A note on the choice of a sample of firms for reliable estimation of sector returns to scale

Michal Černý[1]

Abstract. A sector is defined as a family of firms sharing the same Cobb-Douglas production function. Our aim is to estimate the Cobb-Douglas-based returns to scale of the sector. Limited resources allow us to collect data (stock of production factors and production) from a limited number of firms only. We address the question how the sample of firms, used then for estimation of the sector returns to scale, should be selected to achieve a "good" estimate of the returns to scale. (The estimate is "good" if it has low variance.) We propose a three-step procedure for the sample selection problem, adopting a method from the theory of c-optimal experimental designs. We consider both homoscedastic and heteroscedastic models. We illustrate the approach by examples.

Keywords: sample selection, Cobb-Douglas function, returns to scale, c-optimal design

JEL classification: C81
AMS classification: 62K05, 91B38, 91G70

1 Introduction, definitions and assumptions

Let $\Phi_1, \dots, \Phi_n$ denote production factors. A *firm* is a $(n+1)$-tuple of nonnegative real numbers

$$(y^*, \varphi_1, \dots, \varphi_n), \tag{1}$$

where y^* denotes the level of the firm's output and φ_i denotes the stock of i-th production factor available to the firm.

A *sector* $\mathcal{S}$ is the set

$$\mathcal{S} = \{F_1, \dots, F_N\},$$

where $F_1, \dots, F_N$ are firms. We also use the notation

$$F_j = (y_j^*, \varphi_{1j}, \dots, \varphi_{nj}). \tag{2}$$

We assume that all the firms of the sector $\mathcal{S}$ share a common Cobb-Douglas production function of the form

$$\ln y_j = \beta_0 + \sum_{i=1}^{n} \beta_i \ln \varphi_{ij} + \varepsilon_j, \qquad j = 1, \dots, N, \tag{3}$$

where ε_j are independent $N(0, \sigma^2)$ error terms. In (2) we assume that the value y_j^* is the observed realization of the random variable y_j.

Returns to scale of the sector $\mathcal{S}$ is the number $r := \sum_{i=1}^{n} \beta_i$. Recall that the returns to scale are

$$\left.\begin{array}{r}\text{constant}\\ \text{increasing}\\ \text{decreasing}\end{array}\right\} \text{ iff } \left\{\begin{array}{l} r = 1,\\ r > 1,\\ r < 1.\end{array}\right.$$

[1]University of Economics Prague, Department of Econometrics, Winston Churchill Square 4, 13067 Prague, Czech Republic, cernym@vse.cz

1.1 The problem

Our aim is to measure the number r of the sector $\mathcal{S}$. Of course, due to the presence of the error terms ε_j, we can never measure r exactly. Therefore we are interested in an *estimate* $\widehat{r}$ of r. We shall use the standard estimator $\widehat{r} = \sum_{i=1}^{n} \widehat{\beta}_i$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \dots, \widehat{\beta}_n)^{\mathrm{T}}$ is the standard OLS estimator of (3). Then we can, for example, test the null hypothesis

$$r = 1 \tag{4}$$

using the standard t-test or F-test.

Assume that N, the size of the sector, is large. In order to obtain as precise estimates of r as possible, it is desirable to collect data (2) for all firms $F_1, \dots, F_N$. However, this process is usually costly. Usually only limited resources are available to us; with these resources we are able to collect data from a limited number of firms only. We have arrived at the main question of the paper: *assume that we are able to collect data from only $m \ll N$ firms. Which firms from the sector $\mathcal{S}$ should be included in the selected sample $\mathcal{S}'$ (of cardinality m) in order the value $\widehat{r}$, estimated from the sample $\mathcal{S}'$, be as precise as possible?*

The relevance of the question is motivated by the following example.

1.2 Example

Assume that $n = 2$ and $\Phi_1 =$ *labor* and $\Phi_2 =$ *capital stock*. Assume that the sector $\mathcal{S}$ of $N = 12$ firms is governed by the model (3) with

$$\beta_0 = 0, \quad \beta_1 = 0.5, \quad \beta_2 = 0.6, \quad \sigma = 0.1.$$

Then $r > 1$ and the returns to scale of the sector $\mathcal{S}$ are increasing.

