The Stata Journal (2017)
17, Number 1, pp. 39–55

Fitting endogenous stochastic frontier models in Stata

Mustafa U. Karakaplan
Georgetown University
Department of Economics
Washington, DC
mukarakaplan@yahoo.com

Abstract. In this article, I introduce `sfkk`, a new command for fitting endogenous stochastic frontier models. `sfkk` provides estimators for the parameters of a linear model with a disturbance assumed to be a mixture of two components: a measure of inefficiency that is strictly nonnegative and a two-sided error term from a symmetric distribution. `sfkk` can handle endogenous variables in the frontier or the inefficiency, and the `sfkk` estimates outperform the standard frontier estimates that ignore endogeneity.

Keywords: st0466, sfkk, endogeneity, endogenous stochastic frontier models, production frontier, cost frontier, endogenous inefficiency

1 Introduction

Stochastic frontier models constitute a popular subfield of econometrics. They were introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) and further developed by many other researchers. Kumbhakar and Lovell (2000) provide an extended review of stochastic frontier models, and the literature has many empirical examples from various fields such as agriculture, aviation, banking, education, energy, and health.

Standard estimators of the stochastic frontier models estimate the parameters of a linear model with a disturbance composed of two components: a measure of inefficiency that is strictly nonnegative and a two-sided error term with a symmetric distribution. The `frontier` command provides estimators and options to fit these models. However, these standard estimators do not handle endogeneity in the model, which would exist if the determinants of the frontier or inefficiency are correlated to the two-sided error term.

Stata researchers provide a few commands that solve similar econometric issues. For example, Petrin, Poi, and Levinsohn (2004) introduced the `levpet` command, which estimates production functions with intermediate inputs as proxies to control for unobservable productivity shocks using the methodology of Levinsohn and Petrin (2003). Yasar, Raciborski, and Poi (2008) introduced the `opreg` command, which estimates production functions with selection bias or simultaneity by implementing the three-stage algorithm of Olley and Pakes (1996). While these methodologies and commands are useful for analyzing certain economic scenarios, Mutter et al. (2013) emphasize that a

complete method for handling endogeneity in stochastic frontier models is not available. Empirical studies such as Gronberg et al. (2015) apply pseudoinstrumental variable methodologies to deal with endogeneity. To address this need in the literature, Karakaplan and Kutlu (2013) offer a practical maximum-likelihood-based approach that can control for the endogeneity in the frontier or inefficiency, or both depending on the research questions. In this article, I introduce `sfkk`, a new command for fitting endogenous stochastic frontier models in the style of Karakaplan and Kutlu (2013).

2 The estimator

Karakaplan and Kutlu (2013) consider a stochastic frontier model with endogenous explanatory variables in the frontier and inefficiency functions and present the following estimator, which outperforms standard estimators that ignore the endogeneity in the model,

$$
\begin{aligned}
\ln L(\theta) &= \ln L_{y|x}(\theta) + \ln L_x(\theta) \qquad (1)\\
\ln L_{y|x}(\theta) &= \sum_{i=1}^{n}\left\{\ln 2 - \frac{1}{2}\ln\sigma_i^2 + \ln\phi\left(\frac{e_i}{\sigma_i}\right) + \ln\Phi\left(\frac{-s\lambda_i e_i}{\sigma_i}\right)\right\}\\
&= \sum_{i=1}^{n}\left\{\frac{\ln(2/\pi) - \ln\sigma_i^2 - (e_i^2/\sigma_i^2)}{2} + \ln\Phi\left(\frac{-s\lambda_i e_i}{\sigma_i}\right)\right\}\\
\ln L_x(\theta) &= \sum_{i=1}^{n}\left(\frac{-p\times\ln 2\pi - \ln(|\Omega|) - \epsilon_i'\Omega^{-1}\epsilon_i}{2}\right)\\
e_i &= y_i - x_{1i}'\beta - \frac{\sigma_{wi}}{\sigma_{cw}}\eta'\left(x_i - Z_i\delta\right)\\
\epsilon_i &= x_i - Z_i\delta\\
\sigma_i^2 &= \sigma_{wi}^2 + \sigma_{ui}^2\\
\lambda_i &= \frac{\sigma_{ui}}{\sigma_{wi}}
\end{aligned}
$$

