William Greene
Stern School of Business, New York University

Assignment 8
Stochastic Frontier Models

This assignment will use two data sets, the airlines data and the Spanish dairy data used in Assignment 1.

I. Airlines data

This part of the assignment is based on the airlines data contained in airlines.lpj. Though they are a panel of 25 firms observed over varying numbers of years, for present purposes, we will treat them as a cross section.

1. Deterministic frontier models – using OLS

Although the maximum likelihood estimator of the normal – half normal stochastic frontier model is now standard, researchers still examine the corrected least squares and modified least squares estimators occasionally. One interpretation is that the frontier is deterministic – there is no 'noise' component, only inefficiency. This is consistent with the deterministic frontier model

$$y(i) = \alpha + \beta' x(i) \ - \ u(i)$$

a. The <u>**corrected least squares**</u> (COLS) estimator is computed by fitting the frontier model by ordinary least squares, then computing u(i) = [maximum e(i)] - e(i) where e(i) is the OLS residuals. These corrected residuals provide the interesting information about inefficiency in the data. Using the airlines data, let **X = one,lf,le,ll,lp** (i.e., omit materials), fit the production function model (**LQ** is the log of the output variable). Compute the residuals. (You can just add **;Res = e** to your **REGRESS** command. Now, compute the COLS residuals with

```
REGRESS ; Lhs = lq ; Rhs = x ; Res = e $
CALC ; Maxe = Max(e) $
CREATE ; uicols = Maxe – e $
```

Use

```
DSTAT ; Rhs = uicols $
```

to obtain the mean inefficiency suggested by this method. (Note, data are in logs, so the mean is roughly the percentage. For example, if the sample mean is 0.30, this suggests 30% inefficiency on average.) Are the results plausible? Look at the minimum and maximum as well.

This is all automated. You can use

```
FRONTIER    ; Lhs = lq ; Rhs = x ; Model = cols ; Techeff = eucols $
KERNEL ; Rhs = eucols $
```

Give these commands. Note the results provided for analyzing the frontier function

b. The modified least squares (MOLS) estimator is also a deterministic frontier estimator (at least at this point). In this computation, we 'estimate' ui with

Est.u(i) = Est.[E[e(i)]] – e(i).

In order to compute the expectation, we need to assume a distribution. (Note by construction, the OLS residuals have mean zero, so the sample mean won't help.) Suppose we assume the distribution of u(i) is exponential with parameter θ. Then, the sample standard deviation of the OLS residuals is an estimator of θ, and our estimates of u(i) can be θ - e(i). You can obtain the results using

CALC ; theta = sdv(e) $
CREATE ; uimols = theta – e $
DSTAT ; Rhs = uimols $

Now use **DSTAT** to examine your results. What is the estimate of average inefficiency. Note that the minimum value is negative, implying that at least one firm/year is 'superefficient.' This is an unfortunate shortcoming of this method of estimating the model. What is at work even more than the estimation method is the assumption of a deterministic frontier model. The MOLS method can also be applied to the stochastic frontier model (see Section 2.4.8 in the Greene survey). Since maximum likelihood estimation is so easy, and is more efficient, we will leave these corrected and modified OLS estimators at this point.

2. Stochastic frontier model

Still using the airlines data, fit a stochastic frontier model. Retain the Cobb-Douglas assumption **X = one,lf,le,ll,lp.** Examine your results. Are the coefficient estimates plausible? What are the estimates of σ_v and σ_u. Which appears to be the greater source of variation in log output, inefficiency, u, or noise, v? What is the log likelihood function for the frontier model? The kernel compares the COLS results to these based on the stochastic frontier.

FRONTIER ; Lhs = lq ; Rhs = X ; Techeff = euihn ; Eff=uihn$
KERNEL ; Rhs = euihn, eucols $

The coefficients and variance parameters look mostly ok. The negative output elasticity (also significant) for labor is a problem, however. From the output for the model, we have the estimates of the variance parameters,

```
              Sigma(v)        =       .14824
              Sigma(u)        =       .18661
```

so the "u" part appears to be larger. But, be careful. The standard deviation of u is $[(\pi - 2)/\pi]^{1/2} \sigma_u$ = 0.11249.

