Peer effects and selection effects in youth smoking

Brian Krauth
Simon Fraser University*

January 31, 2003

Abstract

A number of studies have indicated that peer behavior is a highly influential factor in a youth's decision to smoke, and public health policymakers take these results seriously. However, these results are suspect because most fail to account for selection effects. This paper develops an econometric model of youth smoking which incorporates both peer effects and selection effects. 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 peers are substantially less influenced by their peers than reduced-form models suggest.

1 Introduction

Youth smoking is a major concern in public health. The US Centers for Disease Control and Prevention (2002a) estimate that the negative health consequences of smoking have an annual economic cost of $157 billion, including 440,000 premature deaths per year. Governments often spend large sums on reducing tobacco use, with particular focus on reducing youth smoking. In 2002, the US states alone spent over $861 million[1] on tobacco control, in addition to a substantial federal expenditure. As a

*This paper is preliminary and incomplete. Please do not cite or distribute without permission. In particular, all estimates are preliminary - the estimation technique developed and used in this paper is computationally intensive, and current estimates are calculated using a relatively low number of simulations. More accurate estimation involving more simulations is proceeding and will appear in later drafts of this paper.

[1]Author's calculation from CDC (2002b) data.

result, there is a sizable literature in public health, as well as in economics, that investigates the determinants of youth smoking. One of the most robust findings is that peer smoking has strong predictive power for an young person's decision to smoke. For example, a recent review article (Tyas and Pederson 1998, p. 413) states: "peer smoking [is] consistently found to be related to adolescent smoking initiation, maintenance, and intentions."

However, these estimates are suspect because most treat peer behavior as an exogenous variable and fail to account for selection and simultaneity bias. Selection bias arises because individuals tend to choose peers with similar characteristics, while simultaneity bias arises because the choice of each member of a peer group potentially affects the choice of every other member. Both selection and simultaneity induce positive correlation between a person's unobservables and the behavior of his or her peers, leading to an upwardly biased estimate of peer influence. Empirical researchers have been aware of these issues for some time, but have met with limited success in solving them. This paper develops a structural econometric model which explicitly accounts for both selection and simultaneity. Simultaneity is addressed by treating the group outcome as an endogenous variable in the likelihood function, while selection is addressed by explicitly allowing correlation in unobservables between peers. The model is identified by using the between-peer correlation in observables as a proxy for the correlation in unobservables. This strong identifying assumption provides point estimates, but weaker assumptions can be employed to generate interval estimates.

Peer effects are estimated using data from the 1994 Canadian Youth Smoking Survey (YSS). Estimates from a naive probit model suggest a very large peer effect. Estimates from the model with selection and simultaneity indicate a lower but still positive effect.

1.1 Related literature

To be added.

2 Data

The primary data source is the 1994 Youth Smoking Survey (YSS), a national survey of Canadian youth aged 10-19. It has the advantage over the major US-based surveys that survey respondents are asked to report the number of their friends that smoke. It also has information on province of residence, which is important because many important policies such as excise taxes and

restrictions on smoking are set at the provincial level. The YSS has two components, a school-based sample of 10 to 14 year olds, and a household-based sample of 15 to 19 year olds. I restrict attention to the household sample, which has 9,491 observations of which 9,210 are used in estimation.

The measures of own smoking and peer smoking are constructed as follows. As is standard in smoking research, a respondent is defined as a "current smoker" if he or she has smoked at least one cigarette in the past 30 days and over 100 cigarettes in his or her lifetime. Respondents are also asked both how many close friends they have and how many of them smoke. From the raw data, 281 observations are dropped because the respondent reports having no friends, or fails to report his or her own smoking behavior or friends' smoking behavior. Respondents can claim up to 95 close friends; to keep computational time reasonable, those respondents who claim more than 10 close friends are coded as having 10 close friends and the fraction of their close friends who smoke is rounded to the nearest tenth. This recoding affects 1144 observations. For all other variables, missing values are replaced with the sample mean. Summary statistics for the data are reported in Table 1.

Despite efforts by survey collectors to emphasize and ensure confidentiality, underreporting is a common concern with self-reported data on youth smoking. Audit studies (Wagenknecht, Burke, Perkins, Haley and Friedman 1992) which compare self-reported smoking with blood tests find that underreporting is substantial. Although these studies cannot directly provide a measure of the underreporting rate, a rough estimate can be constructed by looking at the relative frequency of self-reported smoking and the fraction of friends that the respondent reports as smokers. In the YSS sample, 23.1% of respondents are current smokers, whereas 42.5% of their friends are smokers. Assuming that each respondent's friends are also in the sample universe, that respondents truthfully report the behavior of their friends, and that smokers and nonsmokers have the same number of friends, this implies that 46% of smokers falsely claim to be nonsmokers. The econometric model described in Section 3 includes an adjustment for underreporting.