where $\theta = (\beta', \eta', \varphi', \delta')'$ is the vector of coefficients; $y = (y_1, y_2, \ldots, y_n)'$ is the vector of dependent variables; $x = (x_1', x_2', \ldots, x_n')'$ is the matrix of endogenous variables in the model; ϕ and Φ denote the standard normal probability density function and the cumulative distribution function, respectively; $s = 1$ (or $s = -1$ for production functions); $y_i = x_{1i}'\beta + v_i - su_i = x_{1i}'\beta + (\sigma_{wi}/\sigma_{cw})\eta'(x_i - Z_i\delta) + w_i - su_i$ is the logarithm of expenditure (or output for production functions) of the ith producer; x_{1i} is a vector of exogenous and endogenous variables; $x_i = Z_i\delta + \epsilon_i$ is a $p \times 1$ vector of all endogenous variables (excluding y_i); $Z_i = I_p \bigotimes z_i'$ and z_i is a $q \times 1$ vector of all exogenous variables; v_i and ϵ_i are two-sided error terms; $u_i = \sigma_u(x_{2i};\varphi_u)u_i^* \geq 0$ is the one-sided error term capturing the inefficiency; x_{2i} is a vector of exogenous and endogenous variables; $u_i^* \sim N^+(0,1)$ is a producer-specific random component; $\sigma_{ui}^2 = \exp(x_{2i}'\varphi_u)$; $w_i = \sigma_{vi}\sqrt{1-\rho'\rho}\widetilde{w}_i = \sigma_{wi}\widetilde{w}_i$; $\sigma_{wi} = \sigma_{cw}\sigma_w(.;\varphi_w)$; $\sigma_{wi}^2 = \exp(x_{3i}'\varphi_w)$; $\sigma_{cw} > 0$ is a function of the constant term; $\sigma_{cw}^2 = \exp(\varphi_{cw})$, where φ_{cw} is the coefficient of the constant

term for $x'_{3i}\varphi_w$; $\widetilde{w}_i \sim N(0,1)$; x_{3i} is a vector of exogenous and endogenous variables that can share the same variables with x_{1i} and x_{2i}; Ω is the variance–covariance matrix of ϵ_i; σ^2_{vi} is the variance of v_i; and ρ is the vector representing the correlation between $\widetilde{\epsilon}_i$ and v_i. The details about the assumptions and how the estimator is derived are presented in Karakaplan and Kutlu (2013).

Moreover, Karakaplan and Kutlu (2013) provide the following formula to predict the efficiency, $\text{EFF}_i = \exp(-u_i)$:

$$E\left\{\exp(-su_i)|e_i\right\}^s = \left\{\frac{1-\Phi(s\sigma_i^* - \mu_i^*/\sigma_i^*)}{1-\Phi(-\mu_i^*/\sigma_i^*)}\exp\left(-s\mu_i^* + \frac{1}{2}\sigma_i^{*2}\right)\right\}^s \tag{2}$$

$$\mu_i^* = \frac{-se_i\sigma_{ui}^2}{\sigma_i^2}$$

$$\sigma_i^{*2} = \frac{\sigma_{wi}^2\sigma_{ui}^2}{\sigma_i^2}$$

Finally, a test for endogeneity is proposed by Karakaplan and Kutlu (2013). In this test, the joint significance of the components of the η term is checked. If the joint significance of the components is rejected, then correction for endogeneity is not necessary, and the model can be fit by traditional frontier models. However, if the components of the η term are jointly significant, then there is endogeneity in the model, and a correction through (1) would be necessary.

3 The sfkk command

Gould, Pitblado, and Poi (2010) provide an excellent guideline for researchers who need to compute maximum likelihood estimators that are not available as prepackaged routines. Following their suggestions and using Stata's powerful `ml` tools, I programmed the `sfkk` command, which can estimate (1) and (1). The `sfkk` package includes three files: `sfkk.ado`, `sfkk_ml.ado`, and `sfkk.sthlp`. `sfkk.ado` provides the main estimation syntax that users access through running `sfkk`. `sfkk_ml.ado` includes the evaluator subroutines that `sfkk` calls for the actual estimation of the parameters. The default subroutine in `sfkk_ml.ado` is a method-d0 evaluator that calculates the overall log likelihood. `sfkk_ml.ado` has another subroutine that is a method-lf0 evaluator called by `sfkk` if the `fast(#)` option is specified. The method-lf0 evaluator speeds up the regression because it does not compute any derivatives. The `fast(#)` option combines this evaluator with a tolerance-based methodology to complete the estimation faster. The postestimation routines, such as predicting the efficiency, testing the endogeneity, and documenting the results, are handled in `sfkk.ado`. Finally, `sfkk.sthlp` is the `sfkk` command's help file, which provides an extended version of the subsections below with further details, such as option abbreviations, stored results, and examples with clickable features.

3.1 Syntax

Estimation syntax

`sfkk` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [`,` `noconstant` `production` `cost`
`endogenous(`*endovarlist*`)` `instruments(`*ivarlist*`)` `exogenous(`*exovarlist*`)`
`leaveout(`*lovarlist*`)` `uhet(`*uvarlist*[`,` `noconstant`]`)` `whet(`*wvarlist*`)`
`initial(`*matname*`)` `delve` `fast(`#`)` `difficult` `technique(`*algorithm_spec*`)`
`iterate(`#`)` `mlmodel(`*model_options*`)` `mlmax(`*maximize_options*`)` `header` `timer`
`beep` `compare` `efficiency(`*effvar*[`,` `replace`]`)` `test` `nicely`
`mldisplay(`*display_options*`)`]

Version syntax

`sfkk, version`

Replay syntax

`sfkk` [`,` `level(`#`)`]

3.2 Options for the estimation syntax

Frontier

`noconstant` suppresses the constant term (intercept) in the frontier.

