`simarwilson`:

DEA based Two-Step Efficiency Analysis

Harald Tauchmann

Friedrich-Alexander-Universität Erlangen-Nürnberg
Professur für Gesundheitsökonomie

June 26, 2015

2015 German Stata Users Group Meeting
Nuremberg, IAB

Efficiency Measurement

- Efficiency measurement industry in empirical research
  - ✓ Thousands of applications
- Two major methodological approaches
  1. Parametric approaches
     - Most important: **stochastic frontier** (SF; Aigner et al., 1977) → `frontier`, `xtfrontier` (real Stata); `sfcross` and `sfpanel` (user written programs implementing additional model variants; Belotti et al., 2013)
  2. Non-parametric approaches
     - Most important: **DEA** (Data Envelopment Analysis; Charnes et al., 1978) → `dea` (user written Stata command implementing most common DEA models; Ji and Lee, 2010)
     - Less often applied: FDH (Free Disposal Hull; Deprins et al., 1984), partial frontier (Cazals et al., 2002; Aragon et al., 2005) → `orderm`, `orderalpha` (user written Stata commands implementing FDH and partial frontier models; Tauchmann, 2012)

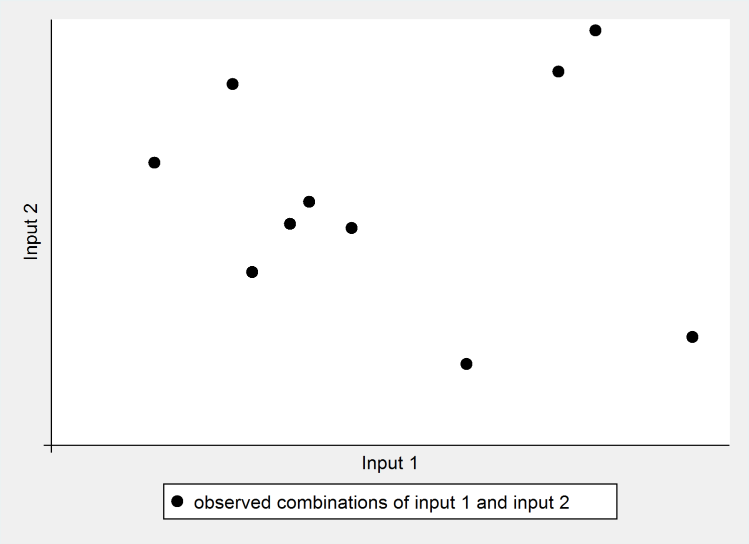
Stochastic Frontier Models

- SF embedded in familiar regression framework
  $y_i = x_i'\beta + \varepsilon_i - \nu_i$ with i indexing DMUs (decision making unit)
- y_i: log-output from production
- x_i: log-inputs to production
- ε_i: conventional normal error
  - ✓ Unexplained heterogeneity in production possibility frontier
- ν_i support on the $[0, \infty)$ interval (exponential, half-normal, truncated normal)
  - ✓ Deviation from production possibility frontier ($\rightarrow$ inefficiency)
- Efficiency measured as $\text{E}(\exp(-\nu_i)|\varepsilon_i - \nu_i)$
- $\text{E}(\nu_i)$ or $\text{Var}(\nu_i)$ can be specified as a function of DMU specific characteristics z_i
- Stochastic Frontier model allows for both
  1. Estimating individual efficiency
  2. Identifying effects DMU characteristics exert on (in)efficiency

