Ethiopia Stochastic Frontiers Analysis Draft 2

Tomas Morley

20 March 2017

Abstract

We investigate the determinants of the maize yield gap in a nationally representative sample of Ethiopian smallholder farmers.

Introduction

Data

The main data source is the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) Ethiopian Socioeconomic survey (ESS). This survey was implemented by the Central Statistical Agency of Ethiopia (CSA) and was funded by the World Bank through a grant from the Bill and Melinda Gates foundation. Although part of a panel, the first wave of the ESS did not record crop harvested quantities for a majority of plots making it unsuitable for this analysis. However, the second wave (2013-2014) of the ESS provides detailed information on crop production, input use, environmental and scocioeconomic variables.

In order to calculate maize yields and input rates, accurate measurements of field areas are required. GPS measurements were recorded for the vast majority of plots in the second wave of the ESS. For those plots lacking a GPS measurement, multiple imputation was used following LSMS-ISA World Bank (2014). Maize yield (`yld`), nitrogen rates (`N`), labour rates (`lab`) and seed rates (`seedha`), are measures of the per hectare application of each variable to a plot. In some cases more than one crop was grown per plot and this is captured in the analysis by the incorporation of variables referring to the number of crops on a field, and whether a field had more than one crop (`crop_count`, `crop_count2`, respectively). Furthermore, variables covering the Ph, soil quality, slope and elevation of the field, and environmental characteristics were also recorded. The longitude and latitude of each household was recorded allowing climate, soil and geographical data to be matched to households from sources including IFPRI, Worlclim, NASA and the Ethiopian Roads Agency amongst others.

Table 1: Descriptive Statistics

Variable	Mean	Median	SD	Skewness	Min	Max
yld	1,748	1,059	2,225	3.302	2.956	18,182
N	25.16	0	55.31	4.632	0	691.3
area	0.167	0.089	0.325	12.08	0	8.125
lab	334.9	81.08	3,918	49.6	0	205,882
seedha	145.4	34.83	1,934	30.85	0	74,738
phdum55_2_70	0.666	1	0.472	-0.706	0	1
crop_count2	0.561	1	0.496	-0.247	0	1
dumoxen	0.62	1	0.485	-0.495	0	1
SOC2	16.68	16.26	4.771	0.284	3.912	37.21
logslope	2.254	2.282	0.799	-0.064	0	4.439
elevation	1,800	1,842	435	-0.869	345	2,909
GGD	7.204	7.135	0.895	0.673	4.922	10.42
AI	6.949	6.7	2.442	0.242	0.765	12.85

Variable	Mean	Median	SD	Skewness	Min	Max
TS	1,137	1,062	355.1	1.006	538	3,026

Estimation procedures

The core estimation method in this paper is the stochastic frontiers (SF) method (Aigner, Lovell, and Schmidt 1977; **???**). This involves specification of the form of a production function and a composite error term that reflects both statistical error in the model and an assymetric inefficiency term.

$$y_i = \alpha + f(x_i, \beta) + \epsilon_i \tag{1}$$

$$\epsilon_i = v_i - u_i \tag{2}$$

$$v_i \sim \mathbb{N}(0, \sigma_v^2) \tag{3}$$

$$u_i \sim \mathbb{N}^+(0, \sigma_u^2) \tag{4}$$

Where y_i is the log of maize yield and x_i is the log of the inputs including the nitrogen, labour and seed rates. The composite error term ϵ_i includes a truncated normal inefficiency term u_i and a statistical error term v_i where u and v are independent. (Aigner, Lovell, and Schmidt 1977) show that the composite error term ϵ_i follows the following density

$$f_\epsilon(\epsilon_i) = \frac{2}{\sigma}\phi(\frac{\epsilon_i}{\sigma})\Phi(-\frac{\lambda\epsilon_i}{\sigma}) \tag{5}$$

Given a suitable functional form $f(.)$ The parameters of the stochastic frontiers model can be estimated using maximum likelihood (ML) estimation. Common production functions that have been used in crop yield models include the Cobb Douglass and translog production functions. The translog is the more flexible of the two and includes the squares and interactions of the inputs. In addition we can append environmental variables such as the slope and elevation of the plot to the chosen production form yielding the following model in the translog case:

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} \sum_{k=1}^{K}\sum_{j=1}^{K} \gamma_{jk} x_{ij} x_{ik} + W_i\theta + \epsilon_i \tag{6}$$

Where W_i are the environmental variables. With a slight modification to the density of ϵ_i we can allow the mean of the inefficiency term u_i to depend on exogenous factors that explain technical inefficiency such as education, age and access to markets.

Results

The resulting SF ML estimates are presented in table 1 for four models. Model 1 shows the coefficient estimates for the core translog production function terms. Model 2 extend this to include the exogenous determinants of inefficiences. Models 3 and 4 do likewise and also include the environmental variables.

Table 2: Translog production function

Variable	model 1	model 2	model 3	model 4
(Intercept)	4.658	4.019	4.636	3.861
logN	0.2	0.213	0.171	0.148
loglab	0.275	0.261	0.296	0.275

Variable	model 1	model 2	model 3	model 4
logseed	0.72	0.55	0.674	0.513
logNsq	0.047	0.028	0.048	0.031
loglabsq	0.018	0.016	0.018	0.015
logseedsq	-0.022	-0.016	-0.014	-0.011
logN:loglab	-0.023	-0.024	-0.024	-0.023
logN:logseed	-0.052	-0.033	-0.05	-0.028
loglab:logseed	-0.047	-0.044	-0.052	-0.045
sigmaSq	2.263	1.896	2.695	2.257
gamma	0.82	0.787	0.837	0.809
logarea		-0.121		-0.116
phdum55_2_70		0.122		0.107
crop_count2		0.419		0.422
dumoxen		0.069		0.074
SOC2		0.017		0.021
logslope		-0.223		-0.224
elevation		0		0
GGD		0.15		0.177
AI		0.035		0.037
TS		0		0
Z_age			0.003	-0.001
Z_sex			-0.257	-0.096
Z_ed_any			0.013	-0.029
Z_title			0.132	0.083
Z_extension			-0.675	-0.747
Z_credit			-0.414	-0.482
Z_dist_market			-0.003	0.001
Z_popEA			0	0
Z_logarea_tot			-0.112	-0.042

The raw coefficients of the exogenous determinants of technical inefficinecy cannot be interpreted as elasticities. However, the marginal effects of each variable can be computed following Kumbhakar & Sun (2013). These are observation specific and a concise estimate of the marginal effects is presented in table 2 in the form of the average partial effects (APE) for models 2 and 4 respectively.

Table 3: Marginal effects (APE) of exogenous determinants of inefficiency

	model 3	model 4
age	0.000	-0.000
sex	-0.027	-0.011
ed_any	0.001	-0.003
title	0.014	0.010
extension	-0.071	-0.088
credit	-0.044	-0.057
dist_market	-0.000	0.000
popEA	0.000	-0.000
logarea_tot	-0.012	-0.005

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
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.

Aigner, Dennis, C.A.Knox Lovell, and Peter Schmidt. 1977. “Formulation and estimation of stochastic frontier production function models.” *Journal of Econometrics* 6 (1). North-Holland: 21–37. doi:10.1016/0304-4076(77)90052-5.