CALC ; List ; SDU = Sqr((pi - 2)/pi) * .18661 $

3. Exponential and Rayleigh frontiers

Using the model specification in problem 2 but with the exponential and Rayliegh models, obtain the JLMS inefficiency estimates. This is done with

```
FRONTIER   ; Lhs = lq ; Rhs = one,lf,le,ll,lp ; Model=Exponential
           ; Techeff = euiexp ; Eff = uiexp  $
FRONTIER   ; Lhs = lq ; Rhs = one,lf,le,ll,lp ; Model=Rayleigh
           ; Techeff = euiray ; Eff = uiray  $
```

Now, use a kernel estimator to examine the estimated inefficiency distribution. What is the sample estimate of the expected value of u? You can use DSTAT to obtain this.

```
KERNEL     ; Rhs = euihn,euiexp, euiray,eucols $
DSTAT      ; Rhs = uihn, uiexp,uiray,uicols,uimols $
```

How do you explain the large difference between the deterministic measures of efficiency and the stochastic frontier estimates of efficiency?

4. Environmental variables

The airlines data contain three variables that might (emphasize might) help to explain the inefficiency in the data, LOAD FACTOR, STAGE – this is the average length of flights, and POINTS which is the number of nodes in the route. There are two (or more, as we will see tomorrow) ways to examine this proposition. First, you obtained estimates of u(i) when you computed the model in 3. You could now regress these on a constant and the three factors noted. Does there appear to be a relationship between u(i) and these variables? A second, indirect way to deal with this possibility is to include the three variables in the model as additional RHS variables. Fit the model with LOADFCTR, POINTS and LSTAGE (log of stage length) included, and save the technical efficiencies as euihnz. (That is, as a new variable.) Now, compare euihn to euihnz. You can use a correlation coefficient. The correlation is quite high, but it does not reveal a striking relationship between the two sets of estimates. Inclusion of the ‘z’ variables moves the heterogeneity from inefficiency into the frontier function itself.

```
FRONTIER ; Lhs = lq ; rhs = x ; techeff = euihn $
REGRESS  ; Lhs = euihn ; rhs = one,loadfctr,lstage,points $
FRONTIER ; Lhs = lq ; rhs = x,loadfctr,lstage,points ; techeff=euihnz $
CALC     ; List ; Cor (euihn, euihnz) $
```

or you could regress UI on a constant and UIA. Or, you could plot UI against UIA using

```
PLOT     ; LHS = euihn ; RHS = euihnz ; Rh2 = euihn $
```

(The red line in the figure is a 45 degree line.) What do you find? Are the two sets of estimates the same? How do the environmental variables affect efficiency? The following plots the average efficiency for the sample against the range of values of load factor.

```
SIMULATE ; Scenario: & loadfctr = .4(.05).95 ; Plot(ci) $
```

5. The dreaded error 315 – wrong skewness

(A vexing problem for stochastic frontier modelers) Now, fit the stochastic frontier model in part 4, but include LM the model, as well as ONE,LL,LF,LE,LP. What happened? Can you explain the outcome? What should you do next?

```
FRONTIER ; Lhs = lq ; rhs = x,lm,loadfctr,lstage,points $
```

One thing we might do is add a restriction to the model. Note, we did this earlier, implicitly, by omitting LM from it. Now, instead, we impose constant returns to scale. The ;CML specification in the model below is 'Constrained maximum likelihood,' which is what we need here.

```
FRONTIER ; Lhs = lq ; rhs = x,lm,loadfctr,lstage,points
; cml: lf+lm+le+ll+lp = 1 $
```

