Evaluation Strategies in Labor Economics – An Application to Post-Secondary Education

Inaugural-Dissertation

zur Erlangung

der Würde eines Doktors der Wirtschaftswissenschaften

der Wirtschaftswissenschaftlichen Fakultät

der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Boris Augurzky

aus Heilbronn

Heidelberg, Oktober 2000

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Overview

Identifying causal relationships is of central concern in many applied economic research. For example, political groups might be interested in how union membership affects labor market outcomes of union members. The government might want to know how certain active labor market programs help improve participants' labor market status. Frequently, for the sake of evaluation of such programs, candidate causal variables are simply compared with their supposed effects. However, a simple bivariate comparison might generally lead to misinterpretations and wrong conclusions. Most likely, there might be other – so-called confounding – forces determining both the supposed cause and its effect. For example, participants in an active labor market program might systematically differ from non-participants, say, they have higher motivation or higher program-specific skills, such that they would be more successful in the labor market anyway, even without the program. Thus, the association of participation and success in the labor market might wrongly be interpreted as a causal relationship. Applied econometric research attempts to take account of confounding background variables by multivariate estimation techniques such as the classical linear multivariate regression approach.

Recently, the econometric literature has been incorporating a new alternative statistical technique which does not need to rely intensively on parametric or functional form assumptions (see Heckman, LaLonde & Smith, 1999, and Angrist & Krueger, 1999). Rather, this technique is based on directly matching individuals who are equal in all observable respects with the only exception that one individual has experienced the impact of a potentially causal variable while the other has not. Their difference in the outcome variable under scrutiny will then be attributable to the effect of the intervention. The idea of matching originates in the randomized controlled trial (RCT). In an RCT with binary causal variable, units are randomly assigned to one of two states, either the *treatment* or the *control* state. Then, the treatment effect is estimated by taking the difference between the mean outcome of treated and control units. The estimate is unbiased since the randomization property of the experiment – if implemented appropriately – ensures that, on average, all covariates of treated and control units – be they observed or unobserved – are balanced. In contrast to an RCT, in an observational study treated and untreated units may differ considerably because of their being *self-selected* into the treatment state in lieu of being selected by an exogenous random mechanism.

Matching aims at removing systematic imbalance of covariates in an observational study by selecting controls from the untreated group, the control reservoir, who are "similar" to treated units in all relevant variables. In other words, it aims at constructing an artificial control group. Of course, imbalance in unobservable characteristics cannot be remedied. Insofar, both matching and the classical linear regression model control for *observable* confounding variables. Yet, the difference is that the first technique is non-parametric while the latter interpolates linearly when there are no perfectly equal units. However, note that even a saturated linear model would not necessarily identify the same parameter as matching if heterogeneity in the effect was present (see Angrist, 1998 or Angrist & Krueger, 1999). This is because OLS and matching impose different weighting schemes when averaging over individual effects.

If the relevant variables are of high dimension exact matching in a finite sample is, in all likelihood, impossible. As an alternative, Rosenbaum & Rubin (1983) suggest in their seminal paper to match on the one-dimensional probability of participation in the treatment, the *propensity score*. They show that matching on the propensity score is a valid approach whenever matching on all covariates is valid. However, since the propensity score is unknown, further problems might arise in its estimation. For example, it is often unclear how to specify the selection equation, which variables to include in the estimation, and how to define a propensity score distance. Furthermore, in case of parametric binary choice models – probit or logit – the question arises whether to match on the estimated index, i.e. the probits or logits, that is linear in the covariates or to match on the estimated propensity score that, in order to be located in the unit interval, depends on the covariates in a nonlinear fashion.

In this thesis the matching approach is used to evaluate postsecondary education and to contrast estimation results with conventional ordinary least squares estimation. To this end, data are drawn from the *National Longitudinal Survey of Youth 1979* (NLSY), an American panel data set that started in 1979, comprising young individuals aged between 14 and 22 who have been re-interviewed annually until 1994 and biennially thereafter. In terms of the matching methodology, postsecondary education constitutes the treatment while the control reservoir is made up of individuals having a high school diploma only. All empirical chapters of this thesis will use these data.

It turns out that selection into college, especially into four-year colleges, is extraordinarily strong. Observable variables such as ability test scores and socio-economic background variables are essential determinants affecting the decision to take up a college education. Unfortunately, matching treated and untreated units in such a case is no easy matter. For instance, in the extreme case, if selection were perfectly predictable, matching would be impossible because there would be no unit in the control reservoir having the same characteristics as any treated. Note that also a linear model interpolating the extremes would not be promising either. Consequently, pair matching, i.e. matching exactly one treated and one untreated, as frequently pursued in applied work to evaluate active labor market programs (see e.g. Lechner, 1999), might be inappropriate in the schooling example. It would have to drop great a number of treated units at the high end of the propensity score scale and, thus, it would produce matched pairs which are not anymore representative of the whole treated population. Hence, if the treatment effect is heterogeneous pair matching estimates might be severely biased.

In order to keep as many treated units as possible, one control should be matched to more than one treated person, something which is often referred to as *matching with re-*

placement. Dehejia & Wahba (1998) suggest an algorithm which follows this approach. Usually, however, this algorithm lacks some optimality criterion in that it does not necessarily achieve to minimize the overall distance between treated and control units. What is more, several controls could also be matched to one treated unit to increase statistical precision. In the end, the *full* sample might be stratified into small strata consisting of either one treated and one or more controls or one control and more than one treated unit.

Rosenbaum (1991) suggests an *optimal full matching* algorithm which not only matches all units in the sample but which also manages to minimize the total distance between treated and controls. In general, however, exact matches on the propensity score are not possible and a certain small distance between the matched treated and the control has to be accepted. *Greedy* algorithms address this issue by matching a randomly chosen treated unit to the closest untreated available who will then be removed from the sample. By contrast, the *optimal* algorithm is apt to find the overall minimum distance by reconsidering and possibly rearranging already matched units.

Chapter 2 examines several steps in the practical implementation of the method of matching. Typically, the applied researcher has to make decisions on how to adapt certain parameters in the matching algorithm. In contrast to the simulation study of Gu & Rosenbaum (1993), this chapter performs a sensitivity analysis with data from the NLSY. First, it analyses whether matching on the propensity score or on the linear index is to be preferred. Second, it suggests to use a so-called propensity score caliper approach which ensures that the distance between the treated and control unit does not exceed a certain pre-specified range. Otherwise, arbitrarily large distances might occur. A broad and a narrow caliper width are set against. Third, the question arises how to define distance within calipers. The literature suggests to either use the propensity score distance or the Mahalanobis metric, both of which will be investigated. Fourth, three matching algorithms are compared: optimal full, a greedy full, and a greedy pair matching. Fifth, suitable stratum weighting schemes are built to identify the mean effect of treatment on the treated and the mean effect on a randomly assigned person. The results will be discussed with respect to three measures of success: balance of covariates after matching, variance of the matching estimates, and how systematic treated units are discarded by the algorithms.

Sensitivity of the decision parameters as to the estimated treatment effects appears to be rather modest. Systematic variation in the estimates caused by variation of the distance measures between treated and untreated units or by altering matching algorithms is negligible and statistically insignificant. Moreover, roughly 80% of the initial bias in the observable covariates is removed by full matching algorithms and 87% by pair matching. Yet, mean propensity scores of pair-matched treated individuals are markedly lower than in the original treatment group before matching. If high-propensity score individuals experience a higher effect of their education pair matching estimates are expected to be biased.

Alas, heterogeneity is too weak to unanimously favor full matching since its disadvantages clearly emerge. Full matching estimates are accompanied by relatively large standard errors because a full stratification is far from being as uniform as pair matching. For example, the more strata consist of a large number of treated units sharing only one control the higher standard errors of the estimated mean effect on the treated individual are. It turns out that the specific greedy full algorithm as implemented in this thesis achieves a more uniform stratification than the optimal one. Notwithstanding, in order to attain a more uniform stratification, greedy algorithms can always be replaced by a suitable optimal one when restrictions on the size of the strata are incorporated in the optimization process. Therefore, greedy algorithms should be abandoned. Furthermore, matching on the linear index score turns out to better discriminate between units at the low end of the propensity score scale. As a result, it drops many low-score untreated individuals who would almost all be used in matching on the propensity score.

Chapter 3 is dedicated to specific problems that might arise under strong selection into college as in Chapter 2. If treatment effects are heterogeneous and selection into treatment is exceptionally strong, pair matching is an efficient evaluation strategy. In the heterogeneous case, it is unclear which matching method to prefer. This chapter, however, suggests to concentrate less on the choice of method but, alternatively, to carefully reconsider the selection equation. Some variables might be strong determinants of the selection but exhibit a rather modest impact on the outcome. If they are omitted randomness of the selection process increases or, in other words, some observable self-selection is left to stochastic noise. This will result in a smaller propensity score difference between treatment and comparison group and, consequently, matching will become easier. Only the relevant variables which rule both the selection *and* the outcome have to be balanced.

On the other hand, a consistent estimation of the propensity score might make it necessary to include into the selection equation all the variables that rule the selection process even if they do not determine the outcome. Many applied research emphasizes the importance of consistent estimation of the selection equation. For instance, Lechner (1999, 2000) performs and recommends several specification tests to examine whether a probit model is adequate for describing the selection decision. Heckman, Ichimura & Todd (1997: section 8) choose the predictor variables to maximize the within-sample correct prediction rates of participation. Although a selection process well understood might in itself be an important contribution, it is not the main objective of propensity score matching envisaging to identify the mean effect of treatment. What is to be achieved by propensity score matching is balance of the relevant covariates in order to eliminate selection bias, even if the propensity score is inconsistently estimated. Obviously, there is a trade-off between feasible matching on the one hand and consistent estimation of the propensity score on the other.

To assess this trade-off Chapter 3 performs a simulation study. It turns out that even when matching builds on quite inconsistent propensity score estimates, estimation results of the mean effect of treatment can still be superior, in terms of the mean squared error, to results produced by a consistent propensity score estimation which might separate treated and untreated units too successfully. The findings of this simulation study recommend to only include variables into the selection equation that are highly significant. Variables with low significance levels are obvious candidates for exclusion, even if they might play a role in the outcome equation. Furthermore, if established research suggests that certain variables are irrelevant to the outcome under study, they should solely be included if there are other strong reasons for doing so. In sum, the main criterion to judge the success of matching is how well it balances the relevant covariates. This aim is more likely to be obtained if self-selection is weak or the control reservoir is large.

Chapter 4 concentrates on the evaluation of post-secondary education by the method of matching incorporating the findings of the previous two chapters. A somewhat different concept of *return to education* is introduced, namely the *effect* of college education on earnings, which takes account of the effect of education on labor market experience, as well. Its primary aim is estimation of the effect of the associate's, the bachelor's, and graduate degrees on hourly wages for both men and women during the first ten years after they have finished their college education. Moreover, heterogeneity in the effect ruled by family background and inherent ability will be considered. The results are compared to conventional OLS estimation which allows (i) to verify the linear specification of the earnings equation and (ii) to bring out the determinants why matching and OLS estimates differ. Indeed, there is evidence that matching and OLS deviate particularly when heterogeneity in the effect is substantial. For men, the effect of college education on wages seems to depend significantly on ability and parents' education, while, for women, estimates do not support such clear heterogeneity. At the same time, matching and OLS estimates differ less for women than for men.

Basically, the empirical results are along the lines of the existing literature. Estimates of the effect of college education are larger for women. Individuals who obtained their degree more recently experience a higher effect, i.e. there is evidence in favor of a general increase. Moreover, the effect seems to grow gradually over the first ten years after leaving college. Yet, this growth cannot be attributed to a positive interaction between experience and education but partly to a faster accumulation of labor market experience on the part of college graduates.

Apart from evaluating the effects by the method of matching, Chapter 5 examines the functional form assumption of the typical Mincerian human capital earnings equation. A formal framework recently proposed by Card (1995, 1999) is used to take account of endogeneity of the schooling decision. If the return to education depends positively on inherent earnings abilities, individuals with higher abilities tend to opt for more schooling.

If ability is itself rewarded in the labor market but is not controlled for in regression analyses, coefficient estimates of the return to schooling might be upward biased. Griliches (1977) discusses this classical ability bias. However, this chapter shows that not controlling for ability might additionally bias the estimated *functional form* of the earnings equation. Indeed, this bias leads to returns to education that increase with years of schooling acquired, thus rejecting constant returns as generally assumed in the literature (Becker, 1967, Mincer, 1974). Other empirical studies also implicitly report increasing marginal returns to schooling as years of education rise. This suggests that postsecondary education might work as some magic potion.

However, these findings seem to be prompted by endogeneity of schooling as a result of the optimization behavior. Explicit control for ability, especially for an interaction between ability and schooling, shows that, in fact, the return to education diminishes as more schooling is acquired, especially for men. In particular, the results show that the interaction term is mostly statistically significant, i.e. that heterogeneity in the returns is substantial.

Finally, Chapter 6 treats self-selection on unobservables and compares possible solutions to problems raised by this additional dimension in a numerical simulation study. While it is straightforward to tackle selection on observables, selection on unobservables provides a serious intellectual challenge. Researchers have proposed several alternative strategies to overcome this *identification problem*, by invoking *a priori* information on the process of selection into treatment in an observational study (Heckman & Robb, 1985, Angrist & Krueger, 1999) or by designing an appropriate randomized experiment. In the natural sciences, the RCT has become the method of choice for the evaluation of interventions.

While emphasis in methodological work is on the individual level, practical applications frequently concern the case of group-level or community-based interventions. Implementation of policy measures at the community level is often a matter of necessity. Moreover, analysts might choose a community-level approach to evaluation for reasons of costs. Nothing seems more natural as a methodological approach to the evaluation of these interventions than the translation of the RCT paradigm to the community level. Objects of randomized assignment into treatment and control samples are then entire communities. The possible correlation of outcomes within communities, clusters, or groups might seriously distort conclusions regarding the statistical precision of the results.

Although one might be able to collect data on sizeable numbers of individuals within each community participating in the study, the number of communities is typically limited. Thus, while group-randomized experiments produce unbiased estimates it is difficult to enhance precision. Observational studies, by contrast, typically include a respectable number of communities, yet, they might suffer from the selection problem. Possibly, a biased but more precise estimate from an observational study may yield a lower mean squared error than the corresponding estimate of program impact from a group-randomized experiment. In other words, there might be a serious trade-off to consider in the choice of the evaluation strategy.

Chapter 6 investigates the potential and the limits of experimental and non-experimental approaches to the evaluation problem. In particular, it contrasts the use of instrumental variables as a quasi-experimental technique against the particular background of community-based interventions. In the simulations, trade-off between bias and precision is emphasized by imposing a smaller number of communities in a randomized experiment, and by allowing for a correspondingly larger number of communities in all cases where selection into the program is not controlled by the analyst.

Obviously, standard estimators perform well as long as more or less restrictive assumptions on the selection process are satisfied. The randomized experiment – appropriately implemented – always performs well without imposing strong assumptions. However, its small sample size involves disadvantages, especially at group level. Instrumental variable estimation may be a helpful device to circumvent the small sample problem and may open the field for less costly large scale observational studies, provided that a suitable instrument is available. The simulation results suggest that correlations between instrument and endogenous treatment indicator of around 0.3 to 0.4 can be considered to make up a good instrument if the observational study comprises ten times more observations than a randomized experiment.

IV estimation yields inconsistent estimates, though, if treatment effects are heterogeneous and individuals or groups decide whether to undergo treatment upon their true effects. In this case, IV identifies the mean effect of treatment on compliers, the so-called *local average treatment effect* (LATE), see e.g. ANGRIST, IMBENS & RUBIN (1996). In case of a binary instrument, for example, LATE identifies the mean effect of those individuals who opt for participation in accordance with the value of their instrument. That is, they participate if the instrument takes the value 1 and they do not if it is 0. Note that this parameter might also be policy relevant, for instance, in answering the question whether to install additional treatment sites or not, when proximity to treatment site is a valid instrument.

Chapter 2

Matching the Extremes – A Sensitivity Analysis Based on Real Data

May 1999/October 2000

Abstract. This chapter uses observational data to estimate the effect of a bachelor's degree on earnings for men by the method of matching. The data exhibit an extraordinarily large bias in terms of observable confounding variables between treatment and comparison group. Therefore, an appropriate implementation of the matching technique is crucial. Usually, several *ad hoc* decisions have to be made in advance, e.g. decisions on which distance measure or which matching algorithm to use. Sensitivity of the estimation results with respect to some decisions is investigated. In particular, optimal full matching, a greedy full matching, and a greedy pair matching are compared. Furthermore, a simple extension permitting heterogeneous treatment effects is suggested.

2.1 Introduction

Recently, the statistical technique of matching has found widespread attention in econometrics to evaluate effects of policy interventions or welfare programs (Heckman, Ichimura & Todd, 1997, Kluve, Lehmann & Schmidt, 1999, or Lechner, 1999, 2000), or to estimate labor market impacts of military service (Angrist, 1998). Heckman, LaLonde & Smith (1999) provide a comprehensive overview. The technique rests on matching untreated individuals to treated ones with the same (observable) characteristics, thus generating an artificial counterfactual of the treatment group. In effect, this approach attempts at mimicking a randomized experiment using data from an observational study to estimate the mean effect of treatment.

Unlike ordinary least squares estimation (OLS) matching as a non-parametric technique need not rely on functional form or distributional assumptions. What is more, in contrast with OLS, if the treatment effect is heterogeneous the estimated mean effect of treatment – a weighted average of individual effects – builds on a more appropriate weighting scheme than OLS. Angrist & Krueger (1999) or Angrist (1998) show how in case of heterogeneous treatment effects a saturated linear model estimated by OLS weights the individual effects by the individual variances of the treatment indicator. In contrast, matching weights the individual effects by the probability to participate in treatment which is considered a more appropriate weighting scheme.

Often, applied research using propensity score matching has to make many *ad hoc* decisions at various steps of the implementation. For example, decisions have to be made on the concrete definition of a distance measure between treated and untreated units and on which matching algorithm and weighting scheme to use. In an extensive simulation study, Gu & Rosenbaum (1993) examine several alternatives. This chapter uses real observational data from the *National Longitudinal Survey of Youth 1979* to investigate sensitivity of the estimation results with respect to some crucial decisions, specifically decisions on the distance measure and the algorithm to be employed, and, furthermore, on which propensity score estimate to match on.

The empirical example evaluates college education by estimating the effect of the bachelor's degree for men during the first ten years after graduation from college.[1] It turns out that selection into college is extremely strong such that treatment and comparison group are quite distinct. Hence, matching adequate individuals can be expected to be a serious challenge. As a result, the matching algorithm should take account of potential pitfalls which is why three matching algorithms will be explored: *optimal full matching* proposed by Rosenbaum (1991), a *greedy full matching*, and a *greedy pair matching*.

Pair matching produces a stratification composed of non-overlapping pairs of treated and control units. It drops all untreated individuals who are not matched which might reduce efficiency. More importantly, given a certain distance measure some treated might not find a control which would give rise to biased estimates if the loss of treated individuals were systematic and the treatment effect were heterogeneous. In this case, a *full* matching procedure which uses all treated and all untreated units in the sample might be preferred. In a full matching, one control may be matched to more than one treated person and, likewise, one treated may also be matched to numerous controls. The latter event will occur particularly at the low end of the propensity score scale while the first event will mainly happen at the high end. What is more, in a natural way, full matching provides weighting schemes that permit estimation of the *mean effect of treatment on the treated* as well as the *mean effect on a randomly assigned person*.

Dehejia & Wahba (1998) suggest a solution where controls are allowed to be used more than once in a matching algorithm with replacement. However, their strategy generally produces overlapping strata, i.e. certain individuals might be member of more than one stratum. This makes statistical inference more difficult due to dependencies across strata. In contrast, this study adopts the optimal full stratification strategy which produces non-overlapping strata and achieves to effectively minimize the total distance between treated and untreated units. It will be contrasted to a greedy full matching. "Greedy" means that the algorithm does not necessarily attain the minimum. In addition, the framework presented in Rosenbaum (1995) facilitates to estimate variances and to calculate p-values of the estimated treatment effect. This chapter adjusts this

[1] Chapter 4 extends this analysis to other degrees and to both sexes.

framework to the present example and, in giving the statistical model more structure, suggests a simple extension to allow for a special form of heterogeneous treatment effects.

The remainder of the chapter is organized as follows. The next section presents the methodological framework. Section 3 describes the data, the treatment group and the control reservoir. It specifies the sensitivity parameters and, in particular, explains the matching algorithms. Results are discussed in section 4. Section 5 concludes.

2.2 Methodology

Following Rosenbaum (1995), this section starts with outlining the formal setup for the ideal case of a randomized experiment to which the more general case of an observational study can be reduced under certain assumptions. Assume that N units under observation are being stratified into S strata on the basis of their covariates. Let Z_{si} be a dummy variable indicating whether unit i in stratum s, $s = 1, ..., S$, is randomly assigned to treatment ($Z_{si} = 1$) or not ($Z_{si} = 0$). Each stratum s comprises n_s units, $m_s = \sum_{i=1}^{n_s} Z_{si}$ treated and $n_s - m_s$ controls. Since in this study either one treated unit will be matched to one or more controls or one control to more than one treated, m_s will either be 1 or $n_s - 1$. Furthermore, let $\mathbf{Z}_s = (Z_{s1}, ..., Z_{sn_s})'$ and $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_S')'$. Let the random variable R_{si} be the outcome of unit i in stratum s after treatment and $\mathbf{R}$ be the N-tuple of R_{si} arranged in the same order as $\mathbf{Z}$. If unit si exhibits the same value of R_{si} in both states, treatment and control, the treatment has no effect on that unit. This null hypothesis implies that the response of that unit is fixed, denoted r_{si}, and that the only random variable left is $\mathbf{Z}$.

The mean stratum effect Δ_s is estimated as the difference in the mean outcomes of the treated units and their controls in stratum s

$$\hat{\Delta}_s = \frac{1}{m_s}\mathbf{Z}_s'\mathbf{r}_s - \frac{1}{n_s - m_s}(\mathbf{1} - \mathbf{Z}_s)'\mathbf{r}_s = \frac{n_s}{m_s(n_s - m_s)}(\mathbf{Z}_s'\mathbf{r}_s - m_s\bar{r}_s), \tag{2.1}$$

for all $s = 1, ..., S$, where $\mathbf{1}$ is a suitable vector of ones and $\bar{r}_s$ denotes the mean over the r_{si} in stratum s. The overall mean effect τ is a weighted average of the stratum effects

Δ_s, estimated by

$$\hat{\tau} = \sum_{s=1}^{S} \omega_s \hat{\Delta}_s, \tag{2.2}$$

where ω_s are positive stratum weights summing to one: $\sum_{s=1}^{S} \omega_s = 1$. $\hat{\tau}$ identifies the *mean effect of treatment on the treated* if the stratum weights ω_s are proportional to m_s and provided all treated units are being matched or treated units are not systematically discarded by the matching algorithm. $\hat{\tau}$ identifies the *mean effect of treatment on a randomly assigned person* if the stratum weights are proportional to n_s and if all treated and untreated individuals are being matched (full matching)[2]. Estimates of both parameters will be reported in section 4.

The moments under the null hypothesis of no treatment effect are[3]

$$\mathbb{E}\hat{\Delta}_s = 0, \qquad \mathbb{E}\hat{\tau} = 0,$$

where $\mathbb{E}$ denotes the expectation operator,

$$\begin{aligned} \sigma_s^2 = Var(\hat{\Delta}_s) &= \frac{n_s}{(n_s-1)^2} \sum_{i=1}^{n_s} (r_{si} - \bar{r}_s)^2, & (2.3) \\ Var(\hat{\tau}) &= \sum_{s=1}^{S} \omega_s^2 \sigma_s^2. & (2.4) \end{aligned}$$

The stratum differences $\hat{\Delta}_s$ are mutually independent, and their variances differ across strata. Under very mild assumptions Lindeberg's condition is fulfilled and asymptotic normality of $\hat{\tau}$ is established.

Statistical inference will be based on large sample theory exploiting the moments of the relevant test statistics. Alternatively, it could rest on an exact permutation test. Calculating all feasible permutations of zeroes and ones of the vector $\mathbf{Z}$ and counting how often the test statistic of the permuted data exceeds the sample test statistic (2.2) would produce exact p-values. Though, for a large number of strata such a test would exceed current computer power by far.[4]

[2] Sample weights will also be taken into account in order to identify the US population parameters.

[3] Using the distribution of $\mathbf{Z}_s$ yields $Var(\mathbf{Z}_s'\mathbf{r}_s) = \frac{m_s(n_s-m_s)}{n_s(n_s-1)} \sum_{i=1}^{n_s} (r_{si} - \bar{r}_s)^2$. Moreover, note that $m_s = 1$ or $m_s = n_s - 1$ and that r_{si} is no random variable under the null hypothesis.

[4] Good (1994) provides a practical guide to permutation tests and resampling methods in general.

Observational Studies With Overt Bias

In contrast to a randomized experiment, in an observational study the distribution of the assignment vector $\mathbf{Z}$ is unknown because individuals themselves decide whether to participate in treatment or not. If the treatment and control group differ prior to treatment in ways that matter for the outcome under study an observational study is *biased*. An *overt bias* is one that is produced by observable covariates $\mathbf{X}$ and that, in general, can be controlled using adjustments such as matching.

Assuming that there is only overt bias[5] matching on $\mathbf{X}$ mimics *ex post* a randomized experiment in each stratum defined by $\mathbf{X}$. Thus, the formalism for the randomized experiment outlined above can be applied. Alas, whenever $\mathbf{X}$ is of high dimension exact matching will, in all likelihood, be impossible. Alternatively, Rosenbaum & Rubin (1983) suggest to match on the one-dimensional *propensity score*, i.e. the probability to participate in treatment given $\mathbf{X}$, $p(\mathbf{x}) = \mathbb{P}(Z = 1|\mathbf{X} = \mathbf{x})$, where $\mathbb{P}$ denotes probability. They show that if matching on $\mathbf{X}$ removes overt bias matching on $p(\mathbf{X})$ will do so, too.

Unfortunately, the propensity score is unknown and has to be estimated. In this study, this is done by a probit model. Three objections against the estimation might be raised. First, using estimated instead of true propensity scores gives rise to additional error potentially increasing the variance of the treatment effect estimates. Second, exact matching on the propensity score being itself a continuous variable is not feasible either. Its necessary discretization may induce further errors. Third, the special parametric form of the probit model might be misspecified and estimates of the propensity score might thus be inconsistent.[6] Albeit, recalling that balance of the relevant covariates between treatment and control group is exactly the main property of success, these objections will be of minor concern as long as matching on the estimated propensity score achieves acceptable balance. Chapter 3 investigates this issue in a simulation study and conclude that the specification of the selection equation is, in fact, of minor relevance.[7]

[5] This guarantees that the *conditional independence assumption* formulated in Rubin (1977) is fulfilled.

[6] Note, however, that consistency of the coefficients of the probit model is irrelevant as long as $p(\mathbf{X})$ is estimated consistently.

[7] A method circumventing objections in special cases is described in Rosenbaum (1995: 3.5.1) or in Rosenbaum (1984).

Confidence Intervals and a Test for Heterogeneity

Under the null hypothesis of no treatment effect the variance of $\hat{\tau}$ is given in equation (2.4). Yet, if the null hypothesis is rejected (2.4) is not correct anymore. Assuming a constant treatment effect could easily be coped with by just subtracting it from the estimates such that the null hypothesis expresses again a zero effect. In this study, however, the treatment effect may be heterogeneous varying with certain covariates. Therefore, a simple two-step-model is proposed to address this issue. Alternatively, a direct one-step-model is discussed in appendix B. However, it is not used due to an unfavorable weighting scheme.

Assume that the treatment effect differs across strata, but is constant within strata. The response R_{si} of unit i in stratum s is

$$R_{si} = r_{si} + \Delta_s Z_{si}$$

with r_{si} being the outcome when the treatment has no effect. Δ_s is the stratum effect and is estimated according to equation (2.1) replacing $\mathbf{r}_s$ by $\mathbf{R}_s$. Let the stratum effects Δ_s alter with certain covariates

$$\Delta_s = \tau + \alpha(A_s - \bar{A}) + \beta(F_s - \bar{F}) + \gamma(Y_s - \bar{Y}), \tag{2.5}$$

with $\bar{A} = \sum_{s=1}^{S} \omega_s A_s$ and $A_s = \frac{1}{m_s} \sum_{i=1}^{n_s} A_{si} Z_{si}$, and likewise for $\bar{F}, F_s, \bar{Y}, Y_s$. A_{si} denotes inherent earnings abilities of individual i in stratum s, F_{si} characterizes family background, and Y_{si} is the year in which the college degree is obtained. Family background and ability are often considered not only as main determinants of the acquired amount of schooling but also as determinants of the return to education, see e.g. Card (1999) and Willis (1986). Y_{si} intends to capture a possible time trend in the effects. The variables will be specified in section 3 and in appendix A.

Write

$$\begin{aligned}
\delta &= (\tau, \alpha, \beta, \gamma)', \\
H_s &= (1\, ,\, A_s - \bar{A}\, ,\, F_s - \bar{F}\, ,\, Y_s - \bar{Y}), \\
\hat{\boldsymbol{\Delta}} &= (\hat{\Delta}_1, ..., \hat{\Delta}_S)'.
\end{aligned}$$

The variance of $\hat{\mathbf{\Delta}}$ under the null hypothesis $\delta = \delta_0$ is $V(\delta_0) = Var(\hat{\mathbf{\Delta}}) = diag(\sigma_s^2(\delta_0))_{s=1}^S$ with

$$\sigma_s^2(\delta_0) = \frac{n_s}{(n_s-1)^2} \sum_{i=1}^{n_s} [(R_{si} - H_s\delta_0 \cdot Z_{si}) - (\bar{R}_s - H_s\delta_0 \cdot \bar{Z}_s)]^2 \tag{2.6}$$

depending on the null hypothesis $\delta = \delta_0$.[8] For $\delta_0 = 0$, equation (2.6) reduces to (2.3).

Although heteroskedasticity of $\hat{\mathbf{\Delta}}$ might be a reason for using generalized least squares, weighted ordinary least squares with stratum weights ω_s will be used for estimation to keep control over the weighting scheme. Writing $H = (H_1', ..., H_S')'$ and $\hat{\mathbf{\Delta}} = H\delta + \varepsilon$ with an error term $\varepsilon = \hat{\mathbf{\Delta}} - \mathbf{\Delta}$, and a diagonal $S \times S$ weighting matrix $W = diag(\omega_s)_{s=1}^S$, a consistent estimate of δ is $\hat{\delta} = (H'WH)^{-1}H'W\hat{\mathbf{\Delta}}$, with variance under the null $\delta = \delta_0$

$$\tilde{V}(\delta_0) = Var(\hat{\delta}) = (H'WH)^{-1}H'W\,V(\delta_0)\,WH(H'WH)^{-1}.$$

This variance formula permits calculation of standard errors and p-values. For example, a test against the null hypothesis of no mean treatment effect, $\tau = 0$, would make use of $\hat{\tau}/\sqrt{Var(\hat{\tau})} \overset{as.}{\sim} \mathcal{N}(0,1)$. However, a $(1-\alpha)$-confidence region for δ solving $(\hat{\delta} - \delta_0)'\tilde{V}(\delta_0)^{-1}(\hat{\delta}-\delta_0) \leq \chi^2_{4,1-\alpha}$ for δ_0 would be quite cumbersome.[9] Therefore, in the following application, $\tilde{V}(\delta_0)$ will be replaced by the approximation $\tilde{V}(\hat{\delta})$.

2.3 Practical Implementation

Data

The data are taken from the *National Longitudinal Survey of Youth 1979* (NLSY) administered by the US Bureau of Labor Statistics. The NLSY is a sample of 12,686 youths first interviewed in 1979 when they were aged between 14 and 22 and re-interviewed annually until 1994. A detailed description of the data is given by the NLS Handbook

[8] This is due to the fact that $Var(\hat{\Delta}_s) = Var(\hat{\Delta}_s - \Delta_s)$, $\Delta_s = \mathbb{E}\hat{\Delta}_s$, and

$$\begin{aligned} \hat{\Delta}_s - \Delta_s &= \frac{n_s}{n_s-1}(\mathbf{Z}_s'\mathbf{R}_s - m_s\bar{R}_s) - \Delta_s &= \frac{n_s}{n_s-1}(\mathbf{Z}_s'\mathbf{r}_s + \Delta_s\mathbf{Z}_s'\mathbf{Z}_s - m_s\bar{r}_s - m_s\Delta_s\bar{Z}_s) - \Delta_s \\ &= \frac{n_s}{n_s-1}(\mathbf{Z}_s'\mathbf{r}_s - m_s\bar{r}_s). \end{aligned}$$

The variance of $\frac{n_s}{n_s-1}(\mathbf{Z}_s'\mathbf{r}_s - m_s\bar{r}_s)$ is known to be (2.3). Insert $r_{si} = R_{si} - \Delta_s Z_{si}$ to achieve (2.6).

[9] For instance, in the case of $\beta = \gamma = 0$, the left hand side is a polynomial in τ^4 and α^4.

(1997) and the NLSY79 User's Guide (1997). Data on wages are extracted until 1994 for men. Oversampling of Non-whites and economically disadvantaged Whites suggests the use of sample weights pertaining to 1979 in order to identify the population mean effect of treatment on the treated and on a randomly assigned person.

The treatment period is the time to achieve the bachelor's degree, approximately four years at college and maybe some years out of college as well. The treated individuals are those who obtained the degree and left college immediately thereafter, i.e. who have not tried to continue college but eventually dropped out before achieving a higher degree. Controls are drawn from the pool of individuals with only a high school diploma who never attended college. High school dropouts and individuals with a *general educational development* (GED) are removed from the sample.

The year in which a respondent received his high school diploma marks the beginning of the treatment phase of those who went to college. In turn, the year in which he received his bachelor's degree marks the end. A treated and a control person are supposed to finish high school in the same year and at the same age. The control, then, starts to work and gain labor market experience while the treated is allowed to either go to college straight away, interrupt college for a while, or even start to work a certain time before finally attending college. Note that the estimation strategy pursued here does not identify the *return to education* but the *effect of the college degree* on earnings which also includes indirect effects on labor market experience. Chapter 4 discusses the differences of the two concepts and provides empirical evidence that although college degree holders start with less experience, accumulation of experience after college is faster for college than for high school graduates.

The outcome measure is the hourly rate of pay inflated to 1996 dollars using the US consumer price index and transformed into logarithms. For presentation of the results, the estimate $\hat{\tau}$ will be retransformed to $\exp(\hat{\tau}) - 1$. To eliminate outliers, all values below \$1 are set equal to \$1 and maximum or minimum wages of observations whose wages oscillate enormously across years are removed as well.[10] Socioeconomic background variables,

[10] For example, an hourly wage of \$5 in one year, \$1000 in the second, and again \$5 in the third seems more likely to reflect inconsistencies in the calculation of the hourly wage by the NLSY than

Table 2.1: **Distribution of Estimated Propensity and Index Score.**

Estimated Prop. score	Untreated	Treated	Estimated Index score	Untreated	Treated
[0.0 , 0.1)	946	29	[−4.70 , −3.94)	11	0
[0.1 , 0.2)	150	23	[−3.94 , −3.18)	74	0
[0.2 , 0.3)	80	21	[−3.18 , −2.42)	285	2
[0.3 , 0.4)	56	20	[−2.42 , −1.66)	407	9
[0.4 , 0.5)	29	21	[−1.66 , −0.90)	298	37
[0.5 , 0.6)	33	35	[−0.90 , −0.14)	175	54
[0.6 , 0.7)	15	34	[−0.14 , +0.62)	64	96
[0.7 , 0.8)	20	60	[+0.62 , +1.38)	24	139
[0.8 , 0.9)	9	79	[+1.38 , +2.14)	4	86
[0.9 , 1.0]	4	128	[+2.14 , +2.90]	0	27
Mean score	0.11	0.67		-1.77	0.61
Observations	1342	450		1342	450

Comparison of the number of treated and untreated individuals by propensity score and index score intervals.

information about the high school career, and ability measures play an important role in modeling the selection decision to estimate the propensity score. The NLSY provides ten ability measures, the *Armed Services Vocational Aptitude Battery* scores. Since respondents participated in the tests at different ages the scores are adjusted by regressing the raw scores on age dummies and using the residuals subsequently as explanatory variables, analogous to Blackburn & Neumark (1993). Math scores will be used to describe A_s and parents' education to describe F_s in equation (2.5).

The variables and the probit estimation are presented in appendix A. It successfully separates college and high school graduates which, unfortunately, makes matching at the boundaries a difficult project. Note, however, that this aspect does not favor a linear model either because its *ad hoc* linear interpolations between the extremes would not necessarily be correct. Table 2.1 compares the absolute frequencies of treated and untreated men for certain propensity score and index score intervals. The index or *probits* is $\Phi^{-1}(\hat{p})$,

real fundamental economic changes which is why $1000 would be removed. See e.g. the NLSY79 User's Handbook (1997: p. 266): "... the calculation procedure [...] produces, at times, extremely low and extremely high pay rate values."

where Φ is the cumulative normal density function. It is linear in the matching variables $\mathbf{X}$ and might be better suited to reflect the underlying distribution of the estimated propensity scores.

Using $\hat{p}$ might make individuals at the high end of the propensity score scale look more similar than they actually are and, analogously, make individuals at the low end look more identical. Indeed, for low index scores, there are numerous untreated units in cells with hardly any treated while for low $\hat{p}$ there is still a reasonable number of treated units. Matching on the index will drop countless low-score untreated individuals while matching on the propensity score will keep several of them. In the high end of the distribution, the situation is comparable but less pronounced. Results will be discussed for both matching on the propensity score and on the index. For the sake of brevity, "propensity score" will denote both scores in the main text below if closer specification is not necessary.

Distance Measures

A propensity score caliper approach within cells defined by *race*, *age* and *high school graduation year* is pursued, see e.g. Cochran & Rubin (1973). First, the cells are defined. Only individuals of the same race are matched. Furthermore, the age structure is taken into account: individuals of the same age, one year younger or one year older are permitted to be matched. Similarly, only those who receive their high school degree in the same year, one year earlier or later than the treated may become potential controls. This guarantees that untreated individuals within a stratum share a similar economic environment at the beginning of their treatment phase. Exact matches on *age* and *the year of the high school diploma* would be preferable, but would substantially reduce the number of potential controls.

Second, within these cells a pool of potential controls is generated for each treated by excluding all untreated units who exceed a certain propensity or index score caliper ε. The final decision of who becomes an actual control will then be made by minimizing either the *Mahalanobis* or the propensity score distance. The Mahalanobis distance is a weighted Euclidean distance $d(\mathbf{x}_t, \mathbf{x}_c) = (\mathbf{x}_t - \mathbf{x}_c)' V^{-1} (\mathbf{x}_t - \mathbf{x}_c)$, where $\mathbf{x}_t$ and $\mathbf{x}_c$ are the vectors comprising the observable covariates of the treated and the potential control unit, respectively. V is the pooled covariance matrix of these variables which serves to norm the vectors. If the propensity score is inconsistently estimated the Mahalanobis metric within calipers might help circumvent possible problems. In sum, the distance is

$$d(\mathbf{x}_t, \mathbf{x}_c) = \begin{cases} \infty & \text{if} \quad |p(\mathbf{x}_t) - p(\mathbf{x}_c)| > \varepsilon \\ \\ (\mathbf{x}_t - \mathbf{x}_c)' V^{-1} (\mathbf{x}_t - \mathbf{x}_c) & \\ \text{or} \quad |p(\mathbf{x}_t) - p(\mathbf{x}_c)| & \text{else.} \end{cases} \tag{2.7}$$

An infinite distance indicates that matching is forbidden.[11]

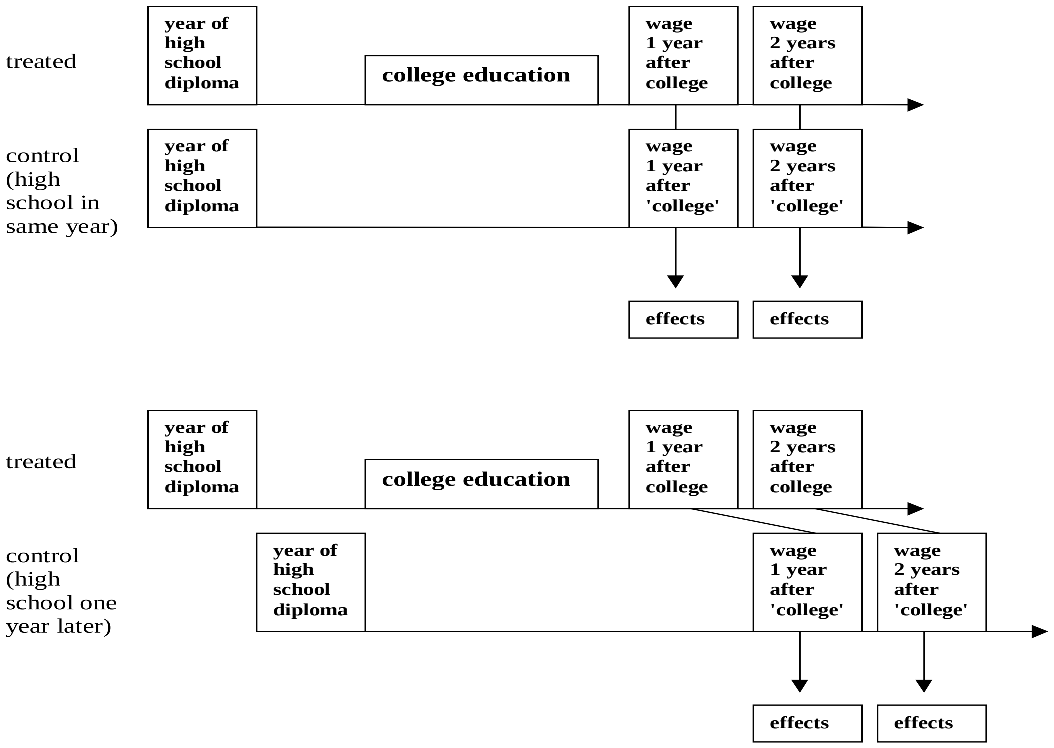
Two different caliper widths ε will be compared, a narrow and a broad one. For propensity score matching, the narrow one will be set equal to 0.05 while the broad one will be 0.10. For index score matching, the respective numbers are 0.30 and 0.60. They are chosen such that both matching on the propensity score and on the index employ an approximately equal number of treated units. Broad calipers allow matching more individuals at the expense of a potentially less favorable balance of covariates. Narrow calipers generate closer similarity of matched units but might have to drop several high- and low-score units. No calipers would have adverse consequences. First, any arbitrarily large distance between treated and controls would then be possible, and, second, matching algorithms would consume substantially more time.[12]

After having constructed the pool of potential controls appropriate wages serving as the counterfactual wages of the treated are assigned. The time span between the year in which the treated unit received his college degree and his high school diploma – the treatment phase – is added to the year in which his potential controls received their high school diploma. The result is considered as the counterfactual year in which his potential controls would have received a college degree. Note that the treatment phase is not necessarily just the years at college because the treated individual might have interrupted education for a while. Figure 2.1 illustrates the procedure. The counterfactual outcome

[11] Matching using the Mahalanobis distance is discussed in Rubin (1980). A comparison of three distance measures is provided in Gu & Rosenbaum (1993). Furthermore, propensity score calipers are discussed in Rosenbaum & Rubin (1985: 3) and Rosenbaum (1989: 3.4).

[12] In each step of the algorithm every treated would have to be compared to the whole control reservoir. Given a caliper, the treated has to be compared to only a small number of suitable untreated units.

Figure 2.1: **Illustration of the Evaluation Procedure.**


treated
year of high school diploma
college education
wage 1 year after college
wage 2 years after college
control (high school in same year)
year of high school diploma
wage 1 year after 'college'
wage 2 years after 'college'
effects
effects
treated
year of high school diploma
college education
wage 1 year after college
wage 2 years after college
control (high school one year later)
year of high school diploma
wage 1 year after 'college'
wage 2 years after 'college'
effects
effects

The first diagram demonstrates the optimal case when treated and control individuals receive their high school diploma in the same year. The second indicates how things change when there is one year difference.

one year after treatment is the wage of the potential control one year after his hypothetical end of college. If wage information is missing the potential control is dropped for that year after treatment but is still used for other years. If the wage of the treated is missing the treated is removed, too. Ten years after college will be examined and each year will be stratified separately such that individuals who are removed in some year due to missing wage information may still be available in other years.

The Matching Algorithms

The final decision regarding the matching procedure is how to implement the chosen matching criteria, in other words, how the distances between treated and controls is minimized. Three algorithms will be compared in this study, one *greedy pair matching* and two *full matching*, *optimal full matching* as proposed by Rosenbaum (1991) and an own *greedy full matching*. Greedy pair matching randomly selects one treated person and chooses – within calipers – the closest untreated as control. Then, the matched pair is removed and a second treated chooses among the remaining control reservoir. The procedure continues until treated units cannot find controls of finite distance anymore. These treated units will then be dropped.

The particular greedy full algorithm used here first shuffles randomly all treated individuals. Then, the first treated is matched to the closest untreated available. This untreated is removed from the control reservoir and the next treated selects the nearest untreated unit. If a treated does not find a control he is taken out for this part of the procedure. After the last treated found his control the first treated starts to search a second control. The algorithm continues until there is no untreated left anymore. By now, some treated have one or more controls and some still none. Those who have none are distributed over the strata consisting of exactly one treated and one control. To this end, the controls of these strata are shuffled randomly and the first control is matched to the closest treated. The next control searches among the remaining treated until there is no one left. Controls may be used more than once.

Although this algorithm attempts to minimize the total distance between treated and their controls it will, in general, not attain the minimum. Rosenbaum (1991) shows in a simple but extreme example how a greedy algorithm might be arbitrarily worse than the optimal. A further unpleasant side effect is that results are different each time the algorithm is used because of the initial random order of records. For illustration, greedy pair and full matching are performed twenty times for each year after college. It turns out that the greedy full algorithm is quite stable and variation over the 20 iterations is negligible. Therefore, standard errors induced by the inherent randomness will only be reported for greedy pair matching.

Optimal full matching circumvents these shortcomings. It attains the overall minimum in that it works backwards and rearranges already matched units if a treated would be better matched to an already matched untreated. In such a case, the first match is broken up and the corresponding treated is again available for matching. Optimal full matching can easily be transformed into a *minimum cost flow problem*[13] (Rosenbaum, 1991).

[13] Bertsekas (1991) discusses *linear network optimization* and provides FORTRAN-algorithms for

In sum, the sensitivity analysis is carried out along five dimensions. First, matching on the propensity and on the index score will be compared. Second, distance within calipers is either defined by the scores or by the Mahalanobis metric. Third, the caliper width is varied. Fourth, the three matching algorithms are compared. Finally, weighting schemes are altered to identify (i) the mean effect of treatment on the treated and (ii) the mean effect on a randomly assigned person. Three measures of success will be discussed: balance of covariates after matching, the variance of the matching estimates, and how systematic treated units are dropped by the algorithms.

2.4 Results

General Remarks

Estimation results of the treatment effects are reported in tables 2.2 to 2.5. For reasons of parsimony, only results for the first, third, fifth, seventh and ninth year after college are shown. The first column of the tables indicates the year after college. Note that the results are not stochastically independent over the years. The first and second columns for the full matching algorithms report estimates of the mean effect of treatment on the treated and on a randomly assigned person, respectively. For greedy pair matching, the two estimates coincide and a supplementary column reports the standard deviations induced by the initial random order of records. They are calculated for the estimates of the effect as well as for its sampling standard errors in parentheses.

