Paper 3596-2019

Modeling Teenage Smoking Behavior with SAS Econometrics® Software

Gunce E. Walton, SAS Institute Inc.

ABSTRACT

Cigarette smoking is harmful to health and is estimated to have a yearly social cost of billions of dollars. Smoking behavior is often established during the teenage years, so understanding why teens start smoking is important in the development of smoking prevention policies. One issue in modeling smoking choice is that factors in a person's choice to smoke are often correlated with unobserved characteristics or individual circumstances. If appropriate statistical techniques are not used, this endogeneity causes biased parameter estimates and incorrect inference. This paper demonstrates how to overcome the problem of endogeneity by using techniques from SAS Econometrics® software.

INTRODUCTION

Social interactions are embedded in individual decisions. When it comes to a teenager's decision to smoke, the potential importance of social interactions has been widely recognized for a long time. Modeling and explaining how the social interactions influence a discrete choice variable, such as smoking, is the main topic of this paper.

The decision to smoke is a binary choice, which can be modeled via a discrete-choice model. This paper uses a probit model with social interactions as explanatory variables of interest. The type of social interactions that are of interest are intragender interactions. In many similar studies, cross-gender interactions have been found to be irrelevant in explaining various discrete choice variables, including the choice to smoke. See Soetevent and Kooreman (2007) for an example. In explaining the teenagers' smoking behavior, the control variables are the teenagers' individual characteristics, time use, income and expenditures, subjective information about norms and values, various behaviors and durable-goods ownership, and information about their parents such as their education and working hours. For more information about the model construction, see the section "Probit Model with Social Interactions."

Including social interactions as explanatory variables results in the problem of endogeneity. Social interactions are affected by unobserved factors that also affect the decision to smoke. Since these factors are unobserved, they are not specified in the model but are absorbed by the error term, thus causing the error term to be correlated with the social interactions. When this is the case, the standard estimation method for the probit model might provide biased estimates, consequently causing wrong inferences. This paper uses tools from SAS Econometrics software to show how to test a probit model for the existence of endogenous explanatory variables and how to estimate that model correctly if the endogeneity exists. For more information about these testing and the estimation methods, see the sections Estimation and Testing for Endogeneity.

The probit model for teenage smoking behavior with social interactions is applied to data from the Dutch National School Youth Survey from the year 2000. When the endogeneity of the social interactions is ignored, the estimates suggest that these interactions are an important determinant of teenagers' smoking behavior. However, when the endogeneity is taken into account in the estimation process, the social interactions are, in fact, insignificant in explaining teenagers' choice to smoke.

PROBIT MODEL WITH SOCIAL INTERACTIONS

An individual's decision to take an action or not depends on his or her net benefit from taking that action. If the net benefit is positive, the individual chooses to take the action; if the net benefit is negative, he or she chooses not to take the action. You can model the net benefit simply as

$$y^* = \boldsymbol{x\beta} + \varepsilon$$

where the variable net benefit, y^*, is a continuous variable and $\boldsymbol{x}$ includes the variables that determine the net benefit. In modeling the teenage smoking behavior, $\boldsymbol{x}$ includes two sets of variables: intragender interactions (boy-boy interactions, γ_{bb}, and girl-girl interactions, γ_{gg}) and the other explanatory variables, $\boldsymbol{x_1}$, that are controlled for (such as, teenagers' individual characteristics, time use, income and expenditures, subjective information about norms and values, various behaviors and durable-goods ownership, and information about their parents such as their education and working hours). Based on $\boldsymbol{x}' = \left[\gamma_{bb}, \gamma_{gg}, \boldsymbol{x_1}\right]'$, the model can be rewritten as

$$y^* = \left[\gamma_{bb}, \gamma_{gg}\right]\boldsymbol{\beta_\gamma} + \boldsymbol{x_1'\beta_1} + \varepsilon$$

The problem with this model is that the variable y^* is typically not observed, and therefore cannot be estimated. However, the sign of y^* (which is also called the *latent* dependent variable) can be observed through the individual's observable decision of whether to take an action. If the individual chooses to take an action, you know that the net benefit is positive; if he or she does not choose to take the action, then the net benefit is negative. Therefore, the latent dependent variable, y^*, is observed as follows:

$$\begin{aligned} y &= 1 \quad \text{if } y^* > 0 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

When the error term, ε, is assumed to have a standard normal distribution, the model for y is called the probit model.

