Journal of Econometrics 51 (1992) 183–189. North-Holland

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.

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

$$y_{it} = \mu_t + \beta d_{it} + \alpha_i + \varepsilon_{it}, \qquad t = 1,2, \tag{1}$$

where i indexes individuals and α_i and ε_{it} are independent random drawings with means 0 and variances σ_α^2 and σ_ε^2, respectively, where α_i and ε_{jt} 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, β, is the treatment effect. No distributional assumptions are made on μ_1 and μ_2, which are consequently treated as fixed unknown parameters. The variable α_i reflects an unobserved individual effect. The importance of this individual effect is indicated by the value of $\rho = \sigma_\alpha^2(\sigma_\alpha^2 + \sigma_\varepsilon^2)^{-1}$ and the overall variance is denoted by $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$. 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_{12}^{c} 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_2^{c}, the number of units with treatment observed in period 2 only is n_2^{t} ('t' for 'treatment') and the number of units observed in both periods with treatment in period 2 by n_{12}^{t}. The corresponding cost parameters are denoted by k_{12}^{c}, k_2^{c}, k_2^{t}, and k_{12}^{t}, respectively. By assumption $k_{12}^{\mathrm{c}} \geq k_2^{\mathrm{c}}$, $k_{12}^{\mathrm{t}} \geq k_2^{\mathrm{t}}$, $k_{12}^{\mathrm{t}} \geq k_{12}^{\mathrm{c}}$, and $k_2^{\mathrm{t}} \geq k_2^{\mathrm{c}}$. Note that it will never be optimal to observe individuals in period 1 only. Finally, we assume that $k_2^{\mathrm{t}}/k_{12}^{\mathrm{t}} > k_2^{\mathrm{c}}/k_{12}^{\mathrm{c}}$ which will typically be the case and which simplifies the presentation somewhat.

The choice of the optimal design can now be formalized as

$$\min_{n_2^{\mathrm{c}}, n_{12}^{\mathrm{c}}, n_2^{\mathrm{t}}, n_{12}^{\mathrm{t}} \geq 0} V\{\hat{\beta}\} \quad \text{s.t.} \quad k_2^{\mathrm{c}} n_2^{\mathrm{c}} + k_{12}^{\mathrm{c}} n_{12}^{\mathrm{c}} + k_2^{\mathrm{t}} n_2^{\mathrm{t}} + k_{12}^{\mathrm{t}} n_{12}^{\mathrm{t}} = B, \tag{2}$$

where $\hat{\beta}$ is the most efficient linear estimator of β and B is some fixed budget. In Aigner and Balestra (1988) it is assumed that $n_2^{\mathrm{c}} = n_2^{\mathrm{t}} = 0$, 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 $k_{12}^{t} = k_{2}^{t}$ 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 μ_1, μ_2, and β it is straightforward to verify that the variance of $\hat{\beta}$ is the lower right element of

$$V\{\hat{\mu}_1, \hat{\mu}_2, \hat{\beta}\} = \sigma^2 \left(n_{12}^{c}(1-\rho^2)^{-1} \begin{pmatrix} 1 & -\rho & 0 \\ -\rho & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + n_2^{c} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right.$$

$$\left. + n_{12}^{t}(1-\rho^2)^{-1} \begin{pmatrix} 1 & -\rho & -\rho \\ -\rho & 1 & 1 \\ -\rho & 1 & 1 \end{pmatrix} + n_2^{t} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \right)^{-1}. \tag{3}$$

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

$$V\{\hat{\beta}\} = \sigma^2 \left\{ \frac{(n_{12}^{c} + n_2^{c})(n_{12}^{t} + n_2^{t})}{n_{12}^{c} + n_{12}^{t} + n_2^{c} + n_2^{t}} + \frac{\rho^2}{1-\rho^2} \frac{n_{12}^{t} n_{12}^{c}}{n_{12}^{t} + n_{12}^{c}} \right\}^{-1}. \tag{4}$$

The optimal panel design, in which $n_2^{c} = 0$ and $n_2^{t} = 0$ *a priori*, and the optimal cross-sectional design, in which $n_{12}^{c} = 0$ and $n_{12}^{t} = 0$ *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 ρ 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 μ_1 and μ_2 only, while the treatment observations yield information on φ_1 and φ_2 only, where we define $\varphi_1 = \mu_1$ and $\varphi_2 = \mu_2 + \beta$. The corresponding efficient estimators are denoted by $\hat{\mu}_1^{c}$, $\hat{\mu}_2^{c}$, $\hat{\varphi}_1^{t}$, and $\hat{\varphi}_2^{t}$. Evidently, because $\hat{\beta}$ is unbiased it has to be of the form

$$\hat{\beta}(\eta) = \hat{\varphi}_2^{t} - \hat{\mu}_2^{c} + \eta(\hat{\varphi}_1^{t} - \hat{\mu}_1^{c}), \tag{5}$$