The third and fourth columns display the number of strata, of treated, and of untreated individuals used for stratification. For pair matching, all three numbers are identical. The last column of the full matching algorithms reports the mean and maximum number of treated units in strata that consist of more than one treated. Large numbers typically increase the standard errors. Furthermore, note that the number of individuals and strata diminishes continuously from the first to the ninth year. This is because many individuals

minimum cost flow problems. Furthermore, there is an *operations research* procedure called *netflow* in SAS for these kinds of problems. Gu & Rosenbaum (1993) examine the performance of optimal full matching in a simulation study.

Table 2.2: **Estimated Effects. Matching on the p.score, Within caliper: p.score.**

	Optimal Full					Greedy Full					Greedy Pair		
	On the	On R'd	S	T	Mean	On the	On R'd	S	T	Mean		Simul.	S
Year	Treated	Assigned		C	Max	Treated	Assigned		C	Max	Effect	Error	Error
Narrow Caliper													
1	-0.007	-0.026	150	287	4.9	0.028	0.027	152	286	4.0	-0.011	0.021	151
	(0.088)	(0.079)		898	27.0	(0.088)	(0.077)		898	27.0	(0.057)	(0.004)	1.34
3	0.188***	0.211**	149	250	4.0	0.175**	0.156**	149	250	3.2	0.188***	0.028	148
	(0.076)	(0.091)		880	16.0	(0.075)	(0.079)		880	15.0	(0.066)	(0.003)	0.97
5	0.201***	0.019	137	230	4.0	0.234***	0.097	137	229	3.2	0.217***	0.026	138
	(0.072)	(0.083)		893	14.0	(0.073)	(0.083)		893	11.0	(0.077)	(0.004)	1.02
7	0.260***	0.216**	123	197	3.7	0.276***	0.222***	123	197	3.2	0.275***	0.032	123
	(0.091)	(0.100)		789	14.0	(0.086)	(0.089)		789	11.0	(0.081)	(0.004)	0.54
9	0.355***	0.244**	93	151	3.9	0.299***	0.234**	92	150	3.1	0.314***	0.059	92
	(0.130)	(0.117)		578	12.0	(0.113)	(0.104)		578	9.0	(0.119)	(0.010)	0.78
Broad Caliper													
1	-0.020	0.059	159	333	5.2	0.006	0.151**	163	332	4.2	-0.022	0.026	164
	(0.090)	(0.090)		1122	28.0	(0.082)	(0.082)		1122	21.0	(0.054)	(0.004)	1.46
3	0.158	0.253**	158	308	5.1	0.164*	0.204***	163	307	4.1	0.206***	0.031	162
	(0.117)	(0.110)		1116	34.0	(0.101)	(0.086)		1116	24.0	(0.067)	(0.004)	1.26
5	0.179**	0.054	148	285	5.2	0.192***	0.205***	150	283	4.0	0.221***	0.035	152
	(0.089)	(0.090)		1084	32.0	(0.075)	(0.081)		1084	20.0	(0.072)	(0.005)	1.56
7	0.228**	0.219**	130	242	4.5	0.256***	0.269***	134	242	3.8	0.278***	0.044	134
	(0.112)	(0.101)		959	32.0	(0.103)	(0.088)		959	21.0	(0.076)	(0.004)	1.10
9	0.362***	0.271***	99	188	4.7	0.361***	0.297***	101	188	3.9	0.327***	0.057	102
	(0.152)	(0.117)		730	27.0	(0.107)	(0.104)		730	16.0	(0.113)	(0.008)	0.85

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on a randomly assigned person. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 2.3: **Estimated Effects. Matching on the p.score, Within caliper: Mahalanobis.**

	Optimal Full					Greedy Full					Greedy Pair		
	On the	On R'd	S	T	Mean	On the	On R'd	S	T	Mean		Simul.	S
Year	Treated	Assigned		C	Max	Treated	Assigned		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.023	0.025	153	287	4.9	0.019	0.010	151	286	3.9	-0.027	0.019	152
	(0.090)	(0.077)		898	27.0	(0.087)	(0.074)		898	27.0	(0.059)	(0.003)	1.07
3	0.214***	0.209***	151	250	3.9	0.172***	0.151**	151	249	3.5	0.143**	0.020	150
	(0.078)	(0.087)		880	17.0	(0.071)	(0.075)		880	16.0	(0.061)	(0.003)	1.14
5	0.219***	0.119	139	230	3.9	0.225***	0.094	137	229	3.4	0.162**	0.030	138
	(0.075)	(0.090)		893	14.0	(0.075)	(0.082)		893	12.0	(0.073)	(0.004)	1.19
7	0.298***	0.258***	123	197	4.9	0.302***	0.238***	122	196	3.4	0.216***	0.028	123
	(0.093)	(0.097)		789	15.0	(0.090)	(0.090)		789	12.0	(0.074)	(0.005)	0.85
9	0.318***	0.286***	91	151	4.2	0.304***	0.253***	92	151	3.2	0.249***	0.041	92
	(0.121)	(0.118)		578	14.0	(0.104)	(0.109)		578	10.0	(0.097)	(0.007)	0.97
Broad Caliper													
1	0.028	0.232***	167	333	5.4	0.002	0.140*	165	333	4.1	-0.003	0.021	166
	(0.086)	(0.097)		1122	25.0	(0.078)	(0.079)		1122	20.0	(0.056)	(0.006)	1.70
3	0.177*	0.248***	166	308	4.9	0.162*	0.196***	168	308	3.9	0.165***	0.037	166
	(0.111)	(0.099)		1116	29.0	(0.094)	(0.082)		1116	21.0	(0.057)	(0.006)	1.62
5	0.174**	0.243***	156	285	4.9	0.216***	0.196***	158	285	4.0	0.142**	0.025	156
	(0.085)	(0.101)		1084	27.0	(0.075)	(0.081)		1084	14.0	(0.069)	(0.004)	0.95
7	0.261***	0.309***	138	242	5.0	0.275***	0.283***	137	242	3.6	0.232***	0.036	137
	(0.107)	(0.104)		959	27.0	(0.086)	(0.085)		959	17.0	(0.064)	(0.004)	0.78
9	0.372***	0.352***	104	188	4.7	0.324***	0.292***	103	187	3.8	0.263***	0.043	103
	(0.142)	(0.125)		730	22.0	(0.114)	(0.104)		730	13.0	(0.090)	(0.007)	1.42

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on a randomly assigned person. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 2.4: **Estimated Effects. Matching on the index, Within caliper: index.**

	Optimal Full					Greedy Full					Greedy Pair		
	On the	On R'd	S	T	Mean	On the	On R'd	S	T	Mean		Simul.	S
Year	Treated	Assigned		C	Max	Treated	Assigned		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.001	0.010	159	300	4.5	0.033	0.024	166	299	3.8	-0.008	0.017	165
	(0.079)	(0.062)		669	19.0	(0.088)	(0.062)		669	19.0	(0.055)	(0.005)	1.65
3	0.199***	0.205***	160	268	4.1	0.230***	0.186***	167	268	3.2	0.232***	0.030	166
	(0.076)	(0.073)		692	18.0	(0.071)	(0.067)		692	15.0	(0.067)	(0.006)	1.67
5	0.202***	0.093	148	249	4.4	0.205***	0.098	151	249	3.5	0.216***	0.033	153
	(0.072)	(0.071)		712	15.0	(0.070)	(0.070)		712	12.0	(0.072)	(0.004)	1.48
7	0.284***	0.249***	130	214	3.8	0.278***	0.246***	135	214	3.3	0.264***	0.034	136
	(0.090)	(0.079)		635	15.0	(0.079)	(0.078)		635	12.0	(0.075)	(0.004)	1.09
9	0.354***	0.239***	99	165	3.9	0.339***	0.262***	102	165	3.0	0.310***	0.040	102
	(0.127)	(0.101)		465	14.0	(0.104)	(0.097)		465	11.0	(0.111)	(0.007)	1.06
Broad Caliper													
1	-0.017	0.019	167	340	5.1	0.077	0.053	189	339	3.9	0.021	0.025	190
	(0.087)	(0.072)		878	25.0	(0.079)	(0.064)		878	18.0	(0.053)	(0.003)	1.31
3	0.162	0.200**	164	322	5.2	0.192**	0.174***	185	322	3.7	0.235***	0.029	184
	(0.116)	(0.087)		865	34.0	(0.090)	(0.068)		865	18.0	(0.064)	(0.006)	2.26
5	0.192**	0.072	153	301	5.4	0.205***	0.113*	173	300	3.8	0.236***	0.024	175
	(0.092)	(0.078)		874	32.0	(0.063)	(0.069)		874	13.0	(0.066)	(0.003)	2.92
7	0.248**	0.245***	134	258	4.9	0.310***	0.276***	151	257	3.4	0.290***	0.044	153
	(0.111)	(0.091)		788	31.0	(0.088)	(0.082)		788	12.0	(0.071)	(0.004)	1.80
9	0.356***	0.253***	104	195	4.8	0.366***	0.274***	118	194	3.2	0.341***	0.051	116
	(0.152)	(0.105)		621	27.0	(0.109)	(0.093)		621	9.0	(0.107)	(0.008)	1.55

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on a randomly assigned person. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 2.5: **Estimated Effects. Matching on the index, Within caliper: Mahalanobis.**

	Optimal Full					Greedy Full					Greedy Pair		
	On the	On R'd	S	T	Mean	On the	On R'd	S	T	Mean		Simul.	S
Year	Treated	Assigned		C	Max	Treated	Assigned		C	Max	Effect	Error	Error
Narrow Caliper													
1	0.046	0.035	173	300	4.8	0.027	0.024	172	300	3.7	-0.014	0.030	172
	(0.083)	(0.063)		669	19.0	(0.081)	(0.060)		669	19.0	(0.054)	(0.005)	1.89
3	0.212***	0.169***	173	268	4.3	0.178***	0.159***	171	268	3.2	0.154**	0.032	172
	(0.074)	(0.067)		692	18.0	(0.067)	(0.062)		692	16.0	(0.058)	(0.004)	1.40
5	0.180***	0.133*	158	249	4.1	0.177***	0.093	158	249	3.2	0.126*	0.027	157
	(0.069)	(0.075)		712	14.0	(0.061)	(0.067)		712	10.0	(0.065)	(0.004)	1.62
7	0.290***	0.257***	141	214	4.3	0.280***	0.244***	140	214	3.1	0.215***	0.037	140
	(0.087)	(0.081)		635	15.0	(0.076)	(0.076)		635	9.0	(0.066)	(0.006)	1.24
9	0.312***	0.282***	106	165	4.0	0.280***	0.252***	104	165	3.1	0.220**	0.042	105
	(0.109)	(0.103)		465	14.0	(0.100)	(0.096)		465	11.0	(0.087)	(0.007)	0.92
Broad Caliper													
1	0.036	0.055	195	340	5.1	0.005	0.040	207	340	3.5	0.023	0.027	202
	(0.074)	(0.068)		878	19.0	(0.068)	(0.062)		878	15.0	(0.053)	(0.003)	2.15
3	0.165*	0.147**	193	322	5.0	0.168**	0.154**	195	322	3.5	0.151***	0.032	198
	(0.092)	(0.072)		865	19.0	(0.087)	(0.065)		865	16.0	(0.052)	(0.003)	1.78
5	0.211***	0.141*	179	301	4.6	0.207***	0.130**	186	301	3.2	0.188***	0.028	185
	(0.075)	(0.079)		874	18.0	(0.059)	(0.068)		874	10.0	(0.062)	(0.003)	1.87
7	0.316***	0.295***	158	258	4.8	0.260***	0.259***	163	254	3.1	0.228***	0.037	162
	(0.098)	(0.091)		788	16.0	(0.072)	(0.077)		788	9.0	(0.066)	(0.006)	1.84
9	0.302***	0.264***	121	195	4.5	0.358***	0.276***	125	195	3.1	0.199**	0.034	123
	(0.118)	(0.104)		621	13.0	(0.095)	(0.093)		621	7.0	(0.089)	(0.007)	1.44

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. The weighting schemes take account of the NLSY sample weights. The first columns of the full matching algorithms show estimates of the mean effect of treatment on the treated while the second show the mean effect on a randomly assigned person. For greedy pair matching, simulation standard deviations are additionally reported in an own column. Columns denoted by S, T, and C display the number of strata, of treated, and of control units, respectively. For pair matching all three numbers are equal, simulation standard deviations for S are reported. Finally, columns titled "Mean" and "Max" show the mean and maximum number of treated units in strata that comprise more than one treated, respectively.

Table 2.6: **Balance of Covariates, Aggregate Measures.**

			Optimal Full					Greedy Full					Greedy Pair				
Match on	Within Caliper	Caliper Width	Mean	-1	Math	Educ	$\Delta\hat{p}$	Mean	-1	Math	Educ	$\Delta\hat{p}$	Mean	-1	Math	Educ	$\Delta\hat{p}$
p.score	p.score	narrow	82	90	98	89	-.09	84	89	98	88	-.09	85	90	97	93	-.24
p.score	p.score	broad	76	89	98	85	-.05	78	87	95	81	-.05	87	91	98	91	-.22
p.score	Mahal	narrow	81	88	96	85	-.09	84	88	98	87	-.09	86	90	97	91	-.24
p.score	Mahal	broad	74	87	98	77	-.05	79	87	94	79	-.05	87	90	97	87	-.22
index	p.score	narrow	81	92	98	91	-.09	85	91	98	88	-.09	87	91	98	92	-.23
index	p.score	broad	75	90	98	84	-.04	83	90	91	79	-.04	89	91	97	87	-.20
index	Mahal	narrow	80	89	98	91	-.09	86	91	97	95	-.09	88	91	98	93	-.22
index	Mahal	broad	80	88	93	77	-.04	85	88	88	77	-.04	85	87	93	85	-.18

The first three columns specify the sensitivity parameters. The first column of each matching algorithm represents mean overall percent bias reduction, the second is the mean reduction when the variable *born in south* is disregarded. The third and fourth display bias reduction in math scores and in parents' education, respectively. The fifth column reports the difference in mean propensity scores between treated units before and after matching.

are not in the sample for the whole nine-year period after college. In 1994, the last year in the panel, some individuals – especially younger ones – are just in their, say, seventh year after college.

The estimates of all tables indicate a clear upward trend in the effects. While the effect of the bachelor's degree (in short BA) on BA holders in the first year after college is not significantly different from zero, it rises up to 35% in the ninth year. Chapter 4 underscores that part of the increase can be explained by the fact that college graduates accumulate experience more quickly after leaving college than high school graduates. However, interaction between labor market experience and schooling does not appear to be existent.

Estimation results for (α, β, γ) are omitted since they would occupy too much space without gaining further insight. The results can easily be summarized as follows. $\hat{\alpha}$ and $\hat{\gamma}$ are almost always positive in the ten years after college, $\hat{\beta}$ oscillates around zero. Nonetheless, never are they statistically significant. Chapter 4 pools all ten years and finds statistical significance for $\hat{\alpha}$ and $\hat{\gamma}$, i.e. math test scores seem to have some positive impact on the effect of a bachelor's degree and individuals who receive their degree more recently experience a higher effect. In contrast, parents' education appears to be negligible.

Table 2.6 is dedicated to the balancing properties of the matching algorithms. Since there are numerous variables and ten years after college, i.e. ten stratifications, some aggregate measures of balance are introduced to facilitate assessment. Tables in appendix C report detailed results. First, an average over all percent bias reductions in each variable and for each year after college is calculated. Then, the weighted average over all ten years is reported under the heading "Mean". The weights correspond to the number of strata in each year. As can be seen in the appendix tables, the matching algorithms, particularly the full ones, face severe problems in balancing the variable *born in south*, actually, they even tend to worsen its balance. Therefore, the columns headed "-1" report the average percent bias reduction disregarding *born in south*.

The third and fourth columns for each algorithm show mean percent bias reductions for the presumably most important single variables *math scores* and *parents' education*. *Math* *scores* exhibit the highest t-value in the probit estimation (appendix A). Also, they are important determinants of wages as documented in other studies (Blackburn & Neumark, 1993, or Murnane, Willett & Levy, 1995). *Parents' education* exhibits the second largest t-value. Finally, the last columns headed $\Delta\hat{p}$ display the propensity score difference between treated individuals before and after matching. A negative sign points to a systematic loss of treated units in the high end of the propensity score distribution and, thus, to a possible bias in the estimates if the treatment effect is heterogeneous.

Specific Comparisons

Optimal Vs. Greedy Full. The greedy full algorithm as constructed in this study achieves to produce a more favorable, i.e. a more uniform, stratification. This is expressed by the mean and maximum number of units in strata consisting of more than one treated which is smaller for greedy matching. The number of strata is slightly larger in the greedy case, especially when calipers are broad. This pattern is more pronounced for index score matching. Though, the estimates are not very distinct. As noted in Gu & Rosenbaum (1993), this might be because greedy and full use the same individuals even though the specific stratification differs. As a result, a good greedy algorithm need not be inferior to the optimal one.[14]

Surprisingly, overall balance is somewhat superior for greedy full matching, too. The main reason is that the optimal one faces severe problems in balancing the variable *born in south*. Disregarding this variable, balancing success is more or less equal.[15] This finding is in line with Gu & Rosenbaum (1993) who observe that when it comes to balance, optimal matching seems to have no advantage over greedy matching. Yet, notice that optimal matching tends to better balance *math scores*.

Full Vs. Pair Matching. Greedy pair matching is performed twenty times. The

[14] Alas, the greedy algorithm as programmed by the author consumes considerably more time than the optimal – a factor between 50 and 100.

[15] A weakly significant interaction between *parents' education* and *born in south* has been included in the probit estimation, but improvements were not attained; other interactions were statistically insignificant. Moreover, exact matching on *born in south* reduced the matched sample size to roughly 80%, though, the number of strata did not diminish much; estimates of the treatment effects increased slightly.

averages and standard deviations over all twenty repetitions are reported in the tables. It produces approximately the same number of strata as greedy full matching, i.e. the effective sample size is constant across algorithms, nevertheless, standard errors are smaller for pair matching. This is because, in contrast with full matching, stratification is most uniform.

In case of pair matching, there is no reason to distinguish between the two treatment effect parameters for two reasons. First, there is only one weighting scheme for pair matching and, second, since the majority of treated and untreated units are not matched, on *a priori* grounds, identification of the respective population parameters is doubtful anyway. These doubts are substantiated when one considers $\Delta\hat{p}$ in table 2.6. As expected, on the one hand, pair matching produces the most favorable balance but, on the other, the loss of treated units with high propensity scores is dramatic.

This systematic loss may well be a reason why pair matching estimates are generally lower than the full matching estimates of the effect on the treated. Though, the differences are small and statistically insignificant. Almost exclusively for Mahalanobis-within-caliper distance are pair matching estimates lower than on-the-treated-effect estimates and sometimes lower than the on-a-randomly-assigned-effect estimates. Thus, the results do not point to strong heterogeneity in the treatment effects. Similarly, coefficient estimates of α – as noted above – are positive in all ten years but almost never significantly so. In such a case pair matching seems to be a superior strategy. However, large variation of the results caused by the inherent randomness of the greedy algorithm should be overcome by using an optimal pair matching approach. Alternatively, restrictions on the stratum sizes might be imposed on full matching approaches in order to achieve a more uniform stratification.

Effect on the Treated Vs. Effect on a Randomly Assigned Person. The effect on a randomly assigned person appears to be lower than the effect on the treated in almost all specifications. The difference is never statistically significant and might therefore be interpreted as only weak evidence in favor of heterogeneous effects. Yet, the results do not contradict the hypothesis that individuals opt for higher education taking account of their expected gains from education *inter alia.* Results of the effect on a randomly assigned person are more or less of the same magnitude as results of the greedy *pair* matching. When Mahalanobis is within-caliper distance greedy pair matching estimates are, on average, lower; when the propensity score distance is chosen they are higher.

Narrow Vs. Broad Calipers. Broad calipers produce more strata because less treated and untreated units have to be dropped. The difference in the number of strata is more pronounced when the Mahalanobis distance is used within calipers. Nonetheless, estimates do not differ systematically and standard errors are not lower for the broad calipers case because the larger amount of strata is offset by a substantially reduced uniformity across strata, especially for optimal full matching. It is not offset for pair matching where standard errors do decrease.

For the full matching algorithms, percent bias reduction is larger for narrow than for broad calipers. For pair matching, the discrepancy is negligible. However, once *born in south* is disregarded, narrow and broad calipers produce an overall balance close to equal. A clear distinction can be made with respect to $\Delta\hat{p}$; narrow calipers put more obstacles on high-score treated individuals in finding an adequate control which is why $\Delta\hat{p}$ is more negative.

Within-Caliper Distance: Score Vs. Mahalanobis. While estimation results of the effect on the treated do not differ much, estimates of the effect on a randomly assigned person are higher using the Mahalanobis distance. The reason for this divergence is unclear. Moreover, the Mahalanobis case tends to supply more strata, though based on the same number of treated and untreated units. This observation is especially evident for index score matching. As a result, standard errors tend to be lower in the Mahalanobis case.

Matching on the Propensity Score Vs. on the Index. The most striking difference is the number of controls used for stratification. Index score matching drops numerous untreated units consistent with table 2.1. For instance, in the first year, index matching utilizes over 200 controls less than propensity score matching. Because of that, it is dubious whether index matching really identifies the effect on a randomly assigned person. Though, a clear distinction between estimates can hardly be established except for the fact that standard errors of the randomly-assigned-effect estimates are lower for index matching. The reason is that the mean number of controls in low-score strata, which typically consist of numerous controls, is smaller for index than for propensity score matching. With regard to balance there seems to be no discrepancy noteworthy.

2.5 Discussion and Conclusion

This chapter addresses the sensitivity with respect to various decisions that have to be made in practical implementations of the method of matching. Observational data from the NLSY79 are employed for illustration. The treatment group comprises individuals who obtained a bachelor's degree while controls are drawn from the pool of individuals with only a high school diploma. It turns out that selection into college is extremely strong. Thus, bias in the relevant covariates prior to treatment is unusually large and matching becomes a serious challenge.

Sensitivity of the decision parameters as to the estimated treatment effects appears to be rather modest. Systematic variation in the estimates caused by variation of the distance measures between treated and untreated units or by altering matching algorithms is minor and statistically insignificant. Therefore, one can generally conclude that the effect of a bachelor's degree on BA holders is fairly low immediately after leaving college but rises during the first ten years after college completion. In the ninth year it approaches 30%. Roughly 80% of the initial bias in the observable covariates is removed by full matching algorithms and 87% by pair matching. However, the latter produces a matched sample which excludes many high-score treated individuals.

A distinction is made between the effect of treatment on the treated, i.e. the BA holder, and the effect on a randomly assigned person. Identification of these two parameters is a matter of applying the appropriate weighting scheme when averaging over single stratum effects. The two parameters might differ if treatment effects are heterogeneous. Results suggest that the mean effect on a treated person is somewhat larger than that on an average person. Yet, since the deviations are statistically imprecise there is only weak evidence in favor of systematic heterogeneity. A full matching of all individuals automatically delivers appropriate weighting schemes while additional distributional information would be necessary for pair matching. Therefore, no distinction is made for the latter. Pair matching estimates tend to be, on average, slightly lower than estimates of the mean effect on the treated. Although results are again very imprecise, they are in line with the overall picture. Pair matching drops countless high-propensity-score treated individuals and, therefore, yields estimates closer to the randomly-assigned-person estimates which put more weight on low-score strata.

Alas, heterogeneity is too weak to unanimously favor full matching since its disadvantages clearly emerge. Full matching estimates are accompanied by relatively large standard errors because the full stratification is far from being as uniform as pair matching. For example, the more strata consist of a large amount of treated units sharing only one control the higher estimated standard errors of the on-the-treated-estimates are. Surprisingly, the greedy full matching algorithm as proposed in this study achieves to produce more uniformity across strata than its optimal counterpart.

However, greedy algorithms, specifically greedy pair, yield estimation results that depend on the random initial order of records. This unpleasant disadvantage can easily be overcome by using an optimal algorithm. The superior uniformity of the special greedy full algorithm used here might also be copied by an optimal procedure when restrictions on the maximum number of units within each stratum would be imposed or when the caliper width would be reduced. See also Ming & Rosenbaum (2000) for a related discussion. Further note that using the Mahalanobis distance within calipers generates more strata from the same number of units than using the propensity score distance, in other words, the first constructs a more uniform stratification.

Furthermore, matching on the linear index score might be preferred since it drops numerous untreated units at the low end of the propensity score scale who would all be used in matching on the propensity score. Dropping them helps generate a more uniform stratification with respect to the estimation of the randomly-assigned-effect. For index matching, however, the distribution of matched controls across strata might not be representative of the initial distribution in the comparison group anymore.

In sum, in finite samples with strong selection into treatment and substantial heterogeneity, there is a trade-off between bias and variance. To remove bias in covariates between treatment and control group and, specifically, to maintain similarity of the matched sample with the initial population, strata tend to be less uniform, thus increasing the variance. On the other hand, a uniform stratification, though, is accompanied by a considerable reduction of the sample size, so bias might be severe. In contrast, from an asymptotic point of view, removing bias would be the strictly recommended strategy. In this study, pair matching has done a good job because heterogeneity does not seem to be very important. Consequently, heterogeneity should be checked in empirical applications with strong selection in order to be able to decide among certain matching algorithms.

Appendix A: The Probit Estimations

Appendix A discusses the estimation of the propensity score by a probit model. Table 2.7 displays the results. The model includes several covariates that reflect socioeconomic background and variables characterizing the high school career. Furthermore, it comprises two ability variables: scores on *math* and *auto and shop information* tests (adjusted for age). The first tend to capture academic while the second tend to capture non-academic abilities. See also the classification in Blackburn & Neumark (1995). Two variables are generated in the following way. *Parents' education* is the mean of the father's and mother's education, it is the mother's if the father's is missing and vice versa. The variable *parents' occupational status* is a binary variable indicating the social status of parents' occupation – high or low – which is the mean of the mother's and father's status. It is only the father's if the mother's is missing and vice versa.

Table 2.7: **Probit Estimation Results.**

Variables	Mean	Coeff.	t-value	P-value
Black	0.263	0.274	1.971	0.049
Hispanic	0.091	0.256	1.443	0.149
Math test scores	-0.442	0.098	15.384	0.000
Auto and shop test scores	4.911	-0.018	-2.857	0.004
Attended private school	0.052	0.432	2.397	0.017
Ever expelled or suspended from school	0.272	-0.536	-4.314	0.000
High school curriculum: college preparatory	0.288	0.972	6.392	0.000
High school curriculum: general program	0.509	0.358	2.439	0.015
Parents' education	11.185	0.154	6.857	0.000
Parents' occup. status high when resp. was 14	0.129	0.432	2.361	0.018
Number of siblings	3.600	-0.065	-2.796	0.005
Born in the south	0.365	0.346	3.333	0.001
Constant	1.000	-3.142	-9.750	0.000
Observations	1792			
$\chi^2(12)$	1046.4			
Overall p-value	0.000			
Pseudo R^2	0.518			

All variables with "yes/no" answers are dummy variables with 1 for "yes" and 0 for "no".

Except for *Hispanic* all variables are statistically significant at conventional levels. *Family income* is not included because (a) there are countless missing observations and (b) it is not significant at a 70%-level. This surprising result might be explained by the fact that other socioeconomic background variables seem to capture already the effect of family income. Furthermore, note that the more variables are included, especially insignificant ones, the more missing observations occur which should be avoided in a data-hungry non-parametric technique such as matching.

Apparently, selection into college is fairly strong confirmed by other studies, too. Ashenfelter & Rouse (1998a) report that (observed and unobserved) family background explains about 60% of the variance in schooling attainment and Murnane, Willett & Levy (1995) assert that math test scores are a strong predictor of subsequent educational attainment.

Appendix B: An Alternative "One-Step-Model"

Estimation of the parameter vector δ as outlined in this chapter is a two-step approach. This appendix concentrates on the mean effect of treatment on the treated and presents a one-step approach to be compared to the two-step approach. It clarifies why the latter is preferred in this chapter.

The Two-Step Approach

In the first step, the stratum effects Δ_s are estimated for all s according to the following equation

$$R_{si} = r_{si} + \Delta_s Z_{si}. \tag{2.8}$$

r_{si} is the outcome of individual i in stratum s if there is no treatment effect. It can be written as the sum of a stratum effect r_s and an individual effect $\tilde{r}_{si}$ with $\sum_{i=1}^{n_s} \tilde{r}_{si} = 0$, thus, $r_{si} = r_s + \tilde{r}_{si}$.

Introduce some useful notation. First, define a *stratum-to-individual-transformation-*

matrix Γ

$$\Gamma = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ & & & & & & \ddots & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix}'$$

of format $N \times S$ that contains 1's in the sth column if individual i, i.e. row i, belongs to stratum s and 0's otherwise. Furthermore, define $S \times 1$-vectors $\mathbf{r}$ and $\mathbf{\Delta}$ consisting of r_s and Δ_s, respectively. Let $\mathbf{R}$ be an $N \times 1$-vector of all individual R_{si} ordered by strata, likewise, $\tilde{\mathbf{r}}$ be the $N \times 1$-vector of $\tilde{r}_{si}$, and, finally, let Z be an $N \times N$-diagonal matrix with diagonal elements Z_{si} ordered by strata. Then, equation (2.8) can be rewritten as

$$\mathbf{R} = \Gamma \mathbf{r} + Z \Gamma \mathbf{\Delta} + \tilde{\mathbf{r}},$$

and

$$\hat{\mathbf{\Delta}} = (\Gamma' Z M_\Gamma Z \Gamma)^{-1} \Gamma' Z M_\Gamma \mathbf{R},$$

with M_Γ being the "residual maker" in each stratum: $M_\Gamma = I_N - \Gamma(\Gamma'\Gamma)^{-1}\Gamma'$, I_N being the $N \times N$-identity matrix. Define $Q = (\Gamma' Z M_\Gamma Z \Gamma)$ which turns out to be an $S \times S$-diagonal matrix with diagonal elements $\frac{m_s(n_s - m_s)}{n_s}$. It can be shown that $\hat{\Delta}_s = \frac{n_s}{m_s(n_s - m_s)}(\mathbf{Z}_s' \mathbf{R}_s - m_s \bar{R}_s)$ reproducing equation (2.1).

In the second step, the estimated stratum effects are regressed on (A_s, F_s, Y_s), and a constant to obtain an estimate for δ. Let (A_s, F_s, Y_s) be measured as deviations from their overall means. The regression is weighted by stratum weights $\omega_s = m_s$, ignoring the NLSY sample weights. Further let W be the $S \times S$-diagonal matrix of weights and H be the $S \times 4$-matrix $(\mathbf{1} \quad \mathbf{A} \quad \mathbf{F} \quad \mathbf{Y})$, then δ is estimated by

$$\hat{\delta} = (H'WH)^{-1} H'W \hat{\mathbf{\Delta}}.$$

The One-Step Approach

A formulation that incorporates all steps in one leads to the following equation

$$R_{si} = r_s + (\tau + \alpha A_s + \beta F_s + \gamma Y_s) Z_{si} + \tilde{r}_{si},$$

alternatively,

$$\mathbf{R} = \Gamma\mathbf{r} + Z\Gamma H\delta + \tilde{\mathbf{r}}.$$

An estimate for δ is

$$\begin{aligned}\hat{\delta}_1 &= (H'\Gamma'ZM_\Gamma Z\Gamma H)^{-1}H'\Gamma'ZM_\Gamma\mathbf{R}\\ &= (H'QH)^{-1}H'Q\hat{\boldsymbol{\Delta}} \qquad \neq \qquad \hat{\hat{\delta}}\end{aligned}$$

Consistency of the estimates. Since $\hat{\hat{\delta}} = (H'WH)^{-1}H'WQ^{-1}\Gamma'ZM_\Gamma(Z\Gamma H\delta + \tilde{\mathbf{r}}) = \delta + (H'WH)^{-1}H'WQ^{-1}\Gamma'Z\tilde{\mathbf{r}}$, $\hat{\hat{\delta}}$ is consistent if $\frac{1}{S}(H'WH)$ does not vanish as S tends to infinity and if $\frac{1}{S}H'WQ^{-1}\Gamma'Z\tilde{\mathbf{r}}$ tends to zero. WQ^{-1} is an $S\times S$-diagonal matrix with elements $\frac{1}{1-m_s/n_s}$ which remain finite provided the relation between treated and control individuals remains finite. Likewise, $\hat{\delta}_1 = (H'QH)^{-1}H'\Gamma'ZM_\Gamma(Z\Gamma H\delta + \tilde{\mathbf{r}}) = \delta + (H'QH)^{-1}H'\Gamma'Z\tilde{\mathbf{r}}$; $\hat{\delta}_1$ is consistent if $\frac{1}{S}(H'QH)$ does not vanish and $\frac{1}{S}H'\Gamma'Z\tilde{\mathbf{r}}$ tends to zero as $S \to \infty$.

Variances. Consider $\hat{\hat{\delta}} - \delta = (H'WH)^{-1}H'WQ^{-1}\Gamma'Z\tilde{\mathbf{r}}$. Note that $\Gamma'Z\tilde{\mathbf{r}}$ is an $S\times 1$-vector of the elements $\sum_{i=1}^{n_s} Z_{si}(r_{si}-r_s)$ and $Var\left(\sum_{i=1}^{n_s} Z_{si}r_{si}\right) = \frac{m_s(n_s-m_s)}{n_s(n_s-1)}\sum_{i=1}^{n_s}(r_{si}-\bar{r}_s)^2$. Thus,

$$Var(\hat{\hat{\delta}}) = (H'WH)^{-1}H'WV(\delta)WH(H'WH)^{-1},$$

with the $S\times S$-diagonal matrix $V(\delta)$ and diagonal elements $\frac{n_s}{m_s(n_s-m_s)(n_s-1)}\sum_{i=1}^{n_s}(r_{si}-r_s)^2$ as already mentioned in equation (2.6). Analogously, the variance of $\hat{\delta}_1$ turns out

$$Var(\hat{\delta}_1) = (H'QH)^{-1}H'QV(\delta)QH(H'QH)^{-1}.$$

Why the Two-Step Estimator Is Preferred? While $\hat{\hat{\delta}}$ weights the strata by the number of its treated m_s, $\hat{\delta}_1$ weights each stratum by $\frac{m_s(n_s-m_s)}{n_s}$. If m_s is either 1 or $n_s - 1$, the weights become $1 - \frac{1}{n_s}$ which increase with the number of individuals n_s in stratum s irrespective of how many treated units there are. However, when focus is on the mean effect of treatment on the treated each stratum should be weighted according to the number of its treated units, consequently, the two-step procedure is to be preferred.

Appendix C: Detailed Balance of Covariates

Tables 2.8 to 2.11 display the balancing properties for all covariates and for all eight sensitivity specifications. They show the means of covariates by treatment status before and after matching. The latter are weighted averages over all stratifications of the ten years after college. The weights correspond to the number of strata in each year. The means are compared by a conventional t-test under the assumption of equal variances in both groups. A “1” indicates that the means are not significantly different. Fractions are due to averaging. Moreover, the percent bias reduction is shown for each variable and as an average over all variables. Since the full matching algorithms face severe problems in balancing the variable *born in south*, the last row displays the average over all variables when it is excluded.

Table 2.8: **Balance of Covariates: Matching on the p.score, Within calipers: p.score.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.57	0.58	1.00	99	0.43	0.43	1.00	100
Index score	-1.77	0.61	0	0.17	0.21	1.00	99	0.16	0.21	1.00	98	-0.28	-0.27	1.00	100
Black	0.30	0.16	0	0.15	0.15	1.00	100	0.15	0.15	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.06	0.06	1.00	100
Age	17.50	17.65	1	17.89	17.85	1.00	70	17.90	17.85	1.00	63	17.82	17.79	1.00	75
Year of high school diploma	79.37	78.74	0	78.57	78.54	1.00	94	78.56	78.54	1.00	95	78.64	78.66	1.00	95
Math test scores	-3.95	10.02	0	8.35	8.36	1.00	98	8.20	8.35	1.00	98	5.25	5.02	1.00	97
Auto+shop test scores	3.88	7.98	0	7.95	7.78	1.00	94	8.02	7.79	1.00	94	7.62	6.54	1.00	73
Attended private school	0.03	0.12	0	0.11	0.09	1.00	80	0.11	0.09	1.00	79	0.08	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.14	0.11	0.71	88	0.16	0.11	0.71	82	0.15	0.14	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.60	0.57	1.00	93	0.62	0.57	1.00	91	0.49	0.46	1.00	94
Curriculum: general	0.59	0.28	0	0.33	0.37	1.00	86	0.31	0.37	1.00	82	0.41	0.44	1.00	88
Highest grades of parents	10.50	13.21	0	12.44	12.59	1.00	89	12.40	12.60	0.55	88	11.93	12.03	1.00	93
Occupation parents high	0.08	0.29	0	0.24	0.22	1.00	94	0.24	0.23	1.00	92	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.80	2.75	1.00	90	2.79	2.75	1.00	91	2.93	2.82	1.00	88
Born in south	0.38	0.33	1	0.25	0.30	0.88	-27	0.27	0.30	1.00	18	0.31	0.34	1.00	17
Mean percent bias reduction							82				84				85
– *born in south* excluded							90				89				90
Broad Caliper															
Propensity score	0.11	0.67	0	0.61	0.62	1.00	98	0.59	0.62	1.00	95	0.43	0.45	1.00	98
Index score	-1.77	0.61	0	0.27	0.37	1.00	96	0.20	0.37	0.31	93	-0.25	-0.22	1.00	98
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.91	17.89	1.00	73	17.97	17.89	1.00	45	17.83	17.82	1.00	79
Year of high school diploma	79.37	78.74	0	78.57	78.50	1.00	89	78.53	78.50	1.00	93	78.63	78.63	1.00	94
Math test scores	-3.95	10.02	0	8.97	9.04	1.00	98	8.34	9.05	0.95	95	5.33	5.24	1.00	98
Auto+shop test scores	3.88	7.98	0	7.14	7.63	1.00	88	7.20	7.65	1.00	89	7.34	6.44	1.00	78
Attended private school	0.03	0.12	0	0.10	0.10	1.00	78	0.11	0.10	1.00	78	0.08	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.13	0.11	0.83	89	0.14	0.11	0.71	86	0.14	0.15	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.65	0.63	1.00	94	0.65	0.63	1.00	94	0.50	0.47	1.00	94
Curriculum: general	0.59	0.28	0	0.29	0.32	1.00	88	0.28	0.32	1.00	84	0.40	0.43	1.00	90
Highest grades of parents	10.50	13.21	0	12.36	12.76	0.37	85	12.25	12.76	0.37	81	11.88	12.10	1.00	91
Occupation parents high	0.08	0.29	0	0.23	0.24	1.00	91	0.24	0.24	1.00	93	0.18	0.17	1.00	93
Number of siblings	3.92	2.64	0	2.79	2.76	1.00	90	2.77	2.76	1.00	91	2.93	2.82	1.00	88
Born in south	0.38	0.33	1	0.23	0.32	0.11	-99	0.25	0.32	0.76	-37	0.32	0.34	1.00	38
Mean percent bias reduction							76				78				87
– *born in south* excluded							89				87				91

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last two rows report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 2.9: **Balance of Covariates: Matching on the p.score, Within calipers: Mahalanobis.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.57	0.58	1.00	98	0.43	0.43	1.00	99
Index score	-1.77	0.61	0	0.17	0.21	1.00	98	0.15	0.21	1.00	98	-0.30	-0.27	1.00	99
Black	0.30	0.16	0	0.15	0.15	1.00	100	0.15	0.15	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.06	0.06	1.00	100
Age	17.50	17.65	1	17.89	17.85	1.00	69	17.92	17.84	1.00	49	17.82	17.77	1.00	63
Year of high school diploma	79.37	78.74	0	78.57	78.54	1.00	94	78.55	78.54	1.00	96	78.63	78.68	1.00	91
Math test scores	-3.95	10.02	0	8.83	8.36	1.00	96	8.19	8.38	1.00	98	5.31	5.04	1.00	97
Auto+shop test scores	3.88	7.98	0	8.44	7.78	1.00	84	7.99	7.77	1.00	93	7.82	6.56	0.98	69
Attended private school	0.03	0.12	0	0.09	0.09	1.00	80	0.11	0.09	1.00	78	0.07	0.08	1.00	83
Expelled or susp. from school	0.33	0.10	0	0.15	0.11	0.71	86	0.15	0.11	0.71	83	0.16	0.14	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.62	0.57	1.00	89	0.61	0.57	1.00	91	0.48	0.46	1.00	95
Curriculum: general	0.59	0.28	0	0.32	0.37	0.88	83	0.32	0.37	1.00	83	0.44	0.44	1.00	94
Highest grades of parents	10.50	13.21	0	12.23	12.59	0.37	85	12.36	12.59	0.45	87	11.82	12.04	1.00	91
Occupation parents high	0.08	0.29	0	0.20	0.22	1.00	89	0.24	0.22	1.00	93	0.17	0.17	1.00	94
Number of siblings	3.92	2.64	0	2.88	2.75	0.78	87	2.85	2.75	1.00	88	2.77	2.82	1.00	93
Born in south	0.38	0.33	1	0.25	0.30	0.88	-10	0.27	0.30	0.88	34	0.31	0.33	1.00	35
Mean percent bias reduction							81				84				86
– *born in south* excluded							88				88				90
Broad Caliper															
Propensity score	0.11	0.67	0	0.60	0.62	1.00	95	0.58	0.62	0.88	93	0.42	0.45	1.00	95
Index score	-1.77	0.61	0	0.21	0.36	0.89	93	0.16	0.36	0.13	91	-0.33	-0.20	1.00	95
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.04	0.04	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.95	17.89	1.00	61	17.97	17.89	1.00	45	17.85	17.82	1.00	68
Year of high school diploma	79.37	78.74	0	78.57	78.50	1.00	88	78.53	78.50	1.00	94	78.63	78.62	1.00	93
Math test scores	-3.95	10.02	0	8.94	9.04	1.00	98	8.24	9.04	0.95	94	4.96	5.36	1.00	97
Auto+shop test scores	3.88	7.98	0	7.73	7.63	1.00	95	7.52	7.63	1.00	96	7.32	6.48	1.00	79
Attended private school	0.03	0.12	0	0.08	0.10	1.00	78	0.11	0.10	0.88	77	0.07	0.08	1.00	82
Expelled or susp. from school	0.33	0.10	0	0.13	0.11	0.83	90	0.14	0.11	0.71	87	0.14	0.15	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.68	0.63	1.00	90	0.64	0.63	1.00	95	0.47	0.48	1.00	96
Curriculum: general	0.59	0.28	0	0.26	0.32	0.76	79	0.29	0.32	1.00	88	0.44	0.43	1.00	93
Highest grades of parents	10.50	13.21	0	12.14	12.76	0.17	77	12.18	12.76	0.36	79	11.75	12.11	0.90	87
Occupation parents high	0.08	0.29	0	0.20	0.24	0.80	80	0.23	0.24	1.00	91	0.17	0.18	1.00	93
Number of siblings	3.92	2.64	0	2.89	2.76	0.89	89	2.80	2.76	1.00	91	2.82	2.81	1.00	93
Born in south	0.38	0.33	1	0.23	0.32	0.00	-92	0.25	0.32	0.77	-35	0.32	0.34	1.00	45
Mean percent bias reduction							74				79				87
– *born in south* excluded							87				87				90

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last two rows report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 2.10: **Balance of Covariates: Matching on the index, Within calipers: index.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.57	0.58	1.00	99	0.56	0.58	1.00	97	0.43	0.44	1.00	98
Index score	-1.77	0.61	0	0.18	0.21	1.00	99	0.14	0.21	1.00	97	-0.26	-0.23	1.00	99
Black	0.30	0.16	0	0.16	0.16	1.00	100	0.16	0.16	1.00	100	0.24	0.24	1.00	100
Hispanic	0.10	0.07	1	0.05	0.05	1.00	100	0.05	0.05	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.90	17.89	1.00	82	17.92	17.89	1.00	74	17.84	17.83	1.00	79
Year of high school diploma	79.37	78.74	0	78.56	78.50	1.00	91	78.57	78.50	1.00	89	78.62	78.61	1.00	94
Math test scores	-3.95	10.02	0	8.34	8.27	1.00	98	8.09	8.27	1.00	98	5.26	5.25	1.00	98
Auto+shop test scores	3.88	7.98	0	7.73	7.58	1.00	92	7.73	7.58	1.00	92	7.40	6.47	1.00	76
Attended private school	0.03	0.12	0	0.10	0.10	1.00	84	0.11	0.10	1.00	82	0.08	0.08	0.99	83
Expelled or susp. from school	0.33	0.10	0	0.13	0.12	0.95	90	0.14	0.12	0.83	90	0.14	0.15	1.00	93
Curriculum: college prepar.	0.16	0.67	0	0.61	0.58	1.00	93	0.61	0.58	1.00	93	0.50	0.47	1.00	93
Curriculum: general	0.59	0.28	0	0.32	0.36	0.88	85	0.32	0.36	1.00	85	0.40	0.44	1.00	88
Highest grades of parents	10.50	13.21	0	12.32	12.53	0.88	91	12.23	12.53	0.45	88	11.88	12.06	1.00	92
Occupation parents high	0.08	0.29	0	0.23	0.23	1.00	95	0.23	0.23	1.00	95	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.80	2.77	1.00	92	2.79	2.77	1.00	91	2.92	2.84	1.00	91
Born in south	0.38	0.33	1	0.24	0.31	0.58	-59	0.28	0.31	1.00	18	0.33	0.35	1.00	41
Mean percent bias reduction							81				85				87
– *born in south* excluded							92				91				91
Broad Caliper															
Propensity score	0.11	0.67	0	0.61	0.63	1.00	96	0.56	0.63	0.06	87	0.43	0.47	0.98	92
Index score	-1.77	0.61	0	0.28	0.38	1.00	96	0.13	0.38	0.00	89	-0.25	-0.13	1.00	95
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.06	0.06	1.00	100	0.06	0.06	1.00	100	0.09	0.09	1.00	100
Age	17.50	17.65	1	17.92	17.91	1.00	78	17.96	17.91	1.00	56	17.81	17.81	1.00	82
Year of high school diploma	79.37	78.74	0	78.55	78.47	1.00	87	78.53	78.48	1.00	92	78.64	78.61	1.00	94
Math test scores	-3.95	10.02	0	9.09	9.19	1.00	98	7.95	9.19	0.55	91	5.48	5.93	1.00	97
Auto+shop test scores	3.88	7.98	0	7.34	7.71	1.00	91	7.84	7.72	1.00	97	7.74	6.75	1.00	76
Attended private school	0.03	0.12	0	0.10	0.10	1.00	76	0.10	0.10	1.00	86	0.08	0.08	0.99	85
Expelled or susp. from school	0.33	0.10	0	0.12	0.11	0.83	88	0.13	0.11	0.76	91	0.14	0.15	1.00	94
Curriculum: college prepar.	0.16	0.67	0	0.66	0.63	1.00	94	0.63	0.63	1.00	97	0.50	0.49	1.00	97
Curriculum: general	0.59	0.28	0	0.29	0.32	1.00	89	0.30	0.32	1.00	93	0.42	0.43	1.00	94
Highest grades of parents	10.50	13.21	0	12.31	12.75	0.29	84	12.18	12.76	0.25	79	11.85	12.17	0.82	87
Occupation parents high	0.08	0.29	0	0.23	0.24	1.00	93	0.23	0.24	1.00	93	0.18	0.18	1.00	94
Number of siblings	3.92	2.64	0	2.77	2.79	1.00	91	2.75	2.78	1.00	92	2.92	2.80	0.99	86
Born in south	0.38	0.33	1	0.22	0.32	0.00	-118	0.27	0.32	1.00	-5	0.33	0.34	1.00	59
Mean percent bias reduction							75				83				89
– *born in south* excluded							90				90				91

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last two rows report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 2.11: **Balance of Covariates: Matching on the index, Within calipers: Mahalanobis.**

	Initially			Optimal Full				Greedy Full				Greedy Pair			
	C	T	t	C	T	t	%	C	T	t	%	C	T	t	%
Narrow Caliper															
Propensity score	0.11	0.67	0	0.56	0.58	1.00	97	0.55	0.58	1.00	95	0.43	0.45	1.00	96
Index score	-1.77	0.61	0	0.15	0.21	1.00	98	0.11	0.21	1.00	96	-0.27	-0.21	1.00	98
Black	0.30	0.16	0	0.16	0.16	1.00	100	0.16	0.16	1.00	100	0.23	0.23	1.00	100
Hispanic	0.10	0.07	1	0.05	0.05	1.00	100	0.05	0.05	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.91	17.89	1.00	79	17.93	17.89	1.00	72	17.85	17.82	1.00	71
Year of high school diploma	79.37	78.74	0	78.60	78.50	1.00	84	78.55	78.50	1.00	92	78.62	78.62	1.00	95
Math test scores	-3.95	10.02	0	8.48	8.27	1.00	98	7.86	8.27	1.00	97	5.39	5.39	1.00	98
Auto+shop test scores	3.88	7.98	0	8.20	7.58	1.00	84	7.76	7.57	1.00	92	7.46	6.51	0.99	76
Attended private school	0.03	0.12	0	0.08	0.10	1.00	83	0.11	0.09	1.00	82	0.08	0.09	1.00	81
Expelled or susp. from school	0.33	0.10	0	0.13	0.12	0.83	91	0.14	0.12	0.82	90	0.15	0.15	1.00	95
Curriculum: college prepar.	0.16	0.67	0	0.63	0.58	0.88	89	0.60	0.58	1.00	95	0.46	0.47	1.00	96
Curriculum: general	0.59	0.28	0	0.30	0.36	0.88	80	0.32	0.37	1.00	86	0.45	0.44	1.00	94
Highest grades of parents	10.50	13.21	0	12.18	12.53	0.37	86	12.22	12.53	0.37	88	11.85	12.07	1.00	92
Occupation parents high	0.08	0.29	0	0.21	0.23	1.00	91	0.23	0.23	1.00	95	0.17	0.18	1.00	93
Number of siblings	3.92	2.64	0	2.89	2.77	1.00	89	2.78	2.77	1.00	93	2.74	2.83	1.00	92
Born in south	0.38	0.33	1	0.26	0.31	0.88	-29	0.28	0.31	1.00	17	0.33	0.35	1.00	48
Mean percent bias reduction							80				86				88
– *born in south* excluded							89				91				91
Broad Caliper															
Propensity score	0.11	0.67	0	0.56	0.63	0.05	87	0.53	0.63	0.00	83	0.42	0.49	0.05	87
Index score	-1.77	0.61	0	0.13	0.38	0.00	89	0.05	0.38	0.00	86	-0.29	-0.07	0.14	91
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.17	0.17	1.00	100	0.22	0.22	1.00	100
Hispanic	0.10	0.07	1	0.06	0.06	1.00	100	0.06	0.06	1.00	100	0.08	0.08	1.00	100
Age	17.50	17.65	1	17.97	17.91	1.00	62	17.98	17.91	1.00	52	17.83	17.80	1.00	76
Year of high school diploma	79.37	78.74	0	78.55	78.47	1.00	88	78.51	78.47	1.00	94	78.64	78.61	1.00	94
Math test scores	-3.95	10.02	0	8.28	9.19	1.00	93	7.52	9.19	0.00	88	5.39	6.31	0.99	93
Auto+shop test scores	3.88	7.98	0	8.40	7.71	1.00	83	8.01	7.72	1.00	92	7.76	6.87	1.00	78
Attended private school	0.03	0.12	0	0.09	0.10	1.00	78	0.10	0.10	1.00	85	0.07	0.09	1.00	77
Expelled or susp. from school	0.33	0.10	0	0.12	0.11	0.88	93	0.14	0.11	0.82	90	0.12	0.15	1.00	87
Curriculum: college prepar.	0.16	0.67	0	0.64	0.63	1.00	99	0.58	0.63	0.88	90	0.43	0.50	0.96	86
Curriculum: general	0.59	0.28	0	0.30	0.32	1.00	94	0.34	0.32	1.00	93	0.48	0.42	0.98	80
Highest grades of parents	10.50	13.21	0	12.12	12.75	0.00	77	12.14	12.75	0.13	77	11.80	12.21	0.59	85
Occupation parents high	0.08	0.29	0	0.21	0.24	1.00	87	0.22	0.24	1.00	93	0.16	0.19	0.98	85
Number of siblings	3.92	2.64	0	2.87	2.79	1.00	91	2.73	2.79	1.00	93	2.83	2.81	1.00	94
Born in south	0.38	0.33	1	0.26	0.32	0.93	-30	0.29	0.32	1.00	37	0.35	0.34	1.00	59
Mean percent bias reduction							80				85				85
– *born in south* excluded							88				88				87

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last two rows report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Chapter 3

The Propensity Score: A Means to An End

March 2000/October 2000

Together with **Christoph M. Schmidt**

Abstract. Propensity score matching is a prominent strategy to reduce imbalance in observational studies. However, if imbalance is considerable and the control reservoir is small, either one has to match one control to several treated units or, alternatively, discard many treated persons. The first strategy tends to increase standard errors of the estimated treatment effects while the second might produce a matched sample that is not anymore representative of the original one. As an alternative approach, this chapter argues to carefully reconsider the selection equation upon which the propensity score estimates are based. Often, all available variables that rule the selection process are included into the selection equation. Yet, it would suffice to concentrate on only those exhibiting a large impact on the outcome under scrutiny, as well. This would introduce more stochastic noise making treatment and comparison group more similar. We assess the advantages and disadvantages of the latter approach in a simulation study.