The probit model is estimated by maximum likelihood estimation. For the probit maximum likelihood estimator (MLE) to be consistent, the assumption that the error, ε, is independent of the explanatory variables, γ_{bb}, γ_{gg}, and $\boldsymbol{x_1}$, is crucial.

Manski (1993, 2000) and others point out the problem of social interactions being endogenous. In other words, the assumption of $E(\varepsilon|\boldsymbol{x}) = 0$ is violated because

$$E\left(\varepsilon\middle|\gamma_{bb}, \gamma_{gg}\right) \neq 0$$

The correlation between the error term and the intragender interactions can occur because of some unobserved factors that affect both the social interactions and the smoking decisions. For example, a teenager's insecurity or desire to be popular are unobserved factors that might affect the teenager's social interaction and decision to smoke. Since these factors are unobserved, they are absorbed by the error term and cause correlation between the error term and the social interactions.

Figure 1 shows this relationship.

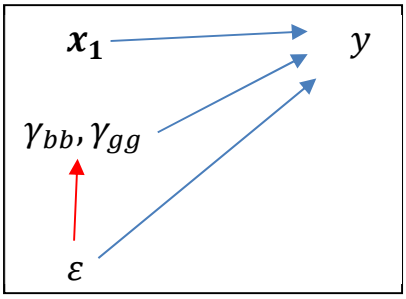

Figure 1. Endogeneity of the Intragender Interactions

The arrows represent direct effects. Any arrow from the error term to an explanatory variable (for example, the red arrow) is undesirable.

Straightforward probit maximum likelihood estimation of a choice model that has endogenous explanatory variables produces inconsistent estimates. Instrumental variables (IV) methods are commonly used to handle the endogeneity problem in linear models. A straightforward generalization of IV methods in nonlinear models, such as a probit model, is unlikely to produce correct results.

ESTIMATION

This paper uses two methods to estimate a probit model with endogenous explanatory variables:

- control function method
- conditional maximum likelihood estimation method

Both methods require a reduced-form model for each endogenous explanatory variable, γ_{bb} and γ_{gg}. These reduced-form models are as follows:

$$\gamma_{bb} = \boldsymbol{z}'\boldsymbol{\delta_{bb}} + \boldsymbol{x_1'}\boldsymbol{\beta_{bb}} + v_{bb}$$
$$\gamma_{gg} = \boldsymbol{z}'\boldsymbol{\delta_{gg}} + \boldsymbol{x_1'}\boldsymbol{\beta_{gg}} + v_{gg}$$

The variables γ_{bb} and γ_{gg} in the model for y are endogenous if the errors ε, v_{bb}, and v_{gg} are correlated. The vector $\boldsymbol{z}$ includes the instrumental variables, which need to be independent of the errors ε, v_{bb}, and v_{gg} and have an important role in explaining γ_{bb} and γ_{gg}. Figure 2 shows how the instrumental variables should be related to the other variables.

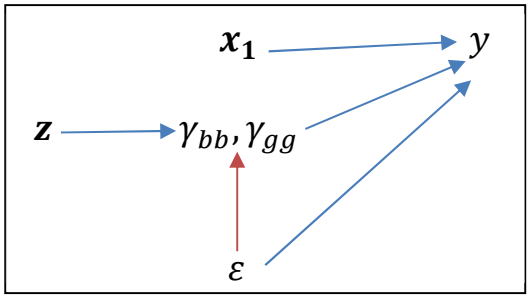

Figure 2. Instrumental Variables in Relation to the Model Variables

In explaining the intragender interactions, cross-gender interactions (γ_{bg} and γ_{gb}), cell phone ownership (y_{cell}), and moped ownership (y_{brom}) are used. Soetevent and Kooreman

(2007) show that the cross-gender interactions are highly insignificant in directly explaining the smoking choice. However, a strong relationship between the cross-gender and intragender interactions is expected. For this reason, the cross-gender interactions are good candidates to be instrumental variables for intragender interactions. Similarly, cell phone ownership and moped ownership do not directly explain the smoking behavior, but they might have a direct effect on the intragender interactions.