`production` specifies that the model to be fit is a production frontier model. The default is `production`.

`cost` specifies that the model to be fit is a cost frontier model. The default is `production`.

Equations

`endogenous(`*endovarlist*`)` specifies that the variables in *endovarlist* be treated as endogenous. By default, `sfkk` assumes the model is exogenous.

`instruments(`*ivarlist*`)` specifies that the variables in *ivarlist* be used as instrumental variables to handle endogeneity. By default, `sfkk` assumes the model is exogenous.

`exogenous(`*exovarlist*`)` specifies that *exovarlist* is the complete list of included exogenous variables. The default for the complete list of included exogenous variables is *indepvars* + *uvarlist* + *wvarlist*. Depending on the model, *exovarlist* can be different from *indepvars* + *uvarlist* + *wvarlist*. For an illustration, please see the `sfkk` help file. `exogenous()` cannot be used with `leaveout()`. The `exogenous()` option is seldom used and can safely be omitted.

`leaveout(`*lovarlist*`)` specifies that the variables in *lovarlist* be removed from the default list of included exogenous variables, which is *indepvars* + *uvarlist* + *wvarlist*. Depending on the model, some variables, such as functions of some included exogenous variables, can be left out of the complete list of included exogenous variables. For an illustration, please see the `sfkk` help file. `leaveout()` cannot be used with `exogenous()`. The `leaveout()` option is seldom used and can safely be omitted.

`uhet(`*uvarlist*[`, noconstant`]`)` specifies the inefficiency component be heteroskedastic, with the variance function depending on a linear combination of *uvarlist*. Specifying `noconstant` suppresses the constant term from the variance function.

`whet(`*wvarlist*`)` specifies that the idiosyncratic error component be heteroskedastic, with the variance function depending on a linear combination of *wvarlist*.

Regression

`initial(`*matname*`)` specifies that *matname* is the initial value matrix.

`delve` provides a regression-based methodology to search for better initial values. The default is to use `ml search`. `delve` is often successful in finding better initial values. Using `delve` is recommended.

`fast(#)` provides a tolerance-based methodology to complete the regression faster. # can be specified to take any value larger than 0. The regression completes faster with larger values of #, but larger values of # result in less accurate findings. Experimenting with various values of # is suggested because different values of # work better with different models. Using `fast()` is recommended to explore the direction of the maximization problem faster. However, to improve the accuracy of the findings, one should avoid using `fast()` once the model is decided and specification is finalized.

`difficult` specifies that the likelihood function is likely to be difficult to maximize because of nonconcave regions. When the message "not concave" appears repeatedly, `ml`'s standard stepping algorithm may not be working well. `difficult` specifies that a different stepping algorithm be used in nonconcave regions. There is no guarantee that `difficult` will work better than the default; sometimes it is better and sometimes it is worse. The `difficult` option should be used only when the default stepper declares convergence and the last iteration is "not concave" or when the default stepper is repeatedly issuing "not concave" messages and producing only tiny improvements in the log likelihood.

`technique(`*algorithm_spec*`)` specifies how the likelihood function is to be maximized. The following algorithms are allowed. For details, see Gould, Pitblado, and Poi (2010).

`technique(nr)` specifies Stata's modified Newton–Raphson algorithm.

`technique(bhhh)` specifies the Berndt–Hall–Hall–Hausman algorithm, which is allowed only with the `fast()` option.

`technique(dfp)` specifies the Davidon–Fletcher–Powell algorithm.

`technique(bfgs)` specifies the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The default is `technique(bfgs)`.

Switching between algorithms is possible by specifying more than one algorithm in the `technique()` option. By default, an algorithm is used for five iterations before switching to the next algorithm. To specify a different number of iterations, include the number after the technique in the option. For example, specifying `technique(bfgs 10 nr 1000)` requests that `sfkk` perform 10 iterations with the BFGS algorithm, followed by 1,000 iterations with the Newton–Raphson algorithm, followed by 10 more iterations with the BFGS algorithm, and so on. The process continues until the convergence or maximum number of iterations is reached.

`iterate(#)` specifies the maximum number of iterations. When the number of iterations equals #, the optimizer stops and presents the current results. If the convergence gets declared before this threshold is reached, the optimizer stops and presents the optimized results. The default is `iterate(16000)`, which is the current value of *maxiter*.

`mlmodel(`*model_options*`)` controls the `ml model` options; it is seldom used.

`mlmax(`*maximize_options*`)` controls the `ml max` options; it is seldom used.

Reporting

`header` displays a summary of the model constraints in the beginning of the regression. `header` provides a way to check the model specifications quickly while the estimation is running or provides a guide to distinguish different regression results that are kept in a single log file.

`timer` displays the total elapsed time `sfkk` took to complete. The total elapsed time is measured from the moment the command is entered to the moment the reporting of all findings is completed.

