Ethiopia Stochastic Frontiers Analysis Draft 2

Tomas Morley

20 March 2017

Abstract

This paper adopts the framework developed by Dijk et al. (2016) to decompose the maize yield gap in Ethiopia, combining information from household survey data and

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 referencing 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 included 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	1,698	1,046	2,140	3.344	2.956	17,978
N	25.26	0	55.18	4.655	0	691.3
area	0.17	0.093	0.327	12.03	0	8.125
lab	194.1	79.17	327.7	3.895	0	2,950
seedha	55.78	34.27	76.81	5.265	0	975.6
phdum55_2_70	0.667	1	0.471	-0.709	0	1
crop_count2	0.564	1	0.496	-0.257	0	1
dumoxen	0.625	1	0.484	-0.515	0	1
SOC2	16.67	16.11	4.77	0.294	3.912	37.21
logslope	2.249	2.272	0.799	-0.057	0	4.439
elevation	1,797	1,841	435.1	-0.87	345	2,909
GGD	7.209	7.135	0.894	0.683	4.922	10.42
AI	6.947	6.644	2.447	0.239	0.765	12.85

Variable	Mean	Median	SD	Skewness	Min	Max
TS	1,137	1,062	355.1	1.013	558	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, nesting the cobb douglass functional form inside its model specification. Which allows a LR test

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{iK}; \beta)E_i \tag{6}$$

For example in the translog case

$$Y_i = exp\Big(\alpha + \sum_{k=1}^{K}\beta_k lnX_{ik} + \sum_{k=1}^{K}\sum_{j=1}^{K}\gamma_{jk} lnX_{ik} lnX_{ij}\Big)E_i \tag{7}$$

Which can then be written in log form

$$y_i = \alpha + \sum_{k=1}^{K}\beta_k x_{ik} + \sum_{k=1}^{K}\sum_{j=1}^{K}\gamma_{jk} x_{ik} x_{ij} + \epsilon_i \tag{8}$$

Where y_i is the log of the dependent variable, and we have replaced lnX with x. The error term ϵ_i is the log of the error term and is comprised of the truncated normal technical inefficiency and the symmetric error term $\epsilon_i = v_i - u_i$ as shown above. Setting the parameters *delta* equal to zero we see that the cobb douglass form is nested within the translog. In addition we can include environmental variables such as the slope and elevation of the plot to the chosen production form yielding the following model in the translog case:

$$y_i = \alpha + \sum_{k=1}^{K}\beta_k x_{ik} + \sum_{k=1}^{K}\sum_{j=1}^{K}\gamma_{jk} x_{ij} x_{ik} + W_i\theta + \epsilon_i \tag{9}$$

This form of the translog production function is suitable for estimation via ML SF using the specification in equation 1 - 4 and the density in equation 5. We are interested in the factors which determine inefficiency and this can be incorporated by ammending equation 5 to allow the mean of the truncated normal distribution (pre-truncated normal?) to depend on exogenous deteminants of inefficiency such as such as the education, age or gender of the farmer or access to markets and extension services. In this case equation 4 becomes

$$u_i \sim \mathbb{N}^+(z_i\delta, \sigma_u^2) \tag{10}$$

Where z_i are the exogenous determinants of inefficiency. Estimation is still carried out by maximum likelihood using the density for the error term $epsilon_i$ incorporating this new term.

We cannot solve for the optimal N analytically and instead we rely on numerical methods.

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)	4.761**	4.221**	4.828**	4.135**
logN	0.224**	0.223**	0.188**	0.149*
loglab	0.329**	0.286**	0.324**	0.281**
logseed	0.569**	0.456**	0.507**	0.392**
logNsq	0.046**	0.027**	0.046**	0.029**
loglabsq	0.018**	0.020**	0.018**	0.019**
logseedsq	0.011	0.009	0.017	0.014
logN:loglab	-0.022**	-0.023**	-0.023**	-0.023**
logN:logseed	-0.059**	-0.036**	-0.053**	-0.026
loglab:logseed	-0.063**	-0.058**	-0.060**	-0.053**
logarea	NA	-0.099**	NA	-0.094**
phdum55_2_70	NA	0.118**	NA	0.106*
crop_count2	NA	0.421**	NA	0.426**
dumoxen	NA	0.065	NA	0.067
SOC2	NA	0.015**	NA	0.020**
logslope	NA	-0.216**	NA	-0.216**
elevation	NA	-0.000**	NA	-0.000**
GGD	NA	0.136**	NA	0.167**
AI	NA	0.035**	NA	0.038**
TS	NA	-0.000**	NA	-0.000**
Z_age	NA	NA	0.002	-0.002
Z_sex	NA	NA	-0.209	-0.047
Z_ed_any	NA	NA	0.030	-0.010
Z_title	NA	NA	0.139	0.112
Z_extension	NA	NA	-0.681**	-0.752**
Z_credit	NA	NA	-0.495*	-0.524**
Z_dist_market	NA	NA	-0.003*	0.002
Z_popEA	NA	NA	-0.000*	-0.000*
Z_logarea_tot	NA	NA	-0.123*	-0.044
sigmaSq	2.254**	1.885**	2.716**	2.241**
gamma	0.822**	0.786**	0.840**	0.809**

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 3 and 4 respectively.

