Ethiopia Stochastic Frontiers Analysis Draft 1

Tomas Morley

9 March 2017

Abstract

Introduction

Econometric methodology

The core estimation method used in this paper is the stochastic frontiers method originating in Meeusen and van Den Broeck 1977 and Aigner, Lovell and Schmidt. This involves the specification of the form of a production function and a composite error term reflecting both statistical error in the model and an assymetric inefficiency term. 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 a formal comparison can be made between the two using a likelihood ratio tests to determine which model is a better fit to the data. Given our preference for the translog we proceed with this functional form in the subsequent analysis.

Therefore the returns from using one additional kg pr hecatre of nitrogen are ...

```
## Error Components Frontier (see Battese & Coelli 1992)
## Inefficiency decreases the endogenous variable (as in a production function)
## The dependent variable is logged
## Iterative ML estimation terminated after 25 iterations:
## log likelihood values and parameters of two successive iterations
## are within the tolerance limit
##
## final maximum likelihood estimates
##                             Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)              7.2034e+00  1.4302e+00   5.0368 4.734e-07 ***
## log(N + 1)               9.4714e-02  1.8799e-01   0.5038 0.6143789
## log(lab + 1)             1.6196e-01  5.6290e-02   2.8772 0.0040121 **
## I(log(N + 1)^2)          2.8124e-02  2.5207e-02   1.1157 0.2645419
## I(log(lab + 1)^2)        1.1301e-02  6.9622e-03   1.6232 0.1045409
## log(slope + 1)          -2.2326e-01  2.6942e-02  -8.2867 < 2.2e-16 ***
## elevation               -1.5823e-04  8.8365e-05  -1.7907 0.0733477 .
## log(area)               -2.1998e-01  2.0802e-02 -10.5747 < 2.2e-16 ***
## SOC                      1.9476e-01  5.0368e-02   3.8667 0.0001103 ***
## I(SOC^2)                -9.6151e-03  3.0147e-03  -3.1894 0.0014257 **
## log(rain_wq)            -3.2123e-01  2.1498e-01  -1.4942 0.1351131
## noN                      1.3109e-01  3.4226e-01   0.3830 0.7017002
## impr                     3.5960e-01  5.8605e-02   6.1360 8.465e-10 ***
## crop_count2              4.8408e-01  4.0714e-02  11.8898 < 2.2e-16 ***
## phdum_gt70              -2.7324e-01  9.5426e-02  -2.8634 0.0041915 **
## phdum55_2_70            -9.4221e-02  7.2407e-02  -1.3013 0.1931646
## GGD                      1.0645e-04  4.7388e-05   2.2463 0.0246834 *
## AI                       7.1170e-06  1.2917e-05   0.5510 0.5816379
## TS                      -5.8940e-05  8.2006e-05  -0.7187 0.4723076
```

```
## log(N + 1):log(lab + 1) -2.6267e-02  8.1556e-03  -3.2207 0.0012787 **
## sigmaSq                  1.9548e+00  9.6555e-02  20.2455 < 2.2e-16 ***
## gamma                    8.1238e-01  2.3326e-02  34.8272 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## log likelihood value: -3262.127
##
## cross-sectional data
## total number of observations = 2381
##
## mean efficiency: 0.4611315
```

```
## Error Components Frontier (see Battese & Coelli 1992)
## Inefficiency decreases the endogenous variable (as in a production function)
## The dependent variable is logged
## Iterative ML estimation terminated after 22 iterations:
## log likelihood values and parameters of two successive iterations
## are within the tolerance limit
##
## final maximum likelihood estimates
##                    Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)     6.8648e+00  1.4066e+00   4.8802 1.060e-06 ***
## log(N + 1)      1.7496e-01  3.5765e-02   4.8918 9.991e-07 ***
## log(lab + 1)    2.1898e-01  1.6614e-02  13.1803 < 2.2e-16 ***
## log(slope + 1) -2.2233e-01  2.7040e-02  -8.2222 < 2.2e-16 ***
## elevation      -1.4764e-04  8.8633e-05  -1.6658 0.0957557 .
## log(area)      -2.3203e-01  1.9728e-02 -11.7617 < 2.2e-16 ***
## SOC             1.9167e-01  5.0521e-02   3.7939 0.0001483 ***
## I(SOC^2)       -9.4256e-03  3.0266e-03  -3.1142 0.0018443 **
## log(rain_wq)   -3.3260e-01  2.1607e-01  -1.5393 0.1237315
## noN             4.4063e-01  1.4024e-01   3.1421 0.0016776 **
## impr            3.7734e-01  5.8533e-02   6.4466 1.144e-10 ***
## crop_count2     4.9749e-01  4.0713e-02  12.2193 < 2.2e-16 ***
## phdum_gt70     -2.6243e-01  9.5600e-02  -2.7451 0.0060491 **
## phdum55_2_70   -8.5446e-02  7.2443e-02  -1.1795 0.2382060
## GGD             1.1346e-04  4.7482e-05   2.3895 0.0168707 *
## AI              7.6772e-06  1.2981e-05   0.5914 0.5542450
## TS             -7.8682e-05  8.2043e-05  -0.9590 0.3375381
## sigmaSq         1.9543e+00  9.6362e-02  20.2810 < 2.2e-16 ***
## gamma           8.0867e-01  2.3614e-02  34.2446 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## log likelihood value: -3269.009
##
## cross-sectional data
## total number of observations = 2381
##
## mean efficiency: 0.4618089
```

```
## Likelihood ratio test
##
## Model 1: TL
## Model 2: CD
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  22 -3262.1
```

```
## 2  19 -3269.0 -3 13.763   0.003246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Endogeneity

Test for endogeneity using ols and a control function first stage

Using the price of nitrogen as the instrument for a simple TL and CD model