`beep` produces a beep when `sfkk` reports all findings. `beep` is useful for multitasking.

`compare` fits the specified model with the exogeneity assumption and displays the regression results after displaying the endogenous model regression results.

`efficiency(`*effvar*`[ , replace ])` generates the production or cost efficiency variable *effvar*`_EN` once the estimation is completed and displays its summary statistics in detail. The option automatically extends any specified variable name *effvar* with `_EN`. If the `compare` option is specified, `efficiency()` also generates *effvar*`_EX`, the production or cost efficiency variable of the exogenous model, and displays its summary statistics. Specifying `replace` replaces the contents of the existing *effvar*`_EN` and *effvar*`_EX` with the new efficiency values from the current model.

`test` provides a method to test the endogeneity in the model. It tests the joint significance of the components of the eta term and reports the findings after displaying

the regression results. For more information about `test`, see Karakaplan and Kutlu (2013).

`nicely` displays the regression results in a single table. `nicely` requires `estout`, a user-written command by Jann (2005), to format some parts of the table, and the `sfkk` table style resembles that of Karakaplan and Kutlu (2013). The `nicely` option checks whether the `estout` package is installed on Stata, and if not, the `nicely` option installs the package. If the `compare` option is specified, `nicely` displays the exogenous and endogenous models with their corresponding equations and statistics side by side in a single table for easy comparison. `nicely` estimates the production or cost efficiency and tests endogeneity and reports them in the table even if the `efficiency(`*effvar*`)` or `test` option is not specified.

`mldisplay(`*display_options*`)` controls the `ml display` options; it is seldom used.

3.3 Options for the version and replay syntax

`version` displays the version of `sfkk` installed on Stata and the program author information. This option can be used only in the version syntax.

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`. This option can be used in the replay syntax or in `mldisplay(`*display_options*`)`.

4 Examples

In this section, I illustrate `sfkk` in three different examples. The first two examples analyze randomly generated datasets in a cost setting and in a production setting. These two datasets are for illustrative purposes, and the results do not represent a specific industry. The last example, however, examines a stochastic cost frontier model with a real dataset that come from the U.S. K–12 education sector. Eta endogeneity test results indicate that all models in the examples suffer from endogeneity problems. Correcting the endogeneity through `sfkk` results in substantially different coefficient estimates and efficiency scores.

4.1 Endogenous stochastic cost frontier example

The first example uses a cross-sectional dataset with 750 observations and fits a cost model in which one of the frontier variables (`z1`) and a variable determining cost inefficiency (`z2`) are endogenous. Two instrumental variables are used (`iv1` and `iv2`) to handle the endogeneity. The `header` option summarizes the model specification.

```
. use http://www.mukarakaplan.com/files/sfkkcost.dta
. sfkk y x1 x2 x3 z1, cost u(z2) en(z1 z2) i(iv1 iv2) header delve compare
> nicely timer

17 Nov 2016 15:22:46

ENDOGENOUS STOCHASTIC COST FRONTIER MODEL (Model EN)

Dependent Variable: y

Frontier Variable(s): Constant x1 x2 x3 z1

U Variable(s): Constant z2

W Variable(s): Constant

Endogenous Variable(s): z1 z2

Excluded Instrument(s): iv1 iv2

Exogenous Variable(s): iv1 iv2 x1 x2 x3

Delving into the problem...
initial:       log likelihood =  -1286.915
rescale:       log likelihood =  -1286.915
rescale eq:    log likelihood = -121.43367
Iteration 0:   log likelihood = -121.43367
Iteration 1:   log likelihood = -63.571795  (backed up)
Iteration 2:   log likelihood = -60.637741  (backed up)
Iteration 3:   log likelihood = -16.951509  (backed up)
Iteration 4:   log likelihood =  15.209068  (backed up)
Iteration 5:   log likelihood =  742.89103  (backed up)
  (output omitted)
Iteration 52:  log likelihood =  1604.6298
Iteration 53:  log likelihood =  1604.6298

Analyzing the exogenous comparison model (Model EX)...
initial:       log likelihood = -1143.5903
alternative:   log likelihood =  -633.5094
rescale:       log likelihood =  -633.5094
rescale eq:    log likelihood =  -358.6175
initial:       log likelihood =  -358.6175
rescale:       log likelihood =  -358.6175
rescale eq:    log likelihood =   21.18479
Iteration 0:   log likelihood =   21.18479
Iteration 1:   log likelihood =  40.199285  (backed up)
  (output omitted)
Iteration 24:  log likelihood =  886.56289
Table: Estimation Results
```

