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

Michal Černý

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

\[
(y^*, \varphi_1, \ldots, \varphi_n),
\]

where \(y^*\) denotes the level of the firm’s output and \(\varphi_i\) denotes the stock of \(i\)-th production factor available to the firm.

A sector \(S\) is the set

\[
S = \{F_1, \ldots, F_N\},
\]

where \(F_1, \ldots, F_N\) are firms. We also use the notation

\[
F_j = (y_j^*, \varphi_{1j}, \ldots, \varphi_{nj}).
\]

We assume that all the firms of the sector \(S\) share a common Cobb-Douglas production function of the form

\[
\ln y_j = \beta_0 + \sum_{i=1}^{n} \beta_i \ln \varphi_{ij} + \varepsilon_j, \quad j = 1, \ldots, N,
\]

where \(\varepsilon_j\) 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 \(S\) is the number \(r := \sum_{i=1}^{n} \beta_i\). Recall that the returns to scale are

\[
\begin{align*}
\text{constant} & \quad \text{iff} \quad r = 1, \\
\text{increasing} & \quad \text{iff} \quad r > 1, \\
\text{decreasing} & \quad \text{iff} \quad r < 1.
\end{align*}
\]
1.1 The problem

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

\[
    r = 1
\]

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 \( F_1, \ldots, F_N \). 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 \( S \) should be included in the selected sample \( S' \) (of cardinality \( m \)) in order the value \( \hat{r} \), estimated from the sample \( S' \), be as precise as possible?

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

1.2 Example

Assume that \( n = 2 \) and \( \Phi_1 = \text{labor} \) and \( \Phi_2 = \text{capital stock} \). Assume that the sector \( S \) 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 \( S \) 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 \( \text{se}(\hat{\beta}_1 + \hat{\beta}_2) \) 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 \( \hat{r} \) being low, the \( t \)-test for the hypothesis \( r = 1 \) is strong. (Recall that the test statistic is of the form \( \frac{\hat{r}-1}{\text{se}(\hat{r})} \))

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

\[
    \text{se}(\hat{\beta}_1 + \hat{\beta}_2) = \sigma \cdot \sqrt{\frac{c^T (X^T X)^{-1} c}{n}}
\]

where \( c = (0, 1, 1)^T \).

We have \( \binom{12}{6} = 924 \) possibilities for the choice of the sample \( S' \) of 6 firms out of 12 total; denote the choices as \( S'_1, \ldots, S'_{924} \). Let \( X_1, \ldots, X_{924} \) denote the corresponding \( X \)-matrices. Define

\[
    \tau_i := \sigma \cdot \sqrt{\frac{c^T (X^T X_i)^{-1} c}{n}}, \quad i = 1, \ldots, 924.
\]

Let the choices \( S'_1, \ldots, S'_{924} \) be ordered in the way that \( \tau_1 \leq \cdots \leq \tau_{924} \). Figure 1 shows values of \( \tau_i \) against \( i \). The best possible choice is

\[
    S'_1 = \{F_1, F_2, F_3, F_6, F_8, F_9\} \quad \text{with} \quad \tau_1 = 0.0435,
\]

while the worst possible choice is

\[
    S'_{924} = \{F_4, F_5, F_7, F_8, F_{11}, F_{12}\} \quad \text{with} \quad \tau_{924} = 0.2064.
\]

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 \( S'_{924} \) we can arrive at an incorrect conclusion that returns to scale are constant. On the other hand, if we choose the sample \( S'_1 \), we have a much higher chance that the \( t \)-test will reject, which is a correct conclusion. In general: the better value \( \tau_i \), 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 \( S \) are likely to contribute to the precision of the estimator of \( \hat{r} \) more than others.
2 Our approach

The question leads us to the theory of optimum experimental designs. Indeed, the sample which minimizes the variance of $\hat{r}$ can be seen as a case of $c$-optimal design: our aim is minimization of $se(c^T \beta)$, where $c = (0, 1, \ldots, 1)^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$.

The problem is that we know nothing about the sector $S$ in advance. We adopt the assumption that we are able to gather information on representants of the sector $S$. Each representant should represent a group of firms in the sector $S$ 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 $S$ 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 $\varphi_{1j}$ denote the capital stock of $j$-th representant and let $\varphi_{2j}$ denote the labor stock of $j$-th representant. Assume that we know that the sector $S$ contains the following groups with the following representants:

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

In our example we will write

$$\mathcal{X} := \left\{ \begin{pmatrix} 1 \\ \ln \varphi_{11} \end{pmatrix}, \ldots, \begin{pmatrix} 1 \\ \ln \varphi_{16} \end{pmatrix} \right\}. \quad (8)$$

The meaning of this set will be explained in the next section.
2.1 Some notions from the theory of 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
\[
y = X\beta + \varepsilon
\]
(9)
with independent disturbances $\varepsilon$, which are homoscedastic with variance $\sigma^2$. We are given a nonzero vector $c$ of parameters and our aim is to select the rows of $X$ in the way that $\text{se}(c^T\beta)$ is minimal. We are restricted by the fact that each row $x^T$ of $X$ must fulfill $x \in \mathcal{X}$. 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 $c^T\beta$.