CONTROL FUNCTION METHOD

This method is the most useful two-step approach; it is attributed to Rivers and Vuong (1988). This method also leads to a simple test for endogeneity of γ_{bb} and γ_{gg}. The driving point of this method is the formalization of the correlation between ε and v_{bb} and v_{gg} as follows:

$$\varepsilon = \theta_1 v_{bb} + \theta_2 v_{gg} + e$$

Based on this structure, the latent model can be written as

$$y^* = [\gamma_{bb}, \gamma_{gg}]\boldsymbol{\beta_\gamma} + \boldsymbol{x_1'\beta_1} + \theta_1 v_{bb} + \theta_2 v_{gg} + e$$

Assuming for the moment that v_{bb} and v_{gg} are observed, then the probit of y on γ_{bb}, γ_{gg}, $\boldsymbol{x}$, v_{bb}, and v_{gg} consistently estimates the probit model parameters.[1] Since v_{bb} and v_{gg} are unknown, they must first be estimated, as in the following steps:

1. Run the ordinary least squares (OLS) regressions γ_{bb} and γ_{gg} on $\boldsymbol{z}$ and $\boldsymbol{x_1}$ and save the residuals, $\hat{v}_{bb}$ and $\hat{v}_{gg}$.
2. Run the probit y on γ_{bb}, γ_{gg}, $\boldsymbol{x_1}$, $\hat{v}_{bb}$, and $\hat{v}_{gg}$ to get consistent estimators of the scaled coefficients $\boldsymbol{\beta_\gamma}$, $\boldsymbol{\beta_1}$, θ_1, and θ_2.

These steps also produce a simple endogeneity test. This testing method is discussed in the section "Testing for Endogeneity."

The control function method is consistent only for continuous endogenous variables. If the endogenous explanatory variable is a noncontinuous variable, then the control function method should not be used for estimation, but it can still be used for testing purposes.

CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION METHOD

This method jointly estimates the structural model and the reduced-form models by the maximum likelihood estimation in one step. The joint density of y, γ_{bb}, and γ_{gg} can be obtained using the conditional densities as follows:

$$f(y, \gamma_{bb}, \gamma_{gg}|\boldsymbol{x}, \boldsymbol{z}) = f(y|\gamma_{bb}, \gamma_{gg}, \boldsymbol{x}, \boldsymbol{z}) f(\gamma_{bb}|\gamma_{gg}, \boldsymbol{x}, \boldsymbol{z}) f(\gamma_{gg}|\boldsymbol{x}, \boldsymbol{z})$$

The conditional density functions on the right-hand side can be specified based on the normality assumptions on the errors (Wooldridge 2010, Ch. 15.7.2). Therefore, maximizing the log of the joint density function summed over the individuals yields the conditional maximum likelihood estimation.

The single-step conditional maximum likelihood method is consistent independent of the nature of the endogenous variables. In addition, the conditional MLE is more efficient than the two-step control function estimator.

[1] The probit MLE estimates scaled model parameters. The form of the scaling is given in Wooldridge (2010). This paper investigates the relative effect of the social interactions to the other explanatory variables on the decision to smoke. Therefore, scaling the estimates does not alter the inference.

TESTING FOR ENDOGENEITY

Whether it is because of omitted variables, measurement error, or simultaneity, endogeneity is caused by the correlation between the error terms of the structural model

$$y^* = [\gamma_{bb}, \gamma_{gg}]\boldsymbol{\beta_\gamma} + \boldsymbol{x_1'\beta_1} + \varepsilon$$
$$y = 1[\, y^* > 0]$$

and the reduced-form models

$$\gamma_{bb} = \boldsymbol{z'\delta_{bb}} + \boldsymbol{x_1'\beta_{bb}} + v_{bb}$$
$$\gamma_{gg} = \boldsymbol{z'\delta_{gg}} + \boldsymbol{x_1'\beta_{gg}} + v_{gg}$$

Therefore, testing to determine whether this correlation is 0 provides an endogeneity test for γ_{bb} and γ_{gg}.