Assume that our resources allow us to gather data from $m = 6$ firms only. We would like to choose the sample of 6 firms in the way that $\mathrm{se}(\widehat{\beta}_1 + \widehat{\beta}_2)$ is minimal, where "se" stands for standard error. In that case, r is estimated with the best possible precision. This is important since the standard error of $\widehat{r}$ being low, the t-test for the hypothesis $r = 1$ is strong. (Recall that the test statistic is of the form $\frac{\widehat{r}-1}{\widehat{\mathrm{se}}(\widehat{r})}$.)

We can write the model (3) in the usual form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$. With this notation we have

$$\mathrm{se}(\widehat{\beta}_1 + \widehat{\beta}_2) = \sigma \cdot \sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}},$$

where $\boldsymbol{c} = (0, 1, 1)^{\mathrm{T}}$.

We have $\binom{12}{6} = 924$ possibilities for the choice of the sample $\mathcal{S}'$ of 6 firms out of 12 total; denote the choices as $\mathcal{S}'_1, \dots, \mathcal{S}'_{924}$. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{924}$ denote the corresponding $\boldsymbol{X}$-matrices. Define

$$\tau_i := \sigma \cdot \sqrt{\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{c}}, \quad i = 1, \dots, 924.$$

Let the choices $\mathcal{S}'_1, \dots, \mathcal{S}'_{924}$ be ordered in the way that $\tau_1 \le \cdots \le \tau_{924}$. Figure 1 shows values of τ_i against i. The best possible choice is

$$\mathcal{S}'_1 = \{F_1, F_2, F_3, F_6, F_8, F_9\} \ \text{ with } \ \tau_1 = 0.0435, \tag{5}$$

while the worst possible choice is

$$\mathcal{S}'_{924} = \{F_4, F_5, F_7, F_8, F_{11}, F_{12}\} \ \text{ with } \ \tau_{924} = 0.2064. \tag{6}$$

In the case (6), t-test for the null hypothesis (4) will probably not reject, though the hypothesis is not true. Hence, with the choice $\mathcal{S}'_{924}$ we can arrive at an incorrect conclusion that returns to scale are constant. On the other hand, if we choose the sample $\mathcal{S}'_1$, we have a much higher chance that the t-test will reject, which is a correct conclusion. In general: the better value τ_i, the stronger the t-test is. And, if we choose the sample of firms "in the best possible way" and the t-test does not reject, we have a strong evidence that $r = 1$ indeed.

This example shows that before we start collecting data, *it is reasonable to ask which firms of the sector $\mathcal{S}$ are likely to contribute to the precision of the estimator of $\widehat{r}$ more than others.*

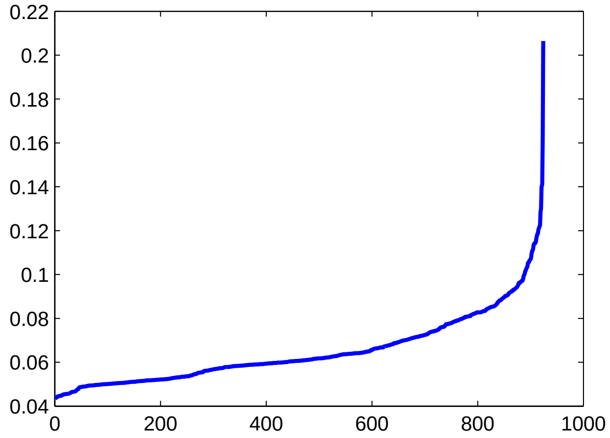

Figure 1 Sequence $\tau_1, \dots, \tau_{924}$.

2 Our approach

The question leads us to the theory of optimum experimental designs. Indeed, the sample which minimizes the variance of $\widehat{r}$ can be seen as a case of $\boldsymbol{c}$-optimal design: our aim is minimization of $\mathrm{se}(\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$, where $\boldsymbol{c} = (0, 1, \dots, 1)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_n)^{\mathrm{T}}$.

The problem is that we know nothing about the sector $\mathcal{S}$ in advance. We adopt the assumption that we are able to gather information on *representants* of the sector $\mathcal{S}$. Each representant should represent a group of firms in the sector $\mathcal{S}$ with similar stock of production factors. (Said more precisely, a representant R should be either a real or fictitious firm such that it is reasonable to expect that in the sector $\mathcal{S}$ there are enough real firms with the stock of production factors similar to R.) Then we restrict ourselves to the representants.

We find an optimal design over the representants; this will give us guidance from which groups of firms it should be suitable to collect final data.