	Model EX		Model EN	
Dep.var: y				
Constant	0.529***	(0.030)	0.225**	(0.081)
x1	-0.084***	(0.016)	-0.042*	(0.020)
x2	0.067**	(0.025)	0.112***	(0.032)
x3	0.058*	(0.029)	0.354***	(0.071)
z1	0.010	(0.021)	0.424***	(0.087)

```
Dep.var: ln(σ²_u)
Constant                -8.137***  (1.018)   -7.710***  (0.580)
z2                       3.701***  (0.964)    4.272***  (0.655)
-------------------------------------------------------------------
Dep.var: ln(σ²_v)
Constant                -5.290***  (0.069)
-------------------------------------------------------------------
Dep.var: ln(σ²_w)
Constant                                   -5.435***  (0.074)
-------------------------------------------------------------------
eta1 (z1)                                  -0.466***  (0.090)
eta2 (z2)                                  -0.075***  (0.022)
-------------------------------------------------------------------
eta Endogeneity Test                        X2=33.99  p=0.000
-------------------------------------------------------------------
Observations                 750                 750
Log Likelihood             886.56              1604.63
Mean Cost Efficiency       0.9760              0.9670
Median Cost Efficiency     0.9816              0.9765
-------------------------------------------------------------------
Notes: Standard errors are in parentheses. Asterisks indicate
significance at the 0.1% (***), 1% (**) and 5% (*) levels.
-------------------------------------------------------------------
```

(*output omitted*)

```
Completed in 0 hour(s), 0 minute(s) and 57 second(s).
```

In the output, the model that ignores endogeneity is `Model EX`, and the model that captures endogeneity is `Model EN`. Individual eta terms of `z1` and `z2` are both significant at the 0.1% level, and the eta endogeneity test result indicates that correction for endogeneity is needed. Looking at the coefficients of the endogenous variables, we see that `z1` is significant in `Model EN`, while it is not significant in `Model EX`, and the difference between the coefficients is substantial. `z2` is significant in both `Model EX` and `Model EN`, but its effect size is larger in `Model EN`. Mean and median cost efficiencies in `Model EN` are slightly less than that in `Model EX`, which shows that in `Model EX`, producers appear more cost efficient than they actually are when endogeneity is properly handled.

4.2 Endogenous stochastic production frontier example

The second example uses a cross-sectional dataset with 500 observations and fits a production model in which one of the frontier variables (`z1`) and a variable determining production inefficiency (`z2`) are endogenous. Two instrumental variables are used (`iv1` and `iv2`) to handle the endogeneity. Notice that in this example, the `compare` and `nicely` options are not specified; instead, the `efficiency()` and `test` options are specified. So the results are presented in raw format with prediction equations but no exogenous comparison model.

```
. use http://www.mukarakaplan.com/files/sfkkprod.dta, clear
. sfkk y x1 x2 z1, prod u(z2) en(z1 z2) i(iv1 iv2) delve header eff(pef) test

17 Nov 2016 15:23:43

ENDOGENOUS STOCHASTIC PRODUCTION FRONTIER MODEL (Model EN)

Dependent Variable: y

Frontier Variable(s): Constant x1 x2 z1

U Variable(s): Constant z2

W Variable(s): Constant

Endogenous Variable(s): z1 z2

Excluded Instrument(s): iv1 iv2

Exogenous Variable(s): iv1 iv2 x1 x2

Delving into the problem...

initial:       log likelihood = -915.26945
rescale:       log likelihood = -674.50035
rescale eq:    log likelihood = -199.47821
Iteration 0:   log likelihood = -199.47821
Iteration 1:   log likelihood = -190.32271  (backed up)
Iteration 2:   log likelihood = -183.28355  (backed up)
Iteration 3:   log likelihood = -164.08851  (backed up)
Iteration 4:   log likelihood =  14.005305  (backed up)
Iteration 5:   log likelihood =  233.97196  (backed up)
  (output omitted)
Iteration 47:  log likelihood =  713.98035
Iteration 48:  log likelihood =  713.98036

Endogenous stochastic prod frontier model with normal/half-normal specification

                                              Number of obs     =        500
                                              Wald chi2(3)      =     129.50
Log likelihood =  713.98036                   Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
frontier_y   |
          x1 |   .1860659   .0314463     5.92   0.000     .1244324    .2476995
          x2 |   .1322843   .0326762     4.05   0.000     .0682402    .1963284
          z1 |  -.7470345   .1112763    -6.71   0.000     -.965132    -.528937
       _cons |   .6314415   .0320355    19.71   0.000     .5686531      .69423
-------------+----------------------------------------------------------------
ivr1_z1      |
         iv1 |   .6650172   .0891591     7.46   0.000     .4902686    .8397658
         iv2 |   .1354525   .0392074     3.45   0.001     .0586074    .2122976
          x1 |  -.0764425   .0370097    -2.07   0.039    -.1489801   -.0039049
          x2 |   .2075451   .0390051     5.32   0.000     .1310965    .2839937
       _cons |  -.0069407   .0351824    -0.20   0.844    -.0758969    .0620155
-------------+----------------------------------------------------------------
eta1_z1      |
       _cons |   .4570557   .1139459     4.01   0.000     .2337257    .6803856
------------------------------------------------------------------------------
```

```
ivr2_z2
         iv1   -.0263354   .0672439    -0.39   0.695    -.1581311    .1054603
         iv2    -.106181   .0203197    -5.23   0.000    -.1460068   -.0663551
          x1    .0587884   .0239589     2.45   0.014     .0118298    .1057469
          x2    -.023259   .0255903    -0.91   0.363     -.073415     .026897
       _cons    .3257763   .0240677    13.54   0.000     .2786045    .3729481
-------------+----------------------------------------------------------------
eta2_z2
       _cons    .6635172   .0568901    11.66   0.000     .5520146    .7750198
-------------+----------------------------------------------------------------
lnsig2u
          z2    8.207562    1.46807     5.59   0.000     5.330198    11.08492
       _cons   -7.095995    .827744    -8.57   0.000    -8.718344   -5.473647
-------------+----------------------------------------------------------------
lnsig2w
       _cons   -4.818588   .1774659   -27.15   0.000    -5.166415   -4.470761
------------------------------------------------------------------------------

                     eta Endogeneity Test
-----------------------------------------------------------

Ho: Correction for endogeneity is not necessary.
Ha: There is endogeneity in the model and correction is needed.

 ( 1)  [eta1_z1]_cons = 0
 ( 2)  [eta2_z2]_cons = 0

           chi2(  2) =  155.14
         Prob > chi2 =    0.0000

Result: Reject Ho at 0.1% level.

    Summary of Model EN Production Efficiency
-------------------------------------------------------

Mean Efficiency         .91521329
Median Efficiency       .9364124
Minimum Efficiency      .4562384
Maximum Efficiency      .98385695
Standard Deviation      .07200863

where
0 = Perfect production inefficiency
1 = Perfect production efficiency
```

