Endogeneity Testing

Tomas Morley

13 maart 2017

Wooldridge (2010) proposes a test for endogeneity following the control function approach. Essentially, the procedure involves regressing the endogenous variable against a suitable instrument in a first stage reduced model, and then incorporating the resiudals from the resuced model in the estimation of the full structural model. The presence of endogeneity is confirmed by the significance of the control function (reduced form residual), as determined by a t-test under the null hypothesis of no exogeneity.

In our case, we explore the relationship between maize yield and nitrogen application at the field level. The possibility that nitrogen is endogenous has been explored by . . . . A suitable instrument for the endogenous variable is the relative price of nitrogen to maize. Both prices are assumed to be exogenous to the biophysical relationship connection maize yields to nitrogen applications. Both prices are given by the market and and while they may affect the decision as to whether or not to produce maize, or how much to produce, they are unlikely to be related to the maize yield response function which is driven by biophysical constraints, crop practices and input availability.

A complication is that in our sample of smallholder maize farmers, many report using no fertilizer and this may influence the choice of reduced form first stage model. Two possible options are a reduced form OLS model or a tobit model. The results for each and the average partial effects (APEs) are displayed in Table 1.

Table 1: First stage results

	OLS	SE	Tobit	SE	APE	BSE
(Intercept)	1.301**	0.191	-1.355*	0.586	-0.506	0.223
relprice	-0.043**	0.005	-0.127**	0.014	-0.048	0.006
sex	0.067	0.076	0.327	0.208	0.122	0.078
age	-0.008**	0.002	-0.016**	0.006	-0.006	0.002
ed_any	-0.056	0.066	-0.029	0.183	-0.011	0.068
logarea_tot	-0.017	0.03	-0.02	0.093	-0.007	0.033
credit	0.279**	0.069	0.607**	0.176	0.227	0.066
extension	2.12**	0.066	4.557**	0.183	1.701	0.061
title	0.086	0.062	0.22	0.17	0.082	0.063
elevation	0.001**	0	0.001**	0	0.001	0
logslope	-0.289**	0.039	-0.669**	0.108	-0.25	0.043
logarea	0.126**	0.024	0.46**	0.073	0.172	0.028
SOC2	0.018**	0.006	0.063**	0.018	0.024	0.007
R-squared	0.453	NA	0.449	NA	NA	NA

The results are broadly similar for those coefficients that are significantly different from zero. Comparing the R-squared from the OLS model to the pseudo R-squared from the Tobit model indicate that there is very little difference between the two in terms of fit. However, it should be noted that the OLS model explicitley maximimizes the R-squared, whereas the tobit model is fit by ML. Nonethless, we find little reason to prefer the one model over the other.

Table 2: Second stage results

	TL-LM	SE	BSE	TL-Tob	SE	BSE
(Intercept)	2.198**	0.467	0.59	2.237**	0.467	0.576
logN	0.254**	0.06	0.06	0.223**	0.059	0.06

	TL-LM	SE	BSE	TL-Tob	SE	BSE
loglab	0.303**	0.06	0.073	0.302**	0.06	0.07
logseed	0.575**	0.077	0.098	0.572**	0.077	0.096
logNsq	0.035**	0.011	0.011	0.028**	0.011	0.011
loglabsq	0.015*	0.006	0.007	0.015*	0.006	0.007
logseedsq	-0.013	0.01	0.012	-0.013	0.01	0.012
logarea	-0.081**	0.024	0.025	-0.08**	0.024	0.025
phdum55_2_70	0.136**	0.042	0.044	0.127**	0.042	0.043
crop_count2	0.475**	0.042	0.042	0.475**	0.042	0.043
dumoxen	0.079	0.042	0.041	0.081*	0.042	0.039
SOC2	0.016**	0.005	0.005	0.016**	0.005	0.004
logslope	-0.205**	0.03	0.031	-0.211**	0.029	0.03
elevation	0*	0	0	0	0	0
GGD	0.229**	0.044	0.053	0.23**	0.044	0.054
AI	0.041**	0.01	0.01	0.039**	0.01	0.01
TS	0**	0	0	0**	0	0
v	-0.07**	0.023	0.024	-0.046**	0.013	0.014
logN:loglab	-0.025**	0.008	0.009	-0.024**	0.008	0.008
logN:logseed	-0.036**	0.013	0.013	-0.031*	0.014	0.014
loglab:logseed	-0.047**	0.014	0.017	-0.047**	0.014	0.018

The full structural model was then estimated by ordinary least squares incorporatoing the control function. The results are reported in table 2. We see that the reported results in table 2 are very close with the exception of the coefficients on the `logN` term and the residual term. Wooldridge (2010) described a method for constructing a generalized residual from a probit model. Here we use a generalized residual from a tobit model constructed as. There are reasons to suspect that the tobit coefficient in the is controlling for the fact that a large number of plots have no nitrogen applied. In this case the response to nitrogen, once this has been accounted for, is higher than estimated in the ols reduced form where. In short the endogeneity that we observe may in fact be more related to a selection issue or data censoring/truncation

$$\hat{v}_{tob} = d_1 \frac{-\phi(-x_i'\beta)}{\Phi(-x_i'\beta)} + d_2(\theta y_i - x_i'\beta)$$

Although we control for a range of environmental variables, several studies have investigated the possibility of endogeneity between nitrogen and maize yield (Liverpool tasie, smale and mason, etc etc). Endogeneity may occur in the presence of a feedback loop between the error term and the choice of inputs (Amsler 2016). In the stochastic frontiers setting this may involve either the statistical error term, the inefficiency term or both. One way of testing for the presence of endogeneity is to employ a control function approach, estimating a first stage reduced form model for the endogenous variable and using the predicted residuals as a regressor in a second stage full structural model. The null hypothesis that the endogenous variable is exogenous is rejected based on the significance of a standard t or F test. The control function approach is an alternative to two stage least squares estimation which is suitable when using a nonlinear second stage estimation such as in the case of a translog production function. It has the additional advantage that only one control function is required per endogenous inputs, rather than four, corresponding to the four terms involving nitrogen, that would be required in 2SLS setting.