The test of endogeneity in the control function method is simple. A joint test on the coefficients of $\hat{v}_{bb}$, and $\hat{v}_{gg}$ is a valid test of the null hypothesis that γ_{bb} and γ_{gg} are exogenous—that is, $H_0: \theta_1 = \theta_2 = 0$. Rejecting the null hypothesis favors the decision that γ_{bb} and γ_{gg} are endogenous.

In the conditional maximum likelihood estimation method, a joint test on the correlation coefficients between the errors of the reduced-form models and the error of the structural (probit) model, $\rho_{(\gamma_{bb},y)}$ and $\rho_{(\gamma_{gg},y)}$, provides a test of endogeneity. In this case, the null hypothesis is $H_0: \rho_{(\gamma_{bb},y)} = \rho_{(\gamma_{gg},y)} = 0$. Rejecting the null hypothesis raises doubts about the exogeneity of γ_{bb} and γ_{gg}.

EMPIRICAL APPLICATION

THE DATA

The data that are used in this paper come from the Dutch National School Youth Survey (NSYS) from the year 2000. This is the data set that is used by Soetevent and Kooreman (2017) and is made available in the *Journal of Applied Econometrics* data archive.

The data set contains information about the teenagers' individual characteristics, time use, income and expenditures, information about their personal norms and values, and information about their smoking behaviors and cell phone and moped ownership. The data set also contains some information about the parents, such as their education level and working hours.

The social interactions are formalized and measured according to the definitions in Soetevent and Kooreman (2017).

SPECIFICATION OF THE EMPIRICAL MODEL

The model variables are specified as follows:

- The explanatory variables other than the social interactions, $\boldsymbol{x_1}$, are
    - *girl*: Dummy for gender
    - *age*: Teenager's age
    - *nondutch*: Dummy for being Dutch
    - *sngpar*: Dummy for single-parent household
    - *mavo*: Lower education level

    - *havo*: Intermediate education level
    - *vwo*: Higher education level
    - *wtfa*: Working time, father
    - *wtmo*: Working time, mother
    - *cath*: Dummy for being Catholic
    - *prot*: Dummy for being Protestant
- The intragender interactions are
    - *gamma_bb*: Boy-boy interaction
    - *gamma_gg*: Girl-girl interaction
- The instrumental variables, $\boldsymbol{z}$, are
    - *gamma_bg*: Boy-girl interaction
    - *gamma_gb*: Girl-boy interaction
    - *ycell*: Dummy for cell phone
    - *ybrom*: Dummy for moped ownership
- The dependent variable, y, is
    - *ysmoke*: Dummy for cell phone ownership

ESTIMATION AND TESTING

First, the probit model is estimated by using the traditional probit maximum likelihood estimation method, under the assumption that all the explanatory variables are exogenous. For this estimation, you can use the QLIM procedure or the CQLIM procedure in SAS Econometrics. The SAS code for using PROC CQLIM for this estimation is as follows:

```
PROC CQLIM DATA=mylib.dutchdata METHOD=NEWRAP;
   MODEL ysmoke = gamma_bb gamma_gg girl age nondutch sngpar mavo
                  havo vwo wtfa wtmo cath prot / DISCRETE(D=PROBIT);
   TEST gamma_bb=0, gamma_gg=0 / ALL;
RUN;
```

The probit estimation is requested by the DISCRETE(D=PROBIT) option in the MODEL statement. The TEST statement tests the null hypothesis that the intragender interactions are jointly insignificant—that is, $H_0: \gamma_{bb} = \gamma_{gg} = 0$. The ALL option in the TEST statement requests all three test statistics: Wald, likelihood ratio, and Lagrange multiplier. The resulting parameter estimates are summarized in Table 1.

Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr>|t}
Intercept	-3.705953	0.250402	-14.8	<.0001
gamma_bb	0.699087	0.168438	4.15	<.0001
gamma_gg	0.601115	0.167603	3.59	0.0003
girl	0.015837	0.100596	0.16	0.8749
age	0.172726	0.015798	10.93	<.0001
nondutch	-0.249367	0.081738	-3.05	0.0023
sngpar	0.209080	0.069696	3.00	0.0027
mavo	0.147745	0.054158	2.73	0.0064
havo	-0.056131	0.066400	-0.85	0.3979
vwo	-0.192613	0.073486	-2.62	0.0088
wtfa	0.002467	0.001767	1.40	0.1626
wtmo	0.004834	0.001383	3.50	0.0005
cath	-0.197016	0.054632	-3.61	0.0003
prot	-0.112215	0.058140	-1.93	0.0536

Table 1. Parameter Estimates from Probit Estimation

The effect of gender is insignificant. The probability of smoking strongly increases with age. The higher the education level, the smaller the probability that a pupil smokes. Pupils from single-parent households and pupils whose mothers have a paid job have a significantly higher probability of smoking. Pupils who are non-Dutch, Catholic, or Protestant have a lower probability of smoking.

Table 2 shows the output from the test of joint significance of the intragender interactions.

Test Results				
Test	Type	Statistic	Pr > ChiSq	Label
Test 1	Wald	29.81809	<.0001	gamma_bb=0, gamma_gg=0
Test 1	L.R.	30.0347	<.0001	gamma_bb=0, gamma_gg=0
Test 1	L.M.	29.89838	<.0001	gamma_bb=0, gamma_gg=0

Table 2. Joint Significance Test of Intragender Interactions

All three test results indicate that the intragender interactions are highly significant in determining the teenage smoking behavior.

Next, the endogeneity of the intragender interactions is taken into account in the estimation by using the two-step control function method as follows:

1. Run an OLS regression of each endogenous variable on the instrumental variables (cross-gender interactions, γ_{bg} and γ_{gb}, cell phone ownership, y_{cell}, and moped ownership, y_{brom}) and all the other exogenous explanatory variables, and then save the corresponding residuals. The SAS code for this step is shown in the following PROC CQLIM statements:

```
PROC CQLIM DATA=mylib.dutchdata METHOD=NEWRAP;
   MODEL gamma_bb = gamma_bg gamma_gb ycell ybrom girl age nondutch
                    sngpar mavo havo vwo wtfa wtmo cath prot;
   OUTPUT OUT=mylib.resbb RESIDUAL;
   TEST gamma_bg=0, gamma_gb=0, ycell=0, ybrom=0;
RUN;


PROC CQLIM DATA=mylib.dutchdata METHOD=NEWRAP;
   MODEL gamma_gg = gamma_bg gamma_gb ycell ybrom girl age nondutch
                    sngpar mavo havo vwo wtfa wtmo cath prot;
   OUTPUT OUT=mylib.resgg RESIDUAL;
   TEST gamma_bg=0, gamma_gb=0, ycell=0, ybrom=0;
RUN;
```

   Because no option is specified in the MODEL statement, a linear regression model is used by default. The TEST statement tests the joint significance of the instrumental variables in both reduced-form models.

   The output for the reduced-form model for boy-boy interaction is shown in Table 3 and for girl-girl interaction in Table 4.

Parameter Estimates for gamma_bb				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr>|t}**
Intercept	-0.851878	0.015491	-54.99	<.0001
gamma_bg	-0.481053	0.010417	-46.18	<.0001
gamma_gb	-0.046678	0.01006	-4.64	<.0001
ycell	0.005209	0.003179	1.64	0.1014
ybrom	0.001964	0.005282	0.37	0.71
girl	0.591728	0.006257	94.56	<.0001
age	0.017138	0.001021	16.78	<.0001
nondutch	-0.015755	0.004683	-3.36	0.0008
sngpar	0.006458	0.004589	1.41	0.1593
mavo	0.017342	0.003254	5.33	<.0001
havo	0.004093	0.003937	1.04	0.2986
vwo	-0.002386	0.004182	-0.57	0.5683
wtfa	-0.000068922	0.000103	-0.67	0.5039
wtmo	-0.000070148	8.39E-05	-0.84	0.4031
cath	-0.015154	0.003103	-4.88	<.0001
prot	-0.004210	0.003423	-1.23	0.2188
_Sigma	0.107917	0.000884	122.09	<.0001