```
##
## Call:
## lm(formula = logyld ~ logN + loglab + v2, data = db1[-first_stageCD$na.action,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9526 -0.6351  0.0917  0.7104  3.3868
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.28501    0.08060  65.568  < 2e-16 ***
## logN         0.06751    0.02279   2.962 0.003086 **
## loglab       0.33459    0.01647  20.317  < 2e-16 ***
## v2           0.08788    0.02616   3.359 0.000795 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 2392 degrees of freedom
## Multiple R-squared:  0.1886, Adjusted R-squared:  0.1875
## F-statistic: 185.3 on 3 and 2392 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = logN ~ Pn, data = new_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2576 -1.1627 -0.3535  1.6091  5.3790
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.39925    0.11365   38.71   <2e-16 ***
## Pn          -0.04855    0.00176  -27.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.696 on 2394 degrees of freedom
## Multiple R-squared:  0.2412, Adjusted R-squared:  0.2408
## F-statistic: 760.8 on 1 and 2394 DF,  p-value: < 2.2e-16
```

```
##
## Call:
```

```
## lm(formula = logyld ~ logN + logNsq + loglab + loglabsq + logNloglab +
##     v2, data = new_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9189 -0.6099  0.0885  0.6861  3.4491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.425457   0.142417  38.096  < 2e-16 ***
## logN        -0.089893   0.063475  -1.416 0.156849
## logNsq       0.074702   0.011562   6.461 1.26e-10 ***
## loglab       0.228870   0.064117   3.570 0.000365 ***
## loglabsq     0.017445   0.007508   2.324 0.020226 *
## logNloglab  -0.041329   0.009167  -4.508 6.85e-06 ***
## v2           0.103283   0.025956   3.979 7.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 2389 degrees of freedom
## Multiple R-squared:  0.2085, Adjusted R-squared:  0.2065
## F-statistic: 104.9 on 6 and 2389 DF,  p-value: < 2.2e-16
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = new_dat, statistic = refitTL, R = 500)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1*  5.42545695 -0.0204132746 0.203645914
## t2* -0.08989309  0.0018179355 0.069463503
## t3*  0.07470207 -0.0001813097 0.011465811
## t4*  0.22887042  0.0091326361 0.087515723
## t5*  0.01744535 -0.0008524619 0.009548251
## t6* -0.04132880 -0.0003110388 0.009243331
## t7*  0.10328261 -0.0001904999 0.027013630
```

Test for endogeneity using ols and a control function first stage

Using the price of nitrogen as the instrument for a full TL and CD model