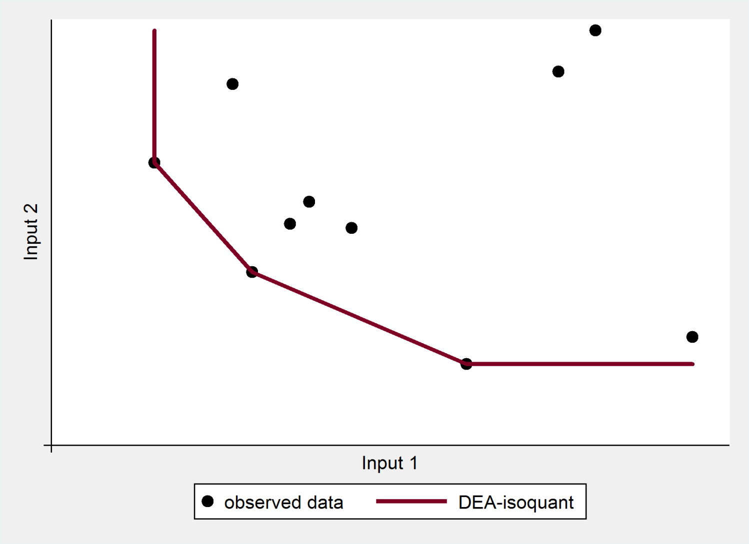
Data Envelopment Analysis

- DEA not a regression model
- Estimation of production possibility frontier by non-parametrically enveloping a given sample of data
- Major advantages as compared to SF-Models
  - ✓ No distributional assumptions required
  - ✓ Straight forward modeling of multi-output processes ($\rightarrow$ no cost-efficiency approach required)
  - ✓ Not a causal model ($\rightarrow$ endogeneity of inputs no issue)
- Various different DEA variants available
  - ✓ Assumptions about frontier ($\rightarrow$ return to scale)
  - ✓ Efficient counterpart of observed DMU at frontier ($\rightarrow$ orientation, treatment of slacks)
- Solving linear program yields eff. score θ_i for each DMU i
  1. $\theta_i^{in} \in (0, 1]$: possible prop. input reduction (input orient.)
  2. $\theta_i^{out} \in [0, \infty)$: possible prop. output increase (output orient.)