3 Model

The econometric model is similar in spirit to the canonical model of discrete choice with social interaction effects due to Brock and Durlauf (2001). 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 members, where n_g may vary across groups. Within each group, individuals are indexed by i, so that the pair i, g identifies an individual. Each member of a 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 incremental utility from smoking is given by:

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

where $x_{i,g}$ is a vector of individual-level exogenous variables, z_g is a vector of group-level exogenous variables, $\epsilon_{i,g}$ is an unobserved individual-level term, and $\bar{s}_{i,g}$ is the fraction of the other group members that smoke:

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

Because of conflict between the average self-reported smoking rate and reported smoking rates of friends, it is necessary to define a separate variable for a person's self-reported behavior and treat the person's actual behavior as an unobservable. Because the data can provide little guidance on the determinants of false reports, I model underreporting as simply as possible. Let $r_{i,g}$ indicate whether a person self-reports as a smoker. Then

$$r_{i,g} = \begin{cases} s_{i,g} & \text{with probability } p_r \\ 0 & \text{with probability } 1 - p_r \end{cases} \tag{3}$$

where p_r is a parameter. Note that conditional on $s_{i,g}$, $r_{i,g}$ is independent of all other variables. While differences in underreporting across socioeconomic groups may be of interest, the data simply cannot address that issue.

3.1 Equilibrium

Given the exogenous variables, the endogenous variables are given by 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 take on the values associated with the lowest-activity Nash equilibrium for the given exogenous variables. The low-activity Nash equilibrium is the steady state of a dynamic process in which agents start as nonsmokers and update their choices myopically in response to the previous-period choices of their peers. Given that youth are

born as nonsmokers and rarely quit smoking until adulthood, such a dynamic process may be a reasonable model of behavior. Regardless of the relative merits of the low-activity rule, Monte Carlo results reported in Krauth (2002) imply that for low or moderate peer effects, misspecification in the equilibrium selection rule does not have a large impact on the resulting estimates.

3.2 Selection on observables and unobservables

The distributional assumptions for exogenous variables are similar in spirit to those in the standard probit model. In particular, for all individuals i, g we have $\epsilon_{i,g}|x_{i,g} \sim N(0,1)$. Unlike a simple probit model, there is correlation in exogenous variables between members of the same peer group. This feature in the model allows for the possibility that young people are more likely to make friends with people of similar backgrounds and preferences. Most previous analysis of peer effects in youth smoking implicitly assumes random selection of friends so that there is no correlation in characteristics.

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. Instead, selection is modeled through a reduced-form correlation in exogenous variables between members of the same peer group. This correlation can be interpreted as the outcome of a more complex process of group formation. Specifically, the joint distribution of the individual-level exogenous variables is multivariate normal across the group members. For a group of size $n_g = 3$:

$$\begin{bmatrix} \beta x_{1,g} \\ \beta x_{2,g} \\ \beta x_{3,g} \\ \epsilon_{1,g} \\ \epsilon_{2,g} \\ \epsilon_{3,g} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \\ \mu \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho_x\sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \rho_x\sigma^2 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_\epsilon & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & 1 & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & \rho_\epsilon & 1 \end{bmatrix} \right) \tag{4}$$

with the distribution being defined similarly for other values of n_g. As is standard in probit models, the utility function is normalized so that ϵ has a mean of zero and a variance of one. It is also standard for ϵ_i to be independent of x_i. In the same spirit, it is also assumed that (x_i is independent of ϵ_j for all i, j). Finally, the distribution is symmetric across group members for the obvious reason that the ordering of group members is arbitrary.

In order to find point estimates, it is necessary to impose an additional restriction on ρ_ϵ, the between-peer correlation in unob-

servables. The primary restriction used here is that the correlation is the same i.e., that $\rho_\epsilon = \rho_x$. The idea of using the degree of selection on observables as a proxy for the degree of selection on unobservables has been used by Altonji, Elder, and Taber (2000) to correct for selection effects in measuring the effect of attending a Catholic school. They 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 particularly salient to potential friends. Fortunately, the model can also be estimated under alternative restrictions on ρ_ϵ, including interval restrictions. Section 5.2 reports point estimates under the equal-correlation assumption, while Section 5.3 reports estimates under alternative assumptions.