3.1 Introduction

In contrast to a randomized experiment, in an observational study the treatment and the comparison group usually differ systematically in terms of their observable and unobservable covariates. Yet, appropriate weighting schemes may provide for a convincing evaluation strategy. In particular, balancing all observable covariates by the method of matching allows the identification of the mean effect of treatment if the remaining unobservable covariates are irrelevant. Usually, the number of covariates is high, thus making exact matching – in all likelihood – impossible. Rosenbaum & Rubin (1983) suggest to alternatively balance the one-dimensional propensity score, which is the conditional probability to participate in treatment given all relevant covariates. They show that this strategy, on average, achieves overall balance, thus circumventing the curse of dimensionality.

However, if treatment and comparison group differ to a considerable extent, i.e. if selection into treatment is remarkably strong, achieving an acceptable balance will be difficult. A *full matching* using all treated and untreated units in the sample might produce many strata consisting of one control and more than one treated unit. Generally, one would like to achieve a stratification which is more *uniform.* Uniform stratifications tend to produce smaller standard errors of the matching estimates. See, for instance, Chapter 2 and Dehejia & Wahba (1998) whose matching is far from producing a uniform stratification because treated units with high propensity scores hardly find adequate controls. Alternatively, *pair matching* tends to discard the majority of treated individuals at the high end of the propensity score scale. As a result, it restricts evaluation of the treatment effect to individuals with low and medium propensity scores. If effects are different for different locations on the propensity score scale pair matching estimates will be biased.

This chapter argues to carefully reconsider the selection equation upon which the propensity score estimates are based. It is common practice to include all available variables that might rule the selection process, with the objective of capturing the selection decision precisely. Yet, we will argue in this chapter that, if selection turns out to be extremely strong, one should better concentrate on only those variables with a large impact on both the selection and the outcome under scrutiny. This procedure increases the random part of the participation process – the whole approach rests on sufficient randomness being retained *after* deriving individuals' propensity score. Alas, a consistent estimation of the propensity score might require including into the selection equation variables which rule the selection process but which are excluded from or only play a minor role in the outcome equation.

In contrast to our arguments, current applied research emphasizes the importance of consistent estimation. For instance, Lechner (1999, 2000) performs and recommends several specification tests to examine whether a probit model is adequate for describing the selection decision. Chapter 2 includes into the probit model several variables that might determine the selection. Heckman, Ichimura & Todd (1997: section 8) choose predictor variables to maximize the within-sample correct prediction rates. Although a thorough understanding of the selection process might in itself be an important contribution, it is not the main objective of propensity score matching for identifying the mean effect of treatment. At best, it is a side effect. What is to be achieved by propensity score matching is balance of all relevant covariates as reflected, for example, in Dehejia & Wahba's (1998) pragmatic estimation strategy concerning the selection equation.

To put it otherwise, there is a trade-off between a consistent estimation of the selection equation that probably balances irrelevant variables, too, and a pragmatic – but probably inconsistent – estimation that concentrates on balancing the relevant variables only. We assess this trade-off in a simulation study relying on the mean squared error criterion. The next section discusses matching as an evaluation strategy and, in particular, outlines the idea behind propensity score matching. Section 3 presents the data generating processes and the dimensions of the simulation study while section 4 explains the algorithm used for matching. Section 5 is dedicated to results for some interesting parameter constellations and the last section summarizes the findings and offers recommendations for applied research.

3.2 The Matching Approach

In this section, the framework and the idea of propensity score matching are briefly discussed. Rosenbaum (1995), Heckman, LaLonde & Smith (1999), and Schmidt (1999) provide a thorough overview of estimation strategies via matching. Let R_i^1 denote the potential response of individual i under the treatment state and R_i^0 the potential response if i receives no treatment. Furthermore, let D_i denote a binary variable indicating treatment status, thus, $R_i = D_i R_i^1 + (1 - D_i) R_i^0$ is the observed outcome. This framework has become known as the *potential outcome approach to causality* suggested by Roy (1951), Rubin (1974, 1977), and Holland (1986). It requires that the response of an individual be independent of the decisions of all other individuals. This implies that there are only two potential outcomes, namely R_i^0 and R_i^1, one for the personal state $D_i = 0$, and one for $D_i = 1$, respectively. There are no further potential outcomes depending on the assignment of any other individual. This requirement is often referred to as *stable unit treatment value assumption* (sutva, see Rubin, 1986).

The individual treatment effect is $\delta_i = R_i^1 - R_i^0$ which, however, is not observable since either R_i^1 or R_i^0 is missing. Alternatively, one might focus on the mean effect of treatment on the treated individuals

$$\mathbb{E}(\delta_i | D_i = 1) = \mathbb{E}(R_i^1 | D_i = 1) - \mathbb{E}(R_i^0 | D_i = 1). \tag{3.1}$$

Yet, while the first expectation $\mathbb{E}(R_i^1 | D_i = 1)$ can be identified in the subsample of the treatment group, the counterfactual expectation $\mathbb{E}(R_i^0 | D_i = 1)$ is not identifiable without invoking further assumptions.

Somehow one has to rely on the untreated units ($D_i = 0$) of the comparison group to obtain information on the counterfactual outcome of the treated in the no-treatment state. A simple replacement of $\mathbb{E}(R_i^0 | D_i = 1)$ by $\mathbb{E}(R_i^0 | D_i = 0)$ is unlikely to be the appropriate strategy, though, since treated and untreated units tend to differ considerably in their characteristics that determine the outcome if they themselves select into treatment. An ideal randomized experiment solves this problem, see Heckman (1996) or Schmidt, Baltussen & Sauerborn (1999). It generates a treatment and a control group by a randomization process ensuring exogenous selection into treatment and thus resulting, on average, in balance of all covariates between treatment and control group, in particular those determining outcome.

In contrast, in an observational study, where self-selection into treatment is typically non-negligible, matching tries to mimic *ex post* a randomized experiment by stratifying the sample of treated and untreated units with respect to covariates X_i that rule both the selection into treatment and the outcome under study. Such a stratification eliminates selection bias provided all variables X_i are observed and balanced. In this case, each stratum would represent a separate small randomized experiment and simple differences between treated and controls would provide an unbiased estimate of the treatment effect. This technique does not require linearity, parametric, or distributional assumptions.

Formally, assume that the response R_i^0 is conditionally independent of D_i given X_i yielding $\mathbb{E}(R_i^0|X_i, D_i = 1) = \mathbb{E}(R_i^0|X_i, D_i = 0)$. Moreover, assume $\mathbb{P}(D_i = 0|X_i = x) > 0$ for all x which guarantees that, with positive probability, there are untreated units for each x. The data generating processes of the simulation presented in the next section are such that these requirements for matching will be fulfilled. The conditional mean response of the *treated* under no treatment for a given X can thus be estimated by the conditional mean response of the *untreated* under no treatment. The overall estimated mean effect is the weighted average over all stratum effects. The stratum weights are proportional to the number of treated units in the stratum in order to identify $\mathbb{E}(\delta_i|D_i = 1)$.

However, in a finite sample balancing X is difficult or even impossible if the vector of observables is of high dimension. To escape this curse of dimensionality, Rosenbaum & Rubin (1983) suggest to alternatively use the conditional probability to participate in treatment $p(x) = \mathbb{P}(D_i = 1|X_i = x)$, the *propensity score*, for purposes of stratifying the sample. They show that if R_i^0 is independent of D_i given X_i, R_i^0 and D_i are also independent given $p(X_i)$. Matching treated and untreated units with the same propensity scores and placing them into one stratum means that the decision whether to participate or not is random in such a stratum. The probability of participation in this stratum equals the propensity score. Alas, some disadvantages accompany this strategy. First, the propensity score itself has to be estimated. Second, since it is a continuous variable exact matches will hardly be achieved and a certain distance between treated and untreated units has to be accepted nonetheless. Prominent candidates measuring the distance are the difference in propensity scores or the Mahalanobis metric (Rubin, 1980).

The Idea Behind Propensity Score Matching

Let there be three kinds of covariates X, Y, and Z characterizing individuals. Generally, both potential outcomes and the participation probability depend on all three variables. For reasons of clarity of the argument further assume that Y and Z are binary and let all considerations to follow be conditional on X. In sum, $R^0 = R^0(Y,Z)$, $R^1 = R^1(Y,Z)$, and $p = p(Y,Z)$.

There are four cells

	$Z=0$	$Z=1$
$Y=0$	n_{00}	n_{01}
$Y=1$	n_{10}	n_{11}

each comprising n_{jk} individuals, $j,k \in \{0,1\}$. For the sake of notational convenience, abbreviate cell-wise expectations as follows

$$\begin{aligned} R^1_{jk} &= \mathbb{E}(R^1 | Y=j, Z=k, D=1) \\ R^0_{jk} &= \mathbb{E}(R^0 | Y=j, Z=k, D=1) = \mathbb{E}(R^0 | Y=j, Z=k, D=0), \end{aligned}$$

$\Delta_{jk} = R^1_{jk} - R^0_{jk}$, and p_{jk} denotes the propensity score in the corresponding cell. As a result, the mean effect Δ (conditional on X) can be written

$$\Delta = \frac{1}{n_t} \left(\Delta_{00}\, p_{00}\, n_{00} + \Delta_{01}\, p_{01}\, n_{01} + \Delta_{10}\, p_{10}\, n_{10} + \Delta_{11}\, p_{11}\, n_{11} \right), \tag{3.2}$$

n_t denotes the total number of treated individuals, $n_t = \sum p_{jk} n_{jk}$.

Selection on Z only. If the propensity score merely depends on Z, $p_{00} = p_{10} = p_{.0}$ and $p_{01} = p_{11} = p_{.1}$. This implies that Y can be expected to be already balanced and that cells with the same value of Z can be combined. Defining $n_{.k} = n_{0k} + n_{1k}$ and the effect in the combined cell $\Delta_{.k} = (\Delta_{0k}\, n_{0k} + \Delta_{1k}\, n_{1k})/n_{.k}$, equation (3.2) reduces to

$$\Delta = \frac{1}{n_t}\left(\Delta_{.0}\, p_{.0}\, n_{.0} + \Delta_{.1}\, p_{.1}\, n_{.1}\right). \tag{3.3}$$

The combination of cells that share the same propensity score is the very advantage of propensity score matching with regard to exact covariate matching. On the one hand, this means that individuals with different characteristics might be matched, here with different values of Y. As a result, in finite samples where Y may still be unbalanced the combined-cell-specific estimates of the treatment effect may deviate from the true value. On the other hand, combination of cells avoids that cells comprising only treated or only untreated units have to be dropped. This would give rise to both larger variance of the estimates and possibly a bias if the treatment effect is heterogeneous and the loss of cells is systematic.

Angrist & Hahn (1999) assess this bias-variance trade-off both theoretically and by means of a simulation study. They argue that the very virtue of propensity score estimation emerges when cells are finite. If cell sizes themselves increased beyond all bounds propensity score matching would not be advantageous to exact matching, see Hahn (1998).

Exclusion Restriction of Z. A symmetric special case arises if the outcome does not but the selection does vary with Z. Consequently, cells with the same value of Z could be combined even though they are subject to a different selection process, i.e. their propensity score differs. Analogously to above, it follows that $\Delta_{00} = \Delta_{01} = \Delta_{0.}$ and $\Delta_{10} = \Delta_{11} = \Delta_{1.}$, implying that imbalance of Z has no effect on the estimation of the outcome and that cells with the same value of Y can be combined without loss of information. Let $n_{k.} = n_{k0} + n_{k1}$ and $p_{k.} = (p_{k0}n_{k0} + p_{k1}n_{k1})/n_{k.}$, equation (3.2) can be reduced to

$$\Delta = \frac{1}{n_t}\left(\Delta_{0.}\, p_{0.}\, n_{0.} + \Delta_{1.}\, p_{1.}\, n_{1.}\right). \tag{3.4}$$

If both cases are fulfilled, i.e. the outcome depends on Y and the selection process is ruled by Z only, all four cells can be combined to one and Δ is just the difference between the unconditional responses of treated and untreated persons in the combined cell (merely defined by X). This point reflects the fact that solely covariates which rule both the outcome and the selection into treatment need to be balanced by matching. Consequently, the question is raised whether the propensity score depending on X and Z is the right measure to match upon or whether it might be better replaced by the marginal propensity score depending solely on X. Matching on the latter would not unnecessarily balance Z. Therefore, one could concentrate on the balance of X. This would probably result in a more uniform stratification of the sample. That is, one control would not be matched to an overwhelmingly large number of treated persons.

In other words, omitting irrelevant variables increases randomness of the selection process and diminishes its deterministic part. For example, if selection were completely determined by certain known variables the propensity score of treated units would be 1 and that of untreated 0. Consequently, no reasonable strategy whatsoever would be able to match controls to any given treated person. In contrast, the more variables determining the selection process can be regarded as stochastic noise because their impact on the outcome variable is negligible, the more randomness will enter the process and the easier treated individuals will find adequate controls. One might equate the *Pseudo* R^2 of a probit model as reflecting the degree of the selection determination.

3.3 The Data Generating Processes

As above, let R_i denote the outcome of individual i, $i = 1, ..., n$, and D_i the binary treatment indicator. On average, there will be 150 treated individuals and between 300 and 900 comparison units. The latter number is variable such that finding adequate controls is more or less difficult. The outcome is a linear function of confounding covariates, $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$, an individual treatment effect δ_i, and normally distributed stochastic noise $\varepsilon_i \sim \mathcal{N}(0, 9)$

$$R_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_4 Y_{2i} + \beta_5 Z_{1i} + \beta_6 Z_{2i} + \delta_i D_i + \varepsilon_i. \qquad (3.5)$$

The selection equation depends on the same covariates

$$D_i = \mathbf{1}[\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 Y_{1i} + \alpha_4 Y_{2i} + \alpha_5 Z_{1i} + \alpha_6 Z_{2i} + \eta_i > 0] \qquad (3.6)$$

where $\mathbf{1}$, the indicator function, is 1 if its argument holds and zero otherwise, and $\eta \sim \mathcal{N}(0,1)$ is standard normal.

The coefficients of the Z-variables β_5, β_6 in the outcome equation (3.5) are comparatively small and, likewise, the same is assumed for those of the Y-variables in the selection equation (3.6), α_3, α_4. This means that Y tends to be already partly balanced between treated and untreated units and, furthermore, although Z will be highly unbalanced its impact on the outcome is minor. The X-variables are the strongest predictors of both the outcome and the selection and most effort should therefore be spent on balancing them. The simulation aims at examining the relative performance of the matching estimator when the propensity score is estimated by means of a probit model including all variables (X, Y, Z) and when based on the most relevant variables X only. Furthermore, the treatment effect δ_i depends on i reflecting heterogeneity in the following manner

$$\delta_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 Y_{1i} + \gamma_4 Y_{2i} + \gamma_5 Z_{1i} + \gamma_6 Z_{2i}.$$

Depending on the parameter setting self-selection into treatment plays a more or less important role resulting in more or less severe imbalance of covariates. If Y and Z are of minor relevance, merely X should actively be balanced by matching on $\mathbb{P}(D = 1|X)$. However, $\mathbb{P}(D = 1|X, Y, Z)$ follows a probit specification in accordance with equation (3.6)

$$\mathbb{P}(D = 1|X, Y, Z) = \Phi(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Y_1 + \alpha_4 Y_2 + \alpha_5 Z_1 + \alpha_6 Z_2),$$

where Φ is the cumulative normal density function. Thus, a probit estimation using covariates X, Y, and Z – henceforth called the *full probit* – would yield consistent estimates of individual propensity scores but matching on them would unnecessarily balance Z, as well. On the other hand, a misspecified probit estimation merely on X – henceforth called the *partial probit* – would indeed use only the most relevant variables but might yield inconsistent estimates of $\mathbb{P}(D = 1|X)$. The choice to proceed as if a probit model held might therefore be one reason for bias in estimates of the mean treatment effect.[1]

[1] Note, though, that consistent estimation of the coefficients α in the probit model are not of any interest. Furthermore, see Yatchew & Griliches (1984) for a discussion of specification errors in probit models.

In general, the functional form of $\mathbb{P}(D = 1|X) = \mathbb{E}(\mathbb{E}(D|X, Y, Z)|X)$ does not follow a probit specification

$$\mathbb{P}(D = 1|X) = \int \Phi(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 y_1 + \alpha_4 y_2 + \alpha_5 z_1 + \alpha_6 z_2)\, f_{(Y,Z)|X}(y, z)\, d(y, z)$$

where $f_{(Y,Z)|X}$ is the conditional density of (Y_1, Y_2, Z_1, Z_2) given X.[2] Another source of bias arises if the impact of Y and Z on the outcome and on the selection are not zero.

In consequence, the questions of this chapter are (i) whether neglecting to balance (Y, Z) produces a bias which is offset by a larger variance of the estimates of the full model, and (ii) whether the functional specification error in estimating $\mathbb{P}(D = 1|X)$ by a probit model causes severe problems. We assess the trade-off on the basis of the mean squared error criterion.

The described setup allows to perform simulations along five dimensions. First, the impact of Z on R and of Y on D may be altered. To this end, β_5, β_6 and α_3, α_4 are varied between 0 and 0.1 while the remaining α- and β-coefficients are set equal to 1, and the constant β_0 equals 0. This strategy allows an exploration of the question whether near exclusion restrictions carry the same implications as genuine exclusion restrictions. Second, the average number of comparison units in the sample is gradually increased from 300 to 900 while the average number of treated is fixed at 150 by accordingly adjusting the constant α_0. Thereby, we address the issue by how much the described trade-off is altered as more and more comparison observations become available.

Third, the deterministic part of the selection equation is successively weakened which means that all α-coefficients except for α_0 are simultaneously reduced until they reach 25% of their original value. This shows how the degree of selection determination influences the stratification results. Fourth, effects δ_i may be homogeneous or heterogeneous corresponding to whether $\gamma = (1, 0, 0, 0, 0, 0, 0)$ or $\gamma = (0.5, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)$. The homogeneous case presents an interesting benchmark to compare the full and partial probit model. Pair matching might be an unbiased and a more efficient evaluation strat-

[2] Since it is not easy to solve the integral analytically the true values are calculated by ways of an auxiliary Monte Carlo simulation: 200 times adequate (Y, Z)'s are generated and $\mathbb{P}(D = 1|X, Y = y, Z = z)$ is calculated for each iteration inserting the given $(Y, Z) = (y, z)$. The mean over all iterations is an approximation to $\mathbb{P}(D = 1|X)$.

egy than full matching when effects are homogeneous. Yet, in this study, the choice of the matching algorithm will not be explored.

Finally, the distribution of $(X_1, X_2, Y_1, Y_2, Z_1, Z_2)$ is varied. In a *basic model* all six variables are independently and identically (iid) standard normal implying that omission of (Y, Z) from the probit model does not bias propensity score estimates because the omitted variables are perfectly absorbed in normally distributed stochastic noise. To avoid this favorable aspect, Z will alternatively be distributed in an odd fashion. Several alternatives have been investigated but those maintaining independence between Z and X and reducing to an exchange of the distribution of Z have been unable to produce biased propensity score estimates.[3]

Apparently, the probit model seems quite insensitive to misspecification of the error distribution as far as the overall fit is concerned and coefficients are of no interest. Yet, as soon as independence of X and Z is abandoned omission of Z leads to heteroskedastic errors of the selection equation and to arbitrarily large biased propensity score estimates, up to estimates that are almost constant for all values of X. One specification that is presented below – called *alternative model* – defines (X_1, X_2, Y_1, Y_2) as *iid uniformly* distributed random variables with mean zero and variance one. In contrast, Z_j will follow the functional form

$$Z_j = U_j \exp(-\mu X_j), \qquad j = 1, 2, \tag{3.7}$$

where U_j is a uniform random variable in the unit interval and $\mu = 1.35$. In addition, Z_j is standardized to have mean zero and variance one in each iteration of the simulation. This is necessary to ensure that selection due to Z is normalized and comparable to the basic model.[4] Furthermore, interactions between Z and X are introduced into the selection equation (3.6) such that it becomes

$$\begin{aligned} D_i &= \mathbf{1}[\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 Y_{1i} + \alpha_4 Y_{2i} + \alpha_5 Z_{1i} + \alpha_6 Z_{2i} + \\ &\qquad \alpha_7 X_{1i} Z_{1i} + \alpha_8 X_{1i} Z_{2i} + \alpha_9 X_{2i} Z_{1i} + \alpha_{10} X_{2i} Z_{2i} + \eta_i > 0] \end{aligned}$$

[3] Even very asymmetric strange densities of Z failed to generate inconsistencies.

[4] If Z has high variance it will strongly determine selection. To normalize its impact with respect to the basic model the variance is required to be 1.

Table 3.1: **The Simulation Setup.**

	Distribution		Parameters	
Variable	Basic	Alternative*	Outcome	Selection
Constant	–	–	1	α_0 (adjusted)
X_1	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_1 = 1$	$\alpha_1 = 1$
X_2	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_2 = 1$	$\alpha_2 = 1$
X_1X_2	–	–	$\beta_{12} \in \{0,1\}$	0
Y_1	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_3 = 1$	$\alpha_3 \in \{0, 0.05, 0.10\}$
Y_2	$\mathcal{N}(0,1)$	$\mathcal{U}[-0.5,0.5]$	$\beta_4 = 1$	$\alpha_4 \in \{0, 0.05, 0.10\}$
Y_1Y_2	–	–	$\beta_{34} \in \{0,1\}$	0
Z_1	$\mathcal{N}(0,1)$	$U_1 \exp(-\mu X_1)$	$\beta_5 \in \{0, 0.05, 0.10\}$	$\alpha_5 = 1$
Z_2	$\mathcal{N}(0,1)$	$U_1 \exp(-\mu X_1)$	$\beta_6 \in \{0, 0.05, 0.10\}$	$\alpha_6 = 1$
Z_1Z_2	–	–	$\beta_{56} \in \{0,1\}$	0
X_1Z_1	–	–	0	$\alpha_7 \in \{0,1\}$
X_1Z_2	–	–	0	$\alpha_8 \in \{0,1\}$
X_2Z_1	–	–	0	$\alpha_9 \in \{0,1\}$
X_2Z_2	–	–	0	$\alpha_{10} \in \{0,1\}$
D_i	–	–	δ_i	see below
U_1	–	$\mathcal{U}[0,1]$	0	0
U_2	–	$\mathcal{U}[0,1]$	0	0
ε_i	$\mathcal{N}(0,9)$	$\mathcal{N}(0,9)$	1	1
η_i	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	1	1

Size of the control reservoir $\in \{300, 600, 900\}$

Size of the treatment group: 300

Importance of the deterministic part $\in \{0.25, 0.50, 0.75, 1.00\}$

$\delta_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 Y_{1i} + \gamma_4 Y_{2i} + \gamma_5 Z_{1i} + \gamma_6 Z_{2i}$ and

$\gamma \in \{(1,0,0,0,0,0,0), (0.5, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)\}$

$\mu = 1.35$

* Furthermore, all variables are standardized to have mean zero and variance 1.

Omission of Z might lead to severe misspecification problems which, however, can substantially be alleviated by adding higher order terms of X into the probit specification. The conditional expectation of Z_j given X_1, X_2 is a function of X_1, X_2

$$\mathbb{E}(Z_j | X_1, X_2) = f(X_1, X_2). \tag{3.8}$$

Hence, inclusion of higher order terms of (X_1, X_2) approximates a Taylor expansion of $f(X_1, X_2)$ such that, again, almost only the stochastic part of Z will be absorbed by the error term of the model. Three alternative probit models will therefore be specified to demonstrate this issue. The first model consists of linear terms in X only, the second one includes an interaction X_1X_2, and the third one further adds quadratic terms in X.

Other interesting features consider (i) whether asymmetry of the parameters $(\beta_1, \beta_2) = (0.5, 2)$ or (ii) whether interaction terms in the outcome equation as follows

$$\begin{aligned} R &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_3 Y_1 + \beta_4 Y_2 + \beta_{34} Y_1 Y_2 \\ &+ \beta_5 Z_1 + \beta_6 Z_2 + \beta_{56} Z_1 Z_2 + \varepsilon \end{aligned} \tag{3.9}$$

might cause additional problems. To keep the presentation of the alternative model simple only a certain parameter constellation of the basic model will be considered more closely: a medium impact of Y on R and Z on D, i.e. with coefficients $\alpha_3 = \alpha_4 = \beta_5 = \beta_6 = 0.05$, a medium size of the control reservoir (600), and a selection determination of 0.75. The setup is summarized in table 3.1.

3.4 The Matching Algorithm

Consider the basic specification retaining independence between X and Z, with Z having no impact on the outcome R, and Y none on selection D but all other α and β-coefficients are 1, and, furthermore, where there are 600 comparison units. This constellation already motivates the use of the special matching algorithm presented below. The columns under the heading *full probit* of table 3.2 compare the absolute frequencies of treated and untreated individuals by propensity score intervals. Obviously, the distribution is very unfavorable for matching at the boundaries. In effect, the full probit model successfully separates the treated from the untreated. Unfortunately, high predictive ability of the model implies difficulties in finding adequate controls for high propensity score treated individuals. The picture improves substantially if Z (and Y) are omitted from the selection equation. Estimation results of the partial probit are presented in the last two columns of the table. Apparently, the difference in the distributions of the estimated propensity scores for treated and untreated is less extreme than in the full probit. Therefore, matching can be expected to be much easier.

Table 3.2: **Distribution of Treated and Untreated Individuals.**

Estimated	Full Probit		Partial Probit	
Propensity score	untreated	treated	untreated	treated
$0.0 \leq \hat{p} < 0.1$	459.58	6.80	293.39	12.41
$0.1 \leq \hat{p} < 0.2$	49.59	8.69	135.02	23.67
$0.2 \leq \hat{p} < 0.3$	29.39	9.77	75.58	24.86
$0.3 \leq \hat{p} < 0.4$	19.88	9.86	45.08	23.71
$0.4 \leq \hat{p} < 0.5$	14.29	10.74	25.49	20.54
$0.5 \leq \hat{p} < 0.6$	10.11	12.48	14.16	16.63
$0.6 \leq \hat{p} < 0.7$	7.18	13.47	6.93	12.65
$0.7 \leq \hat{p} < 0.8$	5.22	15.83	3.01	9.11
$0.8 \leq \hat{p} < 0.9$	3.04	18.90	1.01	4.98
$0.9 \leq \hat{p} \leq 1.0$	1.46	43.50	0.07	1.70
Mean propensity score	0.09	0.65	0.15	0.38
Observations	600	150	600	150

The means are averages over 100 iterations. Comparison of number of treated and untreated individuals by certain propensity score intervals.

After estimation of individual propensity scores a distance between treated and untreated individuals has to be defined because exact matching on the continuous score is impossible. Here a propensity score caliper approach is pursued (Cochran & Rubin, 1973). A small pool of potential controls is generated for each treated unit by excluding all untreated units whose propensity score distance to the chosen treated exceeds a certain caliper ε. Within the caliper, the distances from treated individual to potential control is defined in terms of the *Mahalanobis metric* based on variables W consisting of the estimated propensity score and all matching covariates, either (X, Y, Z) or X for the full or partial specification, respectively. It is a weighted Euclidean distance $d(w_t, w_c) = (w_t - w_c)'V^{-1}(w_t - w_c)$, where indices t, c represent the treated and the potential control units, respectively. V is the pooled covariance matrix of W which serves to norm the vectors. In sum, the distance is

$$d(w_t, w_c) = \begin{cases} \infty & \text{if } |p_t - p_c| > \varepsilon \\ (w_t - w_c)'V^{-1}(w_t - w_c) & \text{else.} \end{cases} \tag{3.10}$$

An infinite distance indicates that matching is forbidden.

Table 3.3: **Specification of Caliper Width ε.**

Basic Model							Alternative Model		
Selection	Full Probit			Partial Probit			Full	Partial Probit	
Determ.	300	600	900	300	600	900	Probit	Order	
0.25	.030	.020	.010	.015	.010	.005	.03	1	.002
0.50	.040	.030	.020	.030	.015	.010		2	.010
0.75	.050	.040	.030	.040	.020	.010		3	.010
1.00	.060	.050	.040	.050	.025	.010			

The first column of the basic model presents the factors which the α-coefficients of the selection equation are multiplied with. The next columns headed by the size of the control reservoir display the critical ε. The first column of the alternative model shows the caliper width used in the full probit, the second shows whether no interactions (1), interactions (2), and additionally squares (3) are included in the partial probit, and the last displays ε.

Matching using the Mahalanobis distance is discussed in Rubin (1980). Gu & Rosenbaum (1993) perform simulations to compare three distance measures. Furthermore, propensity score calipers are discussed in Rosenbaum & Rubin (1985) and Rosenbaum (1989). Calipers help substantially reduce the number of potential controls and, thus, considerably accelerate the matching algorithm and, what is more, they prevent that too distant individuals are being matched. The critical ε is chosen such that there are enough but not too many potential controls in the vicinity of each treated which otherwise would considerably slow down the algorithm without improving results. Table 3.3 summarizes the choices of the critical ε. The results may depend on the choice of ε. A small ε will come with a loss of many treated (and untreated) individuals. On the other hand, however, it increases similarity of the matched units.

The final decision is how to implement the chosen matching criteria, in other words, how the distances between treated units and controls is to be minimized. A stratification producing small strata is preferable in order to ensure that the distance between the units within a stratum is not too large and stratum members are very similar to each other. This yields strata with either one treated and one or more controls or one control and more than one treated unit. It turns out that strata with very high propensity scores contain more than one treated and strata with low scores consist of a large number of controls.

In this study, *optimal full matching* as proposed by Rosenbaum (1991) is implemented. It minimizes the overall distances between treated and controls in that it works backwards and rearranges already matched units if an unmatched treated would better be matched to an already used untreated. In such a case, the existing match is broken up and its treated is available for matching again.[5] The strata will be non-overlapping, i.e. individuals are not members of more than one stratum, which facilitates the calculation of variances.[6] Optimal full matching can easily be transformed into a *minimum cost flow problem*, a special case of *linear network optimization*.[7]

Matching produces different strata in terms of number of treated and controls per stratum. Some might be very extreme comprising numerous treated units and only one control. It is they who substantially increase the variance of the estimated mean effect of treatment on the treated. On the other hand, strata with one treated but countless controls will work in the opposite direction but receive less weight. Therefore, an aggregate measure assessing the uniformity of a given stratification with respect to a benchmark stratification is helpful. To this end, suppose all estimated stratum treatment effects have the same variance, the following formula measures *variance inflation* due to unfavorable stratification[8]

$$\frac{1}{(\sum_{s=1}^{S} m_s)^2} \sum_{s=1}^{S} \frac{m_s^2}{(1-1/n_s)^2}$$

where m_s indicates the number of treated units and n_s the number of all individuals in stratum $s = 1, ..., S$.

In order to make the formula meaningful it ought to be compared to a benchmark stratification which is defined as follows. Let all treated units get their own stratum with exactly one control. Therefore, redefine $\tilde{m}_{\tilde{s}} = 1$ and $\tilde{n}_{\tilde{s}} = 2$ for all $\tilde{s} = 1, ..., \tilde{S}$ with $\tilde{S} = \sum_{s=1}^{S} m_s$, yielding a variance inflation of $4/\sum_{s=1}^{S} m_s$. The ratio of the two

[5] This is in contrast to so-called *greedy* algorithms which do not generally achieve a minimum, see Rosenbaum (1991).

[6] Statistical inference is described in Rosenbaum (1995) and adapted to this setup in Chapter 2. However, non-overlapping strata are not necessary if different techniques are used, see Quade (1981) or Heckman, Ichimura & Todd (1998).

[7] Bertsekas (1991) discusses *linear network optimization* and provides FORTRAN-algorithms for minimum cost flow problems. Furthermore, there is an *operations research* procedure called *netflow* in SAS for these kinds of problems.

[8] See Chapter 2 or Rosenbaum (1995) for the deduction of the general variance formula.

expressions yields a *relative variance inflation factor* denoted κ^2

$$\kappa^2 = \frac{1}{4\sum_{s=1}^{S} m_s} \sum_{s=1}^{S} \frac{m_s^2}{(1-1/n_s)^2}. \tag{3.11}$$

For example, pair matching produces $\kappa = 1$, 1-k-matching, i.e. one treated and k controls share a common stratum, leads to $\kappa = 0.5\,(1+1/k)$, k-1-matching has $\kappa = (k+1)/(2\sqrt{k})$. Note that the benchmark stratification can in general never be achieved since all treated who are used in the optimal stratification would have to find an own control. This would only be possible if there are no high propensity score treated units or else if several high propensity score treated individuals were matched to medium score controls which is either ruled out by a caliper approach or which otherwise would compare the incomparable. As such, κ incorporates neither the balance of covariates after matching nor how many treated units remain unmatched but only the uniformity of the stratification.

As outlined in the introduction, pair matching might be more efficient than full matching. What is more, if the treatment effect is homogeneous pair matching estimates are unbiased. Nevertheless, pair matching is disregarded in this study even in the case of homogeneous effects. The principal aim is to shed more light on the estimation of the propensity score when selection is strong. The homogeneous case is for illustrative purposes only and serves as a valuable benchmark.

Finally, matching should produce balance of all important covariates implying that at least their means for treated and controls be approximately equal. Therefore, to verify balance, simple t-tests of the hypothesis of equal means under equal variances are performed for each of the six variables $j = 1, ..., 6$. If the null hypothesis cannot be rejected at a 5% significance level let $t_j = 1$ and zero otherwise. Then, for an overall measure of balance, define the *aggregate balance* τ as

$$\tau = \frac{\sum_{j=1}^{6} \beta_j\, t_j}{\sum_{j=1}^{6} \beta_j}, \tag{3.12}$$

the β's being the coefficients of the outcome equation (3.5). Weighting by β takes into account that imbalance of the less important variables Z would cause less problems than that of X and Y.[9]

[9] Percent bias reduction has also been examined. Yet, results were quite unsatisfactory because a

3.5 Results

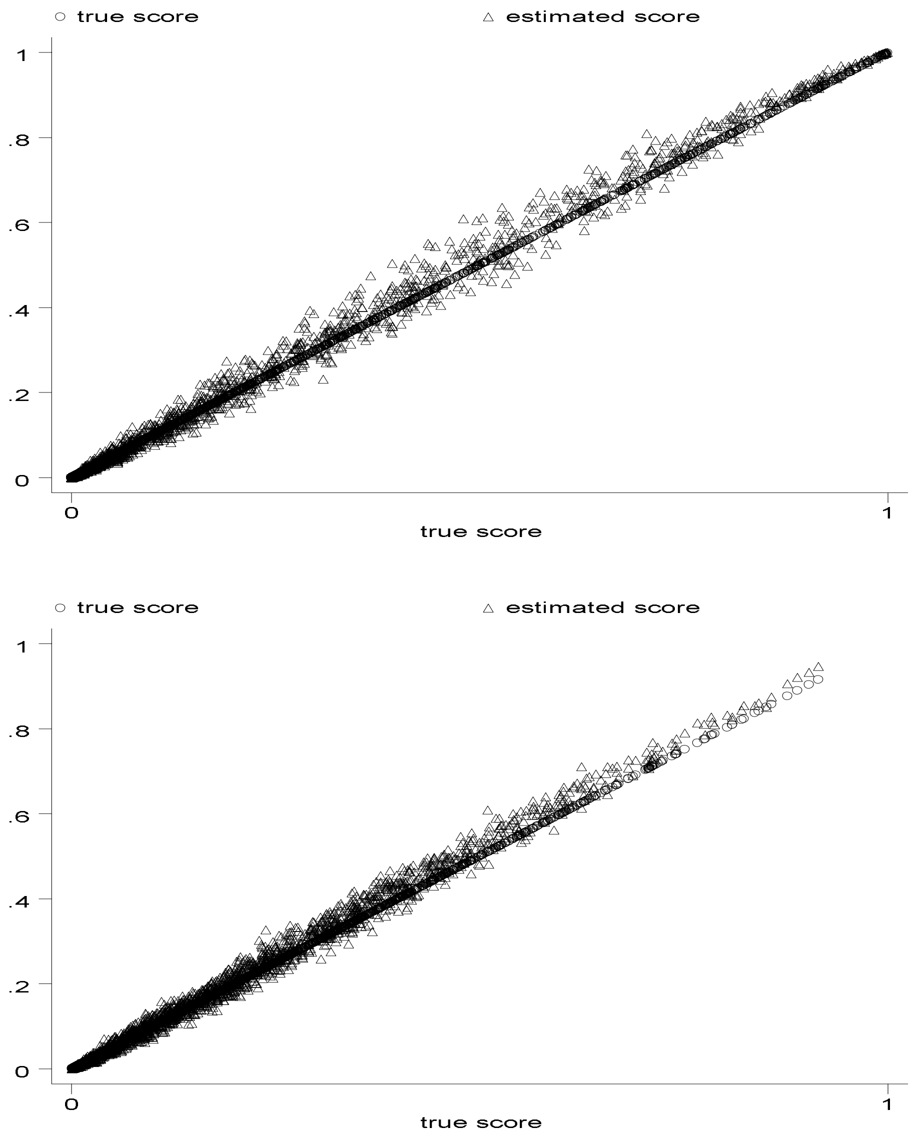
Each simulation is performed 100 times and mean estimation results over all iterations are presented and discussed for the parameter constellations mentioned above. Variability across simulations is reflected by simulation standard errors which, however, are not presented in the tables below for reasons of clarity. Figure 3.1 shows propensity score estimation results of the basic model with true and estimated scores on the vertical axis and the true ones on the horizontal axis. The data are taken from the constellation where the critical coefficients of Y and Z are 0.1, with 600 untreated individuals, and with selection determination of 0.75.

Apparently, the estimates of both the full and the partial specification are unbiased. Tables 3.4 and 3.5 go into the estimation and stratification results of the full and partial model. The first three columns characterize the simulation scenario. The first column reports the values of the coefficients $\beta_5, \beta_6, \alpha_3$, and α_4, the second the size of the control reservoir, and the third shows the factor the α-coefficients of the selection equation are multiplied with. The lower this factor the larger the randomness of the selection process and the less severe self-selection is. The next four columns report the bias and the RMSE in the homogeneous and the heterogeneous case. The remaining columns are self-explanatory.

The most striking result is that matching on the propensity score estimated by the full probit model produces almost always unbiased estimates of the mean effect of treatment on the treated while the bias of the partial probit matching rises to roughly 40% when the impacts of Y and Z are largest. Nevertheless, root mean squared errors of the latter are markedly lower when selection determination is highest. As selection determination successively increases, the full model puts an increasingly heavy burden upon treated individuals in finding appropriate controls, a fact reflected in the diminishing number of strata and the growing number of lost treated individuals. For instance, full probit matching ends with roughly 57 strata if selection determination is highest and the control

negative percent bias reduction is basically unbounded. If balance before matching is already given the denominator in the formula is close to zero. On the other hand, percent bias reduction is at most +100%. Therefore, the mean reduction turned out to be rather low in each single iteration.

Figure 3.1: **Basic Model, Full Probit** and **Partial Probit**



True and estimated propensity scores on the vertical axis versus true scores on the horizontal. The figures represent *one* iteration of the simulation study, the full model on the top, the partial one on the bottom.

reservoir is smallest. By contrast, partial probit matching still produces around 101 strata under these circumstances. That is, stratification in the latter case is more uniform as can also be seen from its lower value of κ which never surpasses 0.92 whereas full probit matching even surpasses $\kappa = 1$. However, the difference in RMSE decreases for a more extensive control reservoir.

Furthermore, the partial probit estimates are unbiased if the omitted variables Y and

Table 3.4: **Basic Model, Full Probit.**

			Effects				Stratification				
Scenario			Homogeneous		Heterogeneous		N^o of	Lost	κ	ΔP-	Bal.
(a)	(b)	(c)	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
0.00	300	0.25	-0.01	0.37	-0.04	0.50	129.97	3.36	0.84	-0.02	1.00
		0.50	-0.08	0.57	-0.14	0.67	97.15	8.99	0.97	-0.05	0.96
		0.75	0.09	0.79	0.03	0.84	73.58	13.10	1.14	-0.05	0.82
		1.00	0.15	1.15	0.11	1.16	56.49	11.34	1.36	-0.04	0.60
	600	0.25	0.03	0.34	0.01	0.41	143.34	3.22	0.71	-0.03	1.00
		0.50	0.03	0.43	-0.02	0.43	115.43	9.17	0.80	-0.07	0.99
		0.75	-0.01	0.56	-0.05	0.51	89.35	12.04	0.95	-0.06	0.90
		1.00	0.05	0.88	-0.00	0.79	72.28	15.59	1.10	-0.06	0.72
	900	0.25	0.03	0.30	-0.00	0.34	141.79	4.58	0.66	-0.04	1.00
		0.50	-0.00	0.39	-0.05	0.37	122.68	9.96	0.74	-0.08	0.99
		0.75	0.01	0.46	-0.04	0.40	100.08	15.10	0.86	-0.09	0.92
		1.00	-0.01	0.61	-0.06	0.51	81.95	17.82	0.99	-0.09	0.79
0.05	300	0.25	-0.01	0.36	-0.04	0.48	128.32	4.02	0.84	-0.02	1.00
		0.50	0.09	0.52	0.05	0.58	98.90	8.18	0.98	-0.04	0.96
		0.75	0.07	0.78	0.02	0.82	73.53	12.51	1.15	-0.05	0.85
		1.00	0.16	1.15	0.11	1.15	57.15	13.50	1.34	-0.05	0.59
	600	0.25	-0.01	0.35	-0.03	0.42	142.70	3.06	0.71	-0.03	1.00
		0.50	0.02	0.42	-0.03	0.43	116.97	8.19	0.82	-0.06	0.99
		0.75	0.04	0.56	-0.02	0.52	90.68	13.99	0.94	-0.08	0.91
		1.00	-0.02	0.81	-0.08	0.71	72.17	17.17	1.09	-0.07	0.68
	900	0.25	-0.04	0.32	-0.08	0.38	141.98	4.63	0.66	-0.05	1.00
		0.50	0.04	0.38	-0.01	0.35	122.73	10.51	0.74	-0.09	0.99
		0.75	0.09	0.43	0.03	0.36	98.23	14.25	0.86	-0.09	0.92
		1.00	0.10	0.71	0.03	0.57	79.78	18.32	0.99	-0.09	0.79
0.10	300	0.25	0.03	0.38	0.01	0.51	129.60	3.43	0.85	-0.02	1.00
		0.50	0.09	0.52	0.05	0.57	96.04	8.77	0.98	-0.04	0.97
		0.75	0.09	0.80	0.03	0.82	72.43	13.22	1.15	-0.05	0.79
		1.00	0.22	0.93	0.16	0.93	58.69	14.34	1.32	-0.05	0.63
	600	0.25	-0.01	0.36	-0.04	0.45	143.86	3.26	0.71	-0.03	1.00
		0.50	-0.07	0.46	-0.11	0.46	116.51	7.92	0.81	-0.06	0.98
		0.75	0.14	0.55	0.07	0.48	90.84	13.78	0.95	-0.07	0.88
		1.00	0.03	0.69	-0.03	0.61	73.88	18.28	1.07	-0.08	0.75
	900	0.25	0.00	0.30	-0.04	0.34	140.43	4.84	0.66	-0.05	1.00
		0.50	-0.01	0.44	-0.06	0.41	121.32	10.77	0.74	-0.09	1.00
		0.75	0.09	0.48	0.02	0.40	100.14	15.97	0.86	-0.10	0.97
		1.00	0.09	0.68	0.02	0.54	80.22	18.47	0.99	-0.09	0.77

The results are averages over all 100 iterations. The first block represents the scenario: (a) value of the coefficients $\beta_5, \beta_6, \alpha_3, \alpha_4$, (b) size of control reservoir, (c) selection determination. The next block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ.