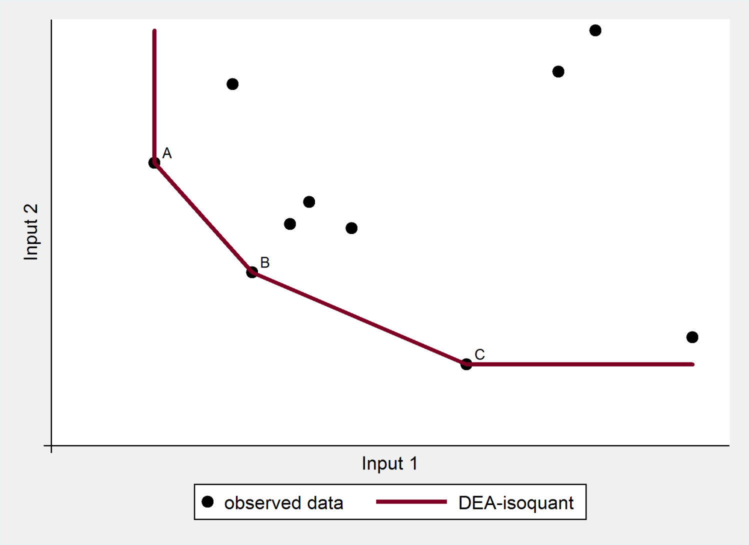
Graphical Illustration of DEA



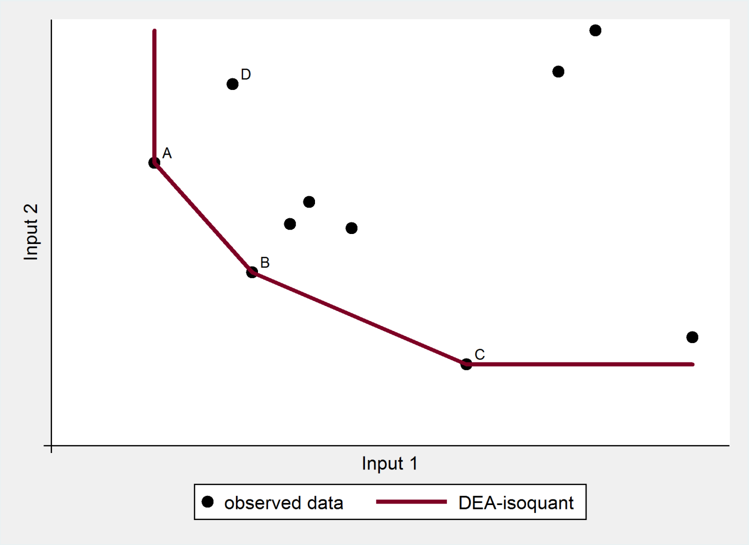
Graphical Illustration of DEA



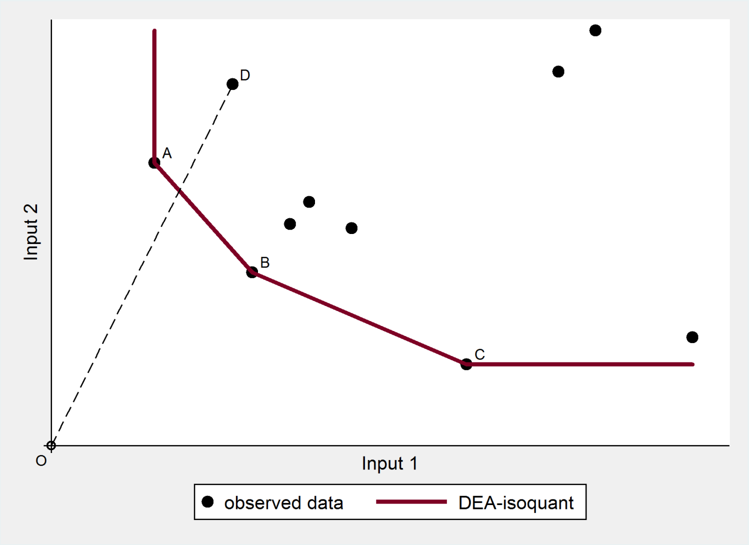
Graphical Illustration of DEA



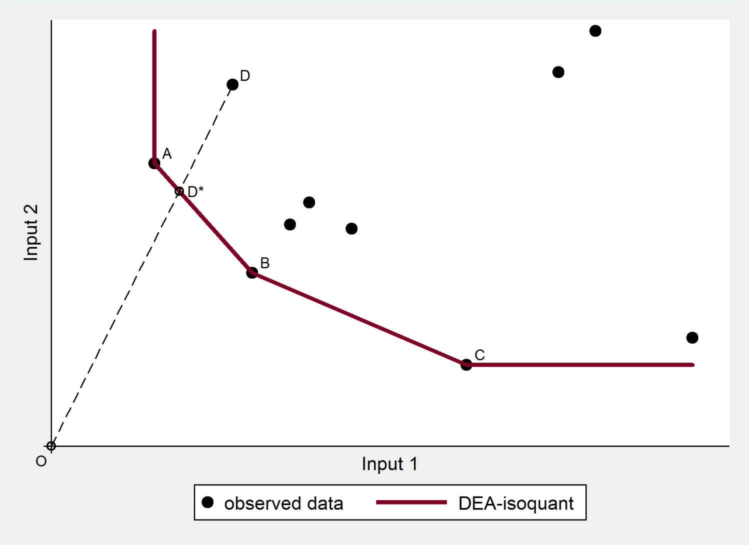
Graphical Illustration of DEA



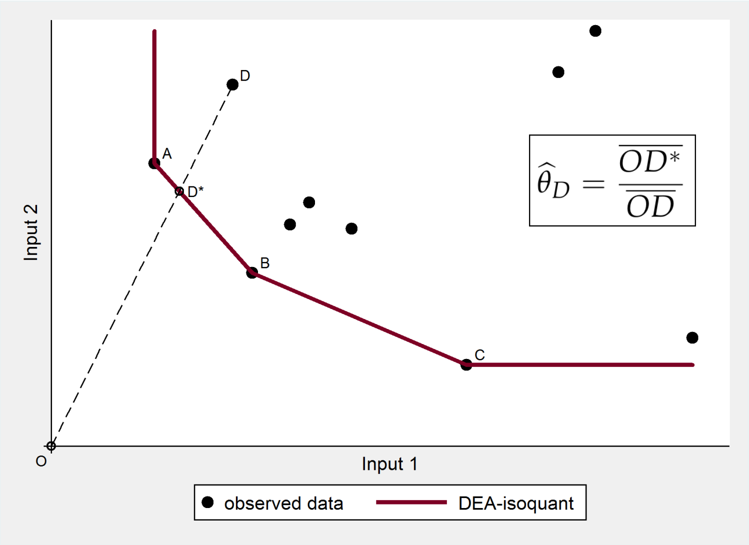
Graphical Illustration of DEA



Graphical Illustration of DEA



Graphical Illustration of DEA



DEA & Explaining Efficiency Differentials

- DEA focussed on **measuring** efficiency
  - ✓ Distance to estimated frontier
  - ✓ Benchmarking major field of applications
- DEA **does not explain** efficiency differentials
- Two-step approach intuitive
  1. Estimating θ_i using DEA ($\rightarrow$ yields certain share M/N of DMUs for which $\widehat{\theta}_i = 1$ holds)
  2. Regressing $\widehat{\theta}_i$ (or transformation of $\widehat{\theta}_i$) on DMU characteristics z_i (OLS, censored regression, ...)
- Numerous applications of such two-step approaches

The argument of Simar and Wilson (2007)

Conventional two-step approaches inappropriate

1. Two-step approaches lack a well defined data generating mechanism
   - ✓ Censored regression model not appropriate
   - ✓ Probability mass at $\theta = 1$ artifact of efficiency measurement by DEA (finite sample problem)
   - ✓ No strictly positive probability for DMU being located on true production possibility frontier ($\neq$ estimated DEA frontier)
2. DEA generates complex (unknown) pattern of correlation between the estimated efficiency scores
   - ✓ $\widehat{\theta}_i$ with $i = 1, \ldots, N$ by construction not independent
   - ✓ Misleading inference based on two-step approaches
   - ✓ Naive bootstrap no solution because of boundary estimation nature of DEA