where η 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)

$$V\{\hat{\beta}\} = \min_{\eta} \sigma^2(1 + 2\rho\eta + \eta^2)\left\{ \frac{d(\eta)n^c + (1 - d(\eta))n_{12}^c}{n^c n_{12}^c} + \frac{d(\eta)n^t + (1 - d(\eta))n_{12}^t}{n^t n_{12}^t} \right\}, \quad (6)$$

where $n^c = n_{12}^c + n_2^c$, $n^t = n_{12}^t + n_2^t$, 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 η, which is straightforward, and in a second step minimize with respect to η. In the appendix we show that for given values of η and a design which is optimal given η, the variance of the estimator $\hat{\beta}(\eta)$ can be written as

$$\begin{aligned} V_{\min}\{\hat{\beta}(\eta)\} &= \sigma^2(1 + 2\rho\eta + \eta^2)B^{-1}f_A^2(\eta) \quad \text{if} \quad d(\eta) < 1 - k_2^t/k_{12}^t, \\ &= \sigma^2(1 + 2\rho\eta + \eta^2)B^{-1}f_B^2(\eta) \quad \text{if} \quad 1 - k_2^t/k_{12}^t \le d(\eta) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad < 1 - k_2^c/k_{12}^c, \\ &= \sigma^2(1 + 2\rho\eta + \eta^2)B^{-1}f_C^2(\eta) \quad \text{if} \quad d(\eta) \ge 1 - k_2^c/k_{12}^c, \end{aligned} \quad (7)$$

where

$$\begin{aligned} f_A(\eta) &= \sqrt{d(\eta)}\left[\sqrt{(k_{12}^t - k_2^t)} + \sqrt{(k_{12}^c - k_2^c)}\right] + \sqrt{(1 - d(\eta))}\left[\sqrt{k_2^t} + \sqrt{k_2^c}\right], \\ f_B(\eta) &= \sqrt{d(\eta)}\sqrt{(k_{12}^c - k_2^c)} + \sqrt{(1 - d(\eta))}\sqrt{k_2^c} + \sqrt{k_{12}^t}, \\ f_C(\eta) &= \sqrt{k_{12}^c} + \sqrt{k_{12}^t}. \end{aligned} \quad (8)$$

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

$$\begin{aligned} G_1 &= \frac{\left\{\sqrt{(k_{12}^c - k_2^c)} + \sqrt{(k_{12}^t - k_2^t)}\right\}^2}{\left\{\sqrt{(k_{12}^c - k_2^c)} + \sqrt{(k_{12}^t - k_2^t)}\right\}^2 + \left\{\sqrt{k_2^c} + \sqrt{k_2^t}\right\}^2}, \\ G_2 &= 1 - k_2^c/k_{12}^c. \end{aligned} \quad (9)$$

Using the fact that the optimal value of η will satisfy $-\rho \le \eta \le 0$ 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 η is negative

for all η, $-\rho \leq \eta \leq 0$ if $\rho^2 < G_1$. This implies that $\eta = 0$ will minimize (7) if this condition on ρ 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 $(1 + 2\rho\eta + \eta^2) f_B^2(\eta)$. This yields $\eta^{\text{opt}} = \tilde{\eta}$. Although no simple analytical expression for the optimal design can be presented (see appendix), the condition on ρ, 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 η is positive on $(-\rho, 0)$ if $\rho^2 \geq G_2$ which implies $\eta^{\text{opt}} = -\rho$. 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	$(n_{12}^{\text{c}} = n_{12}^{\text{t}} = 0)$.
If $G_1 \leq \rho^2 < G_2$:	mixed design	$(n_2^{\text{t}} = 0)$.
If $\rho^2 \geq G_2$:	panel design	$(n_2^{\text{t}} = n_2^{\text{c}} = 0)$.

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 ($k_{12}^{\text{t}} = k_2^{\text{t}}$ and $k_{12}^{\text{c}} = k_2^{\text{c}}$), both G_1 and G_2 are zero and a pure panel data set is optimal for any positive value of ρ. If these costs are however nonnegligible, it may be preferable to have additional cross-sectional control observations (if $G_1 < \rho^2 < G_2$) or to have a pure cross-sectional design (if $\rho^2 < G_1$).

4. A numerical example

As an illustration the optimal design is presented in table 1 for the case where $k_2^{\text{c}} = 1$, $k_{12}^{\text{c}} = 1.5$, $k_2^{\text{t}} = 9$, $k_{12}^{\text{t}} = 9.5$, and varying ρ. 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 1

The optimal design and its relative efficiency with respect to pure cross-sectional and pure panel designs assuming $k_2^c = 1$, $k_{12}^c = 1.5$, $k_2^t = 9$, $k_{12}^t = 9.5$, and $B = 10000$.

