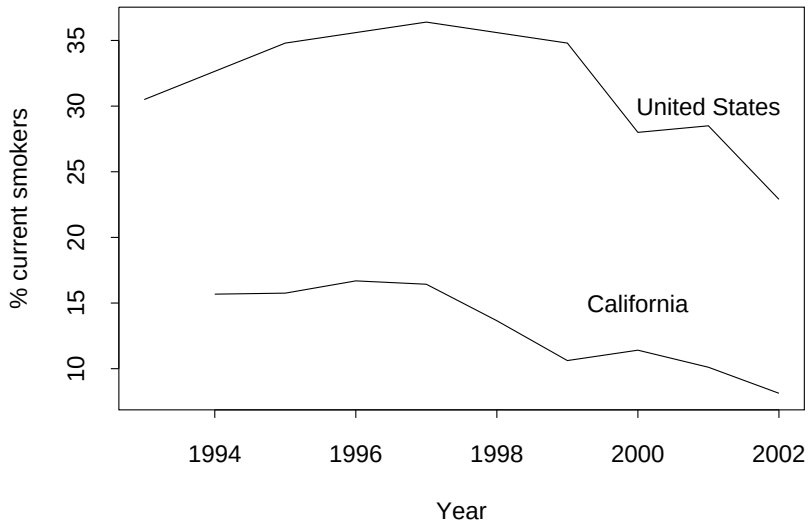

Figure 1: Rates of current smoking among young people in California and entire United States, 1993-2002. California smoking rates based on author's calculations using CYTS, national smoking rates based on CDC estimates.

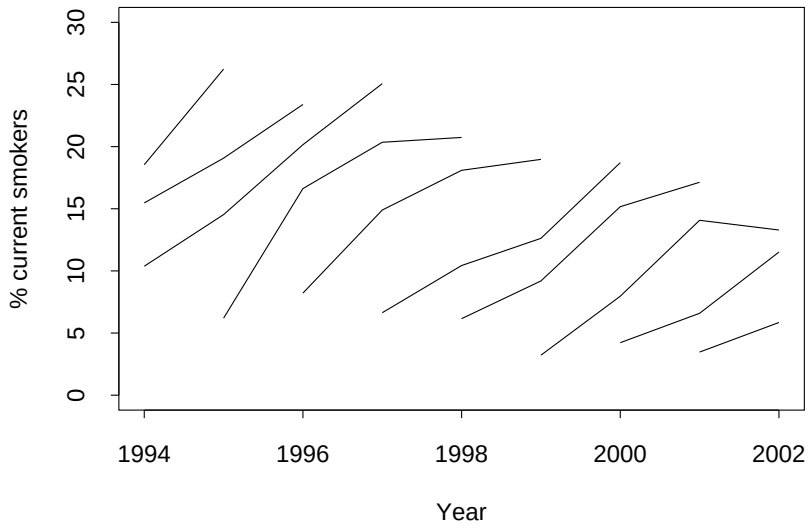

Figure 2: Rates of current smoking among young people in California by year and birth cohort.

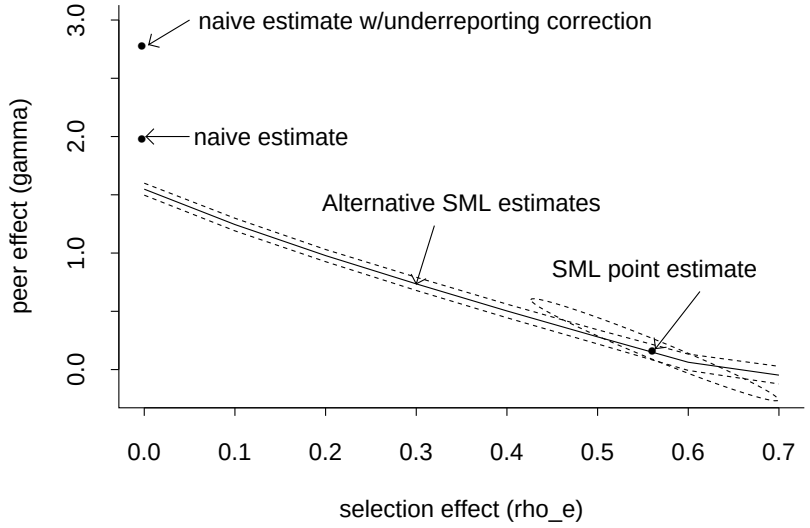

Figure 3: Estimated peer effect (γ) for alternative assumptions about selection effect (ρ_ϵ). Individual points represent point estimate from naive probit model or baseline structural model with equal-correlation assumption. Solid line represents alternative estimates of γ as a function of the assumed value of ρ_ϵ. Dashed lines depict 95% confidence intervals or ellipses.