```
##
## Call:
## lm(formula = logN ~ Pn + loglab + log(slope + 1) + elevation +
##     log(area) + crop_count2 + age + I(age^2) + ed_any, data = db1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9787 -1.0870 -0.2633  1.1986  5.0743
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.9184778  0.4050231   7.206 7.73e-13 ***
## Pn             -0.0435972  0.0017431 -25.011  < 2e-16 ***
## loglab          0.0445800  0.0290010   1.537 0.124381
## log(slope + 1) -0.4364311  0.0437203  -9.982  < 2e-16 ***
## elevation       0.0009088  0.0000882  10.304  < 2e-16 ***
## log(area)       0.2393950  0.0330848   7.236 6.23e-13 ***
## crop_count2     0.3449484  0.0673935   5.118 3.33e-07 ***
## age             0.0388262  0.0136361   2.847 0.004447 **
## I(age^2)       -0.0004449  0.0001332  -3.339 0.000853 ***
## ed_any         -0.1302524  0.0736298  -1.769 0.077020 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.591 on 2362 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.3358, Adjusted R-squared:  0.3333
## F-statistic: 132.7 on 9 and 2362 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = logyld ~ logN + loglab + log(slope + 1) + elevation +
##     logarea + crop_count2 + v2, data = db1[-first_stageCD$na.action,
##     ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.064 -0.602  0.067  0.669  2.996
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.626e+00  1.237e-01  45.492  < 2e-16 ***
## logN            8.641e-02  2.439e-02   3.543 0.000403 ***
## loglab          2.512e-01  1.822e-02  13.791  < 2e-16 ***
## log(slope + 1) -2.136e-01  2.988e-02  -7.150 1.15e-12 ***
## elevation      -1.560e-04  6.385e-05  -2.443 0.014648 *
## logarea        -1.746e-01  2.185e-02  -7.990 2.08e-15 ***
## crop_count2     6.299e-01  4.311e-02  14.611  < 2e-16 ***
## v2              5.522e-02  2.760e-02   2.001 0.045549 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9993 on 2364 degrees of freedom
## Multiple R-squared:  0.2913, Adjusted R-squared:  0.2892
## F-statistic: 138.8 on 7 and 2364 DF,  p-value: < 2.2e-16
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = db1[-first_stageCD$na.action, ], statistic = refitCD,
##     R = 500)
```

```
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1*  5.6261022910 -0.0136743630 1.296260e-01
## t2*  0.0864082047  0.0437968161 1.115489e-02
## t3*  0.2512327468 -0.0011834533 2.119554e-02
## t4* -0.2136271137  0.0244863030 2.998385e-02
## t5* -0.0001559764 -0.0000647167 5.346924e-05
## t6* -0.1745754781 -0.0117515237 2.249500e-02
## t7*  0.6298969299 -0.0151098723 4.105459e-02
## t8*  0.0552176670 -0.0552135789 1.319649e-02
```