Table 4: First stage endogeneity

	OLS	SE	Tobit	SE	APE	BSE
(Intercept)	1.301**	0.191	-1.355*	0.586	-0.506	0.223
relprice	-0.043**	0.005	-0.127**	0.014	-0.048	0.006
sex	0.067	0.076	0.327	0.208	0.122	0.078

	OLS	SE	Tobit	SE	APE	BSE
age	-0.008**	0.002	-0.016**	0.006	-0.006	0.002
ed_any	-0.056	0.066	-0.029	0.183	-0.011	0.068
logarea_tot	-0.017	0.03	-0.02	0.093	-0.007	0.033
credit	0.279**	0.069	0.607**	0.176	0.227	0.066
extension	2.12**	0.066	4.557**	0.183	1.701	0.061
title	0.086	0.062	0.22	0.17	0.082	0.063
elevation	0.001**	0	0.001**	0	0.001	0
logslope	-0.289**	0.039	-0.669**	0.108	-0.25	0.043
logarea	0.126**	0.024	0.46**	0.073	0.172	0.028
SOC2	0.018**	0.006	0.063**	0.018	0.024	0.007
R-squared	0.453	NA	0.449	NA	NA	NA

Table 5: Second stage endogeneity

	TL-LM	SE	BSE	TL-Tob	SE	BSE
(Intercept)	2.198**	0.467	0.59	2.237**	0.467	0.576
logN	0.254**	0.06	0.06	0.223**	0.059	0.06
loglab	0.303**	0.06	0.073	0.302**	0.06	0.07
logseed	0.575**	0.077	0.098	0.572**	0.077	0.096
logNsq	0.035**	0.011	0.011	0.028**	0.011	0.011
loglabsq	0.015*	0.006	0.007	0.015*	0.006	0.007
logseedsq	-0.013	0.01	0.012	-0.013	0.01	0.012
logarea	-0.081**	0.024	0.025	-0.08**	0.024	0.025
phdum55_2_70	0.136**	0.042	0.044	0.127**	0.042	0.043
crop_count2	0.475**	0.042	0.042	0.475**	0.042	0.043
dumoxen	0.079	0.042	0.041	0.081*	0.042	0.039
SOC2	0.016**	0.005	0.005	0.016**	0.005	0.004
logslope	-0.205**	0.03	0.031	-0.211**	0.029	0.03
elevation	0*	0	0	0	0	0
GGD	0.229**	0.044	0.053	0.23**	0.044	0.054
AI	0.041**	0.01	0.01	0.039**	0.01	0.01
TS	0**	0	0	0**	0	0
v	-0.07**	0.023	0.024	-0.046**	0.013	0.014
logN:loglab	-0.025**	0.008	0.009	-0.024**	0.008	0.008
logN:logseed	-0.036**	0.013	0.013	-0.031*	0.014	0.014
loglab:logseed	-0.047**	0.014	0.017	-0.047**	0.014	0.018

Although incorporating the residuals from the reduced form regression tests for endogeneity it does not provide suitable estimation in a stochastic frontiers framework. In other words, knowing that nitrogen is endogenous to the relationship does not take us closer to estimating the frontier function. Instead, a recent survey of methods for incorporating endogeneity in stochastic frontiers models suggests the use of corrected two stage least squares (C2SLS) or Limited information maximum likelihood (LIML) as frameworks for estimating the stochastic frontiers model. We cannot add fitted values for the endogenous variable to the normal COLS or SF ML estimation. Moreover, the composite error term $epsilon_i$ may be correlated with either the statistical noise or the inefficinecy and additional assumptions are required to determine which. determine this based on what could lead to inefficinecy Endogeneity formally refers to the situation where the residuals in the estimation are correlated with an explanatory variable. In this case the explanatory variable is then referred to as the endogenous variable. For our estimation we focus on the potential endogeneity of nitrogen which enters the translog production function in a log level term and a log squared term. As a result we need a minimum of two identifying equations in order to be able to estimate with a control function and remove the endogeneityLIML outperforms 2sls when the instruments are weak or when there are many instruments relative to the number of observations

control functions are used to break the correlation between the endogenous variables and unobservables affecting the response (Wooldrdige 2010) and are often used to handle endogeneity in non-linear models when techniques such as two stage least squares are not applicable. Similar to two stage least squares this approach requires exogenous variables which do not appear in the second stage or structural regression but do appear in the first stage regression. In other words, variables which affect the quantity of nitrogen used, but not the yield response equation.

So due to the squared term on nitrogen we say that this model is nonlinear in the endogenous variable We should be careful to include all terms that appear in the second stage in the first stage, with the exception of the endogenous variable which appears as the dependent variable on the right hand side. When we have a translog (nonlinear function) special attention is needed for identification and choice of instruments. Feedback loop in the sense that you can write yield as a function of nitrogen but you can also write nitrogen as a function of yield. We make the assumption that the endogenous variable is correlated with statistical noise but not with inefficiency. If nitrogen is correlated with the statistical error term then we also expected its square to be correlated with that term too. In the presense of endogeneity the maximum likelihood estimates of the stochastic frontiers model will not be consistent.