Table 3.5: **Basic Model, Partial Probit.**

			Effects				Stratification				
Scenario			Homogeneous		Heterogeneous		N^o of	Lost	κ	ΔP-	Bal.
(a)	(b)	(c)	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
0.00	300	0.25	0.01	0.40	-0.01	0.54	129.07	2.92	0.82	-0.01	0.96
		0.50	-0.05	0.49	-0.07	0.56	117.00	4.36	0.85	-0.02	0.93
		0.75	0.03	0.46	0.01	0.48	107.81	4.24	0.89	-0.01	0.93
		1.00	0.06	0.45	0.05	0.46	101.21	4.46	0.92	-0.01	0.90
	600	0.25	0.04	0.39	0.04	0.47	144.00	2.24	0.69	-0.01	0.95
		0.50	0.05	0.42	0.03	0.42	130.61	5.02	0.71	-0.03	0.94
		0.75	-0.02	0.44	-0.04	0.40	124.04	6.19	0.74	-0.02	0.94
		1.00	-0.01	0.39	-0.02	0.35	120.12	6.36	0.76	-0.02	0.92
	900	0.25	0.00	0.32	-0.01	0.37	143.61	3.16	0.63	-0.02	0.95
		0.50	0.00	0.37	-0.02	0.35	137.36	5.66	0.66	-0.03	0.95
		0.75	0.02	0.35	-0.00	0.30	130.95	8.91	0.69	-0.03	0.94
		1.00	-0.03	0.40	-0.04	0.33	125.78	11.50	0.70	-0.03	0.92
0.05	300	0.25	0.04	0.34	0.04	0.45	128.47	2.95	0.81	-0.01	0.94
		0.50	0.18	0.46	0.19	0.51	116.53	3.52	0.86	-0.01	0.91
		0.75	0.14	0.50	0.13	0.52	108.82	4.67	0.89	-0.01	0.90
		1.00	0.18	0.51	0.16	0.50	101.76	4.48	0.91	-0.01	0.87
	600	0.25	0.04	0.39	0.04	0.47	143.15	2.35	0.68	-0.01	0.94
		0.50	0.12	0.39	0.11	0.39	134.06	5.59	0.72	-0.03	0.91
		0.75	0.17	0.38	0.14	0.34	123.44	6.57	0.74	-0.02	0.91
		1.00	0.18	0.44	0.14	0.38	120.77	6.40	0.75	-0.02	0.87
	900	0.25	0.02	0.31	0.01	0.36	143.88	3.13	0.63	-0.02	0.93
		0.50	0.20	0.42	0.17	0.38	137.84	5.33	0.66	-0.03	0.93
		0.75	0.20	0.42	0.15	0.35	129.33	9.10	0.68	-0.04	0.90
		1.00	0.30	0.48	0.23	0.38	124.34	12.18	0.69	-0.03	0.89
0.10	300	0.25	0.19	0.46	0.24	0.60	129.41	2.85	0.82	-0.01	0.90
		0.50	0.32	0.50	0.34	0.55	115.86	4.19	0.85	-0.02	0.87
		0.75	0.33	0.56	0.32	0.57	107.64	4.46	0.89	-0.01	0.84
		1.00	0.42	0.65	0.41	0.64	103.05	3.53	0.91	-0.01	0.83
	600	0.25	0.14	0.38	0.15	0.46	144.53	2.49	0.68	-0.01	0.92
		0.50	0.24	0.42	0.22	0.40	132.97	4.84	0.71	-0.02	0.86
		0.75	0.41	0.56	0.35	0.49	125.63	6.26	0.74	-0.02	0.84
		1.00	0.44	0.60	0.37	0.51	121.74	6.22	0.76	-0.02	0.82
	900	0.25	0.17	0.35	0.18	0.39	142.11	3.56	0.63	-0.02	0.88
		0.50	0.29	0.48	0.25	0.43	134.63	5.81	0.66	-0.03	0.86
		0.75	0.38	0.52	0.30	0.43	131.07	9.05	0.68	-0.03	0.85
		1.00	0.41	0.57	0.31	0.44	123.81	11.76	0.70	-0.03	0.83

The results are averages over all 100 iterations. The first block represents the scenario: (a) value of the coefficients $\beta_5, \beta_6, \alpha_3, \alpha_4$, (b) size of control reservoir, (c) selection determination. The next block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ.

Z do not have an impact on the selection or outcome equation, respectively. However, there is an increasing bias if their impact increases. Note that there appears to be also a weak upward bias in the full probit model if selection determination is highest, specifically in the homogeneous case. This bias arises due to the remaining imbalance expressed by a τ of around 0.6. An additional bias of opposite direction emerges in the heterogeneous case partly offsetting the initial bias. This is because many high propensity score treated units who tend to experience a higher effect in the heterogeneous case are discarded by the matching algorithm. Furthermore, the RMSE in the heterogeneous case seems to be as large as or larger than in the homogeneous case. Yet, it is smaller for high selection determination and for large control reservoir. This finding might be explained by the additional variability of a heterogeneous δ_i. Since δ_i depends on the observable covariates, its variability diminishes as selection caused by the observables becomes more important.

As far as balancing success is concerned, no strategy surpasses the other in all scenarios. If selection determination is weak full probit always achieves perfect balance. However, its performance diminishes quickly as selection determination is growing. On the other hand, partial probit's balancing success starts worse but does not reduce as fast as full probit's. Part of this finding is explicable by the choice of the caliper width ε. It is wider for strong selection (see table 3.3), hence, treated individuals might choose controls with a relatively low propensity score. For the same reason, τ deteriorates faster in the full than in the partial probit model. However, a constant ε for all scenarios would have produced a large casualty list of treated units in the full model.

In spite of non-constant ε, the full probit loses more treated units such that the relative difference in the *true full* propensity scores between treated individuals before matching and the remaining treated after matching $\Delta P.score$ is more pronounced than in the partial probit. The negative signs show that treated individuals are lost in the high end of the propensity score scale. However, while partial probit matching never exceeds 3%, full probit matching even reaches 10%. Note, however, that the number of lost treated increases with the size of the control reservoir. This counterintuitive result arises because of decreasing caliper widths, see table 3.3.

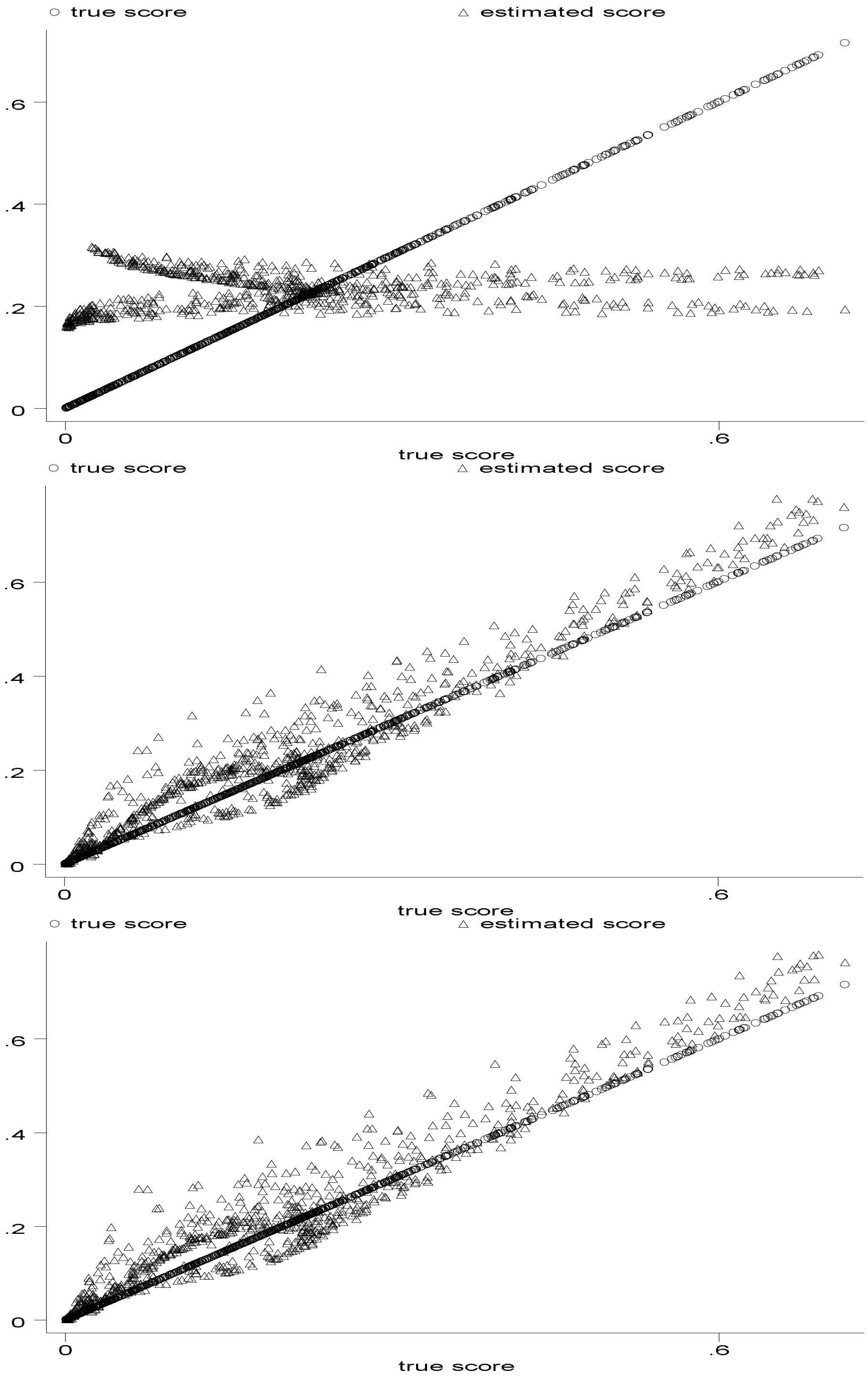
In sum, partial probit produces a better overall performance than full probit for the examined parameter constellations. Alas, if the coefficients of Y and Z grew above the 0.1 considered here, full probit could be expected to be the preferred strategy. Moreover, if there is no strong selection into treatment full probit matching is not at a disadvantage, in contrast, it even sometimes outperforms partial probit. Yet, strong selection as in Dehejia & Wahba (1998) or Chapter 2 calls for a careful assessment of the importance of the variables included in the selection equation.

The basic model seems to be overly optimistic as far as the distributions of Y and Z are concerned. The top panel of figure 3.2 presents propensity score estimates under the alternative partial probit model. Apparently, it underestimates propensity scores for individuals with high $\mathbb{P}(D = 1|X)$ and overestimates for those with low scores. This is in contrast to the next two pictures which present estimated propensity scores built on probit models with higher order terms. Pictures of the full probit are not presented for they are virtually identical to those of the basic model.

One might ask whether the order of treated and untreated units with regard to their estimated biased propensity scores would be similar to the order of individuals in accordance with their true scores $\mathbb{P}(D = 1|X)$. In this case, treated and untreated would hardly change their ranks within the sample. As a result, stratification might be similar to that if the true scores were used for matching and the biased propensity score estimates would not be a source of bias in the matching estimates. However, as illustrated in table 3.6, the mean rank of treated units has diminished considerably for the alternative model with no higher order terms implying that a large number of untreated and treated units must have interchanged their ranks. Alas, once higher order terms are taken into account – particularly interaction between X_1 and X_2 – there is no difference in mean ranks worth mentioning anymore.

Table 3.7 presents simulation results for the alternative model in simulation scenario $(0.05, 600, 0.75)$ for the full and the partial probit. Consider first the partial probit results. Surprisingly, they are still better than the comparable ones of the full probit basic model, though worse than those of the partial probit basic model. Interactions in the *outcome equation* (3.9) lead to an increase of the RMSE and produce a larger bias if no higher order

Figure 3.2: **Alternative Model, Partial Probit With and Without Higher Order Terms**



True and estimated propensity scores on the vertical axis versus true scores on the horizontal. The figures represent *one* iteration of the simulation study. The first picture shows results when no interactions are included; the second contains interactions, the third additionally contains squares in X.

Table 3.6: **Mean Ranks.**

	Full probit		Partial probit	
	untreated	treated	untreated	treated
Basic Model				
True propensity score	313.59	625.75	332.14	550.78
Estimated propensity score	313.28	627.01	331.79	552.19
Alternative Model				
True propensity score	329.87	557.76	335.14	536.65
Estimated, without higher order terms	328.67	562.54	367.40	407.94
Estimated, w/ interactions	328.67	562.54	336.59	530.82
Estimated, w/ inter. & squares	328.67	562.54	335.43	535.44

The results are mean ranks in the treatment and in the comparison group. They are further averaged over all iterations.

terms in the probit model are accounted for. Yet, this pattern disappears once they are included. Asymmetry in the coefficients $(\beta_1, \beta_2) = (0.5, 2)$ instead of $(1, 1)$ of the response equation does not at all alter the results which is why they are omitted. In contrast to the basic model, heterogeneous effects lead to substantially worse estimation results in that biases and RMSEs are markedly larger than in the homogeneous case.

These still surprisingly favorable results in spite of severe misspecifications expressed in the first picture of figure 3.2 might be explained by the fact that within the propensity score calipers the Mahalanobis distance, which is not misspecified, still matches the correct individuals. To explore this hypothesis all results are repeated replacing the Mahalanobis distance by the propensity score distance within calipers. The results are also shown in table 3.7. They are fairly similar to the previous results with one notable exception: the bias and RMSE are markedly larger in case interactions in the response model are introduced but none in the probit model.

For the sake of comparability, the table displays estimates of the full probit model, as well. The most striking result is that it achieves an almost perfect overall balance τ. This unexpected finding, however, may partly be explained by the fact that a considerable number of high propensity score treated units is lost facilitating balancing the variables of

Table 3.7: **Alternative Model.**

	Effects				Stratification				
	Homogeneous		Heterogeneous		N^o of	Lost	κ	ΔP-	Bal.
	Bias	RMSE	Bias	RMSE	strata	tr'd		score	τ
Full Probit									
Mahalanobis distance within calipers									
– no interactions in outcome equation									
	0.08	0.48	0.05	0.80	118.82	13.51	0.74	-0.12	0.98
– with interactions in outcome equation									
	0.11	0.53	0.11	0.87	118.82	13.51	0.74	-0.12	0.98
Propensity score distance within calipers									
– no interactions in outcome equation									
	0.08	0.51	0.06	0.84	113.16	13.51	0.75	-0.12	0.98
– with interactions in outcome equation									
	0.07	0.56	0.04	0.92	113.16	13.51	0.75	-0.12	0.98
Partial Probit									
Mahalanobis distance within calipers									
– no interactions in outcome equation									
1	0.14	0.43	0.23	0.72	134.37	0.50	0.72	-0.00	0.91
2	0.16	0.47	0.24	0.78	128.72	7.82	0.72	-0.05	0.92
3	0.19	0.49	0.30	0.81	128.16	6.58	0.73	-0.04	0.93
– with interactions in outcome equation									
1	-0.33	0.56	-0.55	0.94	134.37	0.50	0.72	-0.00	0.91
2	-0.06	0.48	-0.13	0.81	128.72	7.82	0.72	-0.05	0.92
3	-0.03	0.48	-0.07	0.80	128.16	6.58	0.73	-0.04	0.93
Propensity score distance within calipers									
– no interactions in outcome equation									
1	0.13	0.42	0.21	0.71	140.85	0.50	0.71	-0.00	0.92
2	0.12	0.49	0.17	0.81	127.86	7.82	0.73	-0.05	0.92
3	0.18	0.47	0.27	0.77	126.94	6.58	0.73	-0.04	0.92
– with interactions in outcome equation									
1	-0.80	0.95	-1.34	1.58	140.85	0.50	0.71	-0.00	0.92
2	-0.10	0.52	-0.21	0.88	127.86	7.82	0.73	-0.05	0.92
3	-0.05	0.46	-0.11	0.78	126.94	6.58	0.73	-0.04	0.92

The means are averages over all 100 iterations for scenario (0.05, 600, 0.75) of table 3.5. The first column refers to the partial probit model, 1: no higher order terms, 2: interactions, 3: interactions and squares. The first block reports bias and RMSE for the homogeneous and the heterogeneous case. The last block shows stratification results: the number of strata and of lost treated units, the stratification measure κ, the difference in true propensity scores of treated units after and before matching, and the aggregate balance τ. *Interactions in outcome equation* means that $(\beta_{12}, \beta_{34}, \beta_{56}) = (1, 1, 1)$.

the remaining sample. As a result, the superior balance is accompanied by an unfavorably $\Delta P.score$ of 12% making the matched sample less representative. Similarly, κ is almost as small as in the partial probit model because it merely reports uniformity of the realized stratification given the number of lost treated units. Finally, using a propensity score distance within calipers does not alter the results except for slightly increased RMSE. In sum, the partial probit does not do worse than the full probit even if the partial probit model is severely misspecified. Including higher order terms into the selection equation might be a way to alleviate problems caused by omission of variables which are correlated with the included ones.

3.6 Conclusion

This chapter investigates propensity score matching when selection into treatment is remarkably strong and thus the treatment and comparison group differ considerably in their observable covariates. In such a scenario, matching adequate units is demanding. To alleviate this problem, we suggest to carefully reconsider the selection equation with respect to variables that might play a subordinate role in the outcome equation. Omission of these variables helps increase the randomness of the selection process and reduce the variance of the matching estimates. However, their omission from the selection equation might lead to inconsistent propensity score estimates and hence biased matching estimates. This study assesses the bias-variance trade-off in a simulation resting on the mean squared error criterion.

To this end, we presuppose existence of variables Z which strongly influence the selection decision but which, on the other hand, do not or do only weakly determine the outcome under scrutiny. For a large enough sample size, specification tests of the probit model would then recommend the inclusion of Z to consistently estimate the propensity score. Likewise, we introduce variables Y which are relevant to the outcome but irrelevant to the participation decision. Matching on a propensity score estimate based on Z and Y will balance Z at the expense of balance of the variables most relevant for both the outcome and the selection. Moreover, unnecessary effort is spent to remove small imbalance in the variable Y. In consequence, (i) some treated have to be systematically discarded from the sample because they do not find adequate controls and, (ii) more treated have to share one control, a fact that reduces uniformity of the stratification and thus increases standard errors.

In effect, the results show that matching on inconsistent estimates of the propensity score, i.e. those achieved when Z (and Y) are excluded, produces estimation results of the mean effect of treatment that are often better in terms of the RMSE than those achieved by matching on estimates that rest on all covariates relevant for the selection. This remains true even if Z shows some impact on the outcome as long as this impact is limited. Drake (1993) points to a similar direction in concluding that misspecifying the propensity score results in smaller biases than misspecifying the response model. Therefore, we recommend to only include variables into the selection equation that are highly significant. Variables with low significance levels are obvious candidates for exclusion even if they might play a role in the outcome equation. Moreover, if established research suggests that certain variables Z are irrelevant to the outcome under study they should solely be included into the selection equation if there are other strong reasons for doing so.

If, nevertheless, imbalance of some variables seems to be inacceptable after matching, an additional linear regression adjustment might be pursued with presumably less cost than balancing all the remaining variables in advance. If misspecification of the propensity score seems to be inacceptable, one might additionally take account of statistically significant higher order terms of those variables included in the selection equation. A sensitivity analysis that compares partial models with the full model might be a way to assess different approaches, see e.g. Heckman, Ichimura & Todd (1997: section 13) or Chapter 4. In sum, the main criterion of success for matching remains the balance of the relevant covariates and not the proper estimation of the selection equation. This aim is easily obtained by a full probit model only if selection determination is low and/or the control reservoir is large but in several applied situations it might be better obtained by a partial model.

Chapter 4

Evaluating the Effect of Postsecondary Education

June 1999/October 2000

Abstract. This chapter uses the statistical technique of matching on the propensity score to evaluate the effect of the associate's, the bachelor's, and graduate degrees on hourly wages for men and women during the first ten years after college completion. Moreover, it discusses heterogeneity in the effects ruled by ability and family background. Selection into college education turns out to be extremely strong, notably for the bachelor's and the graduate degrees. As a result, bias in observable covariates prior to matching is immense. An optimal full matching algorithm is implemented to address this issue. Furthermore, sensitivity with respect to the specification of the selection equation is investigated. All results are compared to conventional OLS estimation which allows (i) to verify the linear specification of the earnings equation and (ii) to bring out the determinants of why matching and OLS estimates differ.

4.1 Introduction

The debate on the subject of identification of the returns to education has a long tradition in labor economics. Mincer (1974) specified a theoretical model where log earnings are a linear function of education, labor market experience, and experience square. Numerous studies have estimated the coefficient of the schooling variable by least squares techniques, see Willis (1986), Ashenfelter & Rouse (1998b), and Card (1999) for a comprehensive overview. This chapter investigates the return to college degrees by means of the nonparametric technique of matching. To this end, a somewhat different concept of the return to education is introduced, namely the *effect* of college education on earnings which takes account of the effect of education on labor market experience, as well.

The ideal setup for identifying the effect of schooling on earnings would be a randomized experiment with a treatment group that receives education and a control group that is refused access to education. For obvious reasons, however, such randomization is impossible and one has to rely on *observational studies*. In an observational study, individuals themselves decide whether to participate in treatment or not, thus self-selection into treatment poses a major problem. Matching treated and untreated individuals with respect to all observable variables that both determine the selection into college and exhibit an impact on earnings removes systematic observable differences between the treatment and the comparison group. In balancing these variables matching mimics a randomized experiment provided the relevant variables are observed. Matching on the probability to participate in treatment, the *propensity score*, is an alternative whenever the covariates are high dimensional, see Rosenbaum & Rubin (1983).

This chapter performs matching on the propensity score to evaluate the effect of the associate's, the bachelor's, and graduate degrees on hourly wages for men and women during the first ten years after college completion. Graduate degrees subsume master's, professional, and doctoral degrees. Heterogeneity in the effects ruled by ability and family background is explicitly modeled and discussed. Selection into college education turns out to be extremely strong, notably for the bachelor's and the graduate degrees. As a result, bias in observable covariates prior to matching is immense. The optimal full matching algorithm proposed by Rosenbaum (1991) is implemented to address this issue. A full matching seems to be most suitable when it comes to evaluate the mean effect of treatment on the treated (see also Chapter 2). Furthermore, sensitivity with respect to propensity score estimation as discussed in Chapter 3 is being investigated. All results are compared to conventional OLS estimation which allows (i) to verify the linear specification of the earnings equation and (ii) to bring out the determinants of why matching and OLS estimates deviate.

The structure of the remainder is as follows. The second section elucidates the estimation strategies applied in this chapter. Section 3 describes their practical implementation, specifically the matching algorithm, and presents the data drawn from the *National Longitudinal Survey of Youth 1979*. Finally, results for all three college degrees are presented and discussed in section 4. Section 5 summarizes the findings and concludes. The appendices are dedicated to the estimation of the propensity score, to a more detailed description of the statistical tools, and to estimation results for the graduate degrees.

4.2 Estimation Strategies

Identifying the effect of college education is primarily a question of identifying a causal relationship between earnings and education which requires specifying a counterfactual state of the world. Individuals who opted for post-secondary education, henceforth alternatively called *treatment*, have to be compared to themselves had they not opted for post-secondary education. Given a certain outcome measure, frequently the hourly rate of pay, the difference in outcomes between both states is the individual effect of the treatment. Several strategies have been proposed to address this identification problem, see, for example, Angrist & Krueger (1999) for an overview.

A formal description helps illuminate the idea. Concentrating on the hourly rate of pay as response variable, education can be considered as an investment into human capital that, on average, increases wages. Indeed, there is strong international evidence of a positive correlation between schooling and earnings (see e.g. Psacharopoulos, 1994).

However, there is a long and ongoing debate on whether this correlation is an expression of a causal relationship between education and earnings or whether it is spurious, see the discussion in Griliches & Mason (1972) and Griliches (1977). Confounding variables might substantially mislead the causal interpretation of the education-earnings relationship. For instance, family background might be an important determinant of the choice of the amount of education but might also be a direct component of the earnings equation. The same holds for personal innate earnings abilities which might even more so determine both the selection into higher education and wages. Describing schooling as a dummy variable S taking the value 1 if a certain amount of education is acquired and 0 if not, the earnings equation can be summarized as follows

$$R = f(S, E(S), F, A) \tag{4.1}$$

where F and A denote family background and abilities, respectively, and $E(S)$ represents labor market experience measuring cumulative training on the job as a further investment into human capital. It will itself depend on the schooling decision in that E tends to be lower for a certain point in time if more schooling is acquired.

Prior to the schooling decision there are two potential states

$$R^1 \; = \; f(S = 1, E(S = 1), F, A) \tag{4.2}$$

denoting potential earnings when the amount of education $S = 1$ would be acquired, and

$$R^0 \; = \; f(S = 0, E(S = 0), F, A) \tag{4.3}$$

denoting potential earnings in case $S = 0$ would be chosen. Hence, the individual effect of education equals $R^1 - R^0$. This framework has become known as the *potential outcome approach to causality* suggested by Roy (1951), Rubin (1974, 1977), and Holland (1986). It requires that the response of an individual be independent of the schooling decisions of all other individuals. This implies that there are only two potential outcomes, namely R^0 and R^1, one for the personal state $S = 0$, and one for $S = 1$, respectively. There are no further potential outcomes depending on the assignment of any other individual. This requirement is often referred to as *stable unit treatment value assumption* (sutva, see Rubin, 1986).

Unfortunately, only one of the two potential responses can ever be observed, and the individual treatment effect $R^1 - R^0$ cannot be identified without imposing extraordinarily strong assumptions as, for example, a constant treatment effect for everybody. Therefore, focus will be on the mean effect of education on those who opted for college education

$$\mathbb{E}(R^1 - R^0|S = 1) = \mathbb{E}(R^1|S = 1) - \mathbb{E}(R^0|S = 1).$$

The conditional expectation $\mathbb{E}(R^1|S = 1)$ is identified in the subsample of treated individuals. However, the counterfactual $\mathbb{E}(R^0|S = 1)$ is merely identified when further assumptions are invoked. One is referred to as the *conditional independence assumption.* If all covariates F and A that determine both the outcome under scrutiny and the selection into schooling are known, R^0 is independent of S given (F, A), yielding $\mathbb{E}(R^0|F, A, S = 1) = \mathbb{E}(R^0|F, A, S = 0)$. The conditional mean response of the *educated* $S = 1$, if they had opted for $S = 0$, can thus be inferred for given (F, A) from the conditional mean response of the *less educated* individual $S = 0$, who are observed in schooling level $S = 0$.

However, if some variable in equation (4.1) cannot be observed, which is usually the case for A, the conditional independence assumption might be invalid. The literature suggests three distinct methods to cope with this problem. First, the instrumental variables technique is a prominent – though often statistically imprecise – way to address this issue (see e.g. Angrist & Krueger, 1991, Card, 1995a). A second approach rests on comparing education and earnings levels between individuals who are supposed to be equal with respect to unobserved A. This idea is best implemented in twin studies but suffers particularly from measurement error in the schooling variable (see e.g. Griliches, 1979, Ashenfelter & Krueger, 1994). Third, some observed ability measures as, for instance, scores on mathematical ability tests are explicitly included into equation (4.1). However, these variables might themselves be prone to either endogeneity or measurement error (see e.g. Griliches & Mason, 1972, Griliches, 1977, Blackburn & Neumark, 1995, Murnane, Willett & Levy, 1995, Chapter 5).

This chapter combines the second and third approach but does not consider the special econometric problems that occur with these strategies. It rather focuses on the functional form of f by comparing three estimation methods: a conventional linear model, and two variants of the matching approach, a *pure matching* and a *regression adjusted matching.* Recently, the statistical method of matching has found widespread attention in econometrics, especially for evaluating active labor market programs. For a discussion, see e.g. Heckman, LaLonde & Smith (1999). Rosenbaum (1995) summarizes the statistical literature. Usually, the parameter of interest is *the mean effect of treatment on the treated*, which in this study translates to the mean effect of college education on those individuals who went to college.

Matching

Matching tries to mimic *ex post* a randomized experiment by stratifying the sample of treated and untreated units with respect to the relevant covariates (F, A) that rule the selection into treatment as well as the outcome under study.[1] As a result, matching balances the relevant covariates between treatment and comparison group to achieve comparability. In other words, selection into treatment can be considered to have been random within each stratum defined by (F, A). In contrast, in a true randomized experiment all covariates are *a priori* balanced up to stochastic deviations.

Although the data used in this study provide detailed information on F and A, further complications emerge in practice. It is almost impossible to match individuals with exactly the same covariates whenever (F, A) is of high dimension. To escape this *curse of dimensionality*, Rosenbaum & Rubin (1983) suggest to use the one-dimensional conditional probability to participate in treatment $p(f, a) = \mathbb{P}(S = 1|(F, A) = (f, a))$, the *propensity score*, on which to stratify the sample instead. They show that if R^0 is independent of S given (F, A), R^0 and S are also independent given $p(F, A)$. Matching treated and untreated units with the same propensity score and putting them into one stratum means that the decision whether to go to college or not can be considered as having been random in the stratum. With probability $p(F, A)$ members of a given stra-

[1] In this study, the strata will be non-overlapping, i.e. individuals cannot be in more than one stratum, which facilitates statistical inference. However, this is not necessary if different techniques are used in case of overlapping strata, see Quade (1981) or Heckman, Ichimura & Todd (1998).

tum attend college and with probability $1 - p(F, A)$ they do not. Section 3 discusses the practical problems arising from the stratification specific to this study and appendix B provides a brief framework for statistical inference which is established in Rosenbaum (1995) and adapted to this setup in Chapter 2.

The mean effect of college education is estimated by a weighted average over all stratum effects. The stratum weight corresponds to the number of treated individuals within the stratum multiplied by their sample weights provided by the data set. The resulting estimate will be labeled the *pure matching estimate.* However, it turns out that balance of covariates is not always fully achieved which is why the pure matching estimate might still be biased. Therefore, an additional regression adjustment is made based on the stratum as unit of observation. Let

$$\begin{aligned} R_t &= \alpha_0 + \alpha_1 F_t + \alpha_2 A_t + \delta(F_t, A_t) + \varepsilon_t, \qquad (4.4) \\ R_c &= \alpha_0 + \alpha_1 F_c + \alpha_2 A_c + \varepsilon_c \end{aligned}$$

be the wage of the treated t and the control c, respectively, in a certain stratum. $\delta(F_t, A_t)$ denotes the treatment effect which may vary with F and A.

If overall balance is not achieved, the difference between R_t and R_c keeps on depending systematically on the unbalanced covariates

$$\Delta R = \alpha_1 \Delta F + \alpha_2 \Delta A + \delta(F_t, A_t) + \Delta\varepsilon,$$

with $\Delta F = F_t - F_c$, and analogously ΔA, $\Delta\varepsilon$. Further assume

$$\delta(F_t, A_t) = \delta_0 + \delta_1(F_t - \bar{F}_t) + \delta_2(A_t - \bar{A}_t) \qquad (4.5)$$

where bars denote the respective mean over all *treated* units. The mean effect of college education is thus captured by δ_0. Finally, a regression of ΔR on differences in the covariates and on the level of the treated units' covariates is supposed to remove any bias due to incomplete balance if linearity holds. Note that linear regression after matching might be less prone to functional form misspecifications since linearity is only required to bridge remaining small differences in the variables.[2] What is more, heterogeneity in the effect

[2] A linear Taylor approximation between two points of a well-behaved function is the more accurate the closer two values of the function are.

of college education can be investigated by assessing the statistical significance of δ_1 and δ_2. For this latter reason, pure matching, too, is followed by an additional regression on the level variables F_t and A_t, see also Chapter 2. Since in this study strata will contain either one treated and one or more controls or one control and more than one treated, (R_t, F_t, A_t) and (R_c, F_c, A_c) of equation (4.4) are understood as averages over the treated and untreated individuals within strata, respectively.

OLS Regression

Reconsider equation (4.1). In a linear model, it specializes to

$$R = \gamma_0 + \beta(F, A)\ S + \alpha_1 E(S) + \alpha_2 E^2(S) + \gamma_1 F + \gamma_2 A + \varepsilon. \tag{4.6}$$

Analogously to equation (4.5), the treatment coefficient β depends linearly on F and A. Yet, it does not identify the mean effect of college education, $R^1 - R^0$, for treatment also acts through accumulated experience $E(S)$. Individuals tend to acquire less labor market experience while they attend college. If they do work while being enrolled their acquired experience might be less valuable because many college students work part-time or during vacations without much training on the job. As such, experience acquired while being enrolled at college might differ from experience gained in the labor market after education is completed. Hence, usual work experience $E(S)$ is distinguished from experience acquired while being enrolled at college $E_C(S)$.

Furthermore, since experience is a cumulative measure, missing information in certain years would accumulate at the end of the sample period. Therefore, missing value indicators $mis(S)$ and $mis_C(S)$ for $E(S)$ and $E_C(S)$, respectively, are introduced in equation (4.6) counting the number of years with missing information up to the year under scrutiny. Finally, the experience-earnings profile for high school and college graduates might differ, which is taken into account by adding interaction terms between experience and schooling. In sum, equation (4.6) extends to

$$\begin{aligned} R &= \gamma_0 + \beta(F, A)S + \alpha_1 E(S) + \alpha_2 E^2(S) + \alpha_3 E_C(S) + \alpha_4 E_C^2(S) \\ &+ \alpha_5 S \cdot E(S) + \alpha_6 S \cdot E^2(S) + \alpha_7 S \cdot E_C(S) + \alpha_8 S \cdot E_C^2(S) \end{aligned} \tag{4.7}$$

$$+ \quad \alpha_9 \, mis(S) + \alpha_{10} \, mis_C(S) + \gamma_1 F + \gamma_2 A + \varepsilon.$$

The mean effect of college education can thus be estimated by

$$\begin{aligned} \bar{R}^1 - \bar{R}^0 \quad = & \quad \hat{\beta}(\bar{F}, \bar{A}) + \hat{\alpha}_1 \, (\bar{E}(1) - \bar{E}(0)) + \hat{\alpha}_2 \, (\overline{E^2}(1) - \overline{E^2}(0)) \\ + & \quad \hat{\alpha}_3 \, (\bar{E}_C(1) - \bar{E}_C(0)) + \hat{\alpha}_4 \, (\overline{E_C^2}(1) - \overline{E_C^2}(0)) \\ + & \quad \hat{\alpha}_5 \, \bar{E}(1) + \hat{\alpha}_6 \, \overline{E^2}(1) + \hat{\alpha}_7 \, \bar{E}_C(1) + \hat{\alpha}_8 \, \overline{E_C^2}(1) \qquad (4.8) \\ + & \quad \hat{\alpha}_9 \, (\overline{mis}(1) - \overline{mis}(0)) + \hat{\alpha}_{10} \, (\overline{mis_C}(1) - \overline{mis_C}(0)). \end{aligned}$$

Bars denote means over the *treated* subsample except for the experience variables in the no-treatment state which are averages over all untreated units because the OLS framework does not provide suitable counterfactual averages.

4.3 The Practical Implementation

The Data

The data are taken from the *National Longitudinal Survey of Youth 1979* (NLSY) administered by the US Bureau of Labor Statistics. The NLSY is a sample of 12,686 youths first interviewed in 1979 when they were aged between 14 and 22 and re-interviewed annually until 1994. A detailed description of the data is given by the NLS Handbook (1997) and the NLSY79 User's Guide (1997). For this study, data on wages are extracted until 1994 for men and women; the military subsample is skipped.[3] Oversampling of Non-whites and economically disadvantaged Whites suggests the use of the sample weights of 1979.

The outcome measure is the hourly rate of pay inflated to 1996 dollars using the US consumer price index and transformed into logarithms. To eliminate outliers, all values below $1 are set equal to $1 and maximum or minimum wages of observations whose wages oscillate enormously across years are removed.[4] In particular, the data contain numerous

[3] The self-employed are kept. Kane & Rouse (1995) who also use the NLSY report that their results are not sensitive to the exclusion of self-employed.

[4] For example, an hourly wage of $5 in one year, $1000 in the second, and again $5 in the third seems more likely to reflect inconsistencies in the calculation of the hourly wage by the NLSY than real fundamental economic changes which is why $1000 would be removed. See e.g. the NLSY79 User's

variables about the socioeconomic background, the high school careers of respondents, and labor force status (since 1975). The latter is used to generate a measure of actual experience based on weeks worked per year. What is more, the NLSY provides information on some ability measures collected in 1980 when 94.3% of all respondents participated in tests to update the *Armed Services Vocational Aptitude Battery* (ASVAB) consisting of ten different test scores. Since respondents participated in the tests at different ages the scores are adjusted by regressing the raw scores on age dummies and using the residuals subsequently, analogous to Blackburn & Neumark (1993).

In terms of the formal setup, treated individuals are those who obtained a college degree and left college immediately thereafter. Those who attempted to continue or start college to obtain a (further) degree but eventually dropped out before achieving it are neither considered as treated nor as potential control units and are removed. Potential controls are individuals with only a high school diploma who never attended college. High school dropouts and individuals with a *general educational development* are removed from the sample.

Matching

Three college degrees are evaluated: the *associate's degree* (AA) which is obtained at two-year colleges, the *bachelor's degree* (BA) which is usually obtained after four years at college and the *graduate degrees* (MA) which, for example, include the master's, doctoral and professional degrees. Unfortunately, the number of persons in the latter group remains too low to draw sensible statistical inference.

The year in which respondents receive their high school diploma marks the beginning of the treatment phase of those who went to college.[5] In turn, the year in which the treated units receive their college degree marks the end. An exception is the graduate group which also contains individuals who continued college beyond a graduate degree

Handbook (1997: p. 266): "... the calculation procedure [...] produces, at times, extremely low and extremely high pay rate values."

[5] A considerable number of individuals do not start college right after finishing high school. Roughly 45% wait more than one year. This means that they are compared with their controls for a period that comprises more than merely the time at college.

but then dropped out. These college dropouts are kept in order not to reduce the sample size even more and because this group comprises various degrees of different time lengths anyway. Within a stratum, treated and controls are ideally supposed to finish high school in the same year and to be of the same age and race. After high school, the control starts to work and gain labor market experience while the treated is allowed to either go to college right away, interrupt it for a while or even start to work a certain amount of time before finally attending college. Moreover, individuals of the same stratum should have similar propensity scores.

The propensity score is estimated by ways of a probit model. Although such a parametric approach to modeling the selection equation seems to dilute the idea of matching as being nonparametric, the specification of the selection equation is in fact of minor importance as long as the estimated propensity score achieves to *balance all relevant* covariates. Albeit, it is of major importance which covariates are really required to be included in the selection equation. Chapter 3 suggests not to include all the possibly numerous variables that might determine the selection, even if they are statistically significant, but to consider only those that are relevant for the outcome as well. Above all, in samples with a relatively low number of adequate untreated units there is a trade-off between balancing the most important variables and consistent specification of the selection equation. Balancing irrelevant variables might well be at the expense of balancing the important ones. In this context, "adequate units" means untreated persons with similarly high propensity scores as the treated.

Because of that, this chapter offers two distinct specifications, a *broad* and a *narrow* probit model. The first comprises several variables describing F and A which are statistically significant in the probit estimation while the latter uses only one variable for each F and A, namely parents' education[6] as family background and math test scores of the ASVAB as ability measure. The probit estimations are discussed in appendix A. Yet, matching will be pursued with respect to the *index* or *probits* $\Phi^{-1}(\hat{p})$, where Φ is the cumulative normal density function, in place of the estimated propensity score $\hat{p}$. The index

[6] Parents' education is defined as the mean of father's and mother's education. It is mother's if father's is missing and vice versa.

is linear in (F, A) and might better reflect the diversity of individuals at the boundaries than $\hat{p}$ which is constrained to the unit interval, see also the discussion in Chapter 2.

Table 4.1 presents the distribution of treated and untreated individuals by certain index intervals. Obviously, the probit models for the decision to take up a bachelor's or a graduate degree clearly separates the college from the high school graduates. Unfortunately, this high predictive ability of the probit model also implies that it will be difficult to find enough controls for treated individuals characterized by high propensity scores. The picture differs for members of the group with an AA who are more resembling to the high school graduates; therefore, they will be easier to match. Moreover, although the narrow and broad probit model do not differ much with regard to the estimated mean propensity scores it is evident that the broad model produces an even less favorable prior-distribution than the narrow one. What is more, the broad one has to rely on a smaller sample size for it depends on more covariates coming with more missing observations.

Once the propensity score has been estimated, a distance between treated and untreated individuals has to be defined. Within cells characterized by race, sex, age, and the high school graduation year, a propensity score caliper approach is pursued.[7] Only individuals of the same age, one year younger or one year older are allowed to be matched. Similarly, only those who receive their high school degree in the same year, one year earlier or later than the treated might become potential controls. Exact matches on these variables would be preferable, but would substantially reduce the number of potential controls. Furthermore, only individuals of the same race and sex are matched. Three races are distinguished, Blacks, Hispanics, and Non-black/Non-hispanics, subsequently called Whites. Results are presented separately for men and women, but not for races [8].

Within these cells a pool of potential controls is generated for each treated by excluding all untreated units who exceed an index score caliper ε. The final decision of who becomes an actual control will then be made by minimizing the index score distance. Thus, the

[7] Propensity score calipers are discussed in Rosenbaum & Rubin (1985: 3) and Rosenbaum (1989: 3.4).

[8] Apart from the inacceptable reduction of the sample size if one considered Blacks and Hispanics separately, Ashenfelter & Rouse (1998b) find that there is little variability in the estimates of the return to schooling (annual earnings) by race.

Table 4.1: **Distribution of the Estimated Index Score.**

	AA				BA				MA			
Estimated	Narrow		Broad		Narrow		Broad		Narrow		Broad	
Index	C	T	C	T	C	T	C	T	C	T	C	T
Men												
$[-6.50\ ,\ -5.00)$	0	0	0	0	0	0	0	0	17	0	3	0
$[-5.00\ ,\ -3.50)$	0	0	0	0	11	0	39	0	378	1	320	1
$[-3.50\ ,\ -2.50)$	0	0	0	0	173	3	288	2	511	3	504	4
$[-2.50\ ,\ -1.75)$	163	6	310	7	446	12	408	7	264	5	256	3
$[-1.75\ ,\ -1.00)$	1100	96	829	77	449	42	310	34	163	12	140	10
$[-1.00\ ,\ -0.25)$	228	68	199	64	263	68	188	50	85	16	64	9
$[-0.25\ ,\ +0.50)$	3	3	10	7	113	114	75	86	38	39	36	32
$[+0.50\ ,\ +1.50)$	0	0	0	0	30	208	30	178	6	72	2	67
$[+1.50\ ,\ +2.50)$	0	0	0	0	1	43	4	85	0	22	0	30
$[+2.50\ ,\ +3.50]$	0	0	0	0	0	0	0	8	0	0	0	0
Mean index	-1.37	-1.10	-1.43	-1.04	-1.51	0.32	-1.77	0.61	-2.74	0.39	-2.72	0.59
Mean p. score	0.10	0.15	0.09	0.17	0.13	0.62	0.11	0.67	0.04	0.64	0.04	0.69
Observations	1494	173	1348	155	1396	458	1342	450	1462	170	1325	156
Women												
$[-6.50\ ,\ -5.00)$	0	0	0	0	0	0	0	0	22	0	10	0
$[-5.00\ ,\ -3.50)$	0	0	0	0	9	0	10	0	293	0	296	0
$[-3.50\ ,\ -2.50)$	0	0	7	0	172	2	237	2	524	1	527	1
$[-2.50\ ,\ -1.75)$	230	9	271	6	420	15	444	9	392	7	319	8
$[-1.75\ ,\ -1.00)$	943	96	799	87	508	44	417	34	166	18	152	12
$[-1.00\ ,\ -0.25)$	367	101	318	88	281	84	203	76	105	33	78	24
$[-0.25\ ,\ +0.50)$	18	29	33	36	131	133	76	86	34	31	28	34
$[+0.50\ ,\ +1.50)$	0	0	0	2	33	212	37	169	5	39	7	31
$[+1.50\ ,\ +2.50)$	0	0	0	0	0	33	2	94	0	21	0	27
$[+2.50\ ,\ +3.50]$	0	0	0	0	0	0	0	13	0	0	0	1
Mean index	-1.29	-0.89	-1.32	-0.82	-1.45	0.27	-1.63	0.56	-2.60	0.11	-2.66	0.29
Mean p. score	0.12	0.21	0.12	0.24	0.14	0.61	0.12	0.65	0.05	0.54	0.04	0.57
Observations	1558	235	1428	219	1469	495	1426	483	1541	150	1417	138

Comparison of the number of untreated (C) and treated (T) individuals by certain estimated index score intervals.

distance is defined as

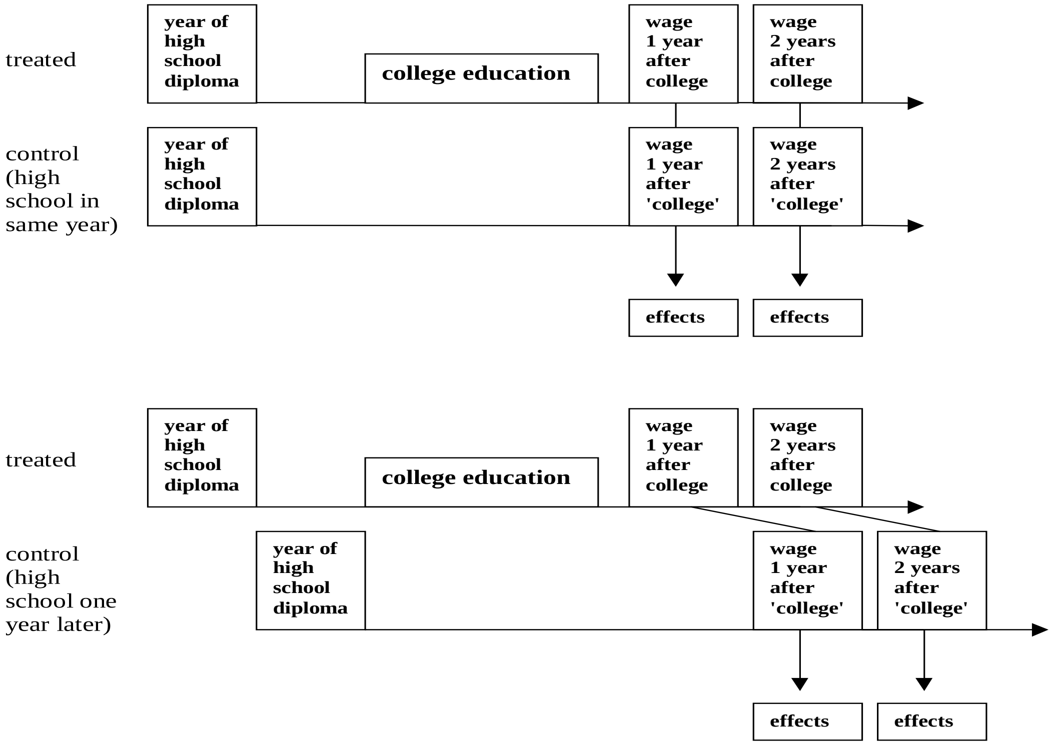
$$d(\mathbf{x}_t, \mathbf{x}_c) = \begin{cases} \infty & \text{if} \quad |p(x_t) - p(x_c)| > \varepsilon \\ \\ |p(x_t) - p(x_c)| & \text{else,} \end{cases} \tag{4.9}$$

where x_t and x_c denote the matching covariates, and $p(\cdot)$ represents the index score for given covariates. An infinite distance indicates that matching is forbidden. The caliper width ε will be set equal to 0.2 for the associate's degree and to 0.4 for the other degrees. The latter caliper width lies close to 0.3 which has been chosen as the narrow width in Chapter 2. A comparatively narrow width is advantageous when it comes to balance of covariates but might drop many treated individuals who do not find controls within their caliper. Yet, results for male BA holders in Chapter 2 indicate that the loss of treated units does not lead to adverse consequences. Note that no calipers might allow arbitrarily large distances between treated and controls, and, moreover, matching algorithms would consume substantially more time.[9]

After having constructed the pool of potential controls appropriate wages serving as the counterfactual wage of the treated person are assigned. The time span between the year in which the treated unit receives the college degree and the high school diploma – the treatment phase – is added to the year in which his or her potential controls receive their high school diploma. The result is considered as the counterfactual year in which his or her potential controls would have received a college degree. Note that the treatment phase is not necessarily just the years at college because the treated individual might have interrupted education for a while. Figure 4.1 illustrates the procedure. The counterfactual outcome one year after treatment is the wage of the potential control one year after his hypothetical end of college. If wage information is missing the potential control is dropped for that year after treatment but is still used for other years. If the wage of the treated is missing the treated is removed, too. Ten years after college will be examined and each year will be stratified separately such that individuals who are removed in some year due to missing wage information may still be available in other years.