Now it works. Is constant returns to scale a reasonable restriction? Keep in mind, the labor coefficient in our results is persistently negative, so in any event, the whole model is suspect. Note that the estimator gives the error 315 message, then estimates the model anyway. At the time that the error occurs, the estimator only has the OLS residuals in hand. It does not know that the constant returns to scale restriction will be imposed later. So, estimation proceeds normally in spite of the warning. After estimating the model, we carry out an experiment to confirm our finding about the least squares residuals.

```
REGRESS  ; Quietly ; Lhs = lq ; rhs = x,lm,loadfctr,lstage,points ; Res = OLS $
REGRESS  ; Quietly ; Lhs = lq ; rhs = x,lm,loadfctr,lstage,points ; Res = CNS_OLS
                   ; cls: lf+lm+le+ll+lp = 1 $
DSTAT    ; Rhs = OLS,CNS_OLS ; NormalityTest $
```

6. Comparing SF and DEA.

Data envelopment is an alternative method of analyzing efficiency. Though we are not spending much class time on this topic, you can get a quick look at how it works, and compare it to the stochastic frontier approach we have been studying. The mechanics of DEA are described in the extract from the LIMDEP manual that is included in your course materials. DEA produces two measures of efficiency, input oriented and output oriented. LIMDEP computes both of these for you. To obtain these for the airlines data, use to obtain the measure from the stochastic frontier model,

```
FRONTIER  ; Lhs = LQ ; Rhs = X ; techeff= euisf $
```

then

```
FRONTIER ; Lhs = output ; Rhs = fuel,eqpt,labor,prop ; alg=dea $
? Pick up the output oriented efficiency from the computation
CREATE   ; euidea=deaeff_o $
```

to obtain the DEA results. Notice that the analysis is based on the levels of the variables. We are not fitting a production function, we are analyzing inputs and outputs. The command produces a summary and computes the two variables that we want to analyze. We will first compare the input oriented measure to what was produced by the stochastic fronteir. The command file for this assignment discusses this further. Three comparisons are provided by

```
DSTAT    ; Rhs = euisf,euidea $
PLOT     ; Lhs=euidea ; Rhs=euisf$
CALC     ; List  ; Cor(euidea,euisf)$
```

The DEA computation has no direct way to incorporate heterogeneity in the computation. Some researchers compute a second step analysis by regressing the estimated efficiencies on the interesting variables. Since some of the efficiency values are 1.0 by construction, some have used a tobit model to account for this. (This is not necessarily a good idea, as the data are not at all generated by a tobit model. But, it has been done.) In a compromise, others have used a truncated regression. Not necessarily a better idea, but it has been done. What do you find?

```
REGRESS  ; lhs = euidea;rhs=one,loadfctr,lstage,points$
```

As an additional exercise, repeat the computations using the input oriented measure, DEAEFF_I.

II. Spanish dairy farms – SFA vs. DEA

The following exercise is similar to exercise 6 above. We compare the results of stochastic frontier estimation with data envelopment analysis

```
SETPANEL ; Group = farm ; Pds=T $
NAMELIST ; (new) ; means=cowsbar,landbar,laborbar,feedbar $
NAMELIST ; factors = cows,land,labor,feed $
CREATE   ; means = Group Mean (factors,pds=t)$
CREATE   ; milkbar=Group Mean (milk,pds=t) $
CREATE   ; yb=log(milkbar)
         ; x1b=log(cowsbar)  ;x2b=log(landbar)
         ; x3b=log(laborbar) ;x4b=log(feedbar)$
CREATE   ; Output = Milkbar/10000 ; Food = feedbar/10000 $
FRONTIER ; If [ year = 98] ; Lhs = output
         ; Rhs = cowsbar,landbar,laborbar,food
         ; Alg=DEA ; List $
FRONTIER ; If[year=98] ;Lhs = yb ; Rhs = one,x1b,x2b,x3b,x4b
         ; techeff = eusf $
DSTAT    ; if[year=98] ; Rhs = eusf,deaeff_o $
PLOT     ; if[year=98] ; Lhs=eusf ; Rhs = deaeff_o ; Rh2=eusf
         ; Title=Stochastic Frontier Efficiency vs. DEA
         ; Grid $
```