We illustrate the approach by example. Let φ_{1j} denote the capital stock of j-th representant and let φ_{2j} denote the labor stock of j-th representant. Assume that we know that the sector $\mathcal{S}$ contains the following groups with the following representants:

group	type	representant
group 1	small capital-intensive firms	$R_1 = (\varphi_{11} = 5,\ \varphi_{21} = 1)$
group 2	small labor-intensive firms	$R_2 = (\varphi_{12} = 1,\ \varphi_{22} = 5)$
group 3	medium capital-intensive firms	$R_3 = (\varphi_{13} = 20,\ \varphi_{23} = 10)$
group 4	medium labor-intensive firms	$R_4 = (\varphi_{14} = 15,\ \varphi_{24} = 22)$
group 5	large capital-intensive firms	$R_5 = (\varphi_{15} = 35,\ \varphi_{25} = 20)$
group 6	large labor-intensive firms	$R_6 = (\varphi_{16} = 20,\ \varphi_{26} = 42)$

(7)

In our example we will write

$$\mathcal{X} := \left\{ \begin{pmatrix} 1 \\ \ln \varphi_{11} \\ \ln \varphi_{21} \end{pmatrix}, \dots, \begin{pmatrix} 1 \\ \ln \varphi_{16} \\ \ln \varphi_{26} \end{pmatrix} \right\}. \tag{8}$$

The meaning of this set will be explained in the next section.

2.1 Some notions from the theory of $\boldsymbol{c}$-optimal designs

In the theory of experimental design, the set $\mathcal{X}$ is usually referred to as *experimental domain.* Its interpretation is as follows. Assume the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{9}$$

with independent disturbances $\boldsymbol{\varepsilon}$, which are homoscedastic with variance σ^2. We are given a nonzero vector $\boldsymbol{c}$ of parameters and our aim is to select the rows of $\boldsymbol{X}$ in the way that $\mathrm{se}(\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ is minimal. We are restricted by the fact that each row $\boldsymbol{x}^{\mathrm{T}}$ of $\boldsymbol{X}$ must fulfill $\boldsymbol{x} \in \mathcal{X}$. Said otherwise, we can make measurements only in the points from the experimental domain $\mathcal{X}$ and our aim is to select those points which minimize the variance of $\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

Assume that $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$ and that we have the regression model (9) with ν observations, where the matrix $\boldsymbol{X}$ is of the form

$$\boldsymbol{X} = (\underbrace{\boldsymbol{x}_1, \boldsymbol{x}_1, \dots, \boldsymbol{x}_1}_{\nu\xi_1 \text{ times}};\ \underbrace{\boldsymbol{x}_2, \boldsymbol{x}_2, \dots, \boldsymbol{x}_2}_{\nu\xi_2 \text{ times}}\ \cdots\ ;\underbrace{\boldsymbol{x}_M, \boldsymbol{x}_M, \dots, \boldsymbol{x}_M}_{\nu\xi_M \text{ times}})^{\mathrm{T}}. \tag{10}$$

The vector $\boldsymbol{\xi} := (\xi_1, \dots, \xi_M)^{\mathrm{T}}$ is called *design* — it simply says that we are making $100\xi_1\%$ observations in the point $\boldsymbol{x}_1$, $100\xi_2\%$ observations in the point $\boldsymbol{x}_2$ etc.

We can define the number $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi})$, called $\boldsymbol{c}$-*variance* of the design $\boldsymbol{\xi}$, implicitly using the equation

$$\mathrm{var}(\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\nu} \cdot \mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}),$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ with $\boldsymbol{X}$ given by (10). (Here, $^{-1}$ might stand for the matrix pseudoinverse.) It is easy to see that the number $\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi})$ does depend on the design $\boldsymbol{\xi}$, *but it depends neither on* σ^2 *nor on the number of observations* ν. Hence it is a good measure of the contribution of the design $\boldsymbol{\xi}$ to the total variance of the estimator $\boldsymbol{c}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

All designs form the simplex $\Sigma := \{\boldsymbol{\xi} : \boldsymbol{\xi} \geq \boldsymbol{0},\ \boldsymbol{1}^{\mathrm{T}}\boldsymbol{\xi} = 1\}$. Our task is to find the design with minimal $\boldsymbol{c}$-variance. Thus we are to solve the optimization problem

$$\min\{\mathrm{var}_{\boldsymbol{c}}(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Sigma\}.$$

Its solution is called $\boldsymbol{c}$-*optimal design.*

Definition 1. The **Elfving set** is the set $\mathcal{E} := \mathrm{convexhull}(\mathcal{X} \cup -\mathcal{X})$, where $-\mathcal{X} = \{-\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$. □

The following theorem, called Elfving's Theorem (see [4]), is a fundamental result in the theory of $\boldsymbol{c}$-optimal designs.