B Notes for referees (not for publication)

This section describes in greater detail the construction of Table 4. Because of differences in estimation methods across studies, and differences in standard reporting methods between researchers in health and in economics, some degree of subjective interpretation is involved. This section describes how each paper was interpreted to generate the estimate in Table 4.

B.1 Alexander et al. 2001

1. Data: Grade 7-12 students in the U.S. National Longitudinal Study of Adolescent Health.
2. Peer group measure: Smoking rate of other students in school (also reported for peer network, best friend, etc., but not in easily translated form).
3. Method: Logistic regression, odds ratios reported.
4. Estimates: Estimated odds ratio associated with a 10% increase in school smoking is 1.73 (Table 2, column 1). This implies that the logit coefficient on school smoking is $\frac{\ln 1.73}{0.10} = 5.481$. Predicted effect of a 25% increase in school smoking is:

$$\begin{aligned} \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.131778) + 5.481 * 0.25) - 0.131778 \\ &= 0.2422265 \end{aligned}$$

B.2 Gaviria and Raphael 2002

1. Data: Grade 10 students in the U.S. National Educational Longitudinal Survey.
2. Peer group measure: Smoking rate of other students in school.
3. Method: Linear probability models, estimated through OLS and IV.
4. Estimates: OLS estimate of LPM coefficient is 0.158 (Table 3, column 4). 2SLS/IV estimate is 0.156 (Table 5, column 4). Given that we have a linear probability model:

$$\begin{aligned} \Delta \Pr(s=1) &= 0.158 * 0.25 \qquad \text{for OLS estimator} \\ &= 0.0395 \\ \Delta \Pr(s=1) &= 0.156 * 0.25 \qquad \text{for 2SLS estimator} \\ &= 0.0390 \end{aligned}$$

B.3 Lloyd-Richardson et al. 2002

1. Data: Grade 7-12 students in the U.S. National Longitudinal Survey of Adolescent Health.
2. Peer group measure: Smoking rate among 3 best friends.
3. Method: Logistic regression, odds ratios reported.
4. Estimates: Odds ratio for 1 peer who smokes vs. 0 peers who smoke was 4.68 (Table 2, column 2), implying logit coefficient of $\ln(4.68) = 1.543$. Effect of a one person (33% in this case) increase in best friend smoking is:

$$\begin{aligned} \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.131778) + 1.543) - 0.131778 \\ &= 0.284 \end{aligned}$$

B.4 Norton et al. 1998

1. Data: Grade 6 and grade 8 students in a U.S. longitudinal study of the DARE anti-drug use program.
2. Peer group measure: Smoking rate of students who had attended the same elementary school. The authors argue that this estimates neighborhood effects rather than school effects, because the students are no longer in elementary school.
3. Method: Probit regression, with and without IV.
4. Estimates: Naive coefficient is 6.31 (Table 3, column 1) in grade 6, 3.31 (Table 3, column 3) in grade 8. 2-stage coefficient is 14 (Table 5, column 3) in grade 6, 3.42 (Table 5, column 4) in grade 8. Effect of a 25% increase in school smoking is:

$$\begin{aligned}
\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 6.31 * 0.25) - 0.131778 \\
&= 0.545 \\
\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 3.31 * 0.25) - 0.131778 \\
&= 0.254 \\
\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 14 * 0.25) - 0.131778 \\
&= 0.860 \\
\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 3.42 * 0.25) - 0.131778 \\
&= 0.264
\end{aligned} \tag{17}$$

B.5 Norton, et al. 2003

1. Data: Grade 9 students in a North Carolina study, grade 5 and 6 students in an Illinois study.
2. Peer group measure: Smoking rate of 3 best friends (North Carolina); school (Illinois).
3. Method: Linear probability models. The researchers also included perceived smoking of peers (as distinct from their actual smoking) as an explanatory variable.
4. Estimates: For North Carolina best friends, LPM coefficient is 0.526 (Table 1, column 1), for Illinois school, LPM coefficient is 0.397 (Table 1, column 2). Effect of a 25% increase in peer smoking is :

$$\begin{aligned}
\Delta \Pr(y=1) &= 0.526 * 0.25 \\
&= .1315 \\
\Delta \Pr(y=1) &= 0.397 * 0.25 \\
&= .0993
\end{aligned}$$

B.6 Powell, et al. 2003

1. Data: High school students in the U.S. Study of Smoking and Tobacco Use Among Young People.
2. Peer group: Smoking rate of other students in school.
3. Method: Probit regression, also probit with instrumental variables (estimated by AGLS). Marginal effects rather than probit coefficients are reported.

