Chapter 8

Conclusion and Topics for Further Research

8.1 Concluding Remarks

The present thesis examined the methods used to evaluate the effectiveness of various social programs within a causal inference framework. Outcomes being evaluated are assumed to be obtained from participants that are self-selected into a particular program in terms of personal, observed or unobserved to the analyst, criteria. This nonrandom nature of participation decision that generates the nonrandom samples of participants (and nonparticipants), leads to biased estimates of the program impacts when applying conventional statistical methods. Thus, alternative procedures had to be considered to deal with nonrandom samples and correct for the selection bias.

Mainly, three evaluation methods have been reviewed. *Randomized experiments* and *Matching methods* are generally thought as the statisticians' way of evaluation. On the other hand, *structural models* are assumed to be applied mostly by econometricians. The advantages and the limitations of these procedures have been outlined extensively in this thesis. Both cases of cross-sectional studies and panel studies have been analysed.

Several aspects are worth emphasizing. An important one is the existence of a multiplicity of parameters of interest in evaluating a social program, which results in part because of the heterogeneous impacts of the program itself. To this extend, some popular estimators that are well-suited for estimating one set of parameters are poorly suited for estimating others as a result of this heterogeneity. In principle, many different parameters are required to answer the interesting evaluation questions.

Econometric estimators allow for this possibility, which is an important insight of the modern literature that challenges traditional approaches to program evaluation.

Evidence that different non-experimental estimators produce different estimates of the same parameter does not indicate that non-experimental methods cannot address the underlying self-selection problem of the data. Instead, different estimates obtained from different estimators simply indicate that different estimators address the selection problem in different ways and that non-random participation in social programs is an important problem that deserves more attention in its own right. Different methods produce the same estimates only if there is no problem of selection bias.

It is also remarkable that there is inherently no method of choice for conducting program evaluations. The choice of an appropriate estimator should be guided by the economics underlying the problem, the data that are available or can be acquired, and the evaluation question being addressed.

As far as the data are concerned, it must be noted that *better data help a lot* to evaluate a particular program. In fact, the more improved the data are, the easier the evaluation process becomes. The best evaluation solution lies in improving the quality of the data on which evaluations are conducted and not in the development of formal econometric methods to circumvent inadequate data.

An essential condition to obtain adequate data is to compare comparable people. Many non-experimental evaluation studies have identified the parameter of interest by comparing observationally different persons using extrapolations based on inappropriate functional forms imposed to make incomparable people comparable. A major advantage of non-parametric, econometric methods, and more specifically matching, for solving the problem of selection bias is that, rigorously applied, they force analysts to compare "similar" individuals.

Finally, the contribution of the experiments in evaluation studies has to be outlined. Although it is a simple, non-parametric method that produces plausible estimates of the program impacts, experiments estimate the effect of the program compared to no program at all when they are used to evaluate the effect of a program for which there are few good substitutes. They are less effective when evaluating ongoing programs, in part because they appear to disrupt established bureaucratic procedures. The threat of disruption leads local bureaucrats to oppose their adoption. To the extend that programs are disrupted, the program evaluated by the method is not

the ongoing program that one seeks to evaluate. The parameter estimated in experimental evaluations is often not likely to be of primary interest to policy makers and researchers, and in any case, it has to be more carefully interpreted than is commonly done in most public discussions. However, if there is no disruption, and the other problems that plague experiments are absent, the evidence from social experiments provides a benchmark for learning about the performance of alternative non-experimental methods.

8.2 Topics for Further Research

Until now, evaluation processes has been focused on evaluation of direct benefits from participating in a social program. It has been straightly indicated that indirect benefits are also of great importance on an evaluation studies. More specifically, programs that are implemented at a national or regional level affect both participants and non-participants. The current practice in the entire "treatment effect" literature is to ignore the indirect effects of programs on non-participants by assuming that they are negligible. This practice can produce substantially misleading estimates of program impacts if indirect benefits are substantial or if Lewis's (1963) assumption, relatively to the approximation of the non-participants outcomes from controls outcomes within a social experiment, is implausible.

To account for the impacts of programs on both participants and non-participants, general equilibrium frameworks are required to be developed and within them to conduct the evaluation process. The problem of indirect effects poses a major challenge to conventional micro-methods that focus only on direct impacts, and demonstrates the need for program evaluations to utilize market-wide data.

Another aspect, not examined yet, is the application of evaluation methods in studies where the outcomes obtained are discrete, that is the dependent variables of both the primary and participation equations are categorical. Although, it seems that a carefully designed randomized experiment or the conduction of Matching Methods may provide answers to such questions by using descriptive measures like medians and modes or performing median regression techniques, structural models approach cannot. The reason for this is that such data do not satisfy the assumption of Normality for the primary's equation error terms. Thus, another challenge for future

research on the evaluation studies area is the development of structural models that can analyze adequately these evaluation problems.

Any conclusions drawn from the various selection models will inevitably depend on their validity. One usual way of assessing the correspondence between the data and the model is by analyzing the residuals. Residuals are used in model evaluation as a visual check on distributional assumptions and to detect any aberrant observations. There is very little literature on residuals tests for the models formulated in this thesis (only a few papers are referred briefly on this subject, see Heckman, Tobias and Vytlacil, 2000) because selection creates a problem for the usual diagnostic tests. Hence, additional work is also needed on this subject.

Generally, evaluation of social programs is not universally applied. Heckman, LaLonde and Smith (1999) mention that important lessons on this rather interesting issue, and specifically on evaluation of government training programs, come from American studies because U.S. government has been much more active in promoting evaluations than have other governments, and the results from the evaluations are often used to expand –or contract– government programs.

In Greece governmental authorities have not organised evaluation studies yet. The lack of relevant data, collected in databases, is possibly one important reason of the perceived inexistence. Another reason is possibly the perceived problems in establishing causality in most applications (see Freedman, 1997, 2002). In any case, we explicitly demonstrate the need for the conduct of such studies in several areas of social life. Illustrative examples are medical applications (e.g. evaluation of a new treatment in clinical trials) or generally the research in the area of Medical Statistics, which is regarded as a modern and continuously developed science field in several EU countries and in US. Also of interest is to study the applications of these methods in educational studies (evaluation in the wages of participants relatively to their educational level), in policy evaluation studies (evaluation of a specific governmental policy to a specific social group) and in workers training programs studies in order to appraise whether a particular program is a worthwhile investment.