	Optimal design				Relative efficiency w.r.t.	
ρ	n_2^c	n_{12}^c	n_2^t	n_{12}^t	pure CS	pure panel
0.0	2500	0	833	0	1.00	1.16
0.1	2500	0	833	0	1.00	1.15
0.2	2500	0	833	0	1.00	1.11
0.3	2500	0	833	0	1.00	1.06
0.4	1142	1079	0	762	1.04	1.02
0.5	516	1541	0	755	1.15	1.01
0.6	0	1896	0	753	1.35	1.00
0.7	0	1896	0	753	1.69	1.00
0.8	0	1896	0	753	2.40	1.00
0.9	0	1896	0	753	4.54	1.00

the optimal design. So, for example, if $\rho = 0.7$, one can save 41% ($100 \times 0.69/1.69$) of the total budget by choosing the optimal panel design instead of the optimal cross-sectional design.

5. Conclusions

In this note we generalized some of the results in Aigner and Balestra (1988) in the fixed time effects model by dropping the assumption that the marginal costs of a repeated observation are negligible. We derived a condition under which the panel design considered by these authors is nevertheless optimal. A mixed design in which only part of the sample is reinterviewed or a pure cross-sectional design may be preferable if the individual effect is small.

Appendix

In this appendix we determine the optimal design for given values of η. The results obtained here are used in section 3 to determine the optimal choice of η. Using (6) we can write

$$V\{\hat{\beta}(\eta)\} = \sigma^2(1 + 2\rho\eta + \eta^2) \times \left\{ \frac{d(\eta)}{n_{12}^c} + \frac{1 - d(\eta)}{n^c} + \frac{d(\eta)}{n_{12}^t} + \frac{1 - d(\eta)}{n^t} \right\}. \quad \text{(A.1)}$$

Defining $k^c = k_{12}^c - k_2^c > 0$ and $k^t = k_{12}^t - k_2^t > 0$, the budget constraint for the minimization problem can be written as

$$k_2^c n^c + k_2^t n^t + k^c n_{12}^c + k^t n_{12}^t = B.$$

If we minimize (A.1) with respect to n_2^c, n_{12}^c, n_2^t, and n_{12}^t for a given value of η without imposing the nonnegativity constraints, we obtain

$$n_{12}^c = (B/\sqrt{k^c})\frac{\sqrt{d(\eta)}}{f_A(\eta)}, \qquad n_{12}^t = (B/\sqrt{k^t})\frac{\sqrt{d(\eta)}}{f_A(\eta)},$$

$$n^c = (B/\sqrt{k_2^c})\frac{\sqrt{(1-d(\eta))}}{f_A(\eta)}, \qquad n^t = (B/\sqrt{k_2^t})\frac{\sqrt{(1-d(\eta))}}{f_A(\eta)}, \tag{A.2}$$

where $f_A(\eta)$ has been defined in section 4. The corresponding minimal variance is $\sigma^2(1+2\rho\eta+\eta^2)B^{-1}f_A^2(\eta)$ as stated in section 4. Note however that the nonnegativity constraints in (2) imply that $n^t - n_{12}^t \geq 0$, which is violated if $d(\eta) > k^t/(k_2^t + k^t)$. If $n_2^t = 0$ is imposed one obtains an optimization problem similar to the previous one, with solution

$$n_{12}^c = (B/\sqrt{k^c})\frac{\sqrt{d(\eta)}}{f_B(\eta)}, \qquad n^t = n_{12}^t = (B/\sqrt{(k_2^t + k^t)})\frac{1}{f_B(\eta)},$$

$$n^c = (B/k_2^c)\frac{\sqrt{(1-d(\eta))}}{f_B(\eta)}. \tag{A.3}$$

This solution is valid and optimal as long as $k^t/(k_2^t + k^t) < d < k^c/(k_2^c + k^c)$. If $d > k^c/(k_2^c + k^c)$ (case C), two nonnegativity constraints are binding and we have to optimize (A.1) imposing $n_2^c = n_2^t = 0$. This yields

$$n_{12}^c = \frac{B/\sqrt{k_{12}^c}}{(\sqrt{k_{12}^t} + \sqrt{k_{12}^c})}, \qquad n_{12}^t = \frac{B/\sqrt{k_{12}^t}}{(\sqrt{k_{12}^t} + \sqrt{k_{12}^c})}. \tag{A.4}$$

The variance of $\hat{\beta}(\eta)$ is minimized with respect to η in the main text. Once the optimal value of η is known, the optimal design is given in (A.2), (A.3), or (A.4), while (7) yields the variance of an efficient estimator based on this design. The optimal value of $\eta = 0$ if $\rho^2 < G_1$ implies that n_{12}^t and n_{12}^c given in (A.2) equal zero, meaning that a pure cross-sectional design is optimal.

References

Aigner, D.J. and P. Balestra, 1988, Optimal experimental design for error components models, Econometrica 56, 955–971.

Hsiao, C., 1986, Analysis of panel data, Econometric Society monograph (Cambridge University Press, Cambridge).