[9] In each step of the algorithm every treated would have to be compared to the whole control reservoir. Given a caliper, the treated has to be compared to only a small number of suitable untreated units.

Figure 4.1: **Illustration of the Evaluation Procedure.**


treated
year of high school diploma
college education
wage 1 year after college
wage 2 years after college
control (high school in same year)
year of high school diploma
wage 1 year after 'college'
wage 2 years after 'college'
effects
effects
treated
year of high school diploma
college education
wage 1 year after college
wage 2 years after college
control (high school one year later)
year of high school diploma
wage 1 year after 'college'
wage 2 years after 'college'
effects
effects

The first diagram demonstrates the optimal case when treated and control individuals receive their high school diploma in the same year. The second indicates how things change when there is one year difference.

Additionally, a model is estimated which pools all years after college, i.e. δ_0 in equation (4.5) remains different in each year but δ_1 and δ_2 are restricted to be time-invariant.[10] Moreover, the year in which the treated individual obtained the college degree has an impact on the effect of college if there is a general (positive or negative) trend in the returns to education in the economy as a whole. Therefore, the treatment effect in equation (4.5) will additionally depend on the year, Y_C, in which the college degree is obtained, so $\delta = \delta(F, A, Y_C)$.

The final decision regarding the matching procedure is that on the implementation of the chosen matching criteria. The question is how the overall distance between treated and controls is minimized. In this study, *optimal full matching* as proposed by Rosenbaum (1991) is used. First, *full* matching means that all treated and, in particular, all untreated individuals who have finite distances to treated units and for whom information on all

[10] Coefficients of the other covariates in OLS estimations are also restricted to be constant over time.

variables is available are used for stratification. The size of the strata is as small as possible to ensure that the distance of the units within a stratum is not too large. This yields strata with either one treated and one or more controls or one control and more than one treated unit. In the end, strata with very high propensity scores tend to contain more than one treated and strata with low scores tend to consist of a large number of controls.

In contrast, *pair* matching that produces strata with exactly one treated and one control would force individuals to find exactly one partner to be matched to, which, in this study, might produce a long casualty list of treated (and also untreated) units who do not find a match within the cells and the propensity score calipers. As a result, the matched sample might be extremely distinct from the original sample and the estimate of the population mean effect might be biased.[11] Moreover, if untreated individuals were dropped efficiency would be reduced. On the other hand, full matching gives each stratum a weight according to the number of treated persons in the stratum in order to identify the mean effect of treatment on the treated. Since a few strata contain many treated units in case of the bachelor's and graduate degrees, the overall variance increases. Under these circumstances, full matching estimates are less biased at the expense of reduced efficiency.

Second, *optimal* matching means that the sum of distances between treated and untreated individuals is effectively minimized. In many applications, a so-called *greedy* procedure is used. In the case of greedy pair matching, for example, a treated unit would be randomly chosen and the closest untreated would be searched in the control reservoir and matched to the treated. The resulting pair would then be removed and the procedure would restart. The outcome of matching would be determined by the random order of records in the sample. Rosenbaum (1991) shows in a simple but extreme example how greedy matching might produce a stratification with an arbitrarily large overall distance. In contrast, an optimal procedure also works backwards and rearranges previously matched units if necessary. It can easily be transformed into a *minimum cost flow*

[11] However, Chapter 2 finds that for male BA holders this bias seems to be relatively small.

problem.[12] Finally, optimal full matching produces non-overlapping strata facilitating statistical inference.

OLS Regression

Since a conventional cross-sectional OLS estimator would be an inappropriate comparison to the matching estimator presented above, individuals and their wages are taken from the stratified samples produced by matching. Then, each year after college can separately be investigated by OLS, as well. The coefficient β in equation (4.7) will additionally depend on the year, Y_C, when the college degree was received and, moreover, equation (4.7) will be augmented by the regressor *year in which the high school diploma was received* to make it comparable to matching. Results presented in the next section report both the mean coefficient estimate $\hat{\beta}(\bar{F}, \bar{A}, \bar{Y}_C)$ for the treatment group and the estimated effect (4.8) resting on two different weighting schemes. The first one utilizes the weights each observation receives in the matching estimation, the second one utilizes the conventional OLS weights. In addition, both schemes are adjusted by NLSY sample weights of 1979. The two weighting schemes help investigate why OLS and matching estimates might deviate.

4.4 Results

First, balancing properties after matching are discussed for all college degrees and for both probit models. They are presented in tables 4.2, 4.3, and 4.4. Since there are ten different stratifications, one for each year after college, means after matching are weighted averages over all these years. The weights correspond to the number of strata. The means are compared by a conventional t-test under the assumption of equal variances in both groups. A “1” indicates that the means are not significantly different. Fractions are due

[12] Bertsekas (1991) discusses *linear network optimization* and provides FORTRAN-algorithms for minimum cost flow problems. Furthermore, there is an *operations research* procedure called *netflow* in SAS for these kinds of problems. Gu & Rosenbaum (1993) examine the performance of optimal full matching in a simulation study.

Table 4.2: **Balance of Covariates, AA.**

	Men							Women						
	Before			After				Before			After			
Year After College	C	T	t	C	T	t	%	C	T	t	C	T	t	%
Narrow Probit Model														
Propensity score	0.10	0.15	0	0.15	0.15	1.00	98	0.12	0.21	0	0.19	0.20	1.00	98
Index score	-1.37	-1.10	0	-1.10	-1.10	1.00	97	-1.29	-0.89	0	-0.95	-0.95	1.00	98
Black	0.29	0.21	1	0.22	0.22	1.00	100	0.24	0.30	1	0.25	0.25	1.00	100
Hispanic	0.13	0.18	1	0.19	0.19	1.00	100	0.14	0.14	1	0.13	0.13	1.00	100
Age	17.56	17.61	1	17.85	17.83	1.00	42	17.84	17.71	1	17.92	17.86	1.00	58
Year of high school diploma	79.42	79.14	1	78.80	78.77	1.00	90	78.86	78.76	1	78.53	78.50	1.00	65
Math test scores	-4.14	1.17	0	1.15	1.23	1.00	98	-4.56	0.84	0	0.59	0.62	1.00	98
Highest grades of parents	10.29	10.94	0	10.89	10.97	1.00	78	10.21	11.13	0	11.05	11.14	1.00	89
Average percent bias reduction							85							85
Broad Probit Model														
Propensity score	0.09	0.17	0	0.16	0.16	1.00	99	0.12	0.24	0	0.22	0.22	1.00	98
Index score	-1.43	-1.04	0	-1.08	-1.07	1.00	98	-1.32	-0.82	0	-0.89	-0.88	1.00	98
Black	0.30	0.22	1	0.23	0.23	1.00	100	0.25	0.31	1	0.25	0.25	1.00	100
Hispanic	0.10	0.14	1	0.13	0.13	1.00	100	0.12	0.12	1	0.12	0.12	1.00	100
Age	17.49	17.65	1	17.89	17.90	1.00	86	17.80	17.68	1	17.90	17.84	1.00	45
Year of high school diploma	79.42	78.98	1	78.70	78.66	1.00	90	78.87	78.79	1	78.55	78.52	1.00	47
Math test scores	-3.97	1.64	0	0.95	1.36	1.00	93	-4.50	1.07	0	0.77	0.74	1.00	96
Auto+shop test scores	3.85	6.94	0	7.37	6.89	1.00	84	-5.31	-3.65	0	-3.43	-3.73	1.00	82
Attended private school	0.03	0.05	1	0.05	0.04	1.00	63	0.04	0.06	1	0.08	0.06	0.94	-1
Expelled or susp. from school	0.33	0.20	0	0.21	0.19	1.00	79	0.18	0.10	0	0.10	0.09	1.00	79
Curriculum: college prepar.	0.16	0.34	0	0.31	0.34	1.00	83	0.16	0.35	0	0.30	0.37	0.89	65
Curriculum: general	0.59	0.52	1	0.56	0.47	0.94	-18	0.60	0.51	1	0.55	0.47	0.76	11
Highest grades of parents	10.49	11.21	0	11.06	11.12	1.00	88	10.32	11.25	0	11.22	11.18	1.00	90
Occupation parents high	0.08	0.15	0	0.14	0.13	1.00	81	0.07	0.13	0	0.12	0.12	1.00	82
Number of siblings	3.92	3.59	1	3.65	3.44	1.00	36	4.02	3.61	1	3.65	3.54	1.00	67
Born in south	0.38	0.28	1	0.29	0.31	1.00	70	0.40	0.35	1	0.33	0.33	1.00	67
Average percent bias reduction							74							67

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last row report the simple average over all single percent bias reductions excluding those of the propensity and index score.

Table 4.3: **Balance of Covariates, BA.**

	Men							Women						
	Before			After				Before			After			
Year After College	C	T	t	C	T	t	%	C	T	t	C	T	t	%
Narrow Probit Model														
Propensity score	0.13	0.61	0	0.55	0.56	1.00	97	0.14	0.59	0	0.57	0.57	1.00	98
Index score	-1.51	0.32	0	0.10	0.15	1.00	97	-1.45	0.27	0	0.17	0.20	1.00	98
Black	0.29	0.16	0	0.15	0.15	1.00	100	0.24	0.19	0	0.17	0.17	1.00	100
Hispanic	0.13	0.09	0	0.06	0.06	1.00	100	0.14	0.07	0	0.05	0.05	1.00	100
Age	17.56	17.63	1	17.73	17.73	1.00	72	17.84	17.82	1	18.10	18.08	1.00	-11
Year of high school diploma	79.36	78.77	0	78.71	78.65	1.00	91	78.83	78.49	0	78.21	78.18	1.00	90
Math test scores	-4.11	9.84	0	9.29	9.08	1.00	98	-4.55	8.10	0	8.17	7.83	1.00	97
Highest grades of parents	10.30	13.06	0	12.13	12.61	0.12	82	10.22	12.84	0	12.36	12.83	0.12	82
Average percent bias reduction							91							76
Broad Probit Model														
Propensity score	0.11	0.67	0	0.59	0.60	1.00	98	0.12	0.65	0	0.60	0.61	1.00	99
Index score	-1.77	0.61	0	0.22	0.27	1.00	98	-1.63	0.56	0	0.31	0.34	1.00	98
Black	0.30	0.16	0	0.17	0.17	1.00	100	0.26	0.19	0	0.18	0.18	1.00	100
Hispanic	0.10	0.07	1	0.06	0.06	1.00	100	0.12	0.06	0	0.03	0.03	1.00	100
Age	17.50	17.65	1	17.90	17.87	1.00	77	17.79	17.79	1	18.03	18.02	1.00	–
Year of high school diploma	79.37	78.74	0	78.57	78.51	1.00	90	78.85	78.51	0	78.32	78.22	1.00	72
Math test scores	-3.95	10.02	0	8.59	8.58	1.00	98	-4.50	8.35	0	7.62	7.67	1.00	98
Auto+shop test scores	3.88	7.98	0	7.55	7.54	1.00	95	-5.31	-1.60	0	-2.32	-1.45	0.79	76
Attended private school	0.03	0.12	0	0.11	0.10	1.00	80	0.04	0.13	0	0.11	0.09	0.88	65
Expelled or susp. from school	0.33	0.10	0	0.12	0.12	0.95	88	0.18	0.06	0	0.04	0.06	1.00	90
Curriculum: college prepar.	0.16	0.67	0	0.63	0.60	1.00	94	0.16	0.59	0	0.56	0.55	1.00	95
Curriculum: general	0.59	0.28	0	0.31	0.35	1.00	87	0.60	0.33	0	0.36	0.37	1.00	91
Highest grades of parents	10.50	13.21	0	12.31	12.59	0.46	89	10.32	12.95	0	12.31	12.72	0.38	84
Occupation parents high	0.08	0.29	0	0.23	0.23	1.00	95	0.07	0.32	0	0.26	0.27	1.00	96
Number of siblings	3.92	2.64	0	2.78	2.76	1.00	91	4.02	2.91	0	3.32	2.94	0.30	64
Born in south	0.38	0.33	1	0.24	0.31	0.88	-49	0.40	0.37	1	0.36	0.36	0.93	-25
Average percent bias reduction							81							77

For reasons of parsimony, weighted averages over all ten years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last row report the simple average over all single percent bias reductions excluding those of the propensity and index score. For women, the bias reduction in *age* could not be calculated because of an almost diminishing denominator.

Table 4.4: **Balance of Covariates, MA.**

	Men							Women						
	Before			After				Before			After			
Year After College	C	T	t	C	T	t	%	C	T	t	C	T	t	%
Narrow Probit Model														
Propensity score	0.04	0.64	0	0.48	0.51	1.00	94	0.05	0.54	0	0.38	0.39	1.00	97
Index score	-2.74	0.39	0	-0.19	-0.07	1.00	96	-2.60	0.11	0	-0.47	-0.41	1.00	98
Black	0.29	0.08	0	0.09	0.09	1.00	100	0.24	0.10	0	0.05	0.05	1.00	100
Hispanic	0.13	0.09	1	0.05	0.05	1.00	100	0.14	0.10	1	0.06	0.06	1.00	100
Age	17.58	18.05	1	17.79	17.85	1.00	86	17.85	18.20	1	18.61	18.60	1.00	91
Year of high school diploma	79.24	78.29	0	78.53	78.48	1.00	93	78.77	78.05	0	77.72	77.66	1.00	92
Math test scores	-4.01	14.00	0	12.21	12.61	1.00	98	-4.51	11.05	0	8.49	8.72	1.00	98
Highest grades of parents	10.31	14.12	0	12.47	12.82	1.00	91	10.22	13.75	0	12.97	13.08	1.00	95
Average percent bias reduction							95							96
Broad Probit Model														
Propensity score	0.04	0.69	0	0.44	0.46	1.00	96	0.04	0.57	0	0.41	0.42	1.00	97
Index score	-2.72	0.59	0	-0.38	-0.29	1.00	97	-2.66	0.29	0	-0.36	-0.29	1.00	97
Black	0.30	0.06	0	0.07	0.07	1.00	100	0.26	0.11	0	0.06	0.06	1.00	100
Hispanic	0.10	0.07	1	0.04	0.04	1.00	100	0.12	0.07	1	0.05	0.05	1.00	100
Age	17.51	17.96	1	18.40	18.28	1.00	73	17.80	18.24	1	18.48	18.45	1.00	94
Year of high school diploma	79.27	78.40	0	78.12	78.08	1.00	91	78.81	77.97	0	77.88	77.77	1.00	86
Math test scores	-3.86	14.20	0	10.26	11.35	1.00	94	-4.47	10.89	0	8.39	8.50	1.00	99
Auto+shop test scores	4.00	9.38	0	7.90	8.95	1.00	78	-5.29	-0.78	0	-0.74	-1.19	1.00	90
Attended private school	0.03	0.14	0	0.07	0.05	1.00	83	0.04	0.13	0	0.10	0.16	1.00	39
Expelled or susp. from school	0.33	0.07	0	0.17	0.15	1.00	88	0.18	0.02	0	0.03	0.03	1.00	95
Curriculum: college prepar.	0.16	0.81	0	0.67	0.67	1.00	97	0.16	0.70	0	0.56	0.61	1.00	92
Curriculum: general	0.58	0.17	0	0.26	0.29	1.00	90	0.60	0.25	0	0.38	0.31	1.00	80
Highest grades of parents	10.51	14.27	0	12.15	12.31	1.00	92	10.32	13.82	0	12.98	13.03	1.00	97
Occupation parents high	0.08	0.41	0	0.31	0.22	0.75	73	0.07	0.35	0	0.26	0.28	1.00	94
Number of siblings	3.91	2.31	0	2.99	2.51	1.00	70	4.02	2.72	0	3.26	2.80	1.00	65
Born in south	0.38	0.27	0	0.17	0.21	1.00	68	0.40	0.33	1	0.32	0.28	1.00	3
Average percent bias reduction							85							81

For reasons of parsimony, weighted averages over all five years after college are shown. Weights correspond to the number of strata in each year after college. C denotes the control or comparison units while T represents treated units, t indicates whether a t-test accepts balance of covariates ($t = 1$), and % represents the percent bias reduction. The last row report the simple average over all single percent bias reductions excluding those of the propensity and index score.

to averaging. Moreover, the percent bias reduction is shown for each variable and as an average over all variables. For the associate's degree, overall bias reduction amounts to 85% for the narrow and to roughly 70% for the broad model. For male BA holders, the reductions are 91% in the narrow and 81% in the broad model; for female BA holders both numbers are roughly 77%. For the graduate degrees, bias reduction amounts to 95% in the narrow and to 83% in the broad model. Apparently, percent bias reduction is larger for higher college degrees. This is because initial biases are markedly more pronounced for these degrees.

Furthermore, although the broad models achieve less overall bias reduction than the narrow models, math scores are sometimes as well balanced in the broad as in the narrow model, and, surprisingly, the broad one achieves a superior balance in parents' education. That is, the broad model is as successful with respect to balancing as the narrow one. The only disadvantage remains that it rests on less observations. In all cases the mean propensity score of matched treated individuals is lower than the original mean of the unmatched indicating that treated individuals at the high end of the propensity score scale have been lost. As expected, this feature is more pronounced in the narrow than in the broad model. Finally, notice that even after matching covariates of control units are on average less favorable than those of treated units. Thus, regression adjustment after matching seems to be a useful tool to further smooth these differences.

Second, estimation results are discussed thoroughly for the associate's and the bachelor's degrees for men and women, and, moreover, for the narrow and the broad probit models. Tables 4.5 to 4.12 present matching and OLS estimates for the first ten years after college completion. Results for the graduate degrees are relegated to the appendix. They are not very reliable due to small sample size and due to a small common support of treatment and comparison group. Since $\hat{\delta}_0$ is rarely close to 0, the estimates reported in the tables are retransformed as $\exp(\hat{\delta}_0) - 1$. For the sake of comparability, equally transformed OLS estimates of Kane & Rouse (1995), who also investigate college degrees using the NLSY, are reported in the tables, too.

All estimations have been repeated replacing the propensity score distance within calipers in equation (4.9) by the Mahalanobis metric. Chapter 2 finds favorable properties of the latter distance measure for male BA holders. However, this could not be generalized. Balance of covariates in the Mahalanobis case turned out to be less advantageous, and the stratification to be even less uniform than in the present version. For that reason, standard errors remained high although more strata were produced from the same number of treated and control units.

Associate's Degree

Tables 4.5 and 4.6 are dedicated to men's results. First note that stratification by optimal full matching has produced almost only 1-k-strata, i.e. strata consisting of one treated and one or more controls. "k" is supposed to indicate that the number of controls is variable but at least 1. This structure is responsible for the relatively low ratio of standard errors between matching and OLS estimates compared to the bachelor's degree, where numerous strata contain more than one treated. Furthermore, the number of strata diminishes over the years. This is because many individuals are not in the sample for the whole ten-year period after college. In 1994, the last year in the panel, some individuals – especially younger ones – are just in their, say, seventh year after college.

The effect of the AA on men's wages seems to increase with years after college for both the pure and the regression adjusted matching. There appears to be no systematic difference between the two estimates. In contrast, OLS *coefficient* estimates resting on the stratum weighting scheme do not show a clear time trend. This picture changes when labor market experience is taken into account to calculate the *effect* of the degree. Then, OLS results are more in line with the matching results: though being smaller, the effect seems to increase, too. Based on the conventional OLS weighting scheme the estimates of the effects are even smaller than those using stratum weights. This suggests that OLS might generally identify a parameter that is different from what matching identifies. However, in comparison with Kane & Rouse (1995) the conventional OLS coefficient estimates are already rather small.

As far as interaction between experience and schooling is concerned, there seems to

Table 4.5: **Treatment Effects, Men, AA, Narrow Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.068	0.065	0.131***	0.067***	0.136**	0.027	132	134	3.0
	(0.056)	(0.056)	(0.038)	(0.026)	(0.060)	(0.045)		1146	3.0
2	0.021	0.025	0.118***	0.019	0.038	-0.020	134	136	3.0
	(0.053)	(0.053)	(0.037)	(0.025)	(0.050)	(0.039)		1146	3.0
3	0.028	0.064	0.040	0.029	0.038	0.008	131	134	3.0
	(0.055)	(0.062)	(0.036)	(0.026)	(0.053)	(0.043)		1090	4.0
4	0.041	0.065	0.083**	0.060**	0.061	0.053	123	126	4.0
	(0.059)	(0.063)	(0.036)	(0.026)	(0.059)	(0.050)		1087	4.0
5	0.076	0.079	0.218***	0.067**	0.061	0.001	114	117	2.5
	(0.065)	(0.070)	(0.043)	(0.029)	(0.058)	(0.047)		1043	3.0
6	0.103*	0.111*	0.163***	0.073***	0.162***	0.061	110	110	2.0
	(0.062)	(0.063)	(0.038)	(0.027)	(0.065)	(0.052)		1055	2.0
7	0.161**	0.115	0.255***	0.126***	0.187***	0.096*	101	103	2.0
	(0.070)	(0.078)	(0.047)	(0.033)	(0.070)	(0.057)		1027	2.0
8	0.111*	0.096	0.132***	0.085***	0.143**	0.092	88	91	2.5
	(0.063)	(0.081)	(0.045)	(0.033)	(0.071)	(0.061)		976	3.0
9	0.172**	0.209**	0.134***	0.151***	0.169**	0.143**	73	75	3.0
	(0.075)	(0.117)	(0.043)	(0.035)	(0.080)	(0.074)		858	3.0
10	0.210***	0.219***	0.187***	0.150***	0.187**	0.120*	64	65	2.0
	(0.078)	(0.079)	(0.044)	(0.035)	(0.084)	(0.075)		778	2.0
Kane & Rouse (1995)					0.230				
					(0.049)				
Heterogeneity, Pooled Model									
Year	0.017***	0.018***	0.006***		0.007				
degree	(0.006)	(0.006)	(0.002)		(0.005)				
Math	-0.006***	-0.005**	-0.007***		-0.004***				
scores	(0.002)	(0.002)	(0.001)		(0.002)				
Educ	0.013*	0.010	0.014***		0.011*				
parents	(0.007)	(0.011)	(0.004)		(0.006)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.2.

Table 4.6: **Treatment Effects, Men, AA, Broad Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.088	0.027	0.099**	0.043	0.112*	-0.010	117	117	–
	(0.071)	(0.065)	(0.045)	(0.031)	(0.064)	(0.047)		946	–
2	-0.010	0.007	0.120***	-0.020	0.057	-0.034	118	118	–
	(0.057)	(0.053)	(0.043)	(0.028)	(0.055)	(0.042)		922	–
3	0.023	0.044	0.063	0.037	0.028	0.016	118	118	–
	(0.063)	(0.066)	(0.042)	(0.031)	(0.057)	(0.048)		935	–
4	0.055	0.065	0.064*	0.065**	0.065	0.052	112	112	–
	(0.061)	(0.069)	(0.040)	(0.029)	(0.064)	(0.053)		933	–
5	0.084	0.097	0.152***	0.071**	0.109*	0.060	102	102	–
	(0.069)	(0.081)	(0.041)	(0.030)	(0.067)	(0.056)		901	–
6	0.100	0.126*	0.195***	0.082***	0.132**	0.061	97	97	–
	(0.073)	(0.077)	(0.045)	(0.031)	(0.068)	(0.056)		888	–
7	0.163**	0.059	0.227***	0.126***	0.158**	0.091	93	93	–
	(0.074)	(0.084)	(0.050)	(0.036)	(0.073)	(0.062)		880	–
8	0.111*	0.164**	0.157***	0.130***	0.087	0.082	83	83	–
	(0.067)	(0.085)	(0.050)	(0.039)	(0.071)	(0.064)		810	–
9	0.126*	0.185**	0.166***	0.102***	0.138*	0.087	68	68	–
	(0.081)	(0.093)	(0.048)	(0.037)	(0.079)	(0.071)		694	–
10	0.086	0.152	0.091*	0.062	0.098	0.079	58	58	–
	(0.106)	(0.124)	(0.050)	(0.040)	(0.086)	(0.079)		622	–
Kane & Rouse (1995)					0.230				
					(0.049)				
Heterogeneity, Pooled Model									
Year	0.018***	0.013*	0.008***		0.016***				
degree	(0.006)	(0.007)	(0.003)		(0.005)				
Math	-0.004	-0.002	-0.003***		-0.002				
scores	(0.002)	(0.003)	(0.001)		(0.002)				
Educ	0.011	0.006	0.016***		0.026***				
parents	(0.009)	(0.016)	(0.004)		(0.007)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.2.

be no evidence for it to be of any importance. If interaction between experience and education is omitted, OLS results hardly change. This means that experience of college graduates is not rewarded any more than experience of high school graduates. Rather, the data reveal that the increasing effect can partly be attributed to a faster accumulation of experience after college on the part of AA holders than of high school graduates. Besides, heterogeneity in the treatment effect is detected for the pooled model, whose results are presented in the last three rows of the tables. Math scores lower the effect of two-year college education while parents' education has only a weakly significantly positive impact. However, for each single year after college – not shown in the tables –, the coefficients are not statistically significant even though the signs almost always coincide with the signs of the pooled models' estimates. A time trend in the effect captured by *year of the college degree* obviously plays an important role confirming rising returns to education as recently reported in the literature, see e.g. Bound & Johnson (1992), Katz & Murphy (1992), or Levy & Murnane (1992).

Results based on the broad model are similar in structure, estimates of the ninth and tenth year tend to be somewhat smaller. As expected, the number of strata diminishes. This is because there are more missing observations when more covariates are included, but obviously not because there are more strata containing at least two treated units. For the bachelor's degree, the latter will be a main reason for reduction of the number of strata in the broad model.

Tables 4.7 and 4.8 present results for women. The most striking difference to the results for men are markedly higher estimates which are even comparable to estimates for male BA holders. Kane & Rouse (1995) attribute the high results for female AA holders to the nursing degree which considerably increases their estimate based on data from the National Longitudinal Survey of the High School Class of 1972. Based on the NLSY their estimate is more or less in accordance with the corresponding OLS estimates in this study. However, returns to college are generally found to be higher for women than for men, see for instance Ashenfelter & Rouse (1998b).

A clear time trend in women's estimates does not emerge. In contrast to men, pure

Table 4.7: **Treatment Effects, Women, AA, Narrow Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.185***	0.233***	0.156***	0.191***	0.206***	0.184***	179	183	3.0
	(0.052)	(0.062)	(0.039)	(0.028)	(0.061)	(0.047)		1212	4.0
2	0.259***	0.264***	0.107***	0.224***	0.219***	0.217***	167	170	2.5
	(0.058)	(0.070)	(0.038)	(0.030)	(0.064)	(0.050)		1193	3.0
3	0.217***	0.171***	0.133***	0.188***	0.129**	0.157***	163	166	2.5
	(0.067)	(0.069)	(0.039)	(0.031)	(0.055)	(0.046)		1187	3.0
4	0.401***	0.413***	0.359***	0.370***	0.320***	0.320***	151	153	3.0
	(0.070)	(0.078)	(0.049)	(0.037)	(0.070)	(0.058)		1179	3.0
5	0.362***	0.342***	0.293***	0.303***	0.236***	0.281***	149	150	2.0
	(0.077)	(0.085)	(0.044)	(0.034)	(0.063)	(0.054)		1122	2.0
6	0.282***	0.247***	0.268***	0.258***	0.244***	0.272***	135	135	–
	(0.069)	(0.073)	(0.040)	(0.032)	(0.068)	(0.060)		1087	–
7	0.311***	0.318***	0.195***	0.276***	0.239***	0.273***	121	122	2.0
	(0.077)	(0.104)	(0.043)	(0.035)	(0.070)	(0.060)		1032	2.0
8	0.179***	0.137*	-0.015	0.158***	0.033	0.163***	98	99	2.0
	(0.073)	(0.082)	(0.036)	(0.033)	(0.060)	(0.059)		930	2.0
9	0.269***	0.279***	0.043	0.251***	0.070	0.248***	85	85	–
	(0.085)	(0.095)	(0.038)	(0.036)	(0.071)	(0.072)		851	–
10	0.298***	0.314***	0.077*	0.290***	0.154**	0.291***	80	80	–
	(0.099)	(0.115)	(0.047)	(0.043)	(0.084)	(0.081)		768	–
Kane & Rouse (1995)					0.206				
					(0.044)				
Heterogeneity, Pooled Model									
Year	-0.005	-0.004	-0.023***		-0.023***				
degree	(0.005)	(0.005)	(0.002)		(0.004)				
Math	0.002	0.003	0.003***		0.001				
scores	(0.002)	(0.002)	(0.001)		(0.002)				
Educ	0.015**	0.014	0.004		0.007				
parents	(0.006)	(0.013)	(0.004)		(0.005)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.2.

Table 4.8: **Treatment Effects, Women, AA, Broad Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.230***	0.242***	0.188***	0.199***	0.173***	0.163***	169	171	2.0
	(0.061)	(0.062)	(0.042)	(0.030)	(0.060)	(0.047)		1051	2.0
2	0.288***	0.290***	0.083**	0.227***	0.176***	0.180***	156	159	2.5
	(0.066)	(0.072)	(0.041)	(0.033)	(0.059)	(0.047)		1008	3.0
3	0.197***	0.145**	0.083**	0.162***	0.128**	0.160***	152	156	2.0
	(0.059)	(0.065)	(0.041)	(0.032)	(0.059)	(0.050)		1009	2.0
4	0.336***	0.382***	0.263***	0.311***	0.263***	0.274***	138	142	2.0
	(0.077)	(0.091)	(0.048)	(0.037)	(0.071)	(0.059)		988	2.0
5	0.335***	0.326***	0.207***	0.273***	0.165***	0.231***	138	139	2.0
	(0.069)	(0.083)	(0.042)	(0.033)	(0.062)	(0.054)		945	2.0
6	0.387***	0.363***	0.231***	0.355***	0.183***	0.285***	124	124	–
	(0.091)	(0.087)	(0.046)	(0.041)	(0.067)	(0.063)		916	–
7	0.403***	0.376***	0.168***	0.327***	0.217***	0.253***	109	110	2.0
	(0.090)	(0.093)	(0.048)	(0.041)	(0.075)	(0.065)		861	2.0
8	0.197**	0.126	-0.015	0.176***	0.013	0.089	91	91	–
	(0.093)	(0.087)	(0.044)	(0.040)	(0.066)	(0.061)		745	–
9	0.197**	0.159*	-0.007	0.162***	0.057	0.134**	80	80	–
	(0.087)	(0.095)	(0.044)	(0.041)	(0.072)	(0.068)		677	–
10	0.242***	0.193**	0.023	0.202***	0.089	0.188***	75	75	–
	(0.100)	(0.107)	(0.050)	(0.045)	(0.081)	(0.076)		631	–
Kane & Rouse (1995)					0.206				
					(0.044)				
Heterogeneity, Pooled Model									
Year	0.005	0.009	-0.016***		-0.018***				
degree	(0.006)	(0.006)	(0.002)		(0.004)				
Math	0.005**	0.005*	0.006***		0.001				
scores	(0.002)	(0.003)	(0.001)		(0.002)				
Educ	0.018**	0.017	-0.001		0.003				
parents	(0.007)	(0.013)	(0.004)		(0.006)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.2.

and adjusted matching as well as OLS effect estimates are all fairly similar. Surprisingly, they are even not lower than the OLS *coefficient* estimates indicating that labor market experience is either negatively rewarded or that women with an AA have more experience than high school graduates. Inspecting the data more closely reveals that although labor market experience is smaller for college-educated women, it is larger when it is augmented by experience acquired while being enrolled at school. This fact is further emphasized by an even higher return to the latter experience. This rather strange pattern might explain why women's estimates of the effect are extraordinarily high. Their OLS coefficient estimates, however, would be more akin to men's AA results. Moreover, heterogeneity in the effects is not supported by the data except for some negative time trend statistically significant only for OLS estimation and except for a significant impact of parents' education in pure matching only. Note that stratification produced again almost always 1-k-strata.

Almost all OLS coefficient estimates of the broad model are slightly smaller than those of the narrow model. OLS effect estimates of the broad model tend to be smaller than matching estimates. Furthermore, conventionally weighted OLS effect estimates are lower than the stratum weighted ones. In contrast to the narrow model, heterogeneity in the effects appears to be driven by the math scores. Finally, the number of strata is only reduced a little indicating that the additional covariates in the broad model do not have a great impact on selection into college.

Bachelor's Degree

Estimation results for male BA holders are summarized in tables 4.9 and 4.10. As for the associate's degree, the effects seem to increase over time reaching 30% to 40%. Regression adjusted matching estimates are, on average, higher than the pure matching ones suggesting that there might still be a certain downward bias in the latter.

OLS coefficient estimates do not show an increase with years after college; however, once labor market experience is taken into account a certain trend in the effects reappears. This means that at the beginning experience is either rewarded more than in late years or that college-educated individuals tend to successively accumulate more labor market

Table 4.9: **Treatment Effects, Men, BA, Narrow Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.028	0.071	0.186**	0.015	0.302***	0.029	206	364	5.3
	(0.092)	(0.090)	(0.081)	(0.024)	(0.087)	(0.036)		971	28.0
2	0.153*	0.217***	0.431***	0.135***	0.420***	0.119***	206	367	5.7
	(0.084)	(0.083)	(0.086)	(0.026)	(0.086)	(0.037)		973	28.0
3	0.209**	0.241**	0.695***	0.240***	0.494***	0.203***	201	352	5.4
	(0.104)	(0.115)	(0.098)	(0.029)	(0.089)	(0.042)		967	26.0
4	0.213**	0.258**	0.568***	0.219***	0.348***	0.188***	205	343	5.5
	(0.107)	(0.123)	(0.075)	(0.027)	(0.074)	(0.039)		959	22.0
5	0.229***	0.275***	0.118	0.487***	0.298***	0.247***	195	322	5.7
	(0.082)	(0.105)	(0.077)	(0.049)	(0.073)	(0.044)		931	24.0
6	0.281***	0.309**	0.399***	0.271***	0.396***	0.228***	183	306	5.8
	(0.122)	(0.154)	(0.067)	(0.032)	(0.080)	(0.045)		896	22.0
7	0.341***	0.419**	0.349***	0.279***	0.339***	0.267***	169	279	5.6
	(0.150)	(0.195)	(0.066)	(0.032)	(0.075)	(0.046)		833	20.0
8	0.420***	0.565***	0.633***	0.600***	0.424***	0.301***	146	247	5.4
	(0.193)	(0.247)	(0.122)	(0.061)	(0.091)	(0.054)		758	19.0
9	0.261**	0.327***	0.434***	0.299***	0.504***	0.293***	118	200	5.6
	(0.132)	(0.125)	(0.091)	(0.042)	(0.110)	(0.061)		649	21.0
10	0.301**	0.321***	0.491***	0.301***	0.416***	0.295***	94	157	5.8
	(0.138)	(0.141)	(0.108)	(0.044)	(0.119)	(0.069)		488	16.0
Kane & Rouse (1995)					0.403				
					(0.043)				
Heterogeneity, Pooled Model									
Year	0.052***	0.050***	0.025***		0.017***				
degree	(0.011)	(0.012)	(0.003)		(0.004)				
Math	0.009***	0.005*	0.008***		0.004***				
scores	(0.004)	(0.003)	(0.001)		(0.001)				
Educ	-0.011	0.026	0.040***		-0.004				
parents	(0.010)	(0.017)	(0.004)		(0.004)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

Table 4.10: **Treatment Effects, Men, BA, Broad Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	-0.001	0.002	0.183**	-0.019	0.326***	-0.030	162	316	4.6
	(0.081)	(0.078)	(0.083)	(0.024)	(0.096)	(0.035)		754	22.0
2	0.160**	0.188**	0.630***	0.190***	0.508***	0.097**	162	307	4.8
	(0.080)	(0.079)	(0.122)	(0.033)	(0.107)	(0.041)		801	20.0
3	0.183**	0.234***	0.288***	0.159***	0.391***	0.214***	163	287	4.4
	(0.079)	(0.087)	(0.082)	(0.029)	(0.093)	(0.045)		756	23.0
4	0.188***	0.238***	0.365***	0.223***	0.385***	0.196***	159	284	4.5
	(0.078)	(0.083)	(0.085)	(0.032)	(0.089)	(0.046)		794	20.0
5	0.202***	0.207***	0.287***	0.204***	0.341***	0.187***	150	267	4.5
	(0.074)	(0.081)	(0.061)	(0.029)	(0.079)	(0.044)		785	20.0
6	0.232***	0.215***	0.275***	0.209***	0.368***	0.223***	145	252	4.7
	(0.085)	(0.078)	(0.064)	(0.032)	(0.085)	(0.050)		720	19.0
7	0.279***	0.292***	0.403***	0.242***	0.363***	0.260***	133	227	3.9
	(0.093)	(0.091)	(0.079)	(0.036)	(0.085)	(0.052)		707	19.0
8	0.288***	0.251**	0.487***	0.293***	0.458***	0.282***	119	206	4.3
	(0.107)	(0.121)	(0.097)	(0.044)	(0.106)	(0.061)		573	19.0
9	0.357***	0.431***	0.380***	0.291***	0.484***	0.286***	102	175	4.2
	(0.134)	(0.154)	(0.103)	(0.051)	(0.119)	(0.067)		529	17.0
10	0.422***	0.477***	0.726***	0.434***	0.512***	0.270***	74	138	4.0
	(0.162)	(0.178)	(0.137)	(0.052)	(0.137)	(0.074)		425	16.0
Kane & Rouse (1995)					0.403				
					(0.043)				
Heterogeneity, Pooled Model									
Year	0.037***	0.043***	0.007**		0.007				
degree	(0.009)	(0.010)	(0.004)		(0.004)				
Math	0.006**	0.009**	0.010***		0.004***				
scores	(0.003)	(0.004)	(0.001)		(0.001)				
Educ	0.005	0.021	0.014***		-0.001				
parents	(0.011)	(0.022)	(0.004)		(0.005)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

experience than high school graduates. A closer inspection of the data reveals that the difference in labor market experience (acquired while not enrolled at college) between BA holders and high school graduates is 3.65 years in the first year after college and decreases monotonically to 2.28 in the tenth year. Thus, the bachelor's degree has a direct effect on the growth of experience which might be explained by a lower probability to get unemployed or by a distinct labor supply behavior of highly educated individuals. A comparison of estimates with and without interaction between experience and schooling exhibits almost no differences. Therefore, they are omitted.

Further note that the OLS estimates of the effect are relatively similar to the matching estimates, but with roughly three to four times lower standard errors. This is because the effective sample size of matching corresponds to the number of strata while OLS relies on all treated and untreated units together. What is more, since stratum effects are weighted according to the number of treated they comprise, variances are larger if the stratification is not very uniform. This is confirmed by the mean and maximum number of treated units in strata with more than one treated. Specifically, the latter reason leads to a high standard error ratio between matching and OLS estimates.

Interestingly, OLS effect estimates resting on the stratum weighting scheme are higher than those resting on the conventional weights. This might confirm the finding in Angrist & Krueger (1999) although OLS in this study does not build on a saturated linear model. They show that due to different weighting schemes matching and a saturated linear model estimated by OLS produce different estimates if the treatment effect is heterogeneous. Indeed, heterogeneity plays an important role in the effect of the bachelor's degree: scores on the math test tend to significantly increase the effect. Yet, parents' education seems to have no clearly directed impact on the effects. In addition, the row labeled "Year degree" shows that there is a clear time trend in the effects of the bachelor's degree in that individuals who obtained their degree more recently experience a higher effect of their education.

All results of the broad model are somewhat smaller than those of the narrow one which indicates that the additional covariates of the broad model might have an additional impact on earnings. Besides, their impact on selection into college is strong which is why the number of strata is substantially reduced. Nevertheless, standard errors of the matching estimates do not increase because stratification of the broad model is more uniform; only those of the OLS estimates are slightly higher than in the narrow model.

Results for women are presented in tables 4.11 and 4.12. Once more they are larger than men's. Alas, in contrast to men, there is no clearly increasing trend in the pure matching effects over the years. There might be some weak positive trend in regression adjusted estimates which tend to be lower than pure matching estimates in early years. OLS *coefficient* estimates exceed the matching estimates by far and do not display a time pattern while, once experience is accounted for, the resulting OLS effects are smaller and appear to increase over time. One reason is, as for men, that labor market experience is accumulated more rapidly by college graduates than by high school graduates. The difference in experience in the first year after college is 3.56 years and diminishes monotonically to 1.25 – even faster than for men.

Moreover, there is no marked difference between OLS effects using stratum weights and effects using conventional weights. This might be explained by the heterogeneity pattern expressed in the last two rows. While math scores exhibit a positive impact on the effect of a bachelor's degree, education of parents seems to have a negative influence. In sum, these two opposing interactions might explain the finding. Furthermore, there is again strong evidence in favor of a time trend in the effects expressed by the row "Year degree". That is, women who received their degree more recently appear to benefit more from their education. Finally, optimal full matching produced a stratification that is hardly more uniform than men's stratification; the mean and maximum number of treated units in strata consisting of more than one treated is only somewhat reduced. However, since there are more treated women more strata are generated.

The broad model produces lower matching but higher OLS coefficient estimates than the narrow model. By contrast, OLS effect estimates are again lower. Alas, they show an increase with years after college that appears somewhat more pronounced than the increase in the matching estimates. Finally, the broad model relies on a smaller sample

Table 4.11: **Treatment Effects, Women, BA, Narrow Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.221***	0.163**	0.399***	0.234***	0.414***	0.208***	242	424	5.2
	(0.085)	(0.086)	(0.087)	(0.032)	(0.084)	(0.042)		1032	30.0
2	0.297***	0.239***	0.508***	0.264***	0.666***	0.293***	234	412	5.2
	(0.087)	(0.101)	(0.076)	(0.028)	(0.093)	(0.044)		1037	28.0
3	0.413***	0.278***	0.712***	0.325***	0.761***	0.368***	230	407	5.2
	(0.099)	(0.109)	(0.088)	(0.031)	(0.094)	(0.046)		1022	26.0
4	0.372***	0.297***	0.783***	0.332***	0.664***	0.342***	222	382	5.2
	(0.099)	(0.100)	(0.082)	(0.031)	(0.090)	(0.048)		984	23.0
5	0.440***	0.365***	0.767***	0.376***	0.607***	0.392***	203	352	4.5
	(0.108)	(0.130)	(0.093)	(0.039)	(0.093)	(0.054)		935	20.0
6	0.515***	0.542***	0.918***	0.465***	0.710***	0.406***	191	345	4.9
	(0.141)	(0.156)	(0.102)	(0.043)	(0.103)	(0.059)		891	25.0
7	0.462***	0.417***	0.733***	0.411***	0.574***	0.492***	185	318	4.7
	(0.142)	(0.139)	(0.095)	(0.041)	(0.094)	(0.062)		825	22.0
8	0.472***	0.424***	0.604***	0.429***	0.628***	0.506***	162	278	5.3
	(0.145)	(0.155)	(0.090)	(0.045)	(0.108)	(0.068)		744	18.0
9	0.569***	0.565***	0.566***	0.599***	0.737***	0.665***	129	205	4.1
	(0.163)	(0.176)	(0.107)	(0.058)	(0.131)	(0.085)		653	10.0
10	0.531***	0.436**	0.292***	0.526***	0.622***	0.515***	104	151	4.0
	(0.188)	(0.215)	(0.105)	(0.063)	(0.142)	(0.087)		554	10.0
Kane & Rouse (1995)					0.392				
					(0.042)				
Heterogeneity, Pooled Model									
Year	0.041***	0.031***	0.011***		0.012***				
degree	(0.008)	(0.008)	(0.003)		(0.004)				
Math	0.011***	0.012***	0.013***		0.011***				
scores	(0.003)	(0.003)	(0.001)		(0.001)				
Educ	-0.018**	-0.042***	-0.030***		-0.006				
parents	(0.008)	(0.013)	(0.004)		(0.004)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

Table 4.12: **Treatment Effects, Women, BA, Broad Probit Model.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
1	0.223***	0.189**	0.364***	0.187***	0.355***	0.149***	201	361	4.8
	(0.087)	(0.087)	(0.097)	(0.034)	(0.088)	(0.041)		806	16.0
2	0.276***	0.205**	0.621***	0.235***	0.588***	0.235***	190	349	5.0
	(0.089)	(0.090)	(0.101)	(0.032)	(0.098)	(0.045)		800	18.0
3	0.335***	0.289***	0.690***	0.276***	0.538***	0.295***	184	341	4.8
	(0.102)	(0.111)	(0.104)	(0.035)	(0.095)	(0.048)		786	18.0
4	0.296***	0.207**	0.775***	0.245***	0.517***	0.235***	176	326	5.1
	(0.097)	(0.101)	(0.102)	(0.033)	(0.094)	(0.049)		756	16.0
5	0.441***	0.468***	0.962***	0.363***	0.536***	0.327***	164	299	5.1
	(0.117)	(0.158)	(0.138)	(0.046)	(0.102)	(0.057)		709	15.0
6	0.530***	0.521***	1.600***	0.448***	0.690***	0.415***	154	287	5.3
	(0.157)	(0.171)	(0.198)	(0.054)	(0.118)	(0.066)		681	17.0
7	0.354***	0.386***	0.851***	0.307***	0.402***	0.354***	141	265	4.9
	(0.128)	(0.157)	(0.132)	(0.044)	(0.100)	(0.065)		677	17.0
8	0.367***	0.326**	0.927***	0.267***	0.453***	0.369***	125	232	4.8
	(0.138)	(0.158)	(0.132)	(0.043)	(0.114)	(0.072)		623	18.0
9	0.664***	0.507***	0.954***	0.599***	0.613***	0.613***	104	179	4.3
	(0.167)	(0.166)	(0.146)	(0.063)	(0.138)	(0.092)		525	17.0
10	0.501***	0.392**	0.528***	0.478***	0.537***	0.492***	80	129	4.8
	(0.222)	(0.191)	(0.133)	(0.063)	(0.161)	(0.100)		442	13.0
Kane & Rouse (1995)					0.392				
					(0.042)				
Heterogeneity, Pooled Model									
Year	0.038***	0.033***	0.006		0.011***				
degree	(0.009)	(0.009)	(0.004)		(0.004)				
Math	0.011***	0.008**	0.006***		0.006***				
scores	(0.004)	(0.004)	(0.001)		(0.002)				
Educ	0.001	-0.027*	-0.027***		-0.009*				
parents	(0.010)	(0.015)	(0.004)		(0.005)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

size due to the reasons already mentioned above. On the other hand, its stratification is unequivocally more uniform than that of the narrow model.