Table 3. Reduced-Form Model for Boy-Boy Interaction Estimates

Parameter Estimates for gamma_gg				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr>|t}**
Intercept	-0.282187	0.01615	-17.47	<.0001
gamma_bg	-0.043410	0.010861	-4	<.0001
gamma_gb	-0.410531	0.010488	-39.14	<.0001
ycell	0.011768	0.003315	3.55	0.0004
ybrom	-0.013897	0.005507	-2.52	0.0116
girl	-0.565948	0.006524	-86.75	<.0001
age	0.018785	0.001065	17.64	<.0001
nondutch	-0.014329	0.004882	-2.93	0.0033
sngpar	0.002503	0.004784	0.52	0.6008
mavo	0.015327	0.003393	4.52	<.0001
havo	-0.009782	0.004105	-2.38	0.0172
vwo	-0.038072	0.00436	-8.73	<.0001
wtfa	0.000104	0.000108	0.97	0.3314
wtmo	0.000013274	8.75E-05	0.15	0.8794
cath	-0.014942	0.003236	-4.62	<.0001
prot	-0.011428	0.003569	-3.2	0.0014
_Sigma	0.112510	0.000922	122.09	<.0001

Table 4. Reduced-Form Model for Girl-Girl Interaction Estimates

The test results show that in both reduced-form models, the instrumental variables are jointly significant in explaining the intragender interactions. The test results are summarized in Table 5 and Table 6.

Test Results				
Test	**Type**	**Statistic**	**Pr > ChiSq**	**Label**
Test 1	Wald	2144.555	<.0001	gamma_bg=0, gamma_gb=0, ycell=0, ybrom=0

Table 5. Test of Instrument Significance in the Model for Boy-Boy Interaction

Test Results				
Test	**Type**	**Statistic**	**Pr > ChiSq**	**Label**
Test 1	Wald	1554.342	<.0001	gamma_bg=0, gamma_gb=0, ycell=0, ybrom=0

Table 6. Test of Instrument Significance in the Model for Girl-Girl Interaction

2. Include the generated residuals in the data set and then run a probit regression that includes the generated residuals as additional explanatory variables, as shown in the following SAS code:

```
DATA mylib.dutchdata_resids1;
   MERGE mylib.dutchdata mylib.resbb;
   RENAME resid=residbb;
RUN;

DATA mylib.dutchdata_resids2;
   MERGE mylib.dutchdata_resids1 mylib.resgg;
   RENAME resid=residgg;
RUN;


PROC CQLIM DATA= mylib.dutchdata_resids2 METHOD=NEWRAP;
   MODEL ysmoke = gamma_bb gamma_gg girl age nondutch sngpar mavo
                  havo vwo wtfa wtmo cath prot
                  Resid_gamma_bb Resid_gamma_gg / DISCRETE;
   TEST gamma_bb=0, gamma_gg=0 / ALL;
   TEST Resid_gamma_bb=0, Resid_gamma_gg=0 / ALL;
RUN;
```

The output from this probit regression is provided in Table 7.