(*output omitted*)

In this example, the eta endogeneity test result rejects the null hypothesis at the 0.1% level, which means that a correction for endogeneity in the model is needed. The coefficient of `z1` in the frontier is negative and significant. Moreover, the coefficient of `z2` in the inefficiency term is positive and significant. If the `compare` option was specified, the results from an exogenous comparison model would show that the coefficient of `z2` is negative and larger in absolute terms if its endogeneity is not handled. This conclusion would also be reflected in production efficiency estimates. The mean production efficiency is 0.915 in `Model EN`, whereas the same statistic is 0.982 in `Model EX`, which is not displayed here. So producers are not as efficient in production as they would appear in a standard frontier model that ignores endogeneity. The `efficiency()` option saves the efficiency scores from `Model EN` as a variable, and when the `compare` option is specified, `efficiency()` would also save the efficiency scores from `Model EX` as a variable. Having these two variables would enable a graphical comparison of the models as shown below.

```
. capture sfkk y x1 x2 z1, prod u(z2) en(z1 z2) i(iv1 iv2) delve
> efficiency(pef, replace) compare
. histogram pef_EX if pef_EX>0.8, w(0.01) freq xtitle("Production efficiency")
> xtick(0.8(0.05)1) xlabel(0.8(0.05)1) ytick(0(25)300) ylabel(0(50)300)
> ytitle("Number of producers") color(gs8) lcolor(gs4) title("Model EX")
> graphregion(color(gs14))
(bin=19, start=.81446136, width=.01)
```

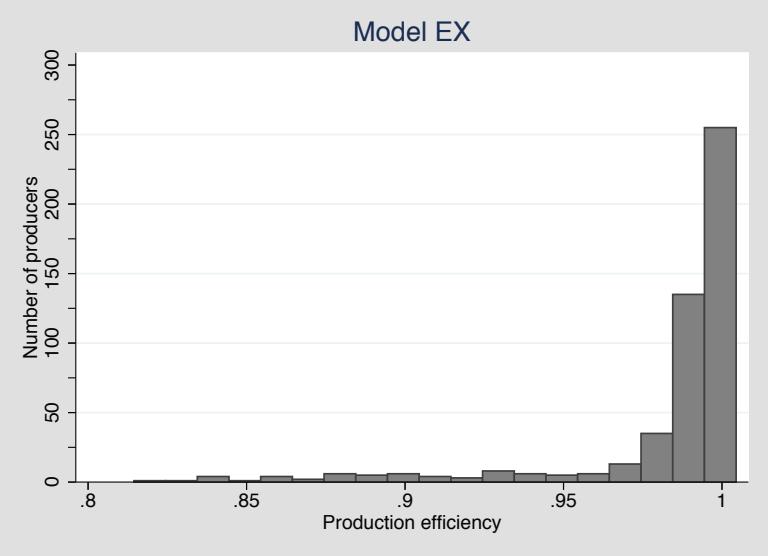

Figure 1. Graphical representation—`Model EX`

```
. histogram pef_EN if pef_EN>0.8, w(0.01) freq xtitle("Production efficiency")
> xtick(0.8(0.05)1) xlabel(0.8(0.05)1) ytick(0(25)300) ylabel(0(50)300)
> ytitle("Number of producers") color(gs8) lcolor(gs4) title("Model EN")
> graphregion(color(gs14))
(bin=19, start=.80048907, width=.01)
```