Graduate Degrees

Owing to small sample size results are only briefly discussed. Tables are relegated to appendix C. The effect of a graduate degree is higher than the effect of an AA or of a BA for both men and women. There is no clear structure in estimation results for men, e.g. OLS effects do not coincide with matching estimates. In contrast, for women, there is still the clear relationship between matching and OLS effect estimates.

There is one interesting result worthy of mention. For men, the OLS effect estimate is larger when the stratum weighting scheme is used in place of conventional OLS weighting. At the same time, heterogeneity is very strong in that math scores and education of parents exhibit a significantly positive impact on the effect. By contrast, for women, the positive impact of math scores and the negative impact of parents' education might weaken overall heterogeneity. This might explain why there is no systematic difference between the stratum weighted and the conventionally weighted OLS effect estimates. This observation is also in line with results discussed above.

As a general remark, hence, although the linear model estimated by OLS is no saturated one, the results do not contradict the theoretical finding that in calculating the mean effect of treatment, matching puts the most weight on individuals most likely to participate in treatment while a saturated OLS estimation puts the most weight on individuals with a participation probability of 1/2 (Angrist & Krueger, 1999). Since those men who are most likely to attend college also gain most from their education, matching estimates for men are higher than OLS estimates. For those women who are most likely to attend college, however, evidence is weak for a larger gain from their education which is why OLS and matching do not differ much.

4.5 Summary and Conclusion

This chapter evaluates college education as to its effects in the labor market. It slightly modifies the concept of the *return to education* frequently used in the literature. While college students are enrolled, high school graduates might acquire labor market experience in the meantime. The control group should therefore comprise individuals with a higher level of labor market experience than the treatment group. The concept of *return to education* would impose equality in the experience levels of treated and controls. Furthermore, after college, there might also be an effect of treatment on the accumulation of labor market experience. Thus, focus is on the effect of college education on earnings.

In contrast to the existing literature the method of matching is used to estimate the effects. It relaxes linearity and parametric assumptions on the model. However, even if alternatively the linear model was augmented by numerous interaction terms up to a fully saturated linear model to account for an arbitrary functional form, Angrist & Krueger (1999) show that it would not necessarily identify the same parameter as matching. This is because matching and OLS implicitly impose a different weighting scheme on observations which might lead to different results if the treatment effect is heterogeneous. To assess the difference, the matching results are compared to conventional OLS estimation.

Indeed, there seems to be evidence that matching and OLS differ systematically when heterogeneity in the effect is substantial. The effect of a BA or MA on men's wages looks as if it depends significantly on ability and parents' education. The effect of a BA or MA on women's wages appears to be positively influenced by math scores, too, but negatively by parents' education. At the same time, matching and OLS estimates differ less for women than for men. The case for AA is inconclusive.

BA or MA recipients are quite distinct from high school graduates, in other words, selection into postsecondary education is extremely strong. This fact makes it very difficult to find adequate controls for each treated unit. In contrast, individuals with an AA are much less self-selected which is why matching AA holders is unburdensome. Under these circumstances, optimal full matching has the advantage for being data-adaptive. It keeps almost all treated individuals of the sample and generates suitable strata in accordance with the necessities of the sample. For example, it produces strata with one treated and a variable number of controls in case of the associate's degree, while, in case of the bachelor's degree, it also produces strata with more than one treated unit. Besides, it minimizes the total distance between treated and control units.

Albeit, matching is accompanied by considerably larger standard errors than OLS which, however, has not come as a surprise. As a nonparametric technique matching is data-hungry; the effective sample size roughly equals the number of strata while for OLS it is the sum of treated and untreated units that matters. As a further disadvantage, full matching comes with a rather non-uniform stratification. Some strata comprise a large number of treated units. As a result, estimated standard errors are inflated.

Moreover, matching on two different propensity score estimates is performed. First, the propensity score is estimated by a narrow probit model based on ability and parents' education only, and by a broad probit model augmented by numerous further socioeconomic indicators and another ability variable. Yet, results of the two models do not differ considerably, with the exception that for some degrees matching estimates of the broad model tend to be slightly lower than estimates of the narrow model. This points to a possible small upward bias in the latter. In consequence, however, one might argue that already the two variables *parents' education* and *math scores* capture the abstract concepts of family background and ability quite well.

The empirical results are along the lines of the existing literature. For men, results obtained by conventionally weighted OLS are similar to results reported in Kane & Rouse (1995: table 3). In contrast, for women, results seem to be larger in this study, specifically for female BA and MA recipients. What is more, female AA holders experience a surprisingly large effect, almost as large as that of male BA holders. Nevertheless, estimates of the effect of college education are generally larger for women than for men, which is confirmed by this study, as well. Individuals who obtained their degree more recently experience a higher effect, i.e. there is some general increase as also witnessed in the literature. Moreover, the effect looks to be increasing during the first ten years after college completion. Yet, this increase cannot be attributed to an interaction between experience and education but partly to a faster accumulation of experience for college graduates.

In sum, the method used in this chapter leads to results that, basically, do not contradict the existing literature which means that the linear approach to the human capital earnings function appears to adequately capture information provided by the data. What is more, due to its stronger assumptions OLS estimates are accompanied by considerably lower standard errors. Yet, OLS and matching might identify distinct parameters if heterogeneity in the effect is systematic in those variables relevant for selection into treatment.

Appendix A: The Probit Estimations

Appendix A discusses the estimation of the propensity score by probit models for all three college degrees. Table 4.13 displays the results. Two models are specified: a *narrow* and a *broad* one. The broad model includes several covariates that reflect socioeconomic background which is condensed into *education of parents* in the narrow model. Furthermore, it comprises two ability variables: scores on *math* and *auto and shop information* tests. The latter is omitted in the narrow model due to its weak explanatory power. Two variables are generated in the following way. *Parents' education* is the mean of the father's and mother's education, it is the mother's if the father's is missing, and vice versa. The variable *occupation parents' high* is a binary variable indicating the social status of the parents' occupation which is the mean of the mother's and father's status. It is only the father's if the mother's is missing, and vice versa.

Although several variables are insignificant for some degrees; in the broad model they are not removed for the corresponding degrees to maintain overall comparability. By contrast, in the narrow model all variables are significant. As expected, the coefficients on *math scores* and *education of parents* increase with higher college degrees. This means that selection into higher degrees is stronger, leading to a more pronounced distinction between recipients of higher college degrees and high school graduates. Thus, matching will be a difficult project for the graduate and bachelor's degrees. This is somewhat alleviated by the narrow model because several variables that rule selection are omitted. As discussed in the main text, the omitted variables seem to have only a minor influence on the outcome under study and usually they are not even included into typical Mincerian human capital earnings equations. Further note that the omission of variables increases the sample size, too.

Other studies also find that selection into college is quite strong. Ashenfelter & Rouse (1998a) report that (observed and unobserved) family background explains about 60% of the variance in schooling attainment and Murnane, Willett & Levy (1995) assert that math test scores are a strong predictor of subsequent educational attainment.

Table 4.13: **Probit Estimation Results.**

	Men			Women		
Variables	AA	BA	MA	AA	BA	MA
Narrow Probit Model						
Black	0.119	0.462***	0.415**	0.598***	0.811***	0.597***
Hispanic	0.464***	0.552***	0.739***	0.530***	0.640***	0.924***
Math test scores	0.042***	0.107***	0.142***	0.056***	0.111***	0.125***
Parents' education	0.043**	0.155***	0.182***	0.060***	0.153***	0.250***
Constant	-1.735***	-2.874***	-4.256***	-1.858***	-2.797***	-4.865***
Observations	1667	1976	1632	1793	2077	1691
$\chi^2(4)$	78.6	943.7	652.4	144.0	952.4	536.8
Overall p-value	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.071	0.426	0.598	0.103	0.406	0.530
Broad Probit Model						
Black	0.207	0.274**	0.031	0.708***	0.841***	0.588**
Hispanic	0.352**	0.256	0.343	0.475***	0.437**	0.667**
Math test scores	0.035***	0.098***	0.117***	0.050***	0.110***	0.117***
Auto+shop test scores	0.005	-0.018***	-0.018*	0.008	-0.007	-0.012
Attended private school	0.096	0.432**	0.204	0.114	0.364**	0.279
Expelled or susp. from school	-0.249**	-0.536***	-0.151	-0.333**	-0.268*	-0.548
Curriculum: college prepar.	0.521***	0.972***	0.829***	0.501***	0.751***	0.867***
Curriculum: general	0.257**	0.358**	0.254	0.240**	0.231*	0.370
Parents' education	0.028	0.154***	0.137***	0.040**	0.078***	0.160***
Occupation parents high	0.503**	0.432**	0.736***	0.377*	1.222***	0.926***
Number of siblings	0.005	-0.065***	-0.035	-0.014	-0.037*	-0.035
Born in south	-0.093	0.346***	0.084	-0.123	0.198**	0.111
Constant	-1.877***	-3.142***	-3.867***	-1.796***	-2.469***	-4.314***
Observations	1503	1792	1481	1647	1909	1555
$\chi^2(12)$	101.7	1046.4	639.0	169.8	1056.0	531.9
Overall p-value	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.102	0.518	0.641	0.132	0.489	0.571

Standard errors are omitted. Stars denote statistical significance in a *two*-sided test, *: 10%, **: 5%, ***: 1%.

Appendix B: Statistical Inference

Matching generates S strata defined by the covariates X_{si} where $s = 1, ..., S$ indicates the stratum and $i = 1, ..., n_s$ the individual in stratum s. Let δ_s be the treatment effect in stratum s. Then, the overall mean effect $\tau = \sum_s \omega_s \delta_s$ is the weighted average of the stratum effects. The weights ω_s are proportional to the number of treated units in stratum s; the number of controls is not taken into account. The variance of $\hat{\delta}_s$ under the null hypothesis of no treatment effect is

$$\sigma_s^2 = Var(\hat{\delta}_s) = \frac{n_s}{(n_s - 1)^2} \sum_{i=1}^{n_s} (r_{si} - \bar{r}_s)^2,$$

where n_s is the number of individuals in stratum s, r_{si} is the log wage of person i, and $\bar{r}_s$ the mean over r_{si} in stratum s. As a result, the variance of $\hat{\tau}$ is $\sum_{s=1}^{S} \omega_s^2 \sigma_s^2$. See Chapter 2 for further details.

In case of a constant treatment effect for all individuals it is easy to construct confidence intervals for the mean effect τ (see Rosenbaum, 1995: chapter 2). Since one advantage of matching is that, by construction, it allows for heterogeneity in the effect and that it weights individual effects appropriately (see Angrist & Krueger, 1999) when calculating the overall mean effect τ, assuming constant effects would impose an unnecessary restriction. On the other hand, unrestricted heterogeneity in the effects leaves too much freedom and makes statistical inference impossible. A compromise solution restricts the variability of the stratum treatment effects δ_s.

To this end, the stratum effect

$$\delta_s = \delta(F_s, A_s, Y_{Cs}) = \delta_0 + \delta_1(F_s - \bar{F}_s) + \delta_2(A_s - \bar{A}_s) + \delta_3(Y_{Cs} - \bar{Y}_{Cs}) \tag{4.10}$$

depends on the education of the parents, F_s, on the math scores, A_s, and on the year in which the respondent obtained the college degree, Y_{Cs}. Since there might be more than one treated unit in a stratum, F_s, A_s, and Y_{Cs} are averages over all treated in such strata. Y_{Cs} takes into account rising returns to education as suggested in the literature, e.g. by Bound & Johnson (1992), Katz & Murphy (1992), or Levy & Murnane (1992).

The model allows to build asymptotic confidence intervals for $\delta = (\delta_0, \delta_1, \delta_2, \delta_3)$ and to perform tests whether the mean effect of treatment δ_0 is positive and whether there is heterogeneity, a test for $(\delta_1, \delta_2, \delta_3)$. The asymptotic variance of the estimate $\hat{\delta}$ is calculated by exploiting the sample variability within strata. A $(1-\alpha)$-confidence region for δ would be obtained solving $(\hat{\delta} - \delta)'\tilde{V}(\delta)^{-1}(\hat{\delta} - \delta) \leq \chi^2_{4,1-\alpha}$ for δ, the parameter under the null hypothesis. $\tilde{V}(\delta)$ denotes the variance of $\hat{\delta}$ which depends on δ.[13] The procedure is outlined in Chapter 2.

It turns out that the parameter estimates for $(\delta_1, \delta_2, \delta_3)$ in the first ten years after college are almost always insignificant although taken together they often exhibit a certain structure. Therefore, further insight might be obtained by requiring time constant $(\delta_1, \delta_2, \delta_3)$ but still allowing a time-variant δ_0. To this end, all ten years are pooled and equation (4.10) is augmented to

$$\delta_{sj} = \delta_{0,1}\, d_{1\,j} + ... + \delta_{0,10}\, d_{10\,j} + \delta_1(F_{sj} - \bar{F}_{sj}) + \delta_2(A_{sj} - \bar{A}_{sj}) + \delta_3(Y_{Csj} - \bar{Y}_{Csj})$$

with $j = 1, ..., 10$ indexing the year after college and $s = 1, ..., S_j$ denoting the stratum of the jth year after college. The indicator variable d_{kj} is one if $k = j$ and zero otherwise.

Appendix C: Results for the Graduate Degrees

Tables 4.14 and 4.15 present estimation results for the graduate degrees for the first five years after college. Late years are omitted because sample size would be too small.

[13] Since statistical inference based on $\tilde{V}(\delta)$ is extremely cumbersome, $\tilde{V}(\delta)$ is replaced by $\tilde{V}(\hat{\delta})$.

Table 4.14: **Treatment Effects, Men, MA.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
Narrow Model									
1	0.454	0.486***	0.494**	0.898***	0.443***	0.332***	51	93	5.2
	(0.342)	(0.212)	(0.233)	(0.111)	(0.198)	(0.086)		374	14.0
2	0.460	0.369*	0.472***	0.359***	0.374**	0.322***	45	79	6.7
	(0.337)	(0.255)	(0.156)	(0.050)	(0.213)	(0.090)		328	13.0
3	0.436***	0.592***	0.941***	0.405***	0.555***	0.364***	44	68	5.0
	(0.145)	(0.220)	(0.203)	(0.046)	(0.241)	(0.092)		300	10.0
4	0.501**	0.214	0.366**	0.249***	0.514**	0.248***	42	65	4.8
	(0.272)	(0.211)	(0.196)	(0.059)	(0.253)	(0.096)		291	10.0
5	0.516***	0.455**	0.286**	0.330***	0.554***	0.476***	37	54	4.4
	(0.235)	(0.230)	(0.150)	(0.049)	(0.245)	(0.109)		275	10.0
Heterogeneity, Pooled Model									
Year	0.088**	0.095**	0.065***		0.016				
degree	(0.038)	(0.041)	(0.008)		(0.011)				
Math	0.032***	0.019**	0.032***		0.026***				
scores	(0.011)	(0.009)	(0.004)		(0.004)				
Educ	0.029	0.078*	0.070***		-0.005				
parents	(0.029)	(0.043)	(0.011)		(0.012)				
Broad Model									
1	0.362**	0.523***	0.686***	0.495***	0.472**	0.401***	41	61	5.0
	(0.170)	(0.238)	(0.179)	(0.063)	(0.224)	(0.105)		345	7.0
2	0.441**	0.449**	0.707***	0.375***	0.471**	0.323***	35	52	4.4
	(0.215)	(0.215)	(0.189)	(0.059)	(0.261)	(0.104)		300	7.0
3	0.524***	0.566**	1.103***	0.560***	0.875***	0.484***	33	47	3.8
	(0.157)	(0.306)	(0.283)	(0.069)	(0.396)	(0.129)		254	5.0
4	0.283	-0.066	0.336	0.114*	0.365	0.098	29	42	3.2
	(0.204)	(0.186)	(0.262)	(0.070)	(0.319)	(0.107)		241	6.0
5	0.498***	0.218*	0.780***	0.501***	0.618**	0.355***	28	37	3.3
	(0.153)	(0.141)	(0.217)	(0.064)	(0.307)	(0.119)		230	4.0
Kane & Rouse (1995)					0.556				
					(0.084)				
Heterogeneity, Pooled Model									
Year	0.024	0.019	0.000		0.016				
degree	(0.018)	(0.020)	(0.009)		(0.013)				
Math	0.031***	0.028***	0.030***		0.026***				
scores	(0.006)	(0.008)	(0.003)		(0.005)				
Educ	-0.003	-0.042	0.008		0.006				
parents	(0.022)	(0.035)	(0.011)		(0.016)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

Table 4.15: **Treatment Effects, Women, MA.**

			OLS				Stratification		
	Matching		Stratum Weighted		Convent. Weighted		S	T	Mean
Year	Pure	Adjusted	Coeff.	Effect	Coeff.	Effect		C	Max
Narrow Model									
1	0.429***	0.473***	0.962***	0.433***	1.158***	0.463***	70	87	3.1
	(0.118)	(0.122)	(0.247)	(0.062)	(0.285)	(0.092)		442	6.0
2	0.443***	0.441***	0.293*	0.437***	0.718***	0.430***	62	75	3.5
	(0.171)	(0.187)	(0.172)	(0.066)	(0.248)	(0.101)		382	6.0
3	0.543***	0.514***	0.784***	0.522***	0.949***	0.620***	52	64	3.4
	(0.137)	(0.145)	(0.219)	(0.065)	(0.294)	(0.124)		364	6.0
4	0.839***	0.807***	1.130***	0.817***	0.930***	0.901***	48	55	2.6
	(0.174)	(0.190)	(0.285)	(0.092)	(0.321)	(0.155)		292	5.0
5	0.787***	0.717***	1.038***	0.751***	1.451***	0.656***	35	40	4.0
	(0.176)	(0.199)	(0.330)	(0.107)	(0.492)	(0.162)		189	6.0
Heterogeneity, Pooled Model									
Year	0.019	0.025	-0.025***		-0.051***				
degree	(0.016)	(0.017)	(0.008)		(0.012)				
Math	0.012**	0.017**	0.009***		0.004				
scores	(0.006)	(0.007)	(0.003)		(0.005)				
Educ	-0.057***	-0.061**	-0.047***		-0.010				
parents	(0.021)	(0.027)	(0.010)		(0.015)				
Broad Model									
1	0.380***	0.452***	1.114***	0.355***	1.124***	0.357***	63	78	4.0
	(0.130)	(0.168)	(0.294)	(0.059)	(0.300)	(0.085)		352	8.0
2	0.331**	0.878***	0.124	0.360***	0.523**	0.399***	52	67	3.1
	(0.185)	(0.254)	(0.182)	(0.070)	(0.261)	(0.113)		308	7.0
3	0.509***	0.617***	1.165***	0.483***	1.023***	0.523***	48	57	3.3
	(0.132)	(0.217)	(0.254)	(0.059)	(0.314)	(0.119)		289	5.0
4	0.952***	1.046***	0.467*	1.012***	0.947***	0.730***	41	46	2.3
	(0.283)	(0.364)	(0.290)	(0.138)	(0.388)	(0.167)		224	3.0
5	0.651***	0.530***	1.065***	0.644***	1.802***	0.562***	30	35	3.5
	(0.180)	(0.163)	(0.394)	(0.107)	(0.700)	(0.187)		128	4.0
Kane & Rouse (1995)					0.532				
					(0.085)				
Heterogeneity, Pooled Model									
Year	0.038**	0.015	-0.021**		-0.041***				
degree	(0.018)	(0.021)	(0.009)		(0.012)				
Math	0.009	0.013	0.006		0.001				
scores	(0.007)	(0.009)	(0.004)		(0.005)				
Educ	-0.036	-0.066**	-0.055***		-0.028*				
parents	(0.022)	(0.030)	(0.011)		(0.016)				

Standard errors are in parentheses. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. *Stratum weighted* denotes the weighting scheme that corresponds to matching and *conventionally weighted* denotes the usual OLS weighting. All weighting takes account of the NLSY sample weights. The last three columns reflect stratification results; S: number of strata, T: number of treated, and C: number of control units. The last three rows report pooled model estimates of variables which might drive heterogeneity. The critical ε is set equal to 0.4.

Chapter 5

Postsecondary Education: The Magic Potion?

October 1999/September 2000

Abstract. Frequently, a log-linear relationship between earnings and years of education is assumed based on the classical human capital earnings equation suggesting constant returns to schooling. This chapter reconsiders this functional relationship employing an extended stylized human capital earnings function. If endogeneity of schooling as a result of optimization behavior is neglected returns to schooling appear to be larger for postsecondary than for high school education. This relationship can be found implicitly in several studies and is confirmed by this chapter as well. However, taking account of endogeneity of schooling leads to returns to education that diminish with more schooling acquired as predicted by the theoretical model.

5.1 Introduction

Frequently, a log-linear relationship between earnings and years of education is assumed based on the classical human capital earnings equation (Becker, 1967, Mincer, 1974). This suggests constant returns to schooling independently of how much schooling is acquired. In addition, Card (1999: fig. 2) presents empirical evidence in favor of a log-linear relationship between earnings and schooling using *Current Population Survey* (CPS) data. This chapter reconsiders the functional relationship using the theoretical framework recently proposed by Card (1995b) and data provided by the *National Longitudinal Survey of Youth 1979* (NLSY). The analysis extends over the years 1989 to 1994 and over both sexes.

Some empirical studies presented below implicitly report increasing marginal returns to schooling as years of education rise suggesting that postsecondary education works as some magic potion. Although this fact is confirmed by this study, these findings seem to be driven by endogeneity of schooling as a result of optimization behavior. If individuals with higher inherent earnings abilities opt for more schooling because their personal return to schooling depends positively on their abilities the relationship between observed schooling and earnings is biased. Yet, explicitly controlling for ability in the manner proposed by the theoretical model shows that, indeed, the return to education diminishes as more schooling is acquired, especially for men. In other words, neglecting ability not only yields classical ability bias in the rate of return to education but might also bias the functional form between earnings and schooling.

Math test scores of the *Armed Services Vocational Aptitude Battery* (ASVAB) provided by the data are used as measures for ability. Alas, they might themselves be prone to endogeneity in that respondents who had already acquired more education than others of the same age while ability tests took place in 1980 might do better in such tests by virtue of their experience with test situations in general. To address this issue schooling is divided into a pre-test and post-test variable as already proposed by Griliches & Mason (1972). Moreover, measurement error in the test scores is tackled by means of the instrumental variables technique.

The following section briefly outlines the theoretical foundation of the human capital model used in this study. Section 3 presents selective empirical evidence in the literature while section 4 discusses evidence from the *National Longitudinal Survey of Youth 1979*. Finally, the last section summarizes the findings.

5.2 An Extended Human Capital Earnings Function

Card (1995b, 1999) develops an analytically tractable version of the human capital earnings function that builds on Becker (1967). To provide a frame of reference his model is briefly presented. Abstracting from labor market experience, let $y(S)$ denote potential earnings after an individual acquires S years of education and let $h(S)$ be an increasing convex function reflecting costs of (or tastes for) schooling. Assume that individuals maximize the utility function $U(S) = \log y(S) - h(S)$ to derive their optimal schooling choice S^*. Individual heterogeneity is modeled by personal differences in the benefits people derive and the costs they face from schooling as follows

$$\begin{aligned} y_i'(S)/y_i(S) &= b_i - k_1 S, & (5.1) \\ h_i'(S) &= r_i + k_2 S & (5.2) \end{aligned}$$

where b_i and r_i are jointly distributed random variables, possibly correlated, and k_1, k_2 are non-negative constants. Equation (5.1) reflects diminishing returns to education while (5.2) mirrors increasing costs. As a result, the optimal schooling choice S_i^* for individual i is

$$S_i^* = (b_i - r_i)/(k_1 + k_2). \tag{5.3}$$

Equation (5.1) implies

$$\log y_i(S) = a_i + b_i S - 0.5 k_1 S^2 \tag{5.4}$$

where a_i is an individual-specific constant of integration. To the extent that a_i and b_i vary across the population, this is a random coefficients model suggesting a concave relationship between schooling and potential log earnings at the individual level. However, endogeneity of schooling owing to the optimization behavior of individuals expressed by equation (5.3) leads to a positive correlation of earnings abilities b_i and years of schooling S_i as $b_i - \bar{b} = \psi(S_i - \bar{S}) + \nu_i$ where $\bar{S}$ represents the population mean of schooling, $\bar{b}$ mean abilities, and ψ is a positive coefficient. Likewise, a_i is related to schooling (via b_i) as $a_i - \bar{a} = \lambda(S_i - \bar{S}) + \varepsilon_i$ with positive λ. Inserting these linear projections of a_i and b_i into (5.4) yields

$$\log y_i(S_i) = (\bar{a} - \lambda\bar{S}) + (\bar{b} + \lambda - \psi\bar{S})S_i + (\psi - 0.5k_1)S_i^2 + \varepsilon_i + \nu_i S_i \tag{5.5}$$

and expected log earnings are a quadratic function of schooling; it is strictly convex if $\psi - 0.5k_1 > 0$, strictly concave if $\psi - 0.5k_1 < 0$, and linear in schooling otherwise. Thus, the observed relationship is convex if there is a strong positive correlation between earnings abilities and schooling.

In this chapter, the model is slightly extended replacing equation (5.2) by

$$h_i'(S) = r_i + k_2 S - k_3 S^2$$

with a non-negative k_3. This extension might reflect the idea that some education, e.g. postgraduate studies, is not acquired solely to increase valuable human capital but also to concentrate on subjects one has a strong personal interest in, to broaden one's horizons etc.; in other words, that education tends to have an additional consumptive character apart from investment in future earnings streams alone. Further suppose that individual borrowing rates $r_i = \bar{r} + \nu_i$ are independent of individual earnings abilities b_i, $\mathbb{E}(\nu_i|b_i) = 0$, then optimization behavior implies

$$b_i = \bar{r} + (k_1 + k_2)S_i - k_3 S_i^2 + \nu_i = \bar{b} + (k_1 + k_2)(S_i - \bar{S}) - k_3(S_i^2 - \overline{S^2}) + \nu_i.$$

Inserted into (5.4) leads to

$$\log y_i(S_i) = (\bar{a} - \lambda\bar{S}) + (\bar{b} + \lambda - (k_1 + k_2)\bar{S} - k_3\overline{S^2})S_i + (0.5k_1 + k_2)S_i^2 - k_3 S_i^3 + \varepsilon_i + \nu_i S_i. \tag{5.6}$$

In contrast to equation (5.5) which allows for an arbitrary quadratic relationship due to possible correlation between b_i and r_i, equation (5.6) requires that the coefficient of the quadratic term be non-negative implying increasing returns but additionally that the return to education diminish again after having reached a certain peak ($k_3 > 0$).

Likewise, Willis (1986, p. 551) discusses a comparable model that produces increasing rates of return to education: Suppose that each individual faces rising borrowing costs as investment into education increases and each individual invests to the point at which the marginal borrowing rate is equal to the own internal rate of return itself depending on personal abilities. If everybody faced the same schedule of borrowing rates, there would be a positive correlation between the return to education and the level of schooling chosen. Individuals with low internal rate of return would leave school earlier, others later.

Modeling Individual Abilities

Reconsider equation (5.4) and assume that the random coefficients a_i and b_i capturing earnings abilities depend on inherent abilities A_i as follows

$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 A_i + \tilde{\varepsilon}_i \\ b_i &= \beta_0 + \beta_1 A_i + \tilde{\nu}_i \end{aligned}$$

with $\mathbb{E}(\tilde{\varepsilon}_i|A_i) = \mathbb{E}(\tilde{\nu}_i|A_i) = 0$ and the mean of A_i normalized to zero. Moreover, let S_i be uncorrelated with $\tilde{\varepsilon}_i$ and $\tilde{\nu}_i$. This is justified by the assumption that individuals themselves merely know their A_i but are ignorant about their $\tilde{\nu}_i$ and therefore choose their optimal schooling level on behalf of their expected ability $\mathbb{E}(b_i|A_i)$. Thus (5.4) becomes

$$\log y_i = \alpha_0 + \alpha_1 A_i + \beta_0 S_i + \beta_1 A_i S_i - 0.5 k_1 S_i^2 + \tilde{\varepsilon}_i + \tilde{\nu}_i S_i, \tag{5.7}$$

and, in contrast to (5.4), the schooling variable in this earnings equation is free of correlation with the residual. Further assume that scores of the math sub-test of the *Armed Services Vocational Aptitude Battery* (ASVAB) provided by the data reflect A_i.[1]

Although the test scores will be adjusted for age they might still be prone to endogeneity as individuals who had acquired already more schooling in 1980 when the ASVAB tests took place might have done better in the tests than they would have without their above-average education. To take account of this possible problem schooling is divided into

[1] Other scores out of the ten different scores in the ASVAB might be used, as well. However, there are no substantial changes compared to results produced by math scores. Other studies also rely on math scores, see e.g. Murnane, Willett & Levy (1995) or Kjellström (1999).

two parts, one part capturing education obtained until 1980 and another one capturing education acquired beyond 1980 similar to Griliches & Mason (1972) who divided their schooling variable in one before and one after military service during which their ability tests were performed. The idea is that the test scores are not influenced by schooling acquired after the tests.

Formally, let S_{1i} denote education acquired before ASVAB tests took place and S_{2i} education after that date. Further let A_i^* denote inherent ability which does not change for any given individual during life-time[2] and which, unfortunately, is unobservable to the analyst. What can be measured instead are test scores A_i after individuals have already acquired a certain amount of education. Suppose education makes it easier to solve test problems because of a general experience in how to cope with test situations and because some knowledge acquired at school helps solve test problems more quickly. Thus, participants with much education tend to fare better in ability tests than they would do with less education.

Abstracting from any confounding variables X_i the following equations summarize the idea

$$\begin{aligned} \log y_i &= \alpha_0 + \alpha_1 A_i^* + \beta_0 S_i + \beta_1 A_i^* S_i + \beta_2 S_i^2 + u \qquad (5.8) \\ A_i &= A_i^* + \delta S_{1i}. \end{aligned}$$

The last equation rests on the assumption that ability tests A_i would measure A_i^* without error if everybody had the same level of education at the time the tests took place, i.e. measurement error in test scores is ruled out. Instrumental variables regressions presented in section 4 will address the additional issue of measurement error. Regression on A as in (5.7) leads to biased estimates. Alternatively, replacing A_i^* by $A_i - \delta S_{1i}$ yields the following regression equation

$$\log y_i = \alpha_0 + \alpha_1 A_i + \beta_0 S_i + \beta_1 A_i S_i + \beta_2 S_i^2 - \alpha_1 \delta S_{1i} - \beta_1 \delta S_{1i} S_i + u. \qquad (5.9)$$

Hence, the coefficient estimates of S_i and S_i^2 should identify β_0 and β_2, respectively.

[2] It might depend on age which, however, is already controlled for.

5.3 Received Evidence

Usually, omission of ability is considered to yield a bias in the estimates of the return to education, the classical ability bias. In addition, as outlined above, it might also bias the functional relationship between schooling and earnings from concavity to convexity. Unfortunately, most studies *a priori* assume constant returns and thus make it impossible to examine this relationship. Nevertheless, some report more detailed estimation results that suggest increasing rather than constant or diminishing returns to education even though this was not necessarily their principal aim.

For example, Blackburn & Neumark (1993: table 2), in a certain specification of their model, mention an estimate of the return to high school education that is significantly *lower* than that to college (about 25%) based on NLSY data. Yet, they have not pursued this issue any further. Cawley et al. (1996) find that returns to education for white collar workers are significantly higher than those for blue collar workers who have usually acquired less education. By investigating the high school premium using PSID data of 1976 to 1981 for men, Weiss' (1988) specification of schooling as a cubic polynomial yields estimates which are larger for higher levels of schooling. Table 5.1 shows own calculations based on his coefficient estimates.[3] Note, however, that Weiss reports strong evidence in favor of a procyclical additional high school premium of 7% in counties with unemployment rates of around 6%. If added to the twelfth schooling year, this would disturb the strict monotonicity of the estimates.

Table 5.1: **Return to Education for Men According to Weiss.**

Years of Schooling	10	12	14	16	18	Mean HS Premium
Estimates of the Return	0.041	0.054	0.075	0.102	0.137	0.070

Own calculations based on Weiss (1988: table 9). Standard errors cannot be imputed. The last column reports mean high school premium.

In a quantile regression approach based on CPS data Buchinsky (1994) reports substantial heterogeneity in the returns to education with higher returns for upper quantiles in his restricted one-group model. His more flexible 16-group-model leads to the general

[3] Variances cannot be imputed because of missing covariance information. Alas, each coefficient estimate in Weiss' specification itself is statistically significant.

Table 5.2: **Return to Education for Men According to Park.**

Years of Schooling		≤ 12	13, 14	15	16	> 16	HS Premium
Year	1979	0.057	0.066	0.010	0.071	0.046	0.026
	1981	0.061	0.071	0.009	0.051	0.037	0.021
	1983	0.057	0.079	-0.010	0.019	0.010	0.058
	1985	0.061	0.086	0.023	0.065	0.048	0.077
	1987	0.058	0.096	0.010	0.045	0.047	0.077
	1988	0.043	0.084	0.037	0.091	0.043	0.094
	1989	0.040	0.098	0.016	0.124	0.096	0.095
	1990	0.044	0.108	-0.013	0.081	0.087	0.096
	1991	0.056	0.106	0.008	0.089	0.088	0.078

Own calculations based on Park (1994: equation 7). Standard errors cannot be imputed. The last column reports high school premium.

point that “the return to college education is higher, in general, than for high school graduation at every quantile and for all experience groups”.[4] Moreover, Kane & Rouse (1995: table 2) investigating labor market returns to two-year and four-year colleges report estimates based on the National Longitudinal Survey of the high school class of 1972 using hourly rate of pay which suggest increasing returns if ability is not controlled for, but constant returns once it is controlled for. Yet, they do not discuss this finding any further.

Park (1994) explicitly examines the functional form of the earnings equation with respect to schooling using data for men from the CPS for the years 1979 to 1991. Although he concludes that linearity may be maintained except for a peculiar deviation at the fifteenth year of schooling it is illuminating to calculate marginal returns based on estimates from his broadest model (his equation 7). Table 5.2 presents own calculations for some years. Although standard errors cannot be calculated due to missing covariance information, the estimates suggest that returns are low for individuals with merely high school education. They are larger for undergraduate education but diminish again for individuals with more than 16 years of schooling. This pattern would be in line with equation (5.6) except for the high school premium and the dip at the fifteenth schooling year.

Altonji & Dunn (1996) investigate the impact of family characteristics and IQ scores

[4] However, this is not the case for high school dropouts.

Table 5.3: **Return to Education According to** Altonji & Dunn.

Schooling Increment	10 to 12	12 to 14	14 to 16
Return to the Increment			
Men, Fixed Effects	0.047	0.097	0.132
Men, No Fixed Effects	0.087	0.097	0.096
Women, Fixed Effects	0.122	0.125	0.112
Women, No Fixed Effects	0.135	0.154	0.151

Own calculations based on Altonji & Dunn (1996: footnote 23). Standard errors cannot be imputed. The estimates represent returns to two years of education.

on the return to education. Their baseline specification of the earnings function includes a cubic polynomial in schooling and excludes all IQ measures and parents' education interaction terms (see their footnote 23). They obtain estimated returns as shown in table 5.3. Their results are based on data of the Young Men and Young Women cohort of the NLS during 1966 and 1981 for men, and 1968 and 1988 for women. Their preferred approach – a fixed effects analysis based on sibling differences – yields increasing returns for men and constant returns for women. The increase seems to disappear for the analysis without fixed effects. Finally, Ashenfelter & Rouse (1998b: figure 1) present a simple graph based on CPS data from 1993 which indicates low, almost zero, returns for low educated workers and positive returns for workers with 11 or more years of education.

The next section aims at replicating the findings of this section using the NLSY and shows how increasing returns disappear if ability is controlled for in the right manner.

5.4 Evidence from the NLSY

The Data

The data are taken from the *National Longitudinal Survey of Youth 1979* (NLSY) administered by the US Bureau of Labor Statistics. The NLSY is a sample of 12,686 youths first interviewed in 1979 when they were aged between 14 and 22 and re-interviewed annually until 1994. A detailed description of the data is given in the NLS Handbook (1997) and the NLSY79 User's Guide (1997).

In 1989 respondents were aged 24 to 32 and most had finished their education. Beginning in that year, returns to education are estimated for all subsequent years until 1994, each year taken as a single cross-section. Variables that change their values after 1989, in particular educational attainment or labor force experience, are updated each year. Men and women are examined separately, but races are pooled. Individuals who are enrolled at school or at college in the year under scrutiny are removed from the sample, yet, the self-employed are kept.[5] Oversampling of Blacks, Hispanics, and economically disadvantaged Whites suggests the use of sample weights provided by the NLSY for each year. The outcome measure is the hourly rate of pay inflated to 1996 dollars using the US consumer price index. Outliers in wages are removed, i.e. observations with an hourly wage above $1000 are deleted and wages below $1 are set equal to $1. Furthermore, extraordinarily large changes in wages between two subsequent years are smoothed by removing the local outliers, as well.

The data contain numerous variables describing socioeconomic background, the high school career, and labor force status (since 1975) used to generate a measure of actual experience based on weeks worked per year. What is more, the NLSY provides information on ten ability measures collected in 1980 when 94.3% of all respondents participated in tests to update the *Armed Services Vocational Aptitude Battery* (ASVAB). Since respondents participated in the tests at different ages the scores are adjusted by regressing the raw scores on age dummies and using the residuals subsequently as explanatory variables in the wage equation analogous to Blackburn & Neumark (1993). A descriptive summary of the variables used in this study is provided in table 5.4 weighted by the sample weights of the NLSY. Two different types of variables in addition to the standard ones, education and experience, will be considered: background variables that determine earnings apart from investment into human capital and ASVAB test scores adjusted by age.

[5] Kane & Rouse (1995), who also use the NLSY, report that their results are not sensitive to the exclusion of self-employed.

Table 5.4: **Description of Variables.**

	Men		Women	
	mean	std.dev.	mean	std.dev.
Log hourly wage in 1989	7.055	0.500	6.806	0.538
Age in 1979	17.731	2.348	17.639	2.288
Years of education (in 1989)	13.005	2.337	13.228	2.138
Years of education before ASVAB tests	11.153	1.874	11.299	1.797
Years of education after ASVAB tests (in 1989)	1.852	2.092	1.929	2.120
Experience in years (in 1989)	8.260	2.723	7.376	2.895
Background Variables				
Black	0.130	0.337	0.140	0.347
Hispanic	0.060	0.237	0.054	0.227
Lived in urban area 1987	0.775	0.417	0.781	0.414
Lived in north-east 1987	0.194	0.396	0.201	0.401
Lived in north-central 1987	0.303	0.460	0.277	0.447
Lived in south 1987	0.334	0.472	0.357	0.479
Lived in an area with high unempl. 1987 (1 to 6)	2.880	0.891	2.915	0.911
Physical height in 1985 (inches)	70.441	2.887	64.527	2.708
Health limit begun under age 18	0.080	0.272	0.096	0.295
Married (in 1989)	0.531	0.499	0.564	0.496
Member of a union (in 1989)	0.137	0.344	0.087	0.282
ASVAB Scores, Adjusted for Age				
Paragraph comprehension	1.456	10.493	4.232	9.077
Word knowledge	2.537	10.086	3.624	8.974
Math knowledge	2.375	10.087	2.115	9.462
Arithmetic reasoning	3.624	9.958	1.633	9.191
General science	4.113	10.044	1.365	8.762
Auto and shop information	7.604	9.563	-2.092	6.561
Numerical operations	0.951	10.133	4.147	8.935
Electronic Information	5.971	9.813	-0.530	8.002
Mechanical Comprehension	6.189	9.929	-0.661	7.856
Coding speed	-0.225	9.407	4.695	9.071
Number of observations	3413		3305	

Means and standard deviations for variables in 1989. Observations are weighted by the NLSY sample weights.

Estimation Results

A model of earnings as broad as possible will be specified in advance and tested against polynomial specifications as proposed in equation (5.6) in order to distinguish between a linear, quadratic, and a cubic relationship. Ability will not yet be controlled for. The most general model possible uses dummy variables for each education level nesting the polynomial models. It maintains the additive structure of the human capital earnings function: $\log y(S) = f(S) + X\beta + \varepsilon$.[6] Individuals are distributed across 17 schooling categories $S \in \{4, 5, ..., 20\}$ as shown in the appendix table 5.10. Alas, owing to small cell size observations with less than seven years of schooling are completely removed. Using incremental dummy variables $\mathbf{1}(S \geq i)$, where $\mathbf{1}$ is the indicator function, the functional relationship between schooling and earnings is modeled as

$$f(S) = \sum_{i=7}^{S} \delta_i \, \mathbf{1}(S \geq i), \qquad S \in \{7, 8, ..., 20\}.$$

Thus, the marginal return to an additional year of schooling at level $S = s$ equals δ_{s+1}.[7]

Tables 5.5 and 5.6 present estimation results for men and women, respectively, also controlling for the background variables. Both for men and women estimates vary considerably, yet, a certain pattern seems to emerge. Returns to the earlier years of education are not significantly different from zero, the signs of the point estimates are sometimes positive sometimes negative. For men, only the last high school year (11 to 12 years) yields positive significant results for almost all years, indicating a potential high school premium. The estimates remain high until the last two years when they are again insignificant and sometimes negative. The peak of the return to education seems to be achieved during college education with a possible premium for graduation from college at age 16 close to Park's (1994) results. Women's results are similar but with generally higher estimates than men's and no clear premium structure after certain schooling levels. Their returns are positive after high school education and remain so even unto the 19th grade with a peak around 14 to 16 and another at 19.

This general specification is tested against the null hypothesis of linearity implying $\delta_8 = ... = \delta_{20}$, against the null of a quadratic relationship, and against a cubic relationship. Results of F-tests are reported in the respective panels of the tables. Linearity is clearly rejected in all six years for men and for women alike. The less restrictive quadratic

[6] Nonparametric estimation of f by means of a partially linear additive model as outlined in Hastie & Tibshirani (1990), for instance, would be an alternative way to proceed. However, since S takes on only discrete values numerous ties would be produced although a conventional nonparametric smoother would generally consume less degrees of freedom for reasonable smoothing parameters than the dummy variables approach.

[7] Actually, the return is $\exp(\delta_{s+1}) - 1$ which is approximately δ_{s+1} for small values.

Table 5.5: **Local Returns to Education, Men.**

Education	1989	1990	1991	1992	1993	1994	Pooled
7 to 8 years	0.068	-0.010	-0.173	-0.050	0.037	0.239	0.005
8 to 9 years	-0.027	0.005	0.020	-0.081	-0.046	-0.079	-0.030
9 to 10 years	-0.049	0.090	-0.005	0.022	0.037	-0.027	0.009
10 to 11 years	0.008	-0.072	0.049	-0.002	-0.000	0.008	0.001
11 to 12 years	0.103***	0.100***	0.033	0.119***	0.072	0.119***	0.091***
12 to 13 years	0.084***	0.131***	0.128***	0.114***	0.103***	0.139***	0.118***
13 to 14 years	0.068*	0.037	0.059	0.046	0.094**	0.017	0.051***
14 to 15 years	0.084*	0.059	0.083	0.071	0.107*	0.093*	0.081***
15 to 16 years	0.122***	0.130***	0.095**	0.169***	0.116**	0.150***	0.129***
16 to 17 years	-0.006	0.080	0.041	0.048	-0.030	-0.040	0.013
17 to 18 years	0.036	0.038	0.027	0.071	0.155**	0.187***	0.088***
18 to 19 years	0.133	0.017	0.045	-0.106	-0.083	-0.158*	-0.037
19 to 20 years	-0.043	-0.165*	-0.120	-0.127	0.276***	0.339***	0.059
H_0: Linear specification							
Coefficient	0.066***	0.068***	0.065***	0.071***	0.080***	0.078***	0.071***
F-value	4.238	4.365	4.554	5.977	3.392	4.050	20.157
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H_0: Quadratic specification							
Linear coeff.	-0.017	0.038	0.021	0.049*	-0.014	-0.005	0.015
Quadr. (÷10)	0.030***	0.011	0.016*	0.008	0.034***	0.030***	0.020***
F-value	3.639	4.628	4.706	6.464	2.786	3.701	19.654
P-value	0.000	0.000	0.000	0.000	0.001	0.000	0.000
H_0: Cubic specification							
Linear coeff.	-0.693	-0.717	-0.846	-1.037*	-0.614	-0.532	-0.718***
Quadr. (÷10)	0.550***	0.589	0.680*	0.837	0.487***	0.427***	0.579***
Cubic (÷100)	-0.129***	-0.143***	-0.164***	-0.204***	-0.110***	-0.096***	-0.137***
F-value	0.952	1.275	0.438	0.794	1.328	2.721	3.049
P-value	0.484	0.238	0.929	0.635	0.209	0.002	0.001
Number of obs.	3413	3355	2976	2838	2874	2834	18290

Incremental dummies reflect the return to schooling for each education category between seven and twenty years of schooling. Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors are omitted. Heteroskedasticity is not adjusted for. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%.

Table 5.6: **Local Returns to Education, Women.**

Education	1989	1990	1991	1992	1993	1994	Pooled
7 to 8 years	-0.111	0.124	-0.126	-0.040	0.068	0.112	-0.006
8 to 9 years	0.058	-0.099	0.180	0.007	0.001	-0.088	0.023
9 to 10 years	0.130*	0.148*	0.115	0.036	-0.050	-0.002	0.060
10 to 11 years	-0.117*	-0.077	0.016	0.015	-0.072	-0.077	-0.044
11 to 12 years	0.067	0.089*	-0.052	0.091	0.123*	0.089	0.073***
12 to 13 years	0.095***	0.081***	0.108***	0.058*	0.069**	0.031	0.078***
13 to 14 years	0.082**	0.114***	0.139***	0.162***	0.113***	0.153***	0.126***
14 to 15 years	0.086*	0.075	0.124**	0.106**	0.141***	0.131**	0.105***
15 to 16 years	0.154***	0.170***	0.068	0.145***	0.139***	0.146***	0.137***
16 to 17 years	0.078	0.092*	0.093*	-0.032	-0.069	0.080	0.030
17 to 18 years	-0.019	-0.056	0.068	0.126*	0.123*	-0.004	0.051*
18 to 19 years	0.242**	0.123	0.069	0.208**	0.298***	0.355***	0.219***
19 to 20 years	0.073	0.141	0.038	-0.022	0.038	0.137	0.059
H_0: Linear specification							
Coefficient	0.086***	0.090***	0.093***	0.097***	0.092***	0.099***	0.092***
F-value	4.358	3.538	3.151	3.219	4.752	6.433	18.942
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H_0: Quadratic specification							
Linear coeff.	-0.087**	-0.018	0.006	-0.006	-0.086**	-0.157***	-0.046***
Quadr. (÷10)	0.062***	0.039***	0.031**	0.036***	0.063***	0.090***	0.049***
F-value	2.406	2.916	2.886	2.705	3.110	2.529	12.101
P-value	0.006	0.001	0.001	0.002	0.000	0.004	0.000
H_0: Cubic specification							
Linear coeff.	-0.558**	-0.618	-0.569	-0.630	-0.637**	-0.592***	-0.581***
Quadr. (÷10)	0.422***	0.495***	0.468**	0.509***	0.480***	0.420***	0.455***
Cubic (÷100)	-0.089***	-0.112***	-0.108***	-0.116***	-0.102***	-0.081***	-0.100***
F-value	1.625	1.554	1.743	1.279	2.283	2.030	6.007
P-value	0.093	0.114	0.066	0.236	0.012	0.027	0.000
Number of obs.	3305	3269	2851	2666	2766	2732	17589

Incremental dummies reflect the return to schooling for each education category between seven and twenty years of schooling. Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors are omitted. Heteroskedasticity is not adjusted for. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%.

model is rejected, as well, for men and women, and, what is more, in all years estimates show *increasing* rather than diminishing returns to education. Finally, a cubic relationship between education and earnings as proposed by equation (5.6) is not rejected at conventional levels for men in five years and for women in four years.