4. Estimates: Marginal effect from simple probit was 0.5573 (Table 2, column 1), from AGLS estimator was 0.5385 (Table 2, column 2). Effect of a 25% increase in school smoking is:

$$\begin{aligned}\Delta \Pr(y=1) &= 0.5573 * 0.25 \\ &= 0.139 \\ \Delta \Pr(y=1) &= 0.5385 * 0.25 \\ &= 0.135\end{aligned}$$

B.7 Wang et al. 1995

1. Data: Youth age 14-18 in the U.S. Teenage Attitudes and Practices Survey.
2. Peer group measure: Smoking rate of four same-sex best friends (0 smokers vs. 1-2 smokers vs. 3-4 smokers).
3. Method: Logistic regression, separately for each age, odds ratios reported.
4. Estimates: Odds ratio of 1-2 smokers vs 0 smokers among peers varies from 5.3 (Table 1, column 3) to 10.7 (Table 1, column 2). Therefore logit coefficient (on dummy variable "1-2 smokers among friends") varies from 1.66707 to 2.370244. Effect of an increase in best-friend smoking from zero to one is:

$$\begin{aligned}\Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.131778) + 1.66707) - 0.131778 \\ &= 0.3140311 \\ \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.131778) + 2.370244) - 0.131778 \\ &= 0.4871306\end{aligned}$$

B.8 Krauth 2004

1. Data: Youth age 15-19 in the Canada Youth Smoking Survey.
2. Peer group measure: Smoking rate of close friends (size of peer group self-reported).
3. Method: Naive probit with and without underreporting correction, structural probit.
4. Estimates: Coefficient from naive probit without underreporting correction 2.240 (Table 2, column 1), with correction 3.626 (Table 2, column 2), from structural model 1.251 (Table 2, column 3). Coefficient from structural model with no sorting ($\rho_\epsilon = 0$) is 1.745 (Table 4, column 1). Estimated reporting rate is 0.544. Effect of a 25% increase in peer smoking is: Effect of a 25% increase in peer smoking is:

$$\begin{aligned}\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 2.240 * 0.25) - 0.131778 \\ &= 0.157 \\ \Delta \Pr(y=1) &= 0.544 * \Phi(\Phi^{-1}(0.131778/0.544) + 3.626 * 0.25) - 0.131778 \\ &= 0.185 \\ \Delta \Pr(y=1) &= 0.544 * \Phi(\Phi^{-1}(0.131778/0.544) + 1.251 * 0.25) - 0.131778 \\ &= 0.058 \\ \Delta \Pr(y=1) &= 0.544 * \Phi(\Phi^{-1}(0.131778/0.544) + 1.745 * 0.25) - 0.131778 \\ &= 0.084\end{aligned}$$

B.9 Krauth (this paper)

1. Data: Youth age 14-17 in the 1994-2002 California Youth Tobacco Survey.
2. Peer group measure: Smoking rate of same-sex close friends (size of peer group self-reported).

3. Method: Naive probit with and without underreporting correction, structural probit.

4. Estimates: Coefficient from naive probit without underreporting correction 1.993 (Table 2, column 1), with correction 2.789 (Table 2, column 2), from structural model 0.235 (Table 2, column 5). Coefficient from structural model with no sorting ($\rho_\epsilon = 0$) is 1.548 (Table 3, column 1). Estimated reporting rate is 0.618. Effect of a 25% increase in peer smoking is:

$$\begin{aligned}
\Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.131778) + 1.993 * 0.25) - 0.131778 \\
&= 0.136 \\
\Delta \Pr(y=1) &= 0.618 * \Phi(\Phi^{-1}(0.131778/0.618) + 2.789 * 0.25) - 0.131778 \\
&= 0.153 \\
\Delta \Pr(y=1) &= 0.618 * \Phi(\Phi^{-1}(0.131778/0.618) + 0.235 * 0.25) - 0.131778 \\
&= 0.011 \\
\Delta \Pr(y=1) &= 0.618 * \Phi(\Phi^{-1}(0.131778/0.618) + 1.548 * 0.25) - 0.131778 \\
&= 0.079
\end{aligned}$$