4 Estimation

The model is estimated by the simulated maximum likelihood method developed in Krauth (2002). The vector of parameters to be estimated is $\theta \equiv (p_r, \alpha, \beta, \lambda, \gamma, \sigma, \rho_x)$. Observations are indexed by g; observation g is treated as person 1 in group g. For each observation, simulation is used to estimate the conditional likelihood $\Pr(s_{1,g}, \bar{s}_{1,g} | x_{1,g}, z_g; \theta)$. This conditional likelihood is then logged and added up across all observations to give

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

The parameter vector θ is then varied to maximize this conditional log-likelihood.

Krauth (2002) describes the simulation procedure in detail, and reports the results from various Monte Carlo experiments. These results indicate some basic properties of the SML estimator. First, it eliminates almost all of the bias in the naive probit estimator. Second, deviations from normality in the explanatory variables do not have a substantial effect on parameter estimates. Third, the assumed equilibrium selection rule only has a substantial effect on parameter estimates when the peer effect is large. Third, the variance of the estimator depends strongly and positively on the explanatory power of the individual-level observables. When the other variables have strong explanatory power for the outcome, the peer effect will be more precisely estimated.

5 Results

5.1 Naive estimator

In order to place the results here in context, I first estimate a "naive" probit model. The naive probit model simply treats average peer choice as an exogenous variable, and corresponds to the standard treatment in the literature. As explained previously, the naive estimator of peer effects is biased upwards due to both simultaneity and selection issues.

The results are reported in the first column of Table 2. These results suggest that peer behavior is highly influential in a young person's decision to smoke. To put the coefficient estimate in perspective, consider a representative individual with the median number of friends (5) and characteristics indicating an average (42%) probability of smoking. A one-person increase in the number of friends who smoke implies that his or her probability of smoking will increase to $\Phi(\Phi^{-1}(0.42) + 2.24 * 0.2) \approx 60\%$ (where Φ is the standard normal CDF).

5.2 SML probit: Point estimates

Next, we use the same set of exogenous variables, but account both for the endogeneity of peer choice and the correlation in peer characteristics. The results are reported in the second column of Table 2. Because of constraints on computational time, these estimates are preliminary. As noted earlier, the estimates are calculated for a relatively small number of simulations, implying the correct likelihood function has not been well approximated. In addition, the province-level fixed effect is only included for Quebec, the province with by far the highest smoking rate. Estimation with a sufficient number of simulations, a full set of province-level fixed effects, and bootstrapped standard errors, is in progress and will appear in a future draft of this paper.

These results suggest that the peer effect remains strong at approximately 1.76, but are substantially (22%) weaker than is implied by the naive estimator. The point estimate for the between-peers correlation in characteristics is 0.171, implying that the higher estimated peer effect in the naive probit is due to a combination of simultaneity and mild selection effects.

5.3 SML probit: Interval estimates

The point estimates in Table 2 are calculated under the assumption that the correlation in observables and unobservables is the same. As always, the estimates are only as credible as the prior

assumptions. Krauth (2002) shows that it is not feasible to separately estimate the two correlations, but that it is quite feasible to estimate with some other assumed relationship between them. This can be used to construct interval estimates based on weaker assumptions.

The simplest approach is to impose a value for ρ_ϵ and estimate the remaining model parameters, then repeat for several different candidate values of ρ_ϵ. The results of this exercise are shown graphically in Figure 1. The horizontal axis is the imposed value of ρ_ϵ and the vertical axis is the restricted SML estimate calculated for that value of ρ_ϵ. To ensure comparability across estimates, the same set of random numbers is used for each estimate. The values of ρ_ϵ vary from a lower bound of zero to an upper bound of the point estimate derived from the common-correlation assumption. As the figure shows, a lower assumed value for ρ_ϵ is associated with a higher estimated peer effect. Figure 1 implies a mapping between reader-selected interval assumptions on ρ_ϵ and interval estimates of γ. For example, a reader who suspects the correlation in unobservables may be lower than the correlation in unobservables, but is confident the correlation is nonnegative can calculate an interval estimate $\hat{\gamma} \in [1.76, 2.05]$. A reader who is confident that ρ_ϵ is no more than ρ_x and no less than half of ρ_x can calculate a narrower interval estimate of $\hat{\gamma} \in [1.75, 1.90]$.

Note that even when $\rho_\epsilon = 0$ the estimated peer effect is below that from the naive probit. This is because, even when there is no selection effect, the simultaneity effect can still create substantial bias in the naive probit estimator when the group is small.