Since estimates vary considerably an average over all years is reported in the last column of the tables. To this end, observations of all six years are pooled and the dummy variables model is re-estimated.[8] However, the pooled estimates are for illustrative purposes only because stochastic dependencies among observations in the pooled model are not accounted for, thus exaggerating statistical precision.

Note that actual experience might be endogenous especially for women. Replacing actual by Mincer potential experience, i.e. age - education - 6, does not markedly alter the patterns found in tables 5.5 and 5.6 and points to a polynomial specification of order 3 as well. Moreover, additionally dropping the first and the last education cell, coefficient estimates of the remaining cells virtually remain unchanged. In sum, empirical evidence seems to support low returns to secondary education for individuals who opted for low education, comparatively high returns for the first four college years for college graduates, and again lower returns for postgraduate education. Alas, these findings do not necessarily reject diminishing personal returns to education as stated initially in equation (5.1). Endogeneity of schooling owing to the optimization behavior may well lead to a reduced-form relationship as is found here and stated in equation (5.6). Therefore, ability measures are included next; both of which might drive individual heterogeneity coefficients a_i and b_i.

The first panel of tables 5.7 and 5.8 show estimation results of the basic equation (5.7) for men and women, respectively. The schooling variable is transformed into years of education exceeding the minimum level of seven years. This ensures that the coefficient of ability expresses the return to ability at seven years of education. For men, coefficient estimates of the linear schooling term are positive in all years and those of the quadratic term are mainly negative although not significantly different from zero.

[8] A further model comprising an additional time trend for all coefficients has not produced convincing evidence in favor of a trend. Therefore, the trend is omitted.

Table 5.7: **Estimation Results for Men.**

	1989	1990	1991	1992	1993	1994	Pooled
Basic Model							
Ability (÷100)	0.080	-0.190	-0.289	0.374	0.257	0.268	0.132
Abil*School (÷100)	0.135**	0.162***	0.183***	0.107*	0.115*	0.123*	0.133***
Schooling	0.035*	0.072***	0.066***	0.062***	0.037	0.041*	0.051***
Schooling^2 (÷10)	-0.000	-0.025	-0.024	-0.018	0.008	0.002	-0.009
Homosk., p-value	0.028	0.132	0.045	0.118	0.003	0.000	0.000
Pre-test and Post-test Schooling							
Ability (÷100)	0.080	-0.185	-0.278	0.368	0.319	0.279	0.142
Abil*School (÷100)	0.135**	0.161***	0.182***	0.109*	0.105	0.121*	0.131***
Schooling	0.036	0.076***	0.083***	0.074***	0.047*	0.047*	0.054***
Schooling^2 (÷10)	-0.004	-0.034*	-0.045**	-0.026	-0.013	-0.006	-0.017*
S_1	-0.003	-0.013	-0.037**	-0.021	-0.037*	-0.016	-0.012
$S_1 * S$ (÷10)	0.006	0.021	0.039	0.011	0.051*	0.018	0.020*
Homosk., p-value	0.027	0.129	0.042	0.118	0.003	0.000	0.000
F-test	0.049	0.453	2.134	1.799	1.766	0.340	28.316
P-value	0.952	0.636	0.119	0.166	0.171	0.712	0.000
Basic Model, IV for Ability							
Ability (÷100)	0.410	1.059	1.103	1.726**	0.911	1.198	1.152***
Abil*School (÷100)	0.172	0.016	0.020	-0.029	0.080	0.034	0.039
Schooling	0.030	0.029	0.018	0.019	0.018	0.010	0.018
Schooling^2 (÷10)	-0.009	0.006	0.010	0.010	0.015	0.021	0.011
Homosk., p-value	0.032	0.129	0.045	0.133	0.003	0.000	0.000
Hausman, p-value	0.014	0.071	0.052	0.024	0.127	0.143	0.000
Canonical correl.	0.252	0.248	0.238	0.238	0.244	0.253	0.248
	0.306	0.297	0.299	0.320	0.316	0.336	0.312
Overid., p-value	0.366	0.315	0.979	0.818	0.271	0.427	0.459
Pre-test and Post-test Schooling, IV for Ability							
Ability (÷100)	0.423	1.092	1.194	1.797**	1.042	1.241	1.179***
Abil*School (÷100)	0.170	0.012	0.011	-0.035	0.066	0.029	0.036
Schooling	0.032	0.033	0.033	0.029	0.027	0.016	0.022
Schooling^2 (÷10)	-0.012	-0.003	-0.009	0.004	-0.005	0.013	0.002
S_1	-0.004	-0.014	-0.038**	-0.022	-0.039*	-0.017	-0.014*
$S_1 * S$ (÷10)	0.007	0.022	0.040	0.011	0.051*	0.019	0.020*
Homosk., p-value	0.032	0.128	0.044	0.134	0.004	0.000	0.000
Hausman, p-value	0.032	0.089	0.080	0.038	0.091	0.150	0.000
Canonical correl.	0.253	0.248	0.238	0.239	0.242	0.253	0.247
	0.306	0.297	0.300	0.322	0.318	0.338	0.313
Overid., p-value	0.357	0.284	0.993	0.839	0.354	0.459	0.496
Number of obs.	3413	3355	2976	2838	2874	2834	18290

Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors – adjusted for heteroskedasticity – are omitted. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. F-tests in the second panel test the augmented pre-/post-test schooling model against the basic model.

Table 5.8: **Estimation Results for Women.**

	1989	1990	1991	1992	1993	1994	Pooled
Basic Model							
Ability (÷100)	0.268	-0.178	-0.602	0.450	0.656	0.767*	0.293
Abil*School (÷100)	0.106	0.173***	0.220***	0.107	0.041	0.030	0.105***
Schooling	0.003	0.054**	0.079***	0.041	-0.010	-0.042	0.024**
Schooling^2 (÷10)	0.040*	0.003	-0.012	0.016	0.055**	0.083***	0.029***
Homosk., p-value	0.129	0.159	0.634	0.001	0.041	0.239	0.000
Pre-test and Post-test Schooling							
Ability (÷100)	0.176	-0.227	-0.710	0.394	0.614	0.606	0.232
Abil*School (÷100)	0.111*	0.175***	0.227***	0.105	0.042	0.043	0.108***
Schooling	0.038	0.071**	0.112***	0.070**	0.012	-0.003	0.047***
Schooling^2 (÷10)	0.005	-0.003	-0.044*	-0.006	0.036	0.038	0.005
S_1	-0.074***	-0.027	-0.067***	-0.055**	-0.045*	-0.088***	-0.048***
$S_1 * S$ (÷10)	0.059**	0.001	0.054*	0.033	0.031	0.080***	0.040***
Homosk., p-value	0.132	0.159	0.615	0.001	0.042	0.231	0.000
F-test	12.896	7.041	8.334	10.558	4.673	11.682	51.973
P-value	0.000	0.001	0.000	0.000	0.009	0.000	0.000
Basic Model, IV for Ability							
Ability (÷100)	-0.307	-0.799	-0.919	-0.980	0.389	0.018	-0.402
Abil*School (÷100)	0.204*	0.278**	0.288**	0.314***	0.081	0.144	0.212***
Schooling	0.022	0.076**	0.092***	0.087**	-0.002	-0.018	0.047***
Schooling^2 (÷10)	0.022	-0.017	-0.025	-0.022	0.048*	0.062**	0.009
Homosk., p-value	0.124	0.153	0.610	0.001	0.041	0.214	0.000
Hausman, p-value	0.600	0.510	0.647	0.092	0.926	0.505	0.000
Canonical correl.	0.236	0.247	0.247	0.252	0.252	0.241	0.246
	0.299	0.331	0.319	0.323	0.346	0.332	0.325
Overid., p-value	0.226	0.844	0.764	0.906	0.711	0.017	0.074
Pre-test and Post-test Schooling, IV for Ability							
Ability (÷100)	-0.132	-0.745	-0.815	-0.791	0.510	0.128	-0.310
Abil*School (÷100)	0.187	0.274**	0.276**	0.289**	0.067	0.133	0.202***
Schooling	0.049	0.091**	0.117***	0.110***	0.016	0.014	0.066***
Schooling^2 (÷10)	-0.009	-0.023	-0.053*	-0.042	0.031	0.022	-0.013
S_1	-0.074***	-0.028	-0.065***	-0.058***	-0.044*	-0.087***	-0.049***
$S_1 * S$ (÷10)	0.059**	0.003	0.053*	0.038	0.031	0.080***	0.041***
Homosk., p-value	0.126	0.152	0.593	0.001	0.041	0.202	0.000
Hausman, p-value	0.750	0.664	0.744	0.273	0.874	0.737	0.000
Canonical correl.	0.244	0.252	0.254	0.260	0.258	0.248	0.252
	0.299	0.331	0.320	0.322	0.348	0.333	0.326
Overid., p-value	0.195	0.845	0.724	0.888	0.677	0.016	0.056
Number of obs.	3413	3355	2976	2838	2874	2834	18290

Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors – adjusted for heteroskedasticity – are omitted. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. F-tests in the second panel test the augmented pre-/post-test schooling model against the basic model.

Women's results are in contradiction to what the theoretical model above predicts. The coefficient of squared schooling tends to be positive, even significantly so in three years, rather than negative. Due to possible endogeneity of women's actual experience, potential experience is used instead which leads to more confirmative results shown in the appendix table 5.12. Indeed, then, five of six coefficient estimates of schooling squared are not significantly different from zero (with three of them having a negative sign), only one remains significantly positive. Notice that men's second order coefficients look more confirmative, too, when potential experience is used (see table 5.11). Another reason for women's results might be that the math scores themselves are inappropriate as a measure of inherent ability A_i which will be discussed below.

Interaction between ability and schooling yields positive and significant estimates for men suggesting strong heterogeneity in the returns to education. Alas, for women, estimates provide only weak evidence underscoring heterogeneity. Chapter 4 reports similar results. Heterogeneity plays an important role in other studies, too, e.g. Buchinsky (1994). In particular, heterogeneity caused by variation in personal abilities is detected by Altonji & Dunn's (1996) preferred fixed effects specification. They report significantly positive coefficient estimates of the interaction between education and IQ scores, yet without overall convincing evidence. Furthermore, Blackburn & Neumark (1993: table 4) discover that the importance of the interaction between ability and education has increased significantly during the 1980s. Murnane, Willett & Levy (1995) confirm this observation for men but not for women.

In addition, the tables indicate that the ability-schooling interaction seems to be less pronounced in the early 90s for both sexes. Therefore, findings in Ashenfelter & Rouse (1998b: table A2) who report an insignificantly *negative* interaction coefficient are not in contrast to this study. They used data of the NLSY in 1993, pooled men and women, and controlled for age instead of experience. Taken together, their estimate is comparable to the average of men's and women's interaction estimate of 1993 shown in the appendix tables 5.11 and 5.12.

The last row of the basic model reports p-values of a test whether homoskedasticity is compatible with the data by regressing the square of the estimated residuals on the square of schooling in accordance with equation (5.7). The significance of the coefficient estimate of squared schooling asymptotically equals the significance of a test against homoskedasticity (see e.g. GREENE, 1993: p. 396).[9] To take account of heteroskedastic errors in the main model, OLS covariance estimates are adjusted by the estimated error variances derived from the coefficient estimates of the auxiliary regression. If equation (5.7) is correctly specified errors should be heteroskedastic. This is confirmed for men, yet, for women, the tests indicate that heteroskedasticity is only weak.

The second panel of the tables present estimation results based on equation (5.9) when test scores may depend on schooling attainment before the tests took place. The first six rows report coefficient estimates of the corresponding variables. Interaction between schooling and ability remains positive. Coefficient estimates of S and S^2 appear to confirm diminishing returns to education for men as already in the basic model. Interestingly, for women, estimates of the coefficient of squared schooling do not support increasing returns anymore. Coefficient estimates of S_1 and S_1S yield some contradictory results concerning the sign of δ with some more weight on a negative than on a positive δ. This would mean that more pre-test schooling would have negatively influenced ability test achievements.

Moreover, heteroskedasticity tests produce almost the same p-values as in the basic model. All variance estimates are again adjusted for heteroskedasticity. A comparison of the broader model augmented by pre- and post-test schooling with the basic model by means of an F-test leads to the conclusion that endogeneity of the math scores plays, on average, a minor role for men but is strongly confirmed for women. As before, results based on potential experience are more pronounced than those based on actual experience. Finally, note that the coefficient estimates of the interaction between ability and schooling remain almost unchanged with respect to the basic model.

[9] The variance of the coefficient estimate is adjusted for heteroskedasticity in this auxiliary regression using WHITE's (1980) covariance estimate.

Measurement Error in Test Scores

Math test scores measure certain aspects of abilities or skills. One might argue that they only imperfectly capture inherent earnings abilities A^*. A similar observation can be made for the other ASVAB scores. Suppose they all measure some sort of skills but their errors in capturing true A^* are independent of each other because different skills vary unsystematically around A^*. Then, if some of them have no impact on earnings given math scores they are valid instruments for math scores. Examining some alternatives shows that scores on *general science* and *electronic information* seem to fulfill the requirements. Griliches (1977) already suggested to use one test score as instrument for another provided more than one is available.

Estimations of the basic and the broader model are repeated using the two additional scores as instruments for math scores. Results are reported in the third and fourth panels of the tables. Hausman as well as overidentification tests assess the instrumental variables estimations. Furthermore, canonical correlations between the instruments and the math scores are presented as an indication for instrumental relevance.[10] The lowest canonical correlation amounts to 0.24 for both men and women and overidentification tests indicate that the instruments are valid except for women in the last year 1994.

Unfortunately, in the basic IV model, almost all estimates of the schooling coefficients for men are statistically insignificant rendering a sound interpretation difficult. Nevertheless, Hausman tests point to significant differences between the IV and OLS regressions. For women, coefficient estimates of schooling squared are still not overly convincing underpinned by the Hausman tests indicating that IV is not significantly different from OLS. IV regressions of the broader model resting on the pre-test schooling variable S_1 (fourth panels of the tables) also yield insignificant estimates for men. Women's estimates seem to weakly confirm a negative quadratic relationship at least for earlier years, however, estimates are statistically imprecise. Again, estimation results on coefficients of S_1 and

[10] Canonical correlations are discussed in Bowden & Turkington (1984: Ch. 2) and in Hall, Rudebusch & Wilcox (1996). They can easily be calculated as the square root of the eigenvalues of the matrix $(X'X)^{-1}(X'Z)(Z'Z)^{-1}(Z'X)$ where Z is the matrix of instruments for the possibly endogenous regressors X. If there are only two endogenous variables solely two eigenvalues differ from 1.

S_1S do not really confirm a positive δ, similar to the second panel. Furthermore, replacing actual experience by potential experience leads again to similar but more pronounced results.

5.5 Summary and Conclusion

This chapter departs from an extension of the human capital model recently proposed by Card (1995b). The model incorporates unobserved ability parameters ruling both the absolute level of earnings and the individual return to education. Assuming that individuals are rational agents maximizing the present value of their life-time earnings and that their return to education diminishes as they acquire more education, their optimal schooling level will depend on their personal earnings abilities. Neglecting this relation between education and ability leads to a reduced form with indeterminate functional form linking earnings and schooling, i.e. increasing, constant, or diminishing returns to education.

Data from the NLSY show that the functional relationship without ability controls follows a polynomial of order 3 yielding returns to education that increase with years of schooling acquired but diminish again after passing a certain peak. This pattern is consistent with strong correlation between education and ability and with the hypothesis that education has an additional consumptive character apart from mere investment into future earnings. To a certain extent, similar results can be found in the literature.

Yet, these findings do not give information about the true personal development of the returns to education as long as the ability components are not taken into account. In other words, omission of ability not only yields classical ability bias in estimates of the return to education but might also bias the functional form between log earnings and education. Indeed, controlling for ability in form of math test scores shows that personal returns to education diminish as schooling increases, specifically for men. This finding is only obtained if interaction between schooling and ability is introduced reflecting differing personal returns to education. The results show that the interaction term is statistically significant for men, i.e. heterogeneity in the returns is substantial. However, this observation is in contrast to earlier findings in Griliches (1977) who has not detected interaction terms. For women, heterogeneity seems to play a less important role. These findings are in line with Chapter 4.

Furthermore, endogeneity of the math scores – even adjusted for age – is a source of bias in estimates of rates of return to education for women in that education acquired before ability tests took place might influence the test results. Dividing schooling into pre-test and post-test education addresses this issue. Measurement error in the scores is tackled by conventional IV estimation techniques. Yet, IV estimation does not produce conclusive results.

Diminishing returns are also reported by Murnane, Willett & Tyler (2000: tables 5 and 6) who analyze the *general educational development* (GED). They include math test scores in their regressions and report estimates of the return to post-secondary education that are markedly lower than those to secondary education. The results also confirm recent studies that cope with endogenous schooling by the technique of instrumental variables finding estimates that are usually higher than corresponding OLS estimates (see e.g. Card, 1999). The deviation might be explained by the fact that IV estimates do not necessarily identify the mean effect of education but the effect of education on those individuals who are mostly affected by the chosen instruments, the so-called *local average treatment effect*. Angrist, Imbens & Rubin (1996) discuss this framework; see also Imbens & Angrist (1994) and Angrist & Krueger (1999). In case of systematic heterogeneity the two parameters differ. Since the instruments that are generally used mainly affect low-educated individuals, higher IV estimates might be interpreted as higher returns to early years of schooling than to later years.

In sum, endogeneity of education mainly caused by heterogeneous inherent earnings ability plays an important role. Omission of ability measures might lead to classical ability bias in estimates of the rate of return to education but, as shown in this study, it might also bias the functional form of the human capital earnings equation. Without correctly controlling for ability one might detect increasing rather than diminishing returns and thus one might wrongly wonder whether post-secondary education was some sort of magic potion. Data from the NLSY seem to support this view and does not contradict the theoretical approach this study is based on, particularly for men. Results are less convincing for women. One reason might be endogeneity of female labor market experience due to more complex female participation decisions supported by results using potential instead of actual experience. Therefore, a better understanding of women's optimization behavior appears to be necessary.

Appendix: Additional Results

For reasons of parsimony, the appendix presents detailed OLS regression results solely for the basic model. Table 5.9 presents OLS estimates for men and women. The standard errors, which are not explicitly presented, are adjusted for heteroskedasticity. The regression results in the main text are not sensitive to the exclusion of the background variables. Furthermore, table 5.10 reports absolute frequencies over all schooling categories and, finally, tables 5.11 and 5.12 are analogous to the tables in the main text; they show results if actual experience is replaced by the MINCER potential experience.

Table 5.9: **Detailed Regression Results.**

	1989	1990	1991	1992	1993	1994	Pooled
Men							
Math scores	0.080	-0.190	-0.289	0.374	0.257	0.268	0.132
Math scores * Schooling	0.135**	0.162***	0.183***	0.107*	0.115*	0.123*	0.133***
Schooling	0.035*	0.072***	0.066***	0.062***	0.037	0.041*	0.051***
Schooling squared	-0.000	-0.025	-0.024	-0.018	0.008	0.002	-0.009
Experience	0.039**	0.050***	0.043***	0.038**	0.042**	0.061***	0.047***
Experience squared	-0.000	-0.001	-0.001	-0.000	-0.000	-0.001*	-0.001***
Black	-0.096***	-0.130***	-0.111***	-0.111***	-0.124***	-0.110***	-0.117***
Hispanic	0.032	-0.053**	-0.025	-0.020	-0.025	-0.023	-0.021*
Lived in urban area ’87	0.100***	0.114***	0.116***	0.089***	0.134***	0.085***	0.106***
Lived in north-east ’87	-0.005	-0.008	-0.024	-0.000	-0.019	-0.028	-0.014
Lived in north-cent. ’87	-0.077***	-0.104***	-0.120***	-0.092***	-0.120***	-0.118***	-0.106
Lived in south ’87	-0.081***	-0.100***	-0.096***	-0.092***	-0.085***	-0.123***	-0.096
Area of hi. unempl. ’87	-0.052***	-0.036***	-0.032***	-0.051***	-0.028**	-0.039***	-0.041
Height in 1985 (inches)	0.008**	0.007**	0.006*	0.009***	0.007*	0.010***	0.008
Health limit under 18	-0.118***	-0.090***	-0.065**	-0.107***	-0.098***	-0.099***	-0.099
Married	0.116***	0.111***	0.122***	0.134***	0.154***	0.136***	0.133
Member of a union	0.256***	0.254***	0.235***	0.232***	0.237***	0.233***	0.245
Constant	5.978***	5.842***	5.936***	5.799***	5.862***	5.644***	5.860
Number of observations	3413	3355	2976	2838	2874	2834	18290
Women							
Math scores	0.268	-0.178	-0.602	0.450	0.656	0.767*	0.293
Math scores * Schooling	0.106	0.173***	0.220***	0.107	0.041	0.030	0.105***
Schooling	0.003	0.054**	0.079***	0.041	-0.010	-0.042	0.024**
Schooling squared	0.040*	0.003	-0.012	0.016	0.055**	0.083***	0.029***
Experience	0.044***	0.021	0.025*	0.025**	0.034**	0.035***	0.039***
Experience squared	0.001	0.001*	0.001	0.001	0.001	0.001	0.000
Black	0.002	-0.023	-0.041	-0.006	0.005	0.024	-0.014
Hispanic	0.116***	0.122***	0.075**	0.122***	0.093***	0.114***	0.104***
Lived in urban area ’87	0.053**	0.049*	0.089***	0.082***	0.072**	0.071***	0.072***
Lived in north-east ’87	0.032	0.063*	0.043	0.014	0.097**	0.035	0.051
Lived in north-cent. ’87	-0.084***	-0.098***	-0.125***	-0.098***	-0.038	-0.058*	-0.082
Lived in south ’87	-0.036	-0.044	-0.075**	-0.081***	-0.052	-0.073**	-0.058
Area of hi. unempl. ’87	-0.068***	-0.062***	-0.054***	-0.057***	-0.035***	-0.055***	-0.056
Height in 1985 (inches)	0.009**	0.008**	0.009**	0.010***	0.014***	0.008**	0.010
Health limit under 18	-0.035	-0.054*	0.004	0.009	-0.045	-0.046	-0.034
Married	-0.029	-0.032	-0.033	0.017	-0.032	-0.066***	-0.030
Member of a union	0.230***	0.174***	0.207***	0.228***	0.240***	0.220***	0.219
Constant	5.823***	5.831***	5.583***	5.585***	5.378***	5.875***	5.657
Number of observations	3305	3269	2851	2666	2766	2732	17589

Estimates obtained from the basic model with only one schooling variable and actual experience. Observations are weighted by the NLSY sample weights. Standard errors – adjusted for heteroskedasticity – are omitted. Stars denote statistical significance, *: 10%, **: 5%, ***: 1%.

Table 5.10: **Observations Per Education Cell.**

Years	1989	1990	1991	1992	1993	1994	1989	1990	1991	1992	1993	1994
	Men						*Women*					
4	4	4	4	3	3	1	2	1	0	1	1	0
5	6	6	1	3	2	1	2	3	1	1	2	0
6	15	14	9	7	6	8	7	9	6	7	8	10
7	27	24	18	14	11	9	18	16	14	12	12	14
8	101	95	69	67	61	59	40	45	32	27	24	23
9	167	158	113	107	100	92	88	90	57	55	56	42
10	172	155	130	112	111	100	108	105	73	72	70	78
11	210	200	166	156	147	138	123	120	86	78	80	80
12	1582	1542	1377	1350	1350	1328	1536	1516	1283	1213	1245	1213
13	239	242	227	201	218	217	287	280	276	263	278	278
14	236	229	222	206	220	219	317	309	285	273	275	278
15	102	105	99	91	92	102	152	135	143	121	125	139
16	413	414	372	352	362	355	477	472	409	370	394	382
17	64	73	64	67	65	66	76	82	85	75	88	88
18	48	55	57	57	69	79	53	64	67	63	70	70
19	24	33	31	28	29	28	16	17	19	22	25	26
20	28	30	31	30	39	42	14	18	22	22	24	21
Total	3438	3379	2990	2851	2885	2844	3316	3282	2858	2675	2777	2742

Table 5.11: **Men, Potential Experience.**

	1989	1990	1991	1992	1993	1994	Pooled
Basic Model							
Ability (÷100)	0.375	0.049	-0.011	0.667	0.564	0.587	0.361**
Abil*School (÷100)	0.104*	0.140**	0.155**	0.079	0.089	0.096	0.112***
Schooling	0.068***	0.088***	0.090***	0.094***	0.053**	0.056**	0.068***
Schooling^2 (÷10)	-0.010	-0.026	-0.036*	-0.039**	0.005	-0.003	-0.014*
Homosk., p-value	0.059	0.225	0.089	0.177	0.007	0.000	0.000
Pre-test and Post-test Schooling							
Ability (÷100)	0.359	0.011	0.011	0.671*	0.591	0.570	0.335**
Abil*School (÷100)	0.097	0.136**	0.138**	0.061	0.068	0.086	0.110***
Schooling	0.041	0.060**	0.073***	0.065**	0.034	0.031	0.053***
Schooling^2 (÷10)	-0.026	-0.049*	-0.099***	-0.100***	-0.061**	-0.034	-0.013
S_1	0.026	0.026	-0.003	0.009	-0.004	0.017	0.014*
$S_1 * S$ (÷10)	0.023	0.032	0.099***	0.097***	0.107***	0.050	0.006
Homosk., p-value	0.071	0.187	0.094	0.217	0.009	0.000	0.000
F-test	4.353	6.092	12.090	13.772	10.331	5.662	34.075
P-value	0.013	0.002	0.000	0.000	0.000	0.004	0.000
Basic Model, IV for Ability							
Ability (÷100)	0.863	1.503**	1.551**	2.247***	1.461*	1.738**	1.529***
Abil*School (÷100)	0.126	-0.028	-0.027	-0.093	0.016	-0.023	0.000
Schooling	0.059*	0.040	0.038	0.043	0.025	0.019	0.031**
Schooling^2 (÷10)	-0.017	0.008	0.001	-0.005	0.019	0.021	0.010
Homosk., p-value	0.068	0.238	0.093	0.198	0.008	0.000	0.000
Hausman, p-value	0.005	0.029	0.026	0.016	0.099	0.079	0.000
Canonical correl.	0.250	0.246	0.238	0.240	0.248	0.258	0.248
	0.308	0.299	0.301	0.322	0.317	0.338	0.312
Overid., p-value	0.203	0.129	0.945	0.624	0.288	0.508	0.191
Pre-test and Post-test Schooling, IV for Ability							
Ability (÷100)	0.775	1.367*	1.521*	2.143***	1.474*	1.667**	1.442***
Abil*School (÷100)	0.127	-0.023	-0.039	-0.101	-0.003	-0.028	0.007
Schooling	0.042	0.021	0.030	0.026	0.014	0.002	0.022
Schooling^2 (÷10)	-0.030	-0.018	-0.065**	-0.068**	-0.047	-0.012	0.006
S_1	0.018	0.018	-0.011	0.000	-0.012	0.010	0.009
$S_1 * S$ (÷10)	0.018	0.036	0.103***	0.099***	0.107***	0.052	0.008
Homosk., p-value	0.077	0.196	0.099	0.236	0.011	0.000	0.000
Hausman, p-value	0.023	0.071	0.093	0.080	0.154	0.167	0.000
Canonical correl.	0.244	0.239	0.232	0.233	0.237	0.248	0.243
	0.298	0.292	0.294	0.315	0.311	0.331	0.311
Overid., p-value	0.208	0.129	0.930	0.734	0.347	0.525	0.222
Number of obs.	3413	3355	2976	2838	2874	2834	18290

Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors – adjusted for heteroskedasticity – are omitted. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. F-tests in the second panel test the augmented pre-/post-test schooling model against the basic model.

Table 5.12: **Women, Potential Experience.**

	1989	1990	1991	1992	1993	1994	Pooled
Basic Model							
Ability (÷100)	1.119**	0.498	0.014	1.118**	1.696***	1.606***	1.001***
Abil*School (÷100)	-0.001	0.087	0.146**	0.021	-0.083	-0.059	0.019
Schooling	0.058**	0.090***	0.109***	0.081***	0.059*	-0.003	0.064***
Schooling^2 (÷10)	0.013	-0.010	-0.021	-0.002	0.025	0.070***	0.014
Homosk., p-value	0.302	0.470	0.968	0.017	0.129	0.592	0.011
Pre-test and Post-test Schooling							
Ability (÷100)	1.091**	0.477	0.000	1.103**	1.637***	1.598***	0.991***
Abil*School (÷100)	-0.006	0.090	0.147**	0.019	-0.084	-0.066	0.019
Schooling	0.065*	0.068**	0.099***	0.069**	0.052	0.009	0.060***
Schooling^2 (÷10)	-0.044	0.014	-0.015	-0.006	-0.012	0.028	0.014
S_1	-0.022	0.033	0.015	0.014	0.001	-0.024	0.004
$S_1 * S$ (÷10)	0.085**	-0.039	-0.011	0.005	0.054	0.066*	0.002
Homosk., p-value	0.276	0.458	0.973	0.014	0.124	0.573	0.010
F-test	4.119	0.844	0.197	0.750	3.140	1.850	34.575
P-value	0.016	0.430	0.821	0.472	0.043	0.157	0.000
Basic Model, IV for Ability							
Ability (÷100)	0.595	-0.154	-0.272	-0.342	1.459	0.931	0.368
Abil*School (÷100)	0.064	0.181	0.196	0.226*	-0.053	0.037	0.109**
Schooling	0.073**	0.110***	0.119***	0.123***	0.066*	0.017	0.084***
Schooling^2 (÷10)	0.002	-0.027	-0.030	-0.036	0.020	0.054*	-0.002
Homosk., p-value	0.302	0.457	0.984	0.014	0.130	0.562	0.009
Hausman, p-value	0.748	0.623	0.865	0.125	0.953	0.660	0.000
Canonical correl.	0.236	0.244	0.244	0.248	0.253	0.241	0.247
	0.297	0.330	0.316	0.314	0.344	0.331	0.324
Overid., p-value	0.409	0.972	0.801	0.963	0.742	0.007	0.057
Pre-test and Post-test Schooling, IV for Ability							
Ability (÷100)	0.556	-0.244	-0.329	-0.413	1.395	0.966	0.342
Abil*School (÷100)	0.059	0.191	0.202	0.229*	-0.060	0.023	0.111**
Schooling	0.079**	0.089**	0.110***	0.110***	0.058	0.027	0.080***
Schooling^2 (÷10)	-0.056	-0.004	-0.023	-0.038	-0.017	0.013	-0.003
S_1	-0.021	0.034	0.015	0.017	0.002	-0.024	0.004
$S_1 * S$ (÷10)	0.086**	-0.039	-0.013	0.001	0.055	0.065	0.002
Homosk., p-value	0.274	0.444	0.989	0.012	0.125	0.548	0.008
Hausman, p-value	0.831	0.773	0.947	0.215	0.905	0.842	0.000
Canonical correl.	0.229	0.236	0.240	0.244	0.245	0.235	0.247
	0.296	0.328	0.311	0.307	0.339	0.325	0.324
Overid., p-value	0.316	0.975	0.805	0.962	0.754	0.005	0.057
Number of obs.	3413	3355	2976	2838	2874	2834	18290

Background variables, experience, and its square are controlled. The regressions are weighted by the sample weights. Standard errors – adjusted for heteroskedasticity – are omitted. Stars denote statistical significance in a two-sided test, *: 10%, **: 5%, ***: 1%. F-tests in the second panel test the augmented pre-/post-test schooling model against the basic model.

Chapter 6

The Evaluation of Community-Based Interventions: A Monte Carlo Study

September 1998/September 2000

Together with **Christoph M. Schmidt**

Abstract. The evaluation of interventions such as active labor market policies or medical programs by means of a randomized controlled trial is often considered the gold standard. However, randomized experiments might face severe shortcomings especially if performed at the group level. One such problem is caused by small sample size which might prevent the experiment from developing its fundamental virtue in balancing all relevant covariates. This paper investigates the potential and limits of experimental and non-experimental approaches to the evaluation problem, in particular the use of instrumental variables, in a numerical simulation study, against the particular background of community-based interventions. In our simulations, we emphasize the trade-off between bias and precision by imposing a smaller number of communities whenever we model a randomized experiment, and by allowing for a correspondingly larger number of communities in all cases where selection into the program is not controlled completely by the analyst.

6.1 Program Evaluation: The Perils of Self-Selection

Self-selection is a fundamental obstacle for the evaluation of policy interventions. In the classical case of an individually-based program, for instance a voluntary training program for unemployed workers, potential trainees will in all likelihood base their participation decision on a comparison of their perceived post-intervention outcomes with the cost of undergoing treatment. It will often be the candidates with better schooling, and the more talented or motivated individuals who tend to enter the program. As a consequence of such self-selection, analysts cannot base the assessment of program impact on a simple comparison of mean outcomes between participant and non-participant groups. Whereas it is straightforward how to tackle selection on observables – schooling in the example of the training program –, selection on unobservables – talent, motivation – provides a serious intellectual challenge.

In the overwhelming majority of applications the *mean effect of treatment on the treated*, that is the population average over the individual gains from treatment for all individuals participating in the program, is the principal object of interest. One can easily construct an estimate of the mean outcome after treatment for program participants from observed data. Yet, to perform an appropriate comparison, one has also to construct the average *counterfactual* outcome that trainees would have achieved had they not been trained, a problem of identification.

Observable data alone will not suffice to construct this entity. Researchers have proposed several alternative strategies to overcome this *identification problem*, either by invoking *a priori* information on the process of selection into treatment or other aspects of the program (Heckman & Robb, 1985, Angrist & Krueger, 1999) in a so-called observational study, or by designing an appropriate experiment. In an experiment (the classical reference is Fisher, 1935), participation is still voluntary, but some of the applicants are withheld the treatment. Who receives treatment and who does not is chosen by a random mechanism, allowing the construction of the desired counterfactual as the simple average over randomized-out controls. In the natural sciences this *randomized controlled trial* (RCT) has become the method of choice for the evaluation of interventions.

While emphasis in methodological work is on the individual, practical applications frequently concern the case of group-level or community-based interventions. Implementation of policy measures at the community-level is often a matter of necessity – whenever it would be difficult to treat some individuals in a community while excluding others, evaluation has also to be at the community-level. For instance, the evaluation of an anti-smoking information campaign at schools would require that some schools as a whole be assigned to treatment. Obviously, it would be quite cumbersome, if not impossible, to plan the intervention and its evaluation at the student level. Spill-over effects from the treated to the control students would easily contaminate the experiment. Hence, the units of interest should be groups. Moreover, analysts might choose a community-level approach to evaluation for reasons of costs. In general, interventions relevant to the social sciences often have a community-based character.

Nothing seems more natural as a methodological approach to the evaluation of community-based interventions as the translation of the RCT paradigm to the community level. Objects of randomized assignment into treatment and control samples are then entire communities, while outcomes are typically still measured at the individual level. A comprehensive overview of the theory and practice of such group-randomized trials is Murray (1998). It has long been recognized in the literature in various fields, for instance in the epidemiological literature cited in Murray (1998) and in the economics literature (see e.g. Kloek, 1981, Moulton, 1986, 1990), that the possible correlation of outcomes within communities, clusters, or groups might seriously distort conclusions regarding the statistical precision of the results.

Although one might be able to collect data on sizeable numbers of individuals within each community participating in the study, the number of communities is typically limited. If within-community correlation is substantial the effective number of observations is closely tied to the number of included communities, irrespective of the number of individuals. Thus, although group-randomized experiments implemented appropriately always produce unbiased estimates[1], it is difficult to increase precision.

[1] Contamination of randomized experiments as, for instance, attrition of treated and control units is disregarded in this paper.

Observational studies, by contrast, typically include a respectable number of communities, yet they might suffer from the selection problem. Possibly a biased but more precise estimate from an observational study might yield a lower mean squared error than the corresponding estimate of program impact from a group-randomized experiment. Thus, there might be a serious trade-off to consider in the choice of evaluation strategy (Schmidt, Baltussen & Sauerborn, 1999).

Moreover, the context of community-based interventions makes it unlikely that a randomization study can be conducted at all. Problems preventing the researcher from implementing a group-randomized experiment may be political or ethical in nature, or reflect cost considerations. However, contrary to what many practitioners apparently believe to be the state of the art – either analyze an experiment or rely on simple regression analysis to alleviate some of the disadvantages of observational data – does not properly reflect the spectrum of identification strategies for dealing with observational data. While the economic literature has long emphasized the potential of the instrumental variables method (see e.g. Bowden & Turkington, 1984, Angrist, Imbens & Rubin, 1996, Heckman, 1996), this method has not been prominent in the epidemiological literature (where it has been advocated recently by Schmidt et al., 1999).

This paper investigates the potential and limits of experimental and non-experimental approaches to the evaluation problem, in particular the use of instrumental variables, in a numerical simulation study, against the particular background of community-based interventions. In our simulations, we emphasize the trade-off between bias and precision by imposing a smaller number of communities whenever we model a randomized experiment, and by allowing for a correspondingly larger number of communities in all cases where selection into the program is not controlled completely by the analyst. We specify several variants of selection, on the individual and the community-level, and on the basis of observable and unobservable factors. Specifically, we explore the potential of instrumental variables in approximating the performance of randomized experiments (for a complementary simulation study on instrumental variables at the group-level see Shore-Sheppard, 1996).

The following section formulates the basic evaluation problem and presents several estimation techniques that have been suggested for its solution. The conceptual design of our simulation study is explained in section 3. Section 4 discusses the results with a focus on the assessment of estimator performance, while section 5 concludes.

6.2 Evaluation Strategies

This section provides the formal background for our simulation study by a statement of the evaluation problem and of several solutions suggested in the literature, in particular the method of instrumental variables. Assume that Y is the outcome variable of interest. For notational convenience, subscripts indicating individuals are suppressed. Let Y_0 be the *potential* outcome if the individual would not participate in treatment and Y_1 be the *potential* outcome if the individual would. Note that only one of the potential outcomes is realized for each individual. Furthermore, let $T \in \{0, 1\}$ be a dummy variable indicating whether a unit is treated, $T = 1$, or not, $T = 0$. Under the assumption of independence of potential outcomes from the treatment status of other individuals (SUTVA, Rubin, 1986) the expected effect of treatment on the treated unit can formally be written as

$$\Delta = \mathbb{E}(Y_1 - Y_0 | T = 1). \tag{6.1}$$

While $\mathbb{E}(Y_1|T = 1)$ is easily identified in the subsample of all treated units, there is no way to identify the counterfactual $\mathbb{E}(Y_0|T = 1)$ unless further assumptions are imposed. The least restrictive way to gather information on $\mathbb{E}(Y_0|T = 1)$ is presented in Manski (1990, 1995) who demonstrates how upper and lower bounds for the counterfactual can be obtained on the basis of indisputable *a priori* information. For instance, a dichotomous outcome variable cannot take a value lower than 0 and higher than 1. In this fashion, at least some values of $\mathbb{E}(Y_0|T = 1)$ can be excluded. If one desires point estimation of the counterfactual, however, one cannot avoid either imposing additional assumptions or addressing the issue already at the stage of designing the study.

A randomized experiment is following the second route to solve this problem as follows[2].

[2] Heckman & Smith (1995) discuss the problems of contamination that might arise in randomized

Units who decide to participate in the program, i.e. units with $T = 1$, are randomly assigned to either an experimental *treatment group* or *control group*. The units of the control group are denied treatment and, thus, realize the potential outcome Y_0. It follows that the control group provides an unbiased estimate of the counterfactual $\mathbb{E}(Y_0|T = 1)$.

Yet, while randomization at the individual level is arguably an ideal way to identify causal relationships, randomized experiments usually suffer from low sample sizes if pursued at the group level. On the other hand, observational studies do much better with regard to the sample size but additional assumptions have to be invoked to identify the counterfactual $\mathbb{E}(Y_0|T = 1)$. To this purpose, several estimators have been proposed. Some of them are presented in this section including necessary assumptions that make them valid; abstracting from finite sample variations only population moments are considered.

Cross-Section and Before-After Estimators

A first assessment of an intervention might be based on comparing treated and untreated individuals after treatment occurred. Unfortunately, the mean difference of their outcomes identifies the mean effect of treatment, equation (6.1), only under strong assumptions on the selection process. Formally, $\mathbb{E}(Y_0|T = 0)$ must be a valid substitute for the counterfactual $\mathbb{E}(Y_0|T = 1)$ which requires that treated and untreated individuals be equal with respect to characteristics that rule both the selection process and the outcome equation.

Another straightforward approach to identifying the effect of an intervention rests on the availability of data for a period t' prior to treatment. In this case, the mean outcome before treatment (at time t') is compared with the outcome after the treatment (at time t), $\mathbb{E}(Y_1^t - Y_0^{t'}|T = 1)$. As above, this approach requires equally restrictive assumptions to hold; otherwise following it might cause severe biases. If external disturbances over time and beyond treatment influence the outcome variable of some units these might

experiments. Nonrandom attrition of participants or randomization biases are prominent examples. Such problems are not considered in this paper. In general, analysts stress the advantages of randomization, though. For instance, Burtless (1995) emphasizes the positive aspects of randomized experiments.

wrongly be attributed to the intervention, producing biased estimates. For instance, it is inappropriate to perform a before-after comparison when the economic environment is characterized by cyclical swings that typically affect individuals under study.

Difference-in-Differences Estimator

A combination of the before-after comparison and the cross-section estimator leads to the difference-in-differences approach (d-i-d). It rests on the assumption that – apart from treatment – both the treated and the untreated units experience the same time-varying shocks. Assuming the time trend in the outcome variable is the same for treated and untreated units

$$\mathbb{E}(Y_0^t - Y_0^{t'}|T = 1) = \mathbb{E}(Y_0^t - Y_0^{t'}|T = 0),$$

the before-after comparison of the untreated group $\mathbb{E}(Y_0^t - Y_0^{t'}|T = 0)$ on average reflects exactly the bias inherent in the simple before-after comparison of the treated units. Subtracting this correction term yields

$$\Delta = \mathbb{E}(Y_1^t - Y_0^{t'}|T = 1) - \mathbb{E}(Y_0^t - Y_0^{t'}|T = 0).$$

In other words, d-i-d requires that the difference $(Y_0^t - Y_0^{t'})$ be mean independent of the treatment T. This is violated, e.g., if the decision to participate is determined by the individual pre-treatment outcome $Y_0^{t'}$.[3] The simulation study takes account of such a selection process when opportunity costs reflected by $Y_0^{t'}$ are involved.

Instrumental Variable Estimator

Finally, instrumental variable estimation (IV) is an evaluation strategy that enjoys considerable prominence in economics (see Angrist et al., 1996 and Heckman, 1996). It has been advocated in the epidemiological literature as a possible tool to evaluate community-based interventions by Schmidt et al. (1999). Consider initially the context of a constant-effects model and assume a variable Z exists that (i) is correlated with

[3] If $Y_0^{t'}$ determines selection unobserved stochastic noise in period t' will be unevenly distributed between treated and untreated units but noise in t – if only weakly correlated with that in t' – will again be more evenly distributed. Thus, the difference $(Y_0^t - Y_0^{t'})$ will depend on T.

the endogenous treatment indicator T, but that (ii) does not have a direct influence on the outcome variable Y except through T. This variable is called an *instrument* for T. The IV technique rests on the idea that the covariance between the outcome and the instrument reflects the impact of the endogenous regressor – the parameter of interest Δ – multiplied by the covariance between the regressor and the instrument.

In case of a binary instrument, the IV estimation technique provides a consistent estimator of the mean effect of treatment on the treated, resting on the ratio[4]

$$\Delta = \frac{Cov(Y,Z)}{Cov(T,Z)} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(T|Z=1) - \mathbb{E}(T|Z=0)}.$$

In randomized experiments, T is its own perfect instrumental variable. Since the randomized experiment is by definition independent of the outcome, and individuals perfectly comply with their treatment assignment indicated by the dichotomous indicator Z, the correlation between T and Z is 1 in absolute value (see also Heckman, 1996). Correspondingly, an instrument Z can be interpreted as a variable that is randomly distributed across units but, in contrast to a fully randomized experiment, only imperfectly induces units to behave according to its realized value. In other words, IV estimation is a *quasi-experimental* technique. Although the IV estimator is consistent if the two principal assumptions (i) and (ii) are satisfied, it might be accompanied by large variance in finite samples, especially if the correlation between the instrument and the endogenous variable T is weak (Bound, Jaeger & Baker, 1995).

A subtle issue is added to estimation with instrumental variables if the treatment effect is heterogeneous, though, i.e. if we leave the realm of the constant-effects model. Furthermore, if selection into treatment is based on the individual effects of the treatment, IV does not identify the mean effect of treatment on the treated. Rather, the IV estimator

[4] The second equation is easily verified

$$\begin{aligned} Cov(Y,Z) &= \mathbb{E}(YZ) - \mathbb{E}Y\mathbb{E}Z \\ \mathbb{E}(YZ) &= \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1) \\ \mathbb{E}Y &= \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1) + \mathbb{E}(Y|Z=0)\mathbb{P}(Z=0) \\ \mathbb{E}Z &= \mathbb{P}(Z=1), \end{aligned}$$

Thus, $Cov(Y,Z) = \mathbb{E}(Y|Z=1)\mathbb{P}(Z=1)\underbrace{(1-\mathbb{P}(Z=1))}_{=\mathbb{P}(Z=0)} - \mathbb{E}(Y|Z=0)\mathbb{P}(Z=0)\mathbb{P}(Z=1)$. Likewise, $Cov(T,Z)$ is transformed.

converges to the average treatment effect for all those individuals who are induced by the instrument to enter the treatment but who would have stayed off treatment otherwise. This entity is the so-called *local average treatment effect* (LATE). Recent research typically re-interprets the IV estimate as a LATE, see e.g. Angrist et al. (1996) and Heckman (1997). Its peculiarities are discussed in section 4.

Conditioning on Observables

Whenever researchers succeed in capturing *observable* elements of the process jointly determining outcomes and program participation, they can improve upon their evaluation strategy by conditioning on these observable covariates. In anticipation of the simulation setup implemented below, let X be an explanatory binary variable that takes the values 0 and 1. If self-selection depends in part on the realization of X, then within the subsamples defined by $X = 0$ and $X = 1$, any remaining bias can only reflect the presence of other factors. The selection bias would even disappear completely if selection depended exclusively on X apart from random disturbances.

Then, participation is purely *random* within the two subsamples characterized by $X = 0$ and $X = 1$, i.e. T is independent of (Y_0, Y_1) given X. It follows that the untreated units in the subsamples $\{X = x, T = 0\}$ provide the counterfactual $\mathbb{E}(Y_0|X = x, T = 1)$.[5] The unconditional mean $\mathbb{E}_X\, \mathbb{E}(Y_0|X, T = 1)$ is obtained as weighted average over the conditional means. In sum, if $\mathbb{E}(Y_0|X = x, T = 1) = \mathbb{E}(Y_0|X = x, T = 0)$ it follows

$$\begin{aligned}\mathbb{E}(Y_1 - Y_0|T = 1) &= \sum_{x\in\mathcal{X}} \mathbb{E}(Y_1 - Y_0|X = x, T = 1)\mathbb{P}(X = x) \qquad (6.2)\\ &= \sum_{x\in\mathcal{X}} (\mathbb{E}(Y_1|X = x, T = 1) - \mathbb{E}(Y_0|X = x, T = 0))\mathbb{P}(X = x)\end{aligned}$$

where $\mathcal{X}$ is the set of all possible values of X.