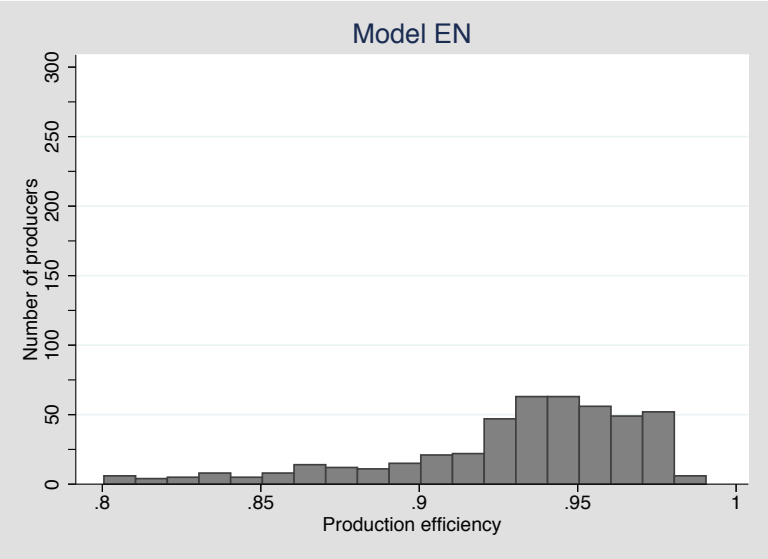

Figure 2. Graphical representation—`Model EN`

4.3 Example from the U.S. K–12 education sector

For this example, I use the main data from the National Center for Education Statistics and New York State Education Department. The cross-sectional dataset consists of 635 traditional public school districts in the 2011–2012 school year. Following the models in Gronberg et al. (2015) and Karakaplan and Kutlu (2015), I set the dependent variable as the natural logarithm of actual current operating expenditures per pupil (`expend`). Cost frontier variables include district enrollment (`enroll`) and the square of district enrollment (`enroll2`) as the output quantity variables, an index of district-level academic performance (`scores`) as the output quality variable, an index of input prices (`prices`) based on the derived prices of instructional material and wages of education personnel,[1] and an index of district-level student characteristics (`body`) that measures the effects of environmental factors such as the percentage of special education students. Cost inefficiency is modeled with a Herfindahl–Hirschman index (`hhi`) of education market concentration ranging between zero and one, with one indicating a monopoly setting. I control for the endogeneity of `scores` and `hhi` by using the number of small

1. The comparable wage index was originally produced by Taylor and Fowler (2006) and is regularly updated at http://bush.tamu.edu/research/faculty/Taylor_CWI/.

streams in a county (`streams`) and the unemployment rate in a county (`unemp`) as the instrumental variables.[2]

In this example, I specify the `compare`, `nicely`, and `header` options in the command line. `Model EX` represents the model that does not control for the endogeneity in the model (comparable with a standard frontier command estimation), and `Model EN` represents the model that handles the endogeneity. The coefficient associated with `scores` is expected to be positive and increasing costs. The coefficient associated with `hhi` is also expected to be positive and increasing cost inefficiency, because authorities in more concentrated markets may be less careful about how they spend their resources. These two coefficients are expected to be downward biased in `Model EX`. Looking at the results in the output, we see that individual eta terms of `hhi` and `scores` are both significant at the 0.1% level and that the eta endogeneity test result shows that correction for endogeneity is needed. As illustrated in the table, the coefficient of `scores` is positive and significant, and the coefficient of `hhi` is positive but not significant in `Model EX`. In `Model EN`, these two coefficients are substantially larger and significant. Because there are differences in the magnitudes and significance of the coefficients in `Model EX` and `Model EN`, controlling for the endogeneity in the model is important. The differences in the mean and median cost efficiencies of `Model EX` and `Model EN` indicate this importance as well.

```
. use http://www.mukarakaplan.com/files/sfkkedu.dta, clear
. sfkk expend enroll enroll2 scores prices body, cost uhet(hhi)
> endogenous(scores hhi) instruments(streams unemp) delve compare nicely
> header beep timer technique(dfp 25 bfgs 25)

17 Nov 2016 15:24:32

ENDOGENOUS STOCHASTIC COST FRONTIER MODEL (Model EN)
Dependent Variable: expend
Frontier Variable(s): Constant enroll enroll2 scores prices body
U Variable(s): Constant hhi
W Variable(s): Constant
Endogenous Variable(s): scores hhi
Excluded Instrument(s): streams unemp
Exogenous Variable(s): streams unemp enroll enroll2 prices body

Delving into the problem...
initial:       log likelihood = -29952.478
rescale:       log likelihood = -23152.856
rescale eq:    log likelihood = -1527.7189
(setting technique to dfp)
Iteration 0:   log likelihood = -1527.7189
Iteration 1:   log likelihood = -978.04977  (backed up)
Iteration 2:   log likelihood = -956.73029  (backed up)
  (output omitted)
Iteration 96:  log likelihood =  949.71336
Iteration 97:  log likelihood =  949.71337
```