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

	model 3	model 4
age	0.000	-0.000
sex	-0.022	-0.006
ed_any	0.003	-0.001
title	0.014	0.013
extension	-0.070	-0.089
credit	-0.051	-0.062
dist_market	-0.000	0.000
popEA	0.000	0.000
logarea_tot	-0.013	-0.005

Figure 3: Potential yield

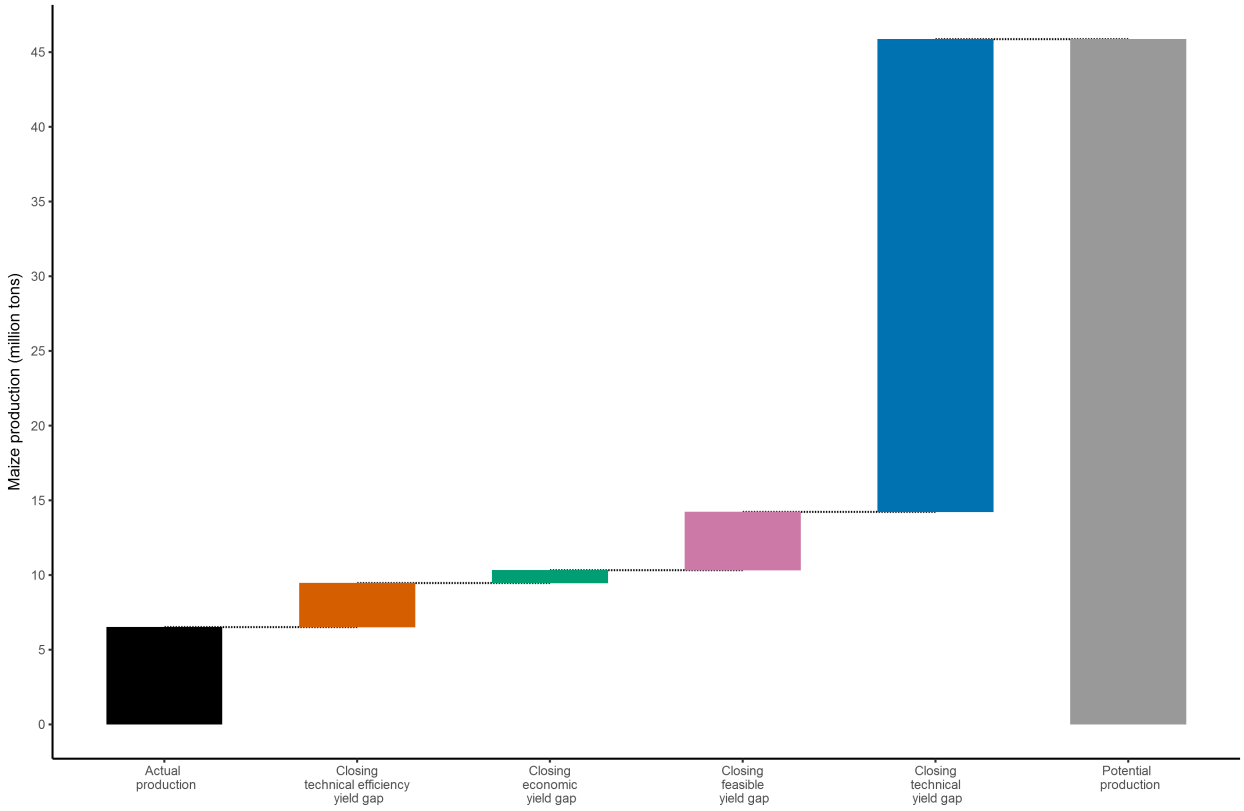

Table 4:

Zone	TEYG	EYG	EUYG	TYG	YG
AFAR	7.367	4.8739	8.951	78.81	100
AMHARA	9.042	1.5376	8.422	81.00	100
BENSHANGULGUMUZ	6.936	3.4431	12.715	76.91	100

Zone	TEYG	EYG	EUYG	TYG	YG
DIRE DAWA	6.420	2.6927	20.151	70.74	100
GAMBELLA	16.870	15.9282	8.015	59.19	100
HARARI	9.867	0.7612	30.205	59.17	100
OROMIYA	6.719	1.5890	10.687	81.00	100
SNNP	7.993	3.5884	10.099	78.32	100
SOMALI	5.572	2.8884	9.561	81.98	100
TIGRAY	8.007	7.6794	3.495	80.82	100
Total	7.741	3.0446	10.273	78.94	100

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.