Since *unobservable* variables might additionally play a role in determining the selection process, conditioning on observables alone might not enable the researcher to avoid selection bias. Yet, at least conditioning on observables might achieve to mitigate the

[5] However, in the extreme case when selection is fully determined by an observable X without any stochastic components left the set $\{X = x, T = 0\}$ is either empty or equals $\{X = x\}$ making it impossible to obtain the counterfactual from untreated individuals.

problem. Because of that, it is recommended whenever possible. In this study, it is very easy to follow this advice due to the binary nature of the observable variables. In practice however, X might be of high dimension, thus, conditioning on all observables would be quite cumbersome or even impossible. This problem is alleviated by imposing a regression model or by conditioning on the *propensity score*, which is the probability of participation given the observable variables. Then, the sample would be stratified into subsamples of units with equal or similar propensity scores and the overall treatment effect estimate would be constructed as a weighted average as in (6.2); the technique is often referred to as *matching*. For further discussion of this topic, see e.g. Rubin (1973, 1974), Rosenbaum & Rubin (1983, 1984, 1985), and Heckman, Ichimura & Todd (1997).

6.3 The Simulation Setup

The simulation is based on a data generating process that consists of two main equations, the outcome and the selection equation, and two time periods, one before and one after treatment. While the outcome equation always combines observable and unobservable characteristics with heterogeneous treatment effects, we consider two conceptually distinct modes of selection into treatment. In one set of experiments, selection into treatment is at the individual level – here we do not expect group level variables to introduce any fundamental difficulties; we also consider situations, though, in which selection into treatment is decided upon at the group level. It is these simulations where we particularly expect new insights to emerge from our simulations.

The outcome Y_{igt} of individual i, $i = 1, ..., n_g$, in group g, $g = 1, ..., G$, at time $t \in \{0, 1\}$ depends linearly on time-invariant individual and group characteristics X_{1ig} and X_{2g}, respectively, which are observable, as well as on unobservable characteristics, ν_{1ig} and ν_{2g}. Furthermore, a variable μ_t captures exogenous time-variant shocks being constant for all individuals in a given time period but displaying an upward time trend. The unobservable variables ε_{1igt} and ε_{2gt} reflect white noise at the individual and the group level. The treatment effect is a sum of an individual effect δ_{1ig} and a group effect δ_{2g} which are both random variables resulting in heterogeneity in the impact of treatment across

Table 6.1: **Variables and Parameters.**

Variable	Comment	Parameter	Comment
Outcome Equation			
Constant	–	α_0	0
X_{1ig}	binomial(1, $\frac{1}{2}$)	α_1	1
X_{2g}	binomial(1, $\frac{1}{2}$)	α_2	1
ν_{1ig}	$\mathcal{N}(0, \frac{1}{4})$	1	–
ν_{2g}	$\mathcal{N}(0, \frac{1}{4})$	1	–
δ_{1ig}	$\mathcal{N}(\frac{1}{2}, \frac{1}{4})$	1	–
δ_{2g}	$\mathcal{N}(\frac{1}{2}, \frac{1}{4})$	1	–
ε_{1igt}	$\mathcal{N}(0, \frac{1}{2})$	$\text{Corr}(\varepsilon_{1ig0}, \varepsilon_{1ig1})$	0.25
ε_{2gt}	$\mathcal{N}(0, \frac{1}{2})$	$\text{Corr}(\varepsilon_{2g0}, \varepsilon_{2g1})$	0.25
μ_t	$\mathcal{N}(\frac{1}{2}t, \frac{1}{16})$	1	–
Cost Equation			
Constant	–	τ_0	such that 50% of the sample participate
Z_{1ig}	binomial(1, $\frac{1}{2}$)	τ_1	suitable for given correlation (Z_1, T)
Z_{2g}	binomial(1, $\frac{1}{2}$)	τ_2	suitable for given correlation (Z_2, T)
η_{ig}	$\mathcal{N}(0, 1)$	1	–

The variables are independently and identically distributed if not mentioned otherwise.

individuals and groups. The dichotomous variable T_{igt} indicates the treatment status. In sum,

$$Y_{igt} = \alpha_0 + \alpha_1 X_{1ig} + \alpha_2 X_{2g} + (\delta_{1ig} + \delta_{2g})T_{igt} + \nu_{1ig} + \nu_{2g} + \mu_t + \varepsilon_{1igt} + \varepsilon_{2gt}. \qquad (6.3)$$

In our simulations X_1 and X_2 are binary variables taking the values 0 and 1 with equal probability. Both treatment effects δ_{1ig} and δ_{2g} follow a normal distribution with mean $\frac{1}{2}$ and variance of $\frac{1}{4}$, the ν's and ε's are distributed normally with mean zero and variance $\frac{1}{4}$ and $\frac{1}{2}$, respectively. Both individual and group ε's are positively correlated over time with value 0.25, and $\mu_0 \sim \mathcal{N}(0, 1/16)$ and $\mu_1 \sim \mathcal{N}(0.5, 1/16)$. The constant α_0 equals 0 while $\alpha_1 = \alpha_2 = 1$. Table 6.1 summarizes parameters and variables.

When selection into treatment is considered to be an individual decision, it is modeled as an optimization process as in Heckman, LaLonde & Smith (1999: ch. 8).

Individuals decide to participate if they expect to gain from treatment and, thus,

$$T_{igt} = \begin{cases} \mathbf{1}[G_{ig} > 0] & : \quad t = 1 \\ 0 & : \quad t = 0. \end{cases} \tag{6.4}$$

The individual net gain G_{ig} represents the difference between benefits and cost of treatment. The benefits comprise all future treatment effects $\delta_{1ig} + \delta_{2g}$ discounted to present value assuming a constant discount factor of 0.1 and constant effects beyond period $t = 1$. The cost of treatment is the sum of opportunity costs and other costs C_{ig} to be specified below. The opportunity costs of undergoing treatment comprise outcome before treatment Y_{ig0} reflecting the presence of observable and unobservable characteristics. Finally, net gains are contaminated by stochastic noise η_{ig}.

It has been recognized in the evaluation literature (see Heckman, 1997) that the information available to individuals at the time of their decision whether to participate in a program is a decisive element of the selection effects to be expected. Specifically, if individuals know their own treatment effect and act upon it, the presence of heterogeneous treatment effects will necessarily lead – *ceteris paribus* – high-impact individuals to be over-represented among the individuals receiving treatment. In consequence, the mean effect of treatment on the treated will exceed the population average of the treatment effects.[6]

On the other hand, individuals acting upon the precise knowledge of their opportunity costs during the treatment period $t = 0$ will – *ceteris paribus* – typically choose to receive treatment if their time-invariant characteristics generate relatively low outcomes in both periods. While observable characteristics are controlled for easily enough, it is the unobservables which create the *selection effects* any successful evaluation strategy has to deal with. We will consider situations in which individuals select treatment on the basis of information on (i) opportunity costs Y_{ig0} and their expectation of treatment $\mathbb{E}(\delta_{1ig} + \delta_{2g})$, on (ii) precise information about both Y_{ig0} and $\delta_{1ig} + \delta_{2g}$, and on (iii) *expected* opportunity costs $\mathbb{E}Y_{ig0}$ and on expected effects $\mathbb{E}(\delta_{1ig} + \delta_{2g})$ conditional on time-invariant

[6] Naturally, as long as the evaluation strategy will be able to identify the mean treatment effect for this subpopulation, this is not a fundamental flaw of the setup, but rather a beneficial consequence of the liberation from a constant-effects model.

characteristics, respectively[7]. These various alternatives of G_{ig} can generally be written as

$$G_{ig} = \mathbb{E}\left(\frac{\delta_{1ig} + \delta_{2g}}{0.1} \middle| \Omega \right) - \mathbb{E}\left(Y_{ig0}|\Omega\right) - C_{ig} + \eta_{ig} \tag{6.5}$$

with $\mathbb{E}(\cdot|\Omega)$ denoting a conditional expectation given information set Ω. This set may contain a subset of all relevant variables, but may also contain all variables, observable and unobservable, rendering the expectation operator unnecessary. Thus, depending on the fineness of Ω either one or even both of the two expectation terms in equation (6.5) may coincide with identity.

Other costs C_{ig} allow the introduction of instrumental variables most naturally. Consider costs being a function of two variables Z_{1ig} and Z_{2g}, where Z_{1ig} is defined at the individual and Z_{2g} at the group level; they take the values 0 and 1 with probability 0.5 each. These variables reflect aspects such as, for example, the distance to the treatment site. In effect, other costs are

$$C_{ig} = \tau_0 - \tau_1 Z_{1ig} - \tau_2 Z_{2g}. \tag{6.6}$$

The constant τ_0 is chosen such that 50% of all units undergo treatment[8] and τ_1 and τ_2 are adapted such that the correlation between the instruments Z_1 and Z_2 and treatment choice T correspond to a given value (see also table 6.1)[9].

Treatment choice is a completely different matter if it is decided upon at the group level. Most importantly, if one of the individuals in a group receives treatment, so do all other members of the group. In our study some democratic majority decision rule is supposed to govern treatment choice. That is, the form of the selection equation (6.4) is retained, albeit with the group specific expected gain G_g as its argument, and thus $T_{ig1} = \mathbf{1}(G_g > 0)$. The gain G_g of a group is simply the sum over all individual expected gains. The same information scenarios arise as under individual treatment choice. Thus, groups join treatment if their aggregate expected gain is positive. Note that summing up individual gains G_{ig} of the group members reduces considerably the importance of all

[7] The timing of treatment choice and outcome realization renders the scenario $\mathbb{E}Y_{ig0}$ and $\delta_{1ig} + \delta_{2g}$ irrelevant.

[8] In fact, this is done by replacing the criterion $G_{ig} > 0$ in equation (6.4) by $G_{ig} > median(G_{ig})$.

[9] The costs might be dependent on the other covariates X and ν, too, but this would complicate the setup without further illuminating the main aspects of the simulation study.

individual level variables as far as selection into treatment is concerned and the group variables clearly dominate the decisions.

Any observational study would proceed along the following lines: take the sample of treated and untreated individuals (with treatment varying within groups or not), and observed individual-level and group-level characteristics (X and Z), under a specific identification assumption, e.g. mean independence of treatment and outcomes conditional on observable X. Alternatively, one might be able to tackle the evaluation problem by design, namely by constructing a randomized controlled trial. In this study, we consider randomized experiments which are performed at the individual and at the group level. Throughout, these experiments are assumed not to be contaminated by attrition or by randomization bias and, throughout, they recruit their volunteers from the pool of individuals (or groups) who are willing to participate, i.e. those with a positive net gain. Irrespective of the level of implementation, these experiments identify the mean effect of treatment on the treated under all combinations of parameters.

However, since randomized experiments, in particular those conducted on the group level, usually suffer from small sample size, the corresponding impact estimates might display a high variance compared to estimates of large scale observational studies: although an experiment achieves to balance all covariates *on average*, it might drastically fail to do so in a particular small sample. To alleviate this problem, our simulated randomized experiments follow the recommendation to stratify samples prior to randomization with respect to observable covariates and then to perform randomization within the strata (Murray, 1998). This procedure ensures that at least observable covariates are balanced.

Nevertheless, the sample size of randomized trials is comparatively small. Thus, the number of groups in randomized experiments is set equal to 20 while the corresponding number in observational studies varies between at least 40 and up to 300. This range is used to investigate the relative performance of estimates produced by observational studies and randomized experiments. In both experimental and observational scenarios, each group consists of 50 individuals. In the field, it typically is the involvement of further communities, not of individuals within communities, which raises the cost of a study.

6.4 Simulation Results

The discussion of results focuses on two main features. First, all estimators of section 2 are presented for different more or less favorable scenarios both at the group and at the individual level and compared with regard to the root mean squared error (RMSE). Apart from root mean squared errors – reported in squared parentheses – variation net of bias, calculated as $\sqrt{RMSE^2 - bias^2}$, will be reported in round parentheses. This serves to assess the importance of the estimator's bias component when switching from one scenario to another. Second, separate more extensive simulations are performed to compare particularly the quasi-experimental technique of IV with fully randomized experiments.

Individual Level Selection

Table 6.2 is dedicated to estimation results when selection into treatment occurs at the individual level. In the basic scenario reported in column (1) only observable variables determine both the outcome and the selection equation. In particular, selection depends on the expected treatment effect $\mathbb{E}(\delta_1 + \delta_2)$, similarly, opportunity costs are captured by $\mathbb{E}(Y_0|X)$, while ν and μ are excluded from the equations. Thus, as documented in column (1), identification problems do not arise except for the simple cross-section estimator that does not control for X and thus misses to control for self-selection. The difference in RMSE between the cross-sectional and the before-after estimator is due mainly to the fact that the first is based on more observations than the latter even though correlated ε's over time help reduce the RMSE of the before-after comparison. Increasing this correlation would successively diminish the RMSE of the before-after comparison.[10] Similarly, the d-i-d estimator is affected by this correlation, too.

On the other hand, IV based on the individual instrument Z_1 suffers from the largest RMSE among all non-experimental estimators owing to its high variance but not to inconsistent estimation; a fact that is common in IV estimation: the lower the correlation

[10] The conditional before-after-comparison is omitted since X-variables are time-constant and thus the conditional and unconditional estimates coincide.

Table 6.2: **Estimation Results, Selection at the Individual Level.**

		ν's	μ's	Indiv.	Indiv.
	Basic	included	included	opport.	treatm.
				costs	effects
Estimators	(1)	(2)	(3)	(4)	(5)
True effect	**0.999**	**0.999**	**0.999**	**1.000**	**1.466**
	(0.040)	(0.040)	(0.039)	(0.039)	(0.039)
Standard Estimators					
Cross-section	**0.430**	**0.033**	**0.032**	**0.012**	**1.233**
	(0.050)	(0.066)	(0.068)	(0.071)	(0.085)
	[0.572]	[0.969]	[0.970]	[0.990]	[0.248]
– controlled for X	**0.999**	**0.458**	**0.456**	**0.360**	**1.326**
	(0.045)	(0.062)	(0.062)	(0.066)	(0.074)
	[0.045]	[0.545]	[0.547]	[0.643]	[0.159]
Before-after	**1.000**	**1.001**	**1.495**	**1.796**	**2.030**
	(0.067)	(0.068)	(0.363)	(0.357)	(0.362)
	[0.067]	[0.068]	[0.614]	[0.872]	[0.671]
Difference-in-differences	**0.999**	**1.000**	**0.998**	**1.586**	**1.605**
	(0.054)	(0.061)	(0.061)	(0.064)	(0.074)
	[0.054]	[0.061]	[0.061]	[0.590]	[0.157]
– controlled for X	**0.999**	**1.000**	**0.998**	**1.638**	**1.605**
	(0.053)	(0.062)	(0.062)	(0.066)	(0.074)
	[0.053]	[0.062]	[0.062]	[0.642]	[0.157]
Instrumental Variables Estimators					
IV Z_1	**0.999**	**0.999**	**1.001**	**1.003**	**1.000**
	(0.087)	(0.101)	(0.103)	(0.102)	(0.102)
	[0.087]	[0.101]	[0.103]	[0.102]	[0.478]
– controlled for X	**0.997**	**1.001**	**1.001**	**1.003**	**0.996**
	(0.090)	(0.100)	(0.101)	(0.096)	(0.091)
	[0.090]	[0.100]	[0.101]	[0.096]	[0.479]
– Corr(Z_1, T)	**0.302**	**0.301**	**0.301**	**0.298**	**0.299**
	(0.010)	(0.011)	(0.011)	(0.011)	(0.013)
IV Z_2	**1.023**	**1.042**	**1.026**	**1.018**	**0.998**
	(0.432)	(0.509)	(0.510)	(0.511)	(0.506)
	[0.433]	[0.511]	[0.510]	[0.512]	[0.689]
– controlled for X	**1.003**	**1.042**	**1.022**	**1.030**	**0.980**
	(0.404)	(0.481)	(0.473)	(0.477)	(0.461)
	[0.404]	[0.483]	[0.474]	[0.478]	[0.670]
– Corr(Z_2, T)	**0.300**	**0.299**	**0.299**	**0.298**	**0.298**
	(0.022)	(0.026)	(0.026)	(0.029)	(0.036)
Experimental Evaluations					
Experiment	**0.999**	**0.996**	**0.998**	**0.998**	**1.465**
	(0.097)	(0.101)	(0.104)	(0.103)	(0.087)
	[0.097]	[0.101]	[0.104]	[0.103]	[0.087]
Stratified experiment	**0.998**	**0.996**	**0.996**	**0.997**	**1.465**
	(0.099)	(0.108)	(0.107)	(0.108)	(0.092)
	[0.099]	[0.108]	[0.107]	[0.108]	[0.092]

Means over all simulation iterations. Root mean squared errors are in square parentheses. Round parentheses show variation net of bias. Number of groups in observational studies: 200, number of iterations: 5000. Controlling for X in the before-after-comparison does not change estimation results and is omitted.

between instrument and endogenous regressor T the larger the variance of the IV estimate. The coefficients of Z_1 and Z_2, τ_1 and τ_2, are chosen such that the correlations are approximately 0.3,[11] the realized values and their standard deviations across simulation iterations are shown in the table. In practice, such correlations are usually considered producing good instruments (Hall, Rudebusch & Wilcox, 1996). Further note that IV estimation controlling for X does not reduce the RMSE which can be explained by the independence of Z and X in this simulation study. Moreover, the IV estimates based on the grouped instrument Z_2 are accompanied by substantially higher variance caused by the intra-group correlation among group members which reduces the effective sample size. The detrimental effects of intra-group correlations on the precision of impact estimates is the topic of a large literature in economics (Kloek, 1981, Moulton, 1986) and epidemiology (e.g. Murray, 1998). Recently, Shore-Sheppard (1996) has extended this discussion to the problem of grouped instruments.

Experimental estimates are reported in the last two rows of the table. Taking into account their small sample size they still perform quite well: compared to the large observational studies consisting of 200 groups or 10000 individuals, the randomized experiments have to rely on only 20 groups or 1000 individuals. Yet, under these circumstances, one would prefer the standard observational approaches and the IV estimate using the individual-level instruments.

Column (2) reports estimates if unobservable characteristics ν enter the outcome equation. Naturally, this does not influence the true effect but poor performance (= low opportunity costs) might mistakenly be attributed to a poor effect of the treatment. Obviously, the cross-section estimator breaks down because it is unable to control for the unobservable ν's. However, since the ν's are time-constant, both the before-after comparison and the d-i-d are entirely unaffected by them, the unobservables just cancel out, and the estimates do not display higher variance. This is in contrast to IV: though it is consistent as a cross-sectional estimator its variance increases to a small extent. The experimental estimator, however, remains basically unaffected and performs as well as the

[11] The values of τ_1 and τ_2 are adapted step by step by performing additional simulations until the correlations take the desired value.

IV estimator.

Including time-variable shocks μ into the outcome equation destroys the before-after comparison (column 3) while the other estimators remain unaffected. Since all individuals experience a higher outcome in the post-treatment compared to the pre-treatment period irrespective of having received treatment or not, the before-after comparison wrongly attributes this general increase to the treatment. D-i-d successfully achieves to correct for this bias by exploiting the before-after comparison of the untreated units who experienced the same upward trend as the treated individuals, thus serving as controls.

The fourth column shows estimates when more severe endogeneity problems are introduced. Opportunity costs are assumed to be captured by individual outcomes before treatment, Y_{ig0}, instead of their conditional expectation, $\mathbb{E}(Y_{ig0}|X, \nu, \mu)$, as above. Consequently, ε_{1ig0} and ε_{2g0} determine selection, too, so they systematically differ between treated and untreated. Since the ε's vary over time and units, the d-i-d estimator cannot difference out the bias caused by them. The problem is the more severe the less is the correlation between ε's before and after the intervention. In the simulation the correlation is set equal to 0.25; a perfect correlation would eliminate the bias in the d-i-d estimate. Note that the first two non-experimental estimators are also negatively affected in this scenario where more unobservables than before rule selection. Specifically, the before-after-comparison suffers from regression to the mean: while ε_{1ig0} and ε_{2g0} determine selection and hence are unevenly distributed across treated and untreated individuals, ε_{1ig1} and ε_{2g1} are more evenly distributed depending on the strength of the correlation ρ. On the other hand, IV still produces estimates of the same quality as before, that is, IV based on Z_1 and the experimental approach are the preferred estimation strategies.

Finally, column (5) presents results of a scenario that additionally assumes that individuals correctly anticipate their individual gain $(\delta_{1ig} + \delta_{2g})$ from participation. Although the assumption is strong, it might be fulfilled if the setup of the program is transparent to the public and people are able to judge well how they succeed in the treatment. In this case, only the most successful individuals undergo treatment and, therefore, the mean effect on the treated increases from 1 to 1.47.

Under this optimization behavior IV breaks down[12]; it does not anymore identify the mean effect of treatment on the treated but the so-called *local average treatment effect* (LATE) which is the effect of treatment on someone who complies with the instrument, i.e. who participates, $T = 1$, if $Z = 1$ and who does not, $T = 0$, if $Z = 0$. In accordance with Angrist et al. (1996) and Imbens & Angrist (1994) further denote *always-takers* as individuals who always undergo treatment irrespective of the realization of their Z and, likewise, *never-takers* as those who never participate.[13] Disregarding the group variables for simplicity *compliers* are characterized by the set

$$\{i : 10\delta_i - Y_{i0} - \tau_0 + \eta_i \leq 0 \quad \text{and} \quad 10\delta_i - Y_{i0} - \tau_0 + \tau_1 + \eta_i > 0\}. \tag{6.7}$$

Since exactly half of the sample undergoes treatment it can be shown that the individual treatment effects of never-takers, compliers, and always-takers are ordered symmetrically around the mean value of δ_i with never-takers at the bottom, compliers in the middle, and always-takers at the top. Thus, under these special circumstances, the mean effect on compliers coincides with the mean effect on a randomly chosen person, namely 1. On the other hand, the mean effect on the treated – the always-takers and compliers who participate – exceeds 1. If the selection criterion were replaced such that exactly 40% of the sample participated, LATE would increase to above 1, if 60% underwent treatment LATE would fall below 1. LATE answers the question of how large the gain from treatment would be if the costs C_{ig} were reduced by τ_1 (or τ_2 in case of Z_2).[14]

Alas, the other non-experimental estimators do improve in this scenario which is due mainly to adverse effects which cancel out some biases. No general pattern underlies this improvement. It is merely an artefact of the special model used here. In this last scenario, solely the randomized experiment is still able to identify the parameter of interest.

Interestingly, the stratified experiments in all scenarios do not do better than the unstratified ones. This is not completely unexpected, since with a sample size of 1000

[12] Notice that the heterogeneity is not caused by observable covariates which could be coped with, but by hidden characteristics only known to the individual.

[13] A fourth category, so-called *defiers* who simply do the opposite of what their Z indicates are ruled out since T is monotonic in Z.

[14] Moreover, the set of compliers (6.7) offers an interesting intuitive interpretation of the relationship between instrumental relevance measured by correlation between Z and T and the variance of the IV estimator. High correlation induces large τ_1 and thus a large set of compliers which, in turn, increases the number of observations IV is based on, i.e. reduces its sample variance.

randomization already balances the two-dimensional observable covariates X. This will be demonstrated to be different in the context of group randomization.

Selection at the Group Level

If treatment occurs at the group level, the effective sample size shrinks because the limited variability of treatment receipt within groups confronts similarly limited variability of observable and unobservable characteristics. In fact, results presented in table 6.3 clearly demonstrate how RMSE's have increased, particularly because the estimator's variances have done so. However, the main pattern of results remains almost unchanged compared to table 6.2. The RMSE's of the cross-section estimator increases whereas that of before-after comparison only slightly rises. This is because a before-after-comparison still works at the individual level since all group variables, which are time-constant, just cancel out and individual level variation caused by ε_{1igt} gains the upper hand again. Therefore, table 6.3 presents an additional before-after-estimator based on data where ε_1 is removed and the variance of ε_2 is increased to 1. Then, the efficiency of this estimator worsens, too.

The difference-in-differences estimator doubled its RMSE and that of its counterpart controlling for X is even three times larger. Controlling for X would reduce bias caused by X, though, it increases the variance because subsamples defined by X might be rather small, specifically at the group level. Albeit, all standard estimators continue to be consistent. Concerning instrumental variables estimation, only Z_2 is a relevant instrument while correlation between Z_1 and T is negligible and therefore results are omitted.[15] Compared to table 6.2 the grouped IV estimates display higher variance which might be attributed to substantially increased variance of the instrumental correlation. If in some iteration of the simulation the correlation happens to be very small, close to zero, this iteration will contribute an extremely high variance to the mean over all iterations, particularly if X is controlled for. Yet, there are no grounds for failure of the IV in identifying the treatment effect. Finally, the experimental estimator's variance quadrupled but achieves to outperform IV. IV estimation and the randomized experiment will be

[15] At the group level, Z_2 dominates the selection equation because individual Z_1's aggregated to the group level are almost equal across all groups and, consequently, do not influence the selection process.

Table 6.3: **Estimation Results, Selection at the Group Level.**

		ν's	μ's	Indiv.	Indiv.
	Basic	included	included	opport.	treatm.
				costs	effects
Estimators	(1)	(2)	(3)	(4)	(5)
True effect	**0.999**	**0.999**	**1.001**	**1.000**	**1.364**
	(0.050)	(0.051)	(0.051)	(0.051)	(0.043)
Standard Estimators					
Cross-section	**0.284**	**-0.057**	**-0.054**	**0.059**	**1.185**
	(0.114)	(0.122)	(0.120)	(0.126)	(0.143)
	[0.724]	[1.064]	[1.062]	[0.949]	[0.229]
– controlled for X	**1.001**	**0.287**	**0.294**	**0.358**	**1.257**
	(0.177)	(0.164)	(0.161)	(0.134)	(0.125)
	[0.177]	[0.731]	[0.725]	[0.655]	[0.165]
Before-after	**0.999**	**0.998**	**1.503**	**1.768**	**1.919**
	(0.087)	(0.087)	(0.366)	(0.363)	(0.363)
	[0.087]	[0.087]	[0.621]	[0.850]	[0.663]
Before-after, no ε_1	**0.999**	**0.997**	**1.503**	**1.874**	**1.941**
	(0.122)	(0.122)	(0.376)	(0.373)	(0.373)
	[0.122]	[0.122]	[0.627]	[0.950]	[0.687]
Difference-in-differences	**0.999**	**1.000**	**1.001**	**1.548**	**1.475**
	(0.123)	(0.122)	(0.124)	(0.117)	(0.123)
	[0.123]	[0.122]	[0.124]	[0.561]	[0.165]
– controlled for X	**1.000**	**1.000**	**1.001**	**1.648**	**1.475**
	(0.217)	(0.173)	(0.177)	(0.133)	(0.124)
	[0.217]	[0.173]	[0.177]	[0.661]	[0.167]
Instrumental Variables Estimators					
IV Z_2	**1.041**	**1.101**	**1.068**	**1.065**	**1.000**
	(0.496)	(0.730)	(0.590)	(0.649)	(0.530)
	[0.497]	[0.737]	[0.594]	[0.652]	[0.643]
– controlled for X	**1.004**	**1.089**	**1.049**	**1.104**	**1.037**
	(0.540)	(0.680)	(0.625)	(0.825)	(4.545)
	[0.540]	[0.686]	[0.627]	[0.831]	[4.557]
– Corr(Z_2, T)	**0.307**	**0.295**	**0.302**	**0.288**	**0.306**
	(0.067)	(0.067)	(0.067)	(0.067)	(0.067)
Experimental Evaluations					
Experiment	**0.998**	**1.007**	**1.006**	**1.006**	**1.362**
	(0.390)	(0.415)	(0.408)	(0.432)	(0.463)
	[0.390]	[0.416]	[0.408]	[0.432]	[0.463]
Stratified experiment	**0.996**	**0.999**	**1.006**	**0.996**	**1.363**
	(0.360)	(0.397)	(0.398)	(0.402)	(0.427)
	[0.360]	[0.397]	[0.398]	[0.402]	[0.427]

Means over all simulation iterations. Root mean squared errors are in square parentheses. Round parentheses show variation net of bias. Due to negligible correlation between Z_1 and T corresponding results are not meaningful and left out. Number of groups in observational studies: 200, number of iterations: 5000.

compared in detail under several settings below. Notice that at the group level – or, in general, in small samples – stratifying the sample prior to randomization produces estimates with lower variance.

As one moves from column (1) to (4) the standard estimators worsen considerably. Note, however, that their RMSE's (in column (4) or (5)) though substantially rising do not exceed those of table 6.2 to a large extent. It is specifically the variation net of bias that has increased at the group level. Compared to the RMSE of the IV estimator the standard estimators still perform quite well. Albeit, their low RMSE's in column (5) should be taken with a grain of salt for different biases tend to cancel out due to special model constellations. This cannot be generalized. As above, in column (5) IV identifies the mean effect on compliers instead of the effect on treated units, yet, its RMSE merely slightly increases with regard to columns (2) to (4).

Exploring the Potential of IV Estimators

Up to this point, results are generated under a certain simulation setup. Neither sample size nor the correlation between instrument and treatment indicator have been varied. For a thorough assessment of the relative performance of IV with respect to pure randomization it is necessary to perform a further simulation that varies these two parameters. The variables and parameters of the scenario reported in the fourth columns of tables 6.2 and 6.3 are selected and fixed. Table 6.4 presents ratios of the root mean squared error of IV and experimental estimates for certain correlations and number of groups.

As expected, IV produces more precise estimates as the correlation and the relative sample size increase. At the individual level, for a reasonable correlation of 0.3, the observational study should comprise ten times as many groups as the randomized experiment to generate a more efficient IV estimator. At the group level, the observational study should be at least 15 times as large as the group level experiment for the same correlation of instrument and treatment participation. Holding the relative sample size of the observational study at ten times as many groups a sufficient instrumental correlation would be around 0.4. This is already a high correlation but not completely utopian in practical

Table 6.4: **IV Versus Experiment.**

	Number of groups in observational study				
Correlation	40	80	120	200	300
Individual Level					
0.105	11.740	5.094	4.003	2.794	2.196
0.199	5.419	2.734	1.964	1.502	1.181
0.301	3.454	1.847	1.354	**0.993**	**0.788**
0.401	2.705	1.355	**0.988**	0.746	0.586
0.502	2.042	1.104	0.800	0.621	0.471
0.601	1.728	**0.915**	0.695	0.511	0.406
0.703	1.469	0.762	0.564	0.418	0.348
0.799	1.214	0.669	0.536	0.380	0.298
0.904	**1.057**	0.579	0.451	0.322	0.267
0.991	0.961	0.521	0.416	0.305	0.226
Group Level					
0.100	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
0.200	$+\infty$	$+\infty$	$+\infty$	$+\infty$	2.030
0.303	$+\infty$	$+\infty$	2.035	1.379	**0.987**
0.403	$+\infty$	1.797	1.233	**0.892**	0.708
0.500	$+\infty$	1.179	**0.914**	0.703	0.569
0.602	1.459	**0.961**	0.723	0.558	0.455
0.703	1.135	0.736	0.609	0.448	0.367
0.802	**0.958**	0.648	0.496	0.388	0.316
0.910	0.818	0.565	0.442	0.324	0.259
0.941	0.789	0.537	0.427	0.309	0.260

The table reports the ratio of root mean squared errors of the IV over the experimental estimate. The number of groups in the experimental setting is 20. Number of iterations: 2000.

applications.

Moreover, note that although the ratios of RMSE's at the group level are infinitely large for low correlations and low sample sizes, they diminish faster than they do at the individual level as correlations rise. Break-even points, i.e. points where the ratios are approximately 1 or less, are bold faced in the table; in general, they are later at the group level than at the individual level indicating that IV suffers more from the grouped structure than a randomized experiment. Only for the lowest number of groups in the first column reaches group-level IV its break-even point earlier. In all other columns the break-even point is reached for a slightly higher correlation.

6.5 Conclusion

This paper performs simulations in order to assess several standard estimation strategies such as the cross-sectional differences between treated and untreated units, before-after comparisons, and, specifically, instrumental variable estimators as a prime example of a quasi-experimental estimation strategy. These are compared to conventional randomized experiments under the assumption that experiments generally suffer from small sample problems. Therefore, they rely on markedly less observations than observational studies.

Standard estimators perform well as long as somewhat restrictive assumptions on the selection process are satisfied. In practical applications, it is typically difficult to justify the applicability of these assumptions. Therefore, randomized controlled experiments often provide the only credible counterfactual control group. However, situations are conceivable – particularly in social sciences – where randomized trials reach their limits. For instance, non-compliance, attrition, or randomization bias are well-known hazards of any experiment.[16] Focus here is rather on the problems caused by the small sample size typical for experiments which might set even more severe limits to evaluation. In this case, randomization might lose its persuasiveness for it cannot be expected to achieve

[16] Since these problems are disregarded in our simulations they show randomized experiments in a favorable light. Other fundamental objections might be of ethical nature since treatments that produce positive effects are withheld the control group.

balance of *all* relevant covariates between the treatment and control group.

Specifically, small sample sizes arise if randomization occurs at the group level and/or if cost considerations prevent analysts from establishing a large scale experiment. Therefore, alternatives should be considered as well. Instrumental variable estimation as a quasi-experimental technique might be a helpful device to circumvent the small sample problem and open the field for less costly large scale observational studies if a good instrument is available. The simulation results suggest that correlations of around 0.3 to 0.4 can be considered to characterize a good instrument if the observational study comprises ten times more observations than a corresponding randomized experiment. In practice, one might even encounter ratios larger than 10 which would thus allow to utilize instruments with lower correlations. Moreover, contaminations of randomized experiments – especially at the group level – would also be avoided in observational studies.

Albeit, IV estimation yields inconsistent estimates in case treatment effects are heterogeneous and individuals or groups decide whether to undergo treatment upon their true effects. In this case, IV identifies the mean effect of treatment on compliers, i.e. the local average treatment effect. Thus, it would answer the question of how large the treatment effect would be if the binary instrument Z were increased from 0 to 1, for example, if more treatment sites were established such that some individuals or groups had a shorter distance to their site. This measure would only affect compliers while always- and never-takers would be unaffected. From this point of view, LATE might give answers to policy relevant questions, too.[17] Nevertheless, it seems fairly unlikely that individuals know their own treatment effects in advance; in contrast, it seems more probable that they have to make their participation decision upon some sort of expected gains.

In sum, if a randomized experiment is infeasible because of practical reasons or because it would not provide enough observations, observational studies are not necessarily a contemptible alternative. They often contain valuable and detailed information that might still help to identify causal relationships. On the other hand, absent randomization bias and systematic attrition or noncompliance, randomized controlled experiments are the

[17] See Angrist (1990), Angrist & Krueger (1991), and Imbens & Angrist (1994) for examples and a formal discussion.

most convincing evaluation approach as long as a sufficient number of units are involved in the trial.

References

Altonji, Joseph G. & Thomas A. Dunn (1996) "The Effects of Family Characteristics on the Return to Education", *Review of Economics and Statistics*, 78: 692-704.

Angrist, Joshua D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80: 313-36.

Angrist, Joshua D. (1998) "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica*, 66: 249-88.

Angrist, Joshua D. & Jinyong Hahn (1999) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *NBER Technical Working Paper* no. 241.

Angrist, Joshua D., Guido W. Imbens & D. B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association* 91: 444-72, with discussion by J.J. Heckman, R.A. Moffitt, J.M. Robins & S. Greenland, and R.P. Rosenbaum.

Angrist, Joshua D. & Alan B. Krueger (1991) "Does Compulsory Schooling Attendance Affect Schooling on Earnings?", *Quarterly Journal of Economics*, 106: 979-1014.

Angrist, Joshua D. & Alan B. Krueger (1999) "Empirical Strategies in Labor Economics", *Handbook of Labor Economics*, vol. 3, edited by Orley Ashenfelter and David Card. New York, NY: North-Holland, 1999.

Ashenfelter, Orley & Alan B. Krueger (1994) "Estimates of the Economic Return to Schooling from a New Sample of Twins", *American Economic Review*, 84: 1157-73.

Ashenfelter, Orley & Cecilia Rouse (1998a) "Income, Schooling and Ability: Evidence from a New Sample of Identical Twins", *Quarterly Journal of Economics*, 113: 253-84.

Ashenfelter, Orley & Cecilia Rouse (1998b) "Schooling, Intelligence and Income in America, Cracks in the Bell Curve", *Industrial Relations Section Working Paper 407*, Princeton University.

Becker, Gary S. (1967) *Human Capital and the Personal Distribution of Income*, University of Michigan Press, Ann Arbor, MI.

Bertsekas, Dimitri B. (1991) *Linear Network Optimization: Algorithms and Codes*, Cambridge MA, MIT Press.

Blackburn, McKinley L. & David B. Neumark (1993) "Omitted-Ability Bias and the Increase in the Return to Schooling", *Journal of Labor Economics*, 11: 521-44.

Blackburn, McKinley L. & David B. Neumark (1995) "Are OLS Estimates of the Return to Schooling Biased Downward? Another Look", *Review of Economics and Statistics*, 77: 217-30.

Bound, John, David A. Jaeger & Regina M. Baker (1995) "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak", *Journal of the American Statistical Association*, 90: 443-50.

Bound, John & George Johnson (1992) "Changes in the Structure of Wages in the 1980s: An Evaluation of Alternative Explanations", *American Economic Review*, 81: 371-92.

Bowden, Roger J. & Darrell A. Turkington (1984) *Instrumental Variables*, Cambridge: Cambridge University Press.

Buchinsky, Moshe (1994) "Changes in the Wage Structure 1963-1987, Application of Quantile regression" *Econometrica*, 62: 405-58.

Burtless, Gary (1995) "The Case for Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, 9: 63-84.

Cochran, W.G. & Donald B. Rubin (1973) "Controlling Bias in Observational Studies: A Review", *Sankhyā*, Series A, 35: 417-46.

Card, David (1995a) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", *National Bureau of Economic Research Working Paper* No. 4483.

Card, David (1995b) "Earnings, Schooling, and Ability Revisited", *Research in Labor Economics*, ed. S. Polachek, Vol. 14 (JAI Press, Greenwich Connecticut): 23-48.

Card, David (1999) "The Causal Effect of Education on Earnings", *Handbook of Labor Economics*, Chapter 30, Volume 3, edited by Orley Ashenfelter and David Card. New York, NY: North-Holland.

Cawley, John, Karen Conneely, James J. Heckman & Edward Vytlacil (1996) "Measuring the Effects of Cognitive Ability", *NBER Working Paper*, 5645.

Dehejia, Rajeev H. & Sadek Wahba (1998) "Propensity Score Matching Methods for Non-Experimental Causal Studies", *NBER Working Paper*, 6829.

Drake, Christiana (1993) "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect", *Biometrics*, 49: 1231-1236.

Fisher, Ronald A. (1935) *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Good, Phillip (1994) *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, New York: Springer Series in Statistics.

Greene, William H. (1993) *Econometric Analysis*, 2nd edition, Englewood Cliffs, NJ: Prentice Hall.

Griliches, Zvi (1977) "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, 45: 1-22.

Griliches, Zvi (1979) "Sibling Models and Data in Economics: Beginnings of a Survey", *Journal of Political Economy* 87: S37-S64.

Griliches, Zvi & William M. Mason (1972) "Education, Income, and Ability", *Journal of Political Economy*, 80: S74-S103.

Gu, Sam X. & Rosenbaum, Paul R. (1993) "Comparison of multivariate matching methods: Structures, distances and algorithms", *Journal of Computational and Graphical Statistics*, 2: 405-20.

Hahn, Jinyong (1998) "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66: 315-31.

Hall, Alastair R., Glenn D. Rudebusch & David W. Wilcox (1996) "Judging Instrument Relevance in Instrumental Variable Estimation", *International Economic Review* 37: 283-98.

Hastie, T. J. & R. J. Tibshirani (1990) *Generalized Additive Models*, London: Chapman and Hall.

Heckman, James J. (1996) "Randomization As An Instrumental Variable", *Review of Economics and Statistics*, 77(2): 336-41.

Heckman, James J. (1997) "Instrumental Variables, A Study of Implicit Behavioral Assumptions Used In Making Program Evaluations", *Journal of Human Resources*, 32: 441-62.

Heckman, James J., Hidehiko Ichimura & Petra Todd (1997) "Matching as an Econ-ometric Evaluation Estimator: Evidence from Evaluating a Job Training Program", *Review of Economic Studies*, 64: 605-54.

Heckman, James J., Hidehiko Ichimura & Petra Todd (1998) "Matching as an Econ-ometric Evaluation Estimator: Theory and Methods", *Review of Economic Studies*, 65: 261-94.

Heckman, James J., Robert J. Lalonde & Jeffrey Smith (1999) "The Economics and Econometrics of Active Labor Market Programs", *Handbook of Labor Economics*, Chapter 31, Volume 3, edited by Orley Ashenfelter and David Card. New York, NY: North-Holland.

Heckman, James J. & R. Robb, Jr. (1985) "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, ed. J. Heckman & B. Singer. New York: Cambridge University Press.

Heckman, James J. & Jeffrey A. Smith (1995) "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, 9: 85-110.

Holland, Paul W. (1986) "Statistics and Causal Inference (with discussion)", *Journal of the American Statistical Association*, 81: 945-70.

Imbens, Guido W. & Joshua D. Angrist (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62: 446-475.

Kane, Thomas J. & Cecilia E. Rouse (1995) "Labor-Market Returns to Two- and Four-Year College", *American Economic Review*, 85: 600-14.

Katz, Lawrence & Kevin M. Murphy (1992) "Changes in Relative Wages, 1963-87: Supply and Demand Factors", *Quarterly Journal of Economics*, 107: 35-78.

Kjellström, Christian (1999) *Essays on Investment in Human Capital*, Swedish Institute for Social Research, no. 36.

Kloek, T. (1981) "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated", *Econometrica*, 49: 205-07.

Kluve, Jochen, Hartmut Lehmann & Christoph M. Schmidt (1999) "Active Labor Market Policies: Human Capital Enhancement, Stigmatization, or Benefit Churning?", *Journal of Comparative Economics*, 27: 61-89.

Lechner, Michael (1999) "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification", *Journal of Business and Economic Statistics*, 17/1: 74-90.

Lechner, Michael (2000) "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources*, 35: 347-75.

Levy, Frank & Richard Murnane (1992) "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations", *Journal of Economic Literature*, 30: 1333-81.

Manski, Charles F. (1990) "Nonparametric Bounds on Treatment Effects", *American Economic Review*, 80(2): 319-23.

Manski, Charles F. (1995) *Identification Problem in the Social Sciences*, Cambridge, MA: Harvard University Press.

Mincer, Jacob (1974) *Schooling, Experience and Earnings*, Columbia University Press, New York.

Ming, Kewei & Paul R. Rosenbaum (2000) "Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls", *Biometrics*, 56: 118-24.

Moulton, Brent R. (1986) "Random Group Effects and the Precision of Regression Estimates", *Journal of Econometrics*, 32: 385-97.

Moulton, Brent R. (1990) "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units", *Review of Economic and Statistics*, 71: 334-38.

Murnane, Richard J., John B. Willett & Frank Levy (1995) "The Growing Importance of Cognitive Skills in Wage Determination", *Review of Economics and Statistics* 77: 251-66.

Murnane, Richard J., John B. Willett & John H. Tyler (2000) "Who Benefits from Obtaining a GED? Evidence from High School and Beyond", *Review of Economics and Statistics*, 82: 23-37.

Murray, David M. (1998) *Design and Analysis of Group-Randomized Trials*, Oxford: Oxford University Press.

NLS Handbook (1997) U.S. Department of Labor, Bureau of Labor Statistics.

NLSY79 User's Guide (1997) Center for Human Resource Research: The Ohio State University.

Park, Jin Heum (1994) "Returns to Schooling: A Peculiar Deviation from Linearity", Working Paper no. 335, Industrial Relations Section, Princeton University.

Psacharopoulos, George (1994) "Returns to Investment in Education: A Global Update", *World Development*, 22: 1325-43.

Quade, Dana (1981) "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38: 597-611.

Rosenbaum, Paul R. (1984) "Conditional Permutation Tests and the Propensity Score in Observational Studies", *Journal of the American Statistical Association*, 79: 565-74.

Rosenbaum, Paul R. (1989) "Optimal Matching for Observational Studies", *Journal of the American Statistical Association*, 84: 1024-32.

Rosenbaum, Paul R. (1991) "A Characterization of Optimal Designs for Observational Studies", *Journal of the Royal Statistical Association*, Series B, 53: 597-610.

Rosenbaum, Paul R. (1995) *Observational Studies*, New York: Springer Series in Statistics.

Rosenbaum, Paul R. & Donald B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70: 41-55.

Rosenbaum, Paul R. & Donald B. Rubin (1985) "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39: 33-38.

Roy, Andrew D. (1951) “Some Thoughts on the Distribution of Earnings”, *Oxford Economic Papers*, 3: 135-46.

Rubin, Donald B. (1973) “Matching to Remove Bias in Observational Studies”, *Biometrics*, 29: 159-83, Printer’s correction (1974), 30: 728.

Rubin, Donald B. (1974) “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”, *Journal of Educational Psychology*, 66: 688-701.

Rubin, Donald B. (1977) “Assignment to Treatment Group on the Basis of a Covariate”, *Journal of Educational Statistics*, 2: 1-26.

Rubin, Donald B. (1980) “Bias Reduction Using Mahalanobis Metric Matching”, *Biometrics*, 36: 293-98.

Rubin, Donald B. (1986) “What Ifs Have Causal Answers?”, *Journal of the American Statistical Association*, 81: 961-62.

Schmidt, Christoph M. (1999) “Knowing What Works: The Case for Rigorous Program Evaluation”, IZA Discussion Paper no. 77.

Schmidt, Christoph M., Rob Baltusen & Rainer Sauerborn (1999) “Evaluation of Community-Based Interventions: Group-Randomization, Limits and Alternatives”, *Discussion paper series* no. 281, Department of Economics, University of Heidelberg.

Shore-Sheppard, Lara (1996) “The Precision of Instrumental Variables Estimates With Grouped Data”, *Industrial Relations Section Working Paper 374*, Princeton University.

Weiss, Andrew (1988) “High School Graduation, Performance, and Wages”, *Journal of Political Economy*, 96: 785-820.

White, H. (1980) “A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity”, *Econometrica*, 48: 817-38.

Willis, Robert (1986) “Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Function”, *Handbook of Labor Economics*, Vol. I, edited by O. Ashenfelter and R. Layards, North Holland, 525-602.

Yatchew, Adonis & Zvi Griliches (1984) “Specification Error in Probit Models”, *Review of Economics and Statistics*, 66: 134-39.

Acknowledgements

It is a pleasure to express my appreciation to those who have influenced this work by their many valuable comments and advises. I want to give special thanks to my supervisor Christoph M. Schmidt for his guidance and his detailed comments on drafts of the chapters which led to substantial improvements, and, furthermore, for giving me impetus to further research. I am greatly indebted to my colleagues and friends for valuable suggestions and discussions on drafts of individual chapter: Martin Biewen, Michael Fertig, Manuel Frondel, Jochen Kluve, and Ralph Würthwein. I owe a debt of gratitude to Bettina Kretz who helped improve my English expression.

Also, I would like to thank the Bureau of Labor Statistics and the Center for Human Resource Research, Ohio State University, for their quick and valuable aid concerning the NLSY data, Joseph Altonji who drew my attention to these data, and Paul Rosenbaum who told me where to find algorithms for minimum cost flow problems. Thanks go to