The Optimal Choice of Controls and Pre-experimental Observations

by
Theo Nijman
and
Marno Verbeek


Reprint Series no. 80
CENTER FOR ECONOMIC RESEARCH

Research Staff
Helmut Bester
Eric van Damme

Board
Helmut Bester
Eric van Damme, director
Arie Kapteyn

Scientific Council
Eduard Bomhoff
Willem Buiter
Jacques Drèze
Theo van de Klundert
Simon Kuipers
Jean-Jacques Laffont
Merton Miller
Stephen Nickell
Pieter Ruys
Jacques Sijben

Erasmus University Rotterdam
Yale University
Université Catholique de Louvain
Tilburg University
Groningen University
Université des Sciences Sociales de Toulouse
University of Chicago
University of Oxford
Tilburg University
Tilburg University

Residential Fellows
Svend Albaek
Pramila Krishnan
Jan Magnus
Eduardo Siandra
Hideo Suehiro

European University Institute
San Francisco State University
Tilburg University
UCLA
Kobe University

Doctoral Students
Roel Beetsma
Hans Bloemen
Sjaak Hurkens
Frank de Jong
Pieter Kop Jansen

Address: Warandelaan 2, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
Phone : +31 13 663050
Telex : 52426 kub nl
Telefax: +31 13 663066
E-mail : "center@htikub5.bitnet"

ISSN 0924-7874

1992
The Optimal Choice of Controls and Pre-experimental Observations

by
Theo Nijman
and
Marno Verbeek


Reprint Series no. 80
The optimal choice of controls and pre-experimental observations

Theo Nijman and Marno Verbeek*

Tilburg University, 5000 LE Tilburg, The Netherlands

Received November 1989, final version received November 1990

In this note we consider the optimal experimental design for cases in which the data may consist of both cross-sectional and panel observations. Our results generalize those of Aigner and Balestra (1988) on the choice of controls and pre-experimental observations by avoiding the assumption that the marginal costs of additional observations on the same unit are negligible. We derive conditions under which the panel design considered by Aigner and Balestra is nevertheless optimal as well as conditions under which reinterviews are inefficient.

1. Introduction

In a recent paper Aigner and Balestra (1988) analyze the trade-off between pre-experimental observations of the treatment group and observations on a contemporaneous control group in social experiments. However, these authors assume throughout their paper that the marginal costs of observing the same unit for an additional time period are negligible compared to set-up costs, in which case a panel data set will always be preferable to a data set in which not all individuals are interviewed in each period. In this note we avoid this restrictive assumption and show how their results generalize in the fixed time effects case.

The plan of this note is as follows. In section 2 the model is outlined and in section 3 the optimal experimental design is derived. Section 4 contains some numerical results, while section 5 concludes. Technical details are deferred to the appendix.

* We thank Arthur van Soest for helpful comments and the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) for financial support.

0304-4076/92/$03.50 © 1992—Elsevier Science Publishers B.V. All rights reserved
2. The model

We consider a two-period model in which some units receive treatment in the second period, while in the first period no treatment is applied to any unit. Our model which includes fixed time effects coincides with the one analyzed by Aigner and Balestra (1988) in their section 5, although they also include random time effects. These are however not identified given the presence of the fixed effects. Thus, the model we consider is

\[ y_{it} = \mu_i + \beta d_{it} + \alpha_i + \varepsilon_{it}, \quad t = 1, 2, \]

where \( i \) indexes individuals and \( \alpha_i \) and \( \varepsilon_{it} \) are independent random drawings with means 0 and variances \( \sigma^2_\alpha \) and \( \sigma^2_\varepsilon \), respectively, where \( \alpha_i \) and \( \varepsilon_{it} \) are independent. The variable \( d_{it} \) is a dummy representing the presence of a treatment, i.e., \( d_{it} = 1 \) if \( t = 2 \) and individual \( i \) receives a treatment, while \( d_{it} = 0 \) otherwise. The parameter of interest, \( \beta \), is the treatment effect. No distributional assumptions are made on \( \mu_1 \) and \( \mu_2 \), which are consequently treated as fixed unknown parameters. The variable \( \alpha_i \) reflects an unobserved individual effect. The importance of this individual effect is indicated by the value of \( \rho = \sigma^2_\alpha (\sigma^2_\alpha + \sigma^2_\varepsilon)^{-1} \) and the overall variance is denoted by \( \sigma^2 = \sigma^2_\alpha + \sigma^2_\varepsilon \). It should be noted that the results for a random time effects model (as analyzed by Aigner and Balestra in their section 4) converge to the results for the fixed time effects model considered here if the budget increases, as can be easily seen from (3.6.20) and (3.3.5) in Hsiao (1986).