```
##
## Call:
## lm(formula = logN ~ Pn + loglab + loglabsq + log(slope + 1) +
##     elevation + logarea + crop_count2 + age + I(age^2) + ed_any,
##     data = db1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9561 -1.0882 -0.2537  1.1967  5.0771
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.4846584  0.4459442   5.572 2.81e-08 ***
## Pn             -0.0433896  0.0017438 -24.882  < 2e-16 ***
## loglab          0.2578708  0.0965924   2.670  0.00764 **
## loglabsq       -0.0274857  0.0118742  -2.315  0.02071 *
## log(slope + 1) -0.4402822  0.0437117 -10.072  < 2e-16 ***
## elevation       0.0009221  0.0000883  10.442  < 2e-16 ***
## logarea         0.2168021  0.0344653   6.290 3.76e-10 ***
## crop_count2     0.3470189  0.0673373   5.153 2.77e-07 ***
## age             0.0382443  0.0136259   2.807  0.00505 **
## I(age^2)       -0.0004376  0.0001331  -3.287  0.00103 **
## ed_any         -0.1282030  0.0735673  -1.743  0.08152 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.59 on 2361 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.3373, Adjusted R-squared:  0.3345
## F-statistic: 120.2 on 10 and 2361 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = logyld ~ logN + loglab + logNsq + loglabsq + logNloglab +
##     +elevation + logarea + crop_count2 + log(slope + 1) + v2,
##     data = db1[-first_stageTL$na.action, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0202 -0.5769  0.0577  0.6571  2.9388
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.641e+00  1.730e-01  32.600  < 2e-16 ***
## logN           -2.742e-03  6.182e-02  -0.044 0.964624
## loglab          2.331e-01  6.120e-02   3.809 0.000143 ***
## logNsq          5.329e-02  1.112e-02   4.792 1.75e-06 ***
## loglabsq        7.494e-03  7.424e-03   1.009 0.312901
## logNloglab     -3.533e-02  8.687e-03  -4.067 4.92e-05 ***
## elevation      -1.535e-04  6.364e-05  -2.412 0.015936 *
## logarea        -1.515e-01  2.270e-02  -6.672 3.14e-11 ***
## crop_count2     6.108e-01  4.295e-02  14.220  < 2e-16 ***
## log(slope + 1) -2.132e-01  2.976e-02  -7.165 1.03e-12 ***
## v2              6.752e-02  2.752e-02   2.454 0.014212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9926 on 2361 degrees of freedom
## Multiple R-squared:  0.3017, Adjusted R-squared:  0.2988
## F-statistic:   102 on 10 and 2361 DF,  p-value: < 2.2e-16
```

Including the ph variables does not give good results in the endogeneity setting. Including the SOC variable does not give good results in the endogeneity setting. Improved seeds are used almost every time nitrogen is used so this dummy is almost like having a nitrogen dummy in the estimation In both cases the SEs are very very large

Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero

The F test suggests that there is multicollinearity. The Nitrogen variable is not significant but we can perform an F test to see whether we can reject the idea that this variable is zero

The endogeneity of nitrogen in yield response curves has been well documented including in (Liverpool tasie, smale and mason, etc etc), and takes the form of a feeback loop between there error term to the choice of inputs, for example ... (back this up with existing literature). As pointed out in Wooldrdige 2010 and Amsler 2016 it is possible to test for the existence of endogeneity in our functional form by including the residuals from a reduced form estimation for the endogenous variable in the full structural equation. The null hypothesis that the variables are eexogenous is rejected based on the significance of a standard t or F test. This test is valid asymptotically. The control function 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. It has the additional advantage that only one control function is required rather than three corresponding to the three terms involving nitrogen that would be required in 2SLS setting. There are several estimation methods that are suitable in this framework including corrected ordinary least squares (COLS) and SF MLE.

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

various instruments have been suggested in the literature

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 endogeneity

LIML 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.

Carry out tests for endogeneity using just the estimates for the translog and the control function by themselves. WITHOUT stochastic frontiers analysis. This will establish whether endogeneity is present, and then we can move on to estimate the stochastic frontiers, firstly with C2SLS and then with LIML if it can be figured out.

We can also do a pooled endogeneity analysis -> with no panel structure. This is basically what Amsler (2016) does with what is otherwise panel data

We use LIML as an alternative to 2SLS -> but under what conditions and why do we use it?

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.

The method described in Aamsler et al 2016 relies on having

Importantly we need to decide whether we think that there is correlation between the endogenous variable, the statistical error term, the inefficiency term or both.

In the presense of endogeneity the maximum likelihood estimates of the stochastic frontiers model will not be consistent.

The translog is a non-linear function and it is therefore not possible to use 2-stage least squares. Instead we rely on the crontrol function approach (Wooldridge 2010) which incorporates the residuals from a first stage regression of the endogeneous variable on instruments exogenous to the dependent variable in the second stage.

Crucially we are assuming that the there is no correlation between the endogeneity and the inefficinecy term. So the only correlation that exists is that between the residual of the first stage and the statistical error

rational for a probit model: was the decision to use Nitrogen also related to other variables that we do not see rational for the tobit model. Was the decision to use Nitrogen related to other variables and is the "size" of that decision also important

Data

The data come from the Ethiopia .... sample of ... rural households. Because both inputs are measured in per acre terms land is not included although we do include the overall landholdings recorded by the household as an additional measure of wealth.

We consider a stochastic frontiers model using data from the Ghana LSMS-ISA survey. Although part of a panel, we ignore the panel structure due to issues with collection for the second year of the panel. The dependent variable in our analysis is the log of maize yield measured in kilograms per hectare. We consider a large number of inputs which can be broken down into household level inputs, plot level inputs and spatial inputs at a low level of aggregation.

results

Discussion

Use a flexible translog function for labour and nitrogen in the frontier proudction analysis.