Assume that $\mathcal{X} = \{x_1, \ldots, x_M\}$ and that we have the regression model (9) with $\nu$ observations, where the matrix $X$ is of the form
\[
X = (x_1, x_1, \ldots, x_1) \cdot \underbrace{(x_2, x_2, \ldots, x_2)}_{\nu \xi_1 \text{ times}} \cdots \underbrace{(x_M, x_M, \ldots, x_M)}_{\nu \xi_M \text{ times}}^T.
\]
(10)
The vector $\xi := (\xi_1, \ldots, \xi_M)^T$ is called design — it simply says that we are making $100\xi_1\%$ observations in the point $x_1$, $100\xi_2\%$ observations in the point $x_2$ etc.

We can define the number $\text{var}_c(\xi)$, called c-variance of the design $\xi$, implicitly using the equation
\[
\text{var}(c^T\beta) = \frac{\sigma^2}{\nu} \cdot \text{var}_c(\xi),
\]
where $\tilde{\beta} = (X^TX)^{-1}X^Ty$ with $X$ given by (10). (Here, $^{-1}$ might stand for the matrix pseudoinverse.) It is easy to see that the number $\text{var}_c(\xi)$ does depend on the design $\xi$, but it depends neither on $\sigma^2$ nor on the number of observations $\nu$. Hence it is a good measure of the contribution of the design $\xi$ to the total variance of the estimator $c^T\beta$.

All designs form the simplex $\Sigma := \{\xi: \xi \geq 0, 1^T\xi = 1\}$. Our task is to find the design with minimal c-variance. Thus we are to solve the optimization problem
\[
\min\{\text{var}_c(\xi): \xi \in \Sigma\}.
\]
Its solution is called c-optimal design.

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

The following theorem, called Elfving’s Theorem (see [4]), is a fundamental result in the theory of c-optimal designs.

**Theorem 1.** Let $c$ be a nonzero vector and let $\mathcal{X} = \{x_1, \ldots, x_M\}$. Let $\omega^* = \max \{\omega \in \mathbb{R} : \omega \cdot c \in \mathcal{E}\}$ and $x^* = \omega^*c$. Let $u_1, \ldots, u_M$ and $v_1, \ldots, v_M$ be nonnegative numbers such that
\[
x^* = \sum_{i=1}^M u_i x_i - \sum_{i=1}^M v_i x_i
\]
and
\[
\sum_{i=1}^M (u_i + v_i) = 1.
\]
Then $(u_1 + v_1, \ldots, u_M + v_M)^T$ is the c-optimal design over $\mathcal{X}$. □

In other words, if we write the point $x^*$ as a convex combination of the points $x_1, \ldots, x_M, -x_1, \ldots, -x_M$, then the coefficients of the convex combination determine the c-optimal design.

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

**Theorem 2.** Let $\Xi = \{x_1, \ldots, x_M\}$. Let $u^*, v^*, \omega^*$ be the solution of the linear program
\[
\max \{\omega \in \mathbb{R} : \Xi(u - v) = \omega \cdot c, \ 1^T(u + v) = 1, \ u \geq 0, \ v \geq 0\}.
\]
(11)
Then $\xi := u^* + v^*$ is the 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

\[
\Xi = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
\ln 5 & \ln 1 & \ln 20 & \ln 35 & \ln 20 & \\
\ln 1 & \ln 5 & \ln 10 & \ln 22 & \ln 20 & \ln 42 \\
\end{pmatrix}
\]

and

\[c = (0, 1, 1)^T.\]

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

\[
\xi_1 = 0.13, \quad \xi_2 = 0.37, \quad \xi_3 = \xi_4 = \xi_5 = 0, \quad \xi_6 = 0.5.
\]

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 \( \sqrt{\varphi_1 j \varphi_2 j} \) (again, \( \varphi_1 j \) denotes the capital stock of \( j \)-th firm and \( \varphi_2 j \) denotes the labor stock of \( j \)-th firm). Then we can write the model (3) in the form

\[
\ln y_j = \beta_0 + \beta_1 \ln \varphi_1 j + \beta_2 \ln \varphi_2 j + \delta_j \sqrt{\varphi_1 j \varphi_2 j},
\]

where \( \delta_j \) are independent and homoscedastic. A simple transformation yields

\[
\frac{\ln y_j}{\sqrt{\varphi_1 j \varphi_2 j}} = \beta_0 \cdot \frac{1}{\sqrt{\varphi_1 j \varphi_2 j}} + \beta_1 \cdot \frac{\ln \varphi_1 j}{\sqrt{\varphi_1 j \varphi_2 j}} + \beta_2 \cdot \frac{\ln \varphi_2 j}{\sqrt{\varphi_1 j \varphi_2 j}} + \delta_j,
\]

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

\[
\Xi = \begin{pmatrix}
1/5 & 1/5 & 1/20 & 1/35 & 1/20 & 1/42 \\
\ln 1/5 & \ln 1/5 & \ln 1/20 & \ln 1/35 & \ln 1/20 & \ln 1/42 \\
\end{pmatrix}
\]

and \( c^T = (0, 1, 1). \) Solution of the linear program (11) yields

\[
\xi_1 = 0.1, \quad \xi_2 = 0.04, \quad \xi_3 = 0.86, \quad \xi_4 = \xi_5 = \xi_6 = 0.
\]

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 $S$ 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 heteroscedasticity is present): find the optimal design $\xi$ using (11).

- **Step 3.** Choose firms according to the design $\xi$.

Acknowledgements

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

References