Theorem 1. *Let* $\boldsymbol{c}$ *be a nonzero vector and let* $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$*. Let* $\omega^* = \max\{\omega \in \mathbb{R} : \omega \cdot \boldsymbol{c} \in \mathcal{E}\}$ *and* $\boldsymbol{x}^* = \omega^*\boldsymbol{c}$*. Let* $u_1, \dots, u_M$ *and* $v_1, \dots, v_M$ *be nonnegative numbers such that*

$$\boldsymbol{x}^* = \sum_{i=1}^{M} u_i\boldsymbol{x}_i - \sum_{i=1}^{M} v_i\boldsymbol{x}_i$$

and

$$\sum_{i=1}^{M}(u_i + v_i) = 1.$$

Then $(u_1 + v_1, \dots, u_M + v_M)^{\mathrm{T}}$ *is the* $\boldsymbol{c}$*-optimal design over* $\mathcal{X}$*.* □

In other words, if we write the point $\boldsymbol{x}^*$ as a convex combination of the points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M, -\boldsymbol{x}_1, \dots, -\boldsymbol{x}_M$, then the coefficients of the convex combination determine the $\boldsymbol{c}$-optimal design.

Harman and Jurík [5] observed that Elfving's Theorem leads to a linear programming problem.

Theorem 2. *Let* $\Xi = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_M)$*. Let* $\boldsymbol{u}^*, \boldsymbol{v}^*, \omega^*$ *be the solution of the linear program*

$$\max\{\omega \in \mathbb{R} :\ \Xi(\boldsymbol{u} - \boldsymbol{v}) = \omega \cdot \boldsymbol{c},\quad \boldsymbol{1}^{\mathrm{T}}(\boldsymbol{u} + \boldsymbol{v}) = 1,\quad \boldsymbol{u} \geq \boldsymbol{0},\quad \boldsymbol{v} \geq \boldsymbol{0}\}. \tag{11}$$

Then $\boldsymbol{\xi} := \boldsymbol{u}^* + \boldsymbol{v}^*$ *is the* $\boldsymbol{c}$*-optimal design.* □

More on the theory of optimal designs can be found in [2], [6], [7]. Computational issues are dealt with in [1], [3].

2.2 The example continued

We now apply Elfving's Theorem to the "experimental domain" $\mathcal{X}$ given by (8). (The form of the model (3) shows why the logarithms are present in (8).) We set

$$\boldsymbol{\Xi} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \ln 5 & \ln 1 & \ln 20 & \ln 15 & \ln 35 & \ln 20 \\ \ln 1 & \ln 5 & \ln 10 & \ln 22 & \ln 20 & \ln 42 \end{pmatrix}$$

and

$$\boldsymbol{c} = (0, 1, 1)^{\mathrm{T}}.$$

Solving the linear program (11) we get the optimal design $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_6)^{\mathrm{T}}$ with

$$\xi_1 = 0.13, \quad \xi_2 = 0.37, \quad \xi_3 = \xi_4 = \xi_5 = 0, \quad \xi_6 = 0.5. \tag{12}$$

This shows that we should compose the sample as follows:

- 13% of the observations should be collected from the group represented by the representant R_1,
- 37% of the observations should be collected from the group represented by the representant R_2,
- 50% of the observations should be collected from the group represented by the representant R_6.

If our budget is limited to, say, $m = 100$ firms, then it is reasonable to collect data from

- 13 small capital-intensive firms,
- 37 small labor-intensive firms and
- 50 large labor-intensive firms.