The Simar and Wilson (2007) Approach

1. Constructing and simulating a ‘sensible’ data generating process
2. Generating artificial iid bootstrap samples from artificial data generating process
3. Construction standard errors and confidence through bootstrapping/simulation

The Simar and Wilson (2007) Procedure

1. Estimate θ_i with $i = 1, \ldots, N$ using **DEA**
2. Fitting $\widehat{\theta}_i = \beta' z_i + \epsilon_i$ using **truncated regression** (ML) ($\rightarrow$ obtain estimates $\widehat{\beta}$ and $\widehat{\sigma}_\varepsilon$)
   - ✓ **Efficient DMUs** j ($\widehat{\theta}_j = 1, j = 1, \ldots, M$) **excluded**
   - ✓ $\epsilon_i \equiv \varepsilon_i + \zeta_i$ with $\zeta_i \equiv \widehat{\theta}_i - \theta_i$
   - ✓ $\widehat{\theta}_i^{in} \in (0, 1]$ (input orient.): right-truncation at 1
   - ✓ $\widehat{\theta}_i^{out} \in [0, \infty)$ (output orient.): left-truncation at 1

The Simar and Wilson (2007) Procedure II

3. **Loop** over the next three steps B times ($b = 1, \ldots, B$)
   3.1 **Draw** ε_i^b from $N(0, \widehat{\sigma}_\varepsilon)$ with **left-truncation** (output orient.) or **right-truncation** (input orient.) at $(1 - \widehat{\beta}' z_i)$ for $i = M + 1, \ldots, N$
   3.2 **Compute** $\theta_i^b = \widehat{\beta}' z_i + \varepsilon_i^b$ for $i = M + 1, \ldots, N$
   3.3 Estimate $\widehat{\beta}^b$ and $\widehat{\sigma}_\varepsilon^b$ by **truncated regression** using the artificial efficiency scores θ_i^b as *lhs*-variable
4. Construct **standard errors** for $\widehat{\beta}$ and $\widehat{\sigma}_\varepsilon$ (conf. interv. for β and σ_ε) from **simulated distribution** of $\widehat{\beta}^b$ and $\widehat{\sigma}_\varepsilon^b$