We assume that four types of observations can be obtained. The number of units observed in periods 1 and 2 not receiving a treatment is denoted by \( n^c_{12} \) where 'c' stands for 'control'. Similarly the number of units not receiving a treatment which are observed in period 2 only is denoted by \( n^c_2 \), the number of units with treatment observed in period 2 only is \( n^t_2 \) ('t' for 'treatment') and the number of units observed in both periods with treatment in period 2 by \( n^t_{12} \). The corresponding cost parameters are denoted by \( k^c_{12}, k^c_2, k^t_2, \) and \( k^t_{12} \), respectively. By assumption \( k^c_{12} \geq k^c_2 \), \( k^t_2 \geq k^t_{12} \), \( k^c_{12} \geq k^c_2 \), and \( k^t_{12} \geq k^t_2 \). Note that it will never be optimal to observe individuals in period 1 only. Finally, we assume that \( k^c_2/k^t_{12} > k^c_2/k^t_{12} \) which will typically be the case and which simplifies the presentation somewhat.

The choice of the optimal design can now be formalized as

\[ \min_{n^c_2, n^c_{12}, n^t_2, n^t_{12} \geq 0} V(\hat{\beta}) \quad \text{s.t.} \quad k^c_2 n^c_2 + k^c_{12} n^c_{12} + k^t_2 n^t_2 + k^t_{12} n^t_{12} = B, \]

where \( \hat{\beta} \) is the most efficient linear estimator of \( \beta \) and \( B \) is some fixed budget. In Aigner and Balestra (1988) it is assumed that \( n^c_2 = n^t_2 = 0 \), which is only a priori optimal if the marginal costs of observing the same unit for an
additional period are negligible compared to set up costs, that is if \( k_{12}^c = k_{2}^c \) and \( k_{12}^c = k_{2}^c \).

3. Determination of the optimal design

Using the fact that the four types of observations referred to in the previous section yield independent information on the parameters \( \mu_1, \mu_2, \) and \( \beta \) it is straightforward to verify that the variance of \( \beta \) is the lower right element of

\[
V\{\hat{\beta}\} = \sigma^2 \left( n_{12}^c (1 - \rho^2) \right) \left( \begin{array}{ccc} 1 & -\rho & 0 \\ -\rho & 1 & 0 \\ 0 & 0 & 0 \end{array} \right) + n_{2}^c \left( \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right) \\
+ n_{12} (1 - \rho^2) \left( \begin{array}{ccc} 1 & -\rho & -\rho \\ -\rho & 1 & 1 \\ -\rho & 1 & 1 \end{array} \right) + n_{2} \left( \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{array} \right) \right)^{-1}.
\]

(3)

Straightforward algebra using partitioned inverses shows that \( V\{\hat{\beta}\} \) can be written as

\[
V\{\hat{\beta}\} = \sigma^2 \left( \frac{(n_{12}^c + n_{2}^c)(n_{12}^c + n_{2}^c)}{n_{12}^c + n_{12}^c + n_{2}^c + n_{2}^c} + \frac{\rho^2}{1 - \rho^2} \frac{n_{12}^c n_{12}^c}{n_{12}^c + n_{12}^c + n_{2}^c} \right)^{-1}.
\]

(4)