2. The topographical data comes from the U.S. Geological Survey Geographic Names Information System. The unemployment rates data come from New York State's Department of Labor.

```
Analyzing the exogenous comparison model (Model EX)...
initial:       log likelihood =  -12515.72
alternative:   log likelihood = -6631.5553
rescale:       log likelihood = -1606.6784
rescale eq:    log likelihood = -1351.0361
initial:       log likelihood = -1351.0361
rescale:       log likelihood = -1351.0361
rescale eq:    log likelihood = -1351.0361
Iteration 0:   log likelihood = -1351.0361
Iteration 1:   log likelihood = -1329.4522  (backed up)
  (output omitted)
Iteration 28:  log likelihood =  296.36882
Table: Estimation Results
```

	Model EX		Model EN	
Dep.var: expend				
Constant	10.880***	(0.305)	11.531***	(0.522)
enroll	-0.531***	(0.080)	-0.823***	(0.156)
enroll2	0.031***	(0.005)	0.048***	(0.010)
scores	0.187***	(0.046)	1.574***	(0.347)
prices	0.656***	(0.042)	0.304**	(0.114)
body	3.432***	(0.294)	6.089***	(0.795)
Dep.var: ln(σ^2_u)				
Constant	-3.685***	(0.288)	-6.939***	(1.073)
hhi	0.373	(0.434)	5.502***	(1.349)
Dep.var: ln(σ^2_v)				
Constant	-4.324***	(0.167)		
Dep.var: ln(σ^2_w)				
Constant			-3.989***	(0.084)
eta1 (scores)			-1.410***	(0.349)
eta2 (hhi)			-0.263***	(0.060)
eta Endogeneity Test			X2=34.64	p=0.000
Observations	635		635	
Log Likelihood	296.37		949.71	
Mean Cost Efficiency	0.8756		0.9394	
Median Cost Efficiency	0.8869		0.9574	

```
Notes: Standard errors are in parentheses. Asterisks indicate
significance at the 0.1% (***), 1% (**) and 5% (*) levels.
```

```
  (output omitted)
Completed in 0 hour(s), 0 minute(s) and 47 second(s).
```

5 Conclusion

In this article, I follow the recent advances in the estimation of endogenous stochastic frontier models presented by Karakaplan and Kutlu (2013) and offer `sfkk`, a new

command to estimate such models in Stata. `sfkk` can handle endogenous variables in the frontier or the inefficiency, and examples show that `sfkk` estimates outperform the standard frontier estimates that ignore endogeneity. `sfkk` provides many options that can become handy for researchers from different fields such as agriculture, aviation, banking, education, energy, and health.

6 Acknowledgments

I thank Levent Kutlu, Isabel Canette, Ben Jann, and Kit Baum for their amazing support.

7 References

Aigner, D. J., C. A. K. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21–37.

Gould, W., J. Pitblado, and B. Poi. 2010. *Maximum Likelihood Estimation with Stata*. 4th ed. College Station, TX: Stata Press.

Gronberg, T. J., D. W. Jansen, M. U. Karakaplan, and L. L. Taylor. 2015. School district consolidation: Market concentration and the scale-efficiency tradeoff. *Southern Economic Journal* 82: 580–597.

Jann, B. 2005. Making regression tables from stored estimates. *Stata Journal* 5: 288–308.

Karakaplan, M. U., and L. Kutlu. 2013. Handling endogeneity in stochastic frontier analysis. http://www.mukarakaplan.com/Karakaplan - EndoSFA.pdf.

———. 2015. Consolidation policies and saving reversals. http://www.mukarakaplan.com/Karakaplan - Reversals.pdf.

Kumbhakar, S. C., and C. A. K. Lovell. 2000. *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.

Levinsohn, J., and A. Petrin. 2003. Estimating production functions using inputs to control for unobservables. *Review of Economic Studies* 70: 317–341.

Meeusen, W., and J. van den Broeck. 1977. Efficiency estimation from Cobb–Douglas production functions with composed error. *International Economic Review* 18: 435–444.

Mutter, R. L., W. H. Greene, W. Spector, M. D. Rosko, and D. B. Mukamel. 2013. Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes. *Journal of Productivity Analysis* 39: 101–110.

Olley, G. S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1297.

Petrin, A., B. P. Poi, and J. Levinsohn. 2004. Production function estimation in Stata using inputs to control for unobservables. *Stata Journal* 4: 113–123.

Taylor, L. L., and W. J. Fowler, Jr. 2006. A comparable wage approach to geographic cost adjustment. Research and Development Report NCES-2006-321, National Center for Education Statistics. https://nces.ed.gov/pubs2006/2006321.pdf.

Yasar, M., R. Raciborski, and B. Poi. 2008. Production function estimation in Stata using the Olley and Pakes method. *Stata Journal* 8: 221–231.

About the author

Mustafa U. Karakaplan has a PhD in Economics from Texas A&M University.