2.3 The heteroscedastic case

In the analysis of production functions it is often reasonable to assume heteroscedasticity. Let us consider an example with a heteroscedasticity model where the standard error of disturbances is proportional to $\sqrt{\varphi_{1j}\varphi_{2j}}$ (again, φ_{1j} denotes the capital stock of j-th firm and φ_{2j} denotes the labor stock of j-th firm). Then we can write the model (3) in the form

$$\ln y_j = \beta_0 + \beta_1 \ln \varphi_{1j} + \beta_2 \ln \varphi_{2j} + \delta_j \sqrt{\varphi_{1j}\varphi_{2j}},$$

where δ_j are independent and homoscedastic. A simple transformation yields

$$\frac{\ln y_j}{\sqrt{\varphi_{1j}\varphi_{2j}}} = \beta_0 \cdot \frac{1}{\sqrt{\varphi_{1j}\varphi_{2j}}} + \beta_1 \cdot \frac{\ln \varphi_{1j}}{\sqrt{\varphi_{1j}\varphi_{2j}}} + \beta_2 \cdot \frac{\ln \varphi_{2j}}{\sqrt{\varphi_{1j}\varphi_{2j}}} + \delta_j,$$

which is a homoscedastic model, and we can apply Elfving's Theorem. Using again the representants from (7), we set

$$\boldsymbol{\Xi} = \begin{pmatrix} \frac{1}{\sqrt{5\cdot 1}} & \frac{1}{\sqrt{1\cdot 5}} & \frac{1}{\sqrt{20\cdot 10}} & \frac{1}{\sqrt{15\cdot 22}} & \frac{1}{\sqrt{35\cdot 20}} & \frac{1}{\sqrt{20\cdot 42}} \\ \frac{\ln 5}{\sqrt{5\cdot 1}} & \frac{\ln 1}{\sqrt{1\cdot 5}} & \frac{\ln 20}{\sqrt{20\cdot 10}} & \frac{\ln 15}{\sqrt{15\cdot 22}} & \frac{\ln 35}{\sqrt{35\cdot 20}} & \frac{\ln 20}{\sqrt{20\cdot 42}} \\ \frac{\ln 1}{\sqrt{5\cdot 1}} & \frac{\ln 5}{\sqrt{1\cdot 5}} & \frac{\ln 10}{\sqrt{20\cdot 10}} & \frac{\ln 22}{\sqrt{15\cdot 22}} & \frac{\ln 20}{\sqrt{35\cdot 20}} & \frac{\ln 42}{\sqrt{20\cdot 42}} \end{pmatrix}$$

and $\boldsymbol{c}^{\mathrm{T}} = (0, 1, 1)$. Solution of the linear program (11) yields

$$\xi_1 = 0.1, \quad \xi_2 = 0.04, \quad \xi_3 = 0.86, \quad \xi_4 = \xi_5 = \xi_6 = 0. \tag{13}$$

So, if we are restricted to $m = 100$ observations, it is reasonable to collect data from

- 10 small-sized capital intensive firms,
- 4 small-sized labor-intensive firms and
- 86 medium-sized capital-intensive firms.

3 Conclusion

The difference between (12) and (13) shows that the homoscedasticity/heteroscedasticity assumption is important. (This is not surprising.) We thus suggest that it could be reasonable to perform the analysis in three steps:

- **Step 1.** Make a rough screening of the sector $\mathcal{S}$ to
  - identify groups of firms and their representants,
  - determine whether heteroscedasticity is present, and if so, estimate a suitable model of heteroscedasticity.
- **Step 2.** Using the data from Step 1, apply the method of Section 2.2 (if heteroscedasticity is not present) or Section 2.3 (if heteroascedasticity is present): find the optimal design $\boldsymbol{\xi}$ using (11).
- **Step 3.** Choose firms according to the design $\boldsymbol{\xi}$.

Acknowledgements

The work was supported by the project P403/12/1947 of the Czech Science Foundation.

References

[1] Antoch, J., Černý, M. and Hladík, M.: On computational complexity of construction of c-optimal linear regression models over finite experimental domains. *Tatra Mountains Mathematical Publications* **5** (2012), 1–10.

[2] Atkinson, A., Donev, A. and Tobias, R.: *Optimum Experimental Designs with SAS*. Oxford University Press, Oxford, 2007.

[3] Černý, M. and Hladík, M.: Two complexity results on c-optimality in experimental design. *Computational Optimization and Applications* **51** (2012), 1397–1408.

[4] Elfving, G.: Optimum allocation in linear regression theory. *Annals of the Institute of Statistical Mathematics* **23** (1952), 255–262.

[5] Harman, R. and Jurík, T.: Computing c-optimal experimental designs using the simplex method of linear programming. *Computational Statistics and Data Analysis* **53** (2008), 247–254.

[6] Pázman, A.: *Foundations of Optimum Experimental Design*. Reidel Publishing Company, Dordrecht, 1986.

[7] Pukelsheim, F. and Rieder, S.: Efficient rounding in approximate designs. *Biometrika* **79** (1992), 763–770.