The optimal panel design, in which \( n_{12}^c = 0 \) and \( n_{2}^c = 0 \) a priori, and the optimal cross-sectional design, in which \( n_{12}^c = 0 \) and \( n_{12}^c = 0 \) a priori, can be determined fairly simply using (4). The general constrained minimization problem is less straightforward since the relevant regime depends in a complicated fashion on \( \rho \) and the cost parameters. Therefore an explicit expression for the optimal design is hard to give. In order to obtain tractable results we choose a different approach. First, note that control units can be used to estimate \( \mu_1 \) and \( \mu_2 \) only, while the treatment observations yield information on \( \varphi_1 \) and \( \varphi_2 \) only, where we define \( \varphi_1 = \mu_1 \) and \( \varphi_2 = \mu_2 + \beta \). The corresponding efficient estimators are denoted by \( \hat{\varphi}_1, \hat{\varphi}_2, \hat{\varphi}_1^c, \) and \( \hat{\varphi}_2^c \). Evidently, because \( \beta \) is unbiased it has to be of the form

\[
\hat{\beta}(\eta) = \hat{\varphi}_2^c - \hat{\varphi}_2^c + \eta(\hat{\varphi}_1^c - \hat{\mu}_2^c).
\]

(5)

where \( \eta \) is a scalar which can be chosen freely in order to minimize the variance of \( \hat{\beta} \). Using expressions for \( V(\hat{\mu}) \) and \( V(\hat{\varphi}) \) similar to (3), one obtains
from (5)

\[
V[\hat{\beta}] = \min_{\eta} \sigma^2 (1 + 2 \rho \eta + \eta^2) \left\{ \frac{d(\eta)n^c + (1 - d(\eta))n_{i2}}{n^c n_{i2}} \right. \\
+ \left. \frac{d(\eta)n^t + (1 - d(\eta))n_{i2}}{n^t n_{i2}} \right\},
\]

(6)

where \( n^c = n_{i2}^c + n_{i2}^s \), \( n^t = n_{i2}^t + n_{i2}^s \), and \( d(\eta) = \eta^2 (1 - \rho^2) / (1 + 2 \rho \eta + \eta^2) \).

Of course the minimal value in (6) will coincide with (4). However, expression (6) can easily be used to obtain the optimal design if we first determine the optimal design for given values of \( \eta \), which is straightforward, and in a second step minimize with respect to \( \eta \). In the appendix we show that for given values of \( \eta \) and a design which is optimal given \( \eta \), the variance of the estimator \( \hat{\beta}(\eta) \) can be written as

\[
V_{\min} [\hat{\beta}(\eta)] = \sigma^2 (1 + 2 \rho \eta + \eta^2) B^{-1} f_A^2(\eta) \quad \text{if} \quad d(\eta) < 1 - k_1^t / k_{i2}^t,
\]

\[
= \sigma^2 (1 + 2 \rho \eta + \eta^2) B^{-1} f_B^2(\eta) \quad \text{if} \quad 1 - k_1^t / k_{i2}^t \leq d(\eta) < 1 - k_1^c / k_{i2}^c.
\]

\[
= \sigma^2 (1 + 2 \rho \eta + \eta^2) B^{-1} f_C^2(\eta) \quad \text{if} \quad d(\eta) \geq 1 - k_1^c / k_{i2}^c,
\]

(7)

where

\[
f_A(\eta) = \sqrt{d(\eta)} \left[ \sqrt{(k_{i2}^t - k_1^t)} + \sqrt{(k_{i2}^c - k_1^c)} \right] + \sqrt{(1 - d(\eta))} \left[ \sqrt{k_1^c} + \sqrt{k_1^t} \right],
\]

\[
f_B(\eta) = \sqrt{d(\eta)} \sqrt{(k_{i2}^c - k_1^c)} + \sqrt{(1 - d(\eta))} \sqrt{k_1^c} + \sqrt{k_1^t},
\]

\[
f_C(\eta) = \sqrt{k_{i2}^c} + \sqrt{k_{i2}^t}.
\]

(8)

The subsequent problem is to minimize this function with respect to \( \eta \). In order to present the results we first of all define

\[
G_1 = \frac{\left\{ \sqrt{(k_{i2}^t - k_1^t)} + \sqrt{(k_{i2}^c - k_1^c)} \right\}^2}{\left\{ \sqrt{(k_{i2}^c - k_1^c)} + \sqrt{(k_{i2}^t - k_1^t)} \right\}^2 + \left\{ \sqrt{k_1^c} + \sqrt{k_1^t} \right\}^2},
\]

\[
G_2 = 1 - k_1^c / k_{i2}^c.
\]

(9)

Using the fact that the optimal value of \( \eta \) will satisfy \(-\rho \leq \eta \leq 0\) where the bounds correspond to a panel and a cross-sectional design, respectively, one can easily show that the derivative of \( V_{\min}(\hat{\beta}(\eta)) \) with respect to \( \eta \) is negative.