Parameter Estimates				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr>|t}**
Intercept	-4.304223	0.316024	-13.62	<.0001
gamma_bb	-0.037946	0.368112	-0.1	0.9179
gamma_gg	-0.497707	0.403553	-1.23	0.2175
girl	-0.151617	0.256008	-0.59	0.5537
age	0.193348	0.016608	11.64	<.0001
nondutch	-0.262658	0.081855	-3.21	0.0013
sngpar	0.219689	0.069784	3.15	0.0016
mavo	0.175542	0.05469	3.21	0.0013
havo	-0.056969	0.066649	-0.85	0.3927
vwo	-0.217639	0.075102	-2.9	0.0038
wtfa	0.002505	0.001771	1.41	0.1571
wtmo	0.004866	0.001387	3.51	0.0005
cath	-0.222942	0.05517	-4.04	<.0001
prot	-0.131623	0.058681	-2.24	0.0249
Resid_gamma_bb	0.981096	0.406789	2.41	0.0159
Resid_gamma_gg	1.334281	0.430740	3.10	0.0020

Table 7. Second-Step Probit Estimates

The estimates on the exogenous variables are very similar to those of the traditional probit model, which are shown in Table 1. However, the coefficients of the boy-boy and girl-girl interactions are insignificant when the control function method is used. These coefficients differ greatly between the outputs of the two estimation methods. When the endogeneity of the social interactions is taken into account in the estimation, these explanatory variables do not seem to explain the smoking decision well. The joint significance test of the social interactions, obtained by the first TEST statement in the PROC CQLIM code, also confirms this. The output of that test is provided in Table 8 with the name Test 1.

The second TEST statement in the PROC CQLIM code is important because it tests the hypothesis $H_0: \gamma_{bb} = \gamma_{gg} = 0$. It is a test of joint significance of the residuals that are included in the model as additional explanatory variables to control for the endogeneity. Therefore, it is a test of endogeneity of the social interactions in the model. Rejecting this test hypothesis indicates that the social interactions are indeed endogenous in the model. The output for Test 2 in Table 8 leads to the conclusion that the social interactions are endogenous.

Test Results				
Test	**Type**	**Statistic**	**Pr > ChiSq**	**Label**
Test 1	Wald	1.630202	0.4426	gamma_bb=0, gamma_gg=0
Test 1	L.R.	1.633119	0.4419	gamma_bb=0, gamma_gg=0
Test 1	L.M.	1.630587	0.4425	gamma_bb=0, gamma_gg=0
Test 2	Wald	18.09549	0.0001	Resid_gamma_bb=0, Resid_gamma_gg=0
Test 2	L.R.	18.02563	0.0001	Resid_gamma_bb=0, Resid_gamma_gg=0
Test 2	L.M.	18.12613	0.0001	Resid_gamma_bb=0, Resid_gamma_gg=0

Table 8. Test 1 Tests the Joint Significance of the Social Interactions, and Test 2 Tests the Endogeneity of the Social interactions

Finally, the model is estimated by using the conditional MLE method. This method jointly estimates the probit model and the two reduced-form models. You can do this by using PROC QLIM as follows:

```
PROC QLIM DATA=dutchdata METHOD=NEWRAP;
   MODEL ysmoke = gamma_bb gamma_gg girl girl age nondutch
                  sngpar mavo havo vwo wtfa wtmo cath prot
                   / DISCRETE;
   MODEL gamma_bb = gamma_bg gamma_gb ycell ybrom girl age
             nondutch sngpar mavo havo vwo wtfa wtmo cath prot;
   MODEL gamma_gg = gamma_bg gamma_gb ycell ybrom girl age
             nondutch sngpar mavo havo vwo wtfa wtmo cath prot;
RUN;
```

When there are multiple MODEL statements, PROC QLIM estimates all the models jointly rather than separately. The first MODEL statement specifies the probit model, and the other two MODEL statements specify the reduced-form models.

The output from this estimation method is provided in Table 9. To confirm that the social interactions are endogenous, you can add the following TEST statement:

```
TEST _Rho.gamma_bb.ysmoke, _Rho.gamma_gg.ysmoke / WALD;
```