The `simarwilson` command

- `simarwilson` implements above procedure in Stata
  - ✓ Except for step 1
  - ✓ Efficiency scores have to be obtained prior to running `simarwilson` ($\rightarrow$ e.g. using `dea`)
    - Implemented procedure is ‘algorithm #1’ (Simar and Wilson, 2007)
    - Alternative (more involved) ‘algorithm #2’ requires looping over DEA
- `simarwilson` requires user written mata modul `RTNORM` Belotti and Ilardi (2010) to drawn from the truncated normal distribution

Syntax of `simarwilson`

`simarwilson` *depvar indepvars* [if] [in], [nounit reps(#) dots level(#)]

- **depvar** is assumed to be an efficiency score estimated in a preceding step. *depvar* needs to be a numeric nonnegative variable.
- **nounit** indicates that *depvar* > 1 holds for inefficient dmus, **unit** indicates that for indicates that *depvar* < 1 holds for inefficient dmus. If *depvar* is is well coded, `simarwilson` recognizes if efficiency scores originate form an input or an output-oriented DEA. Specifying **nounit** is required for poorly coded data or if the data contain superefficient dmus
- With **dots** specified one dot character is displayed for each bootstrap replication
- **reps(#)** specifies the number of bootstrap replications to be performed. The default is 50. For simulating meaningful confidence intervals a much larger number of replications is required
- **level(#)** set confidence level; default is `level(95)`

Application & Data

- Regional efficiency of health care provision in Bavaria
  - ✓ `simarwilson` originates from project analyzing efficiency of nursing homes
  - ✓ Protected data ($\rightarrow$ not well suited for illustrating the command)
- County level data (N = 96) for year 2006
- Output from health production
  - ✓ Regional survival rate ($\rightarrow$ corrected for demographic composition; normalized to national average)
- Input to health production
  1. General practitioners (per 100 000 inhabitants)
  2. Medical specialists (per 100 000 inhabitants)
  3. Hospital beds (per 10 000 inhabitants)

Descriptives for Input & Outputs

```
.        tabstat survival gps specialists beds, columns(statistics) statistics
> (mean sd median min max) format(%7.0g)
```

variable	mean	sd	p50	min	max
survival	1.0075	.08002	.9978	.84532	1.2205
gps	77.829	11.603	74.392	57.792	109.98
specialists	93.127	62.39	66.709	16.497	245.78
beds	66.402	51.808	49.156	1.7513	227.16

- Variables that enter `dea`
- Substantial heterogeneity across counties

Results from DEA

```
.         foreach direction in i o {
  2.                 quietly: dea gps specialists beds = survival, rts(vrs) ort
> (`direction´)
  3.                 mat deascores = r(dearslt)
  4.                 mat deascores = deascores[1...,"theta"]
  5.                 sort dmu
  6.                 cap drop dea1
  7.                 svmat deascores, names(dea)
  8.                 rename dea1 deascore_`direction´
  9.                 gen efficient_`direction´ = deascore_`direction´ == 1
 10. }
options: RTS(VRS) ORT(IN) STAGE(2)
options: RTS(VRS) ORT(OUT) STAGE(2)

. tabstat deascore_i deascore_o efficient_i efficient_o, columns(statistics) st
> atistics(mean sd median min max) format(%7.0g)
```

variable	mean	sd	p50	min	max
deascore_i	.81203	.12388	.82317	.52548	1
deascore_o	1.1421	.09806	1.1424	1	1.3611
efficient_i	.125	.33245	0	0	1
efficient_o	.125	.33245	0	0	1