No. 5  Th. ten Ras and F. van der Ploeg, A statistical approach to the problem of negatives in input-output analysis, Economic Modelling, vol. 6, no. 1, 1989, pp. 2 - 19.


No. 8  Th. van de Klundert and F. van der Ploeg, Wage rigidity and capital mobility in an optimizing model of a small open economy, De Economist 137, nr. 1, 1989, pp. 47 - 75.


for all \( \eta, -\rho \leq \eta \leq 0 \) if \( \rho^2 < G_1 \). This implies that \( \eta = 0 \) will minimize (7) if this condition on \( \rho \) is satisfied. If \( G_1 < \rho^2 < G_2 \) the derivative is negative if \( \eta < \tilde{\eta} \) and positive if \( \eta > \tilde{\eta} \), where \( \tilde{\eta} \) is the minimum of \((1 + 2\rho \eta + \eta^2)f_\theta'(\eta)\). This yields \( \eta^{opt} = \tilde{\eta} \). Although no simple analytical expression for the optimal design can be presented (see appendix), the condition on \( \rho \), the importance of the individual effect, can easily be interpreted. Finally, it can be checked that the derivative of \( V_{min}(\hat{\beta}(\eta)) \) with respect to \( \eta \) is positive on \((-\rho, 0)\) if \( \rho^2 \geq G_2 \) which implies \( \eta^{opt} = -\rho \). From (5) it is clear that \( \eta = 0 \) can only be preferable if the design is cross-sectional. Similarly, \( \eta = -\rho \) can be preferable for a panel design only. Consequently, the optimal design can be summarized as follows:

- If \( \rho^2 < G_1 \): cross-sectional design \( (n_{12}^1 = n_{12}^2 = 0) \).
- If \( G_1 \leq \rho^2 < G_2 \): mixed design \( (n_2^1 = 0) \).
- If \( \rho^2 \geq G_2 \): panel design \( (n_2^1 = n_2^2 = 0) \).

These results are intuitively plausible: if the individual effect is very small, it will be suboptimal to spend money on observing individuals in the first period. If on the other hand the individual effect is very important, it will always be preferable to construct a panel irrespective of the relative cost structure. In the Aigner and Balestra case, where the marginal costs of a repeated observation are negligible \( (k_{ij}^1 = k_{ij}^2 \) and \( k_{ij}^1 = k_{ij}^2 \) ), both \( G_1 \) and \( G_2 \) are zero and a pure panel data set is optimal for any positive value of \( \rho \). If these costs are however nonnegligible, it may be preferable to have additional cross-sectional control observations \( (G_1 < \rho^2 < G_2) \) or to have a pure cross-sectional design \( (\rho^2 < G_1) \).

4. A numerical example

As an illustration the optimal design is presented in table 1 for the case where \( k_{ij}^1 = 1, k_{ij}^2 = 1.5, k_{ij}^1 = 9, k_{ij}^2 = 9.5 \), and varying \( \rho \). The results in the previous section imply that a cross-sectional design is preferable if \( \rho < \sqrt{G_1} = 0.333 \) which is easily checked in the table. Similarly, a pure panel design is preferable if \( \rho > \sqrt{G_2} = 0.577 \). In the final two columns of the table the relative efficiency of the optimal design with respect to the optimal cross-sectional design and Aigner and Balestra's optimal pure panel design are presented. Evidently, the cost savings if the optimal design is used instead of an a priori fixed type of sample design can be substantial.

The relative efficiency measures given in the last column of table 1 can also be interpreted as the relative costs of choosing a suboptimal sample design (with its optimal experimental design) to obtain the same level of accuracy as


<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Authors</th>
<th>Journal/Media</th>
<th>Volume/Issue/Pages</th>
</tr>
</thead>
</table>