6 Conclusion

It has long been known that naive estimation leads to upwardly biased measures of peer effects. For years, empirical researchers have used naive estimation with a few caveats in the absence of credible ways of compensating for simultaneity and sorting. In recent years, the skepticism of economists has led to an increased awareness of selection bias in particular, and a number of valuable papers have used natural experiments to evaluate the strength of peer effects. However, natural experiments are not always available and applied researchers will not avoid asking questions simply because the appropriate natural experiment has not been found. One example of this is the smoking literature in public health; papers continue to be published on peer effects without any attempt to deal with simultaneity and sorting, and these papers continue to inform public policy.

The approach in this paper is complementary to the natural-experiment approach. It can be used in situations where no natural experiment is available, but detailed cross-sectional data is. While it requires more restrictive assumptions than many natural experiments, it involves substantially more realistic assumptions about selection and interactions than are commonly seen in the applied literature.

Under the fairly restrictive assumption of equal-correlation, I find that the estimated peer effect is substantially lower than the naive estimate, but is still quite large. Less restrictive assumptions lead to interval estimates, but the intervals associated with reasonable restrictions are also well below the naive estimate but well above zero. Although preliminary, these empirical results suggest a reasonable middle ground between excessive skepticism and credulity on peer effects.

References

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.

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

Centers for Disease Control and Prevention, "Annual smoking-attributable mortality, years of potential life lost, and economic costs - United States,1995-1999," *Morbidity and Mortality Weekly Report*, 2002, *51*, 300–303.

___ , *Tobacco Control State Highlights 2002: Impact and Opportunity*, US Department of Health and Human Services, 2002.

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

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.

A Tables and Figures

Variable	Mean	Std. Dev.
Current smoker	0.231	0.421
Number of friends	6.140	2.804
Fraction of friends smoking	0.425	0.368
Quebec	0.137	0.344
Age	16.815	1.406
Attending school	0.851	0.356
Performs above average in school	0.287	0.453
Performs below average in school	0.032	0.177
Number of smokers in house	0.828	1.038
Has seen ads for tobacco-sponsored events	0.463	0.487

Table 1: Summary statistics for YSS data.

Variable Name	Naive probit Estimate	SMLE Probit Estimate
Selection effect (ρ_x, ρ_ϵ)	–	0.171
	–	(n.nnn)
Peer effect (γ)	2.240	1.756
	(0.054)	(n.nnn)
Intercept	-3.578	-3.253
	(0.266)	(n.nnn)
Quebec	0.190	0.046
	(0.049)	(n.nnn)
Age on 1/1/94	0.090	0.134
	(0.014)	(n.nnn)
Attending school	-0.329	-0.594
	(0.051)	(n.nnn)
Performs above avg in school	-0.189	-0.568
	(0.045)	(n.nnn)
Performs below avg in school	0.216	0.549
	(0.089)	(n.nnn)
# smokers in house	0.258	0.679
	(0.016)	(n.nnn)
Seen ads for events	0.248	0.284
sponsored by tobacco	(0.036)	(n.nnn)

Table 2: Point estimates for both naive probit and SML probit estimators. Estimated standard errors (not yet available) in parentheses.

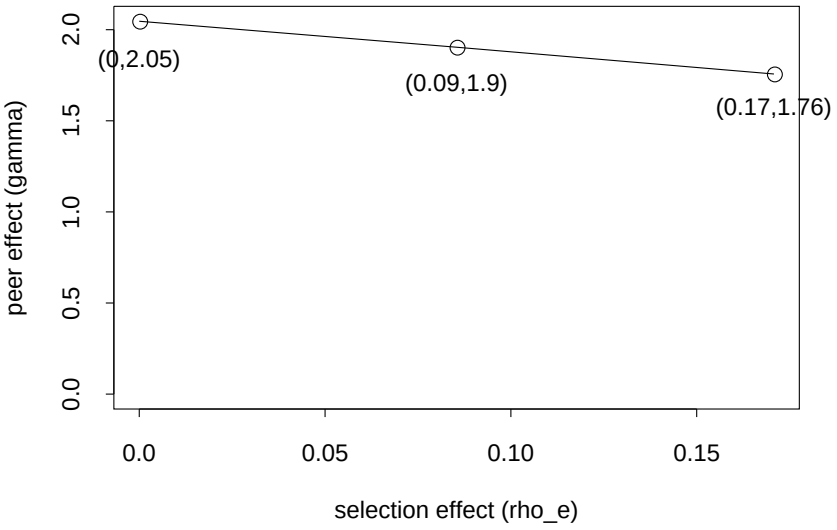

Figure 1: Estimated peer effect for alternative assumptions about selection effect. Peer effect (γ) is estimated under several different restrictions on ρ_ϵ.