Explanatory Variables

- County unemployment rate (*unemployment*)
- Women's share in county population (*female*)
- Indicator for urban county (single town constituting a county, *urban*)
- Share of private hospitals in county hospital beds (*privatehosp*)

```
.        tabstat `reglist´, columns(statistics) statistics(mean sd median min
> max) format(%7.0g)
```

variable	mean	sd	p50	min	max
unemployment	.07069	.02311	.0675	.034	.132
female	.51058	.00881	.50778	.49683	.53575
urban	.26042	.44117	0	0	1
privatehosp	.16448	.29363	0	0	1

(Naive) Censored Regression Analysis

- Estimated input oriented efficiency (*deascore_i*) at lhs

```
.        tobit deascore_i `reglist´, ul(1)

Tobit regression                                  Number of obs   =         96
                                                  LR chi2(4)      =      58.47
                                                  Prob > chi2     =     0.0000
Log likelihood =  62.060332                       Pseudo R2       =    -0.8905

------------------------------------------------------------------------------
  deascore_i |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
unemployment |  -1.498905   .6192545    -2.42   0.017    -2.728798    -.269012
      female |  -4.483155   1.677094    -2.67   0.009    -7.814008   -1.152302
       urban |  -.0657136   .0364946    -1.80   0.075    -.1381952    .0067679
 privatehosp |   .0188993   .0358417     0.53   0.599    -.0522853     .090084
       _cons |   3.226821   .8407472     3.84   0.000     1.557024    4.896618
-------------+----------------------------------------------------------------
      /sigma |   .0976889   .0078113                      .0821749    .1132029
------------------------------------------------------------------------------
  Obs. summary:          0  left-censored observations
                        84     uncensored observations
                        12 right-censored observations at deascore_i>=1
```

Conventional Truncated Regression Analysis

```
.        truncreg deascore_i `reglist´, ul(1)
(note: 12 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood =  102.21542
Iteration 1:   log likelihood =  102.32083
Iteration 2:   log likelihood =  102.32114
Iteration 3:   log likelihood =  102.32114

Truncated regression
Limit:   lower =       -inf                             Number of obs =     84
         upper =          1                             Wald chi2(4)  =  91.26
Log likelihood =  102.32114                             Prob > chi2   = 0.0000
```

deascore_i	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
unemployment	-1.120479	.5156737	-2.17	0.030	-2.13118	-.1097767
female	-5.377884	1.407259	-3.82	0.000	-8.136062	-2.619706
urban	-.0403955	.0289685	-1.39	0.163	-.0971728	.0163818
privatehosp	-.0257407	.0314323	-0.82	0.413	-.0873468	.0358655
_cons	3.634028	.7056814	5.15	0.000	2.250918	5.017138
/sigma	.0753927	.0064815	11.63	0.000	.0626893	.0880962

- Qualitatively similar results as from `tobit`

Simar & Wilson (2007) Procedure

```
.        simarwilson deascore_i `reglist´, reps(500)

Simar & Wilson (2007) truncated regression
DMUs inefficient if deascore_i < unity

Number of obs.               =        96
Number of truncated obs.     =        12
Number of bootstr. reps.     =       500
Wald-test (p-value)          =   5.0e-18
Log-likelihood               =   102.321
```

deascore_i	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deascore_i						
unemployment	-1.120479	.4943829	-2.27	0.023	-2.089451	-.1515059
female	-5.377884	1.289016	-4.17	0.000	-7.904308	-2.85146
urban	-.0403955	.0284727	-1.42	0.156	-.096201	.0154099
privatehosp	-.0257407	.0275043	-0.94	0.349	-.0796481	.0281667
_cons	3.634028	.6554127	5.54	0.000	2.349443	4.918613
sigma						
_cons	.0753927	.0073111	10.31	0.000	.0610632	.0897223