Social interactions are endogenous when the error terms of the probit model and the reduced-form models are correlated, which are indicated by the correlation coefficients $\rho_{(bb,\ ysmoke)}$ and $\rho_{(gg,\ ysmoke)}$. Table 10 provides the output for this test. If you are using the conditional maximum likelihood estimator, the endogeneity test that is described in the

section "Control Function Method" is also available to you directly. You can request this test by specifying the ENDOTEST option in the MODEL statement that specifies the structural equation (in this case, the probit model) as shown in the following code:

```
PROC QLIM DATA=dutchdata METHOD=NEWRAP;
  MODEL ysmoke = gamma_bb gamma_gg girl girl age nondutch
        sngpar mavo havo vwo wtfa wtmo cath prot
      / DISCRETE ENDOTEST(gamma_bb gamma_gg);
  MODEL gamma_bb = gamma_bg gamma_gb ycell ybrom girl age
        nondutch sngpar mavo havo vwo wtfa wtmo cath prot;
  MODEL gamma_gg = gamma_bg gamma_gb ycell ybrom girl age
        nondutch sngpar mavo havo vwo wtfa wtmo cath prot;
RUN;
```

Both testing methods result in rejection of the test hypothesis that the social interactions are exogenous. Once again, the endogeneity of the social interactions in the probit model is confirmed.

Parameter Estimates for ysmoke				
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr>|t|**
Intercept	-4.18664	0.307267	-13.63	<.0001
gamma_bb	0.030278	0.376812	0.08	0.936
gamma_gg	-0.31439	0.411948	-0.76	0.4454
girl	-0.10064	0.26357	-0.38	0.7026
age	0.188295	0.016177	11.64	<.0001
nondutch	-0.25757	0.081088	-3.18	0.0015
sngpar	0.215729	0.069211	3.12	0.0018
mavo	0.169956	0.054121	3.14	0.0017
havo	-0.05566	0.066071	-0.84	0.3995
vwo	-0.21012	0.074425	-2.82	0.0048
wtfa	0.002472	0.001755	1.41	0.159
wtmo	0.004798	0.001375	3.49	0.0005
cath	-0.21695	0.054546	-3.98	<.0001
prot	-0.12715	0.058111	-2.19	0.0287
_Rho.gamma_bb.gamma_gg	-0.061106	0.011539	-5.30	<.0001
_Rho.gamma_bb.ysmoke	0.095325	0.043729	2.18	0.0293
_Rho.gamma_gg.ysmoke	0.141849	0.047638	2.98	0.0029

Table 9. Output of Conditional MLE

Test Results				
Test	**Type**	**Statistic**	**Pr > ChiSq**	**Label**
Test 0	Wald	19.11	<.0001	_Rho.gamma_bb. ysmoke = 0, _Rho.gamma_gg. ysmoke = 0

Table 10. Endogeneity Test in the Joint Model

The estimates of the conditional MLE, shown in Table 9, are very similar to those of the control function method, shown in Table 7. Both methods indicate that the social interactions do not play a significant role in explaining the choice to smoke. Both estimation methods indicate that rather than their social interactions, the teenagers' individual characteristics, time use, income and expenditures, subjective information about norms and values, their education level, and their parents' education and working hours play a much more important role on their decision to smoke.

CONCLUSION

This empirical study estimates a model of the choice to smoke by using a probit model that takes endogeneity of the social interactions into account for data from the Dutch National School Youth Survey (NSYS) in the year 2000. The study confirms that endogeneity exists for social interactions. When endogeneity of social interactions is ignored, you might mistakenly conclude that these variables are important determinants of teenagers' smoking behavior. This paper suggests two estimation methods and shows how to implement them by using PROC CQLIM and PROC QLIM in SAS Econometrics® software. Correct estimation results suggest that the social interactions are, in fact, insignificant in explaining teenagers' choice to smoke. The teenagers' individual characteristics, time use, income and expenditures, subjective information about norms and values, their education level, and their parents' education and working hours play a much more important role in their smoking decision.

REFERENCES

Manski, C. F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*. 60: 531–542.

Manski C. F. 2000. "Economic Analysis of Social interactions." *Journal of Economic Perspectives*. 14: 115–136.

Rivers, D., and Q. H. Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics*, 39: 347–366.

Soetevent, A. R., and P. Kooremen. 2007. "A Discrete-Choice Model with Social Interactions: with an Application to High School Teen Behavior." *Journal of Applied Econometrics*, 22: 599–624.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Gunce Walton
SAS Institute Inc.
(919) 531-2366
gunce.walton@sas.com