- Only standard errors differ from `truncreg` (alg. #1)
- (In this application) just small deviation from `truncreg`

simarwilson: output-oriented

```
.         simarwilson deascore_o `reglist´, reps(500)
warning: all efficiency scores deascore_o outside unit-interval, option unit ch
> angened to nounit

Simar & Wilson (2007) truncated regression
DMUs inefficient if deascore_o > unity

Number of obs.                =        96
Number of truncated obs.      =        12
Number of bootstr. reps.      =       500
Wald-test (p-value)           =   1.1e-08
Log-likelihood                =   108.830
```

deascore_o	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deascore_o						
unemployment	3.315517	.5375734	6.17	0.000	2.261893	4.369142
female	-1.191438	1.384378	-0.86	0.389	-3.90477	1.521893
urban	-.0518028	.0269731	-1.92	0.055	-.1046692	.0010636
privatehosp	-.0301374	.0292012	-1.03	0.302	-.0873707	.0270958
_cons	1.543875	.6942935	2.22	0.026	.1830849	2.904665
sigma						
_cons	.0739952	.0070933	10.43	0.000	.0600927	.0878977

- Results differ from input-oriented analysis
- Estimated effect for *female*, *urban*, and *privatehosp* change direction
- *urban* becomes significant (10% level)

`simarwilson`: output-oriented (inverted score

```
.        gen deascore_oi = 1/deascore_o

.        simarwilson deascore_oi `reglist´, reps(500)

Simar & Wilson (2007) truncated regression
DMUs inefficient if deascore_oi < unity

Number of obs.              =        96
Number of truncated obs.    =        12
Number of bootstr. reps.    =       500
Wald-test (p-value)         =   1.0e-09
Log-likelihood              =   132.557
```

deascore_oi	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deascore_oi						
unemployment	-2.349899	.3703527	-6.35	0.000	-3.075777	-1.624021
female	.8053843	.9966054	0.81	0.419	-1.147926	2.758695
urban	.0374269	.02103	1.78	0.075	-.0037912	.078645
privatehosp	.0247856	.0212903	1.16	0.244	-.0169425	.0665137
_cons	.6116713	.4991997	1.23	0.220	-.3667421	1.590085
sigma						
_cons	.0530765	.0051011	10.40	0.000	.0430784	.0630745

- Results qualitatively equivalent to using not inverted scores at *lhs*

Conclusions

- Using DEA-scores as *lhs*-variable in regression model questionable
- Simar & Wilson (2007) propose procedure that is not ad hoc but has a basis in statistical theory
  - ✓ Very influential in applied efficiency analysis
- `simarwilson` implements the procedure (alg. #1) in Stata
  - ✓ Also implementing alg. #2 worth considering
  - ✓ Complicated by alg. #2 requiring looping over DEA
- In many application results (inference) do not differ much from simple truncated regression

References

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

Aragon, Y., Daouia, A. and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantilebased approach, *Econometric Theory* **21**: 358–389.

Belotti, F., Daidone, S., Ilardi, G. and Atella, V. (2013). Stochastic frontier analysis using stata, *Stata Journal* **13**(4): 719–758.

Belotti, F. and Ilardi, G. (2010). RTNORM: Stata Mata module to produce truncated normal pseudorandom variates, Statistical Software Components, Boston College Department of Economics.

Cazals, C., Florens, J. P. and Simar, L. (2002). Nonparametric frontier estimation: A robust approach, *Journal of Econometrics* **106**: 1–25.

Charnes, A., Cooper, W. W. and Rhodes, E. (1978). Measuring efficiency of decision making units, *European Journal of Operational Research* **2**: 429–444.

Deprins, D., Simar, L. and Tulkens, H. (1984). Measuring labor-efficiency in post offices, *in* M. Marchand, P. Pestieau and H. Tulkens (eds), *The Performance of Public Enterprises: Concepts and Measurement*, Elsevier, Amsterdam, pp. 243–267.

Ji, Y. and Lee, C. (2010). Data envelopment analysis, *Stata Journal* **10**(2): 267–280.

Simar, L. and Wilson, P. W. (2007). Estimation and inference in two-stage semi-parametric models of production processes, *Journal of Econometrics* **136**: 31–64.

Tauchmann, H. (2012). Partial frontier efficiency analysis, *Stata Journal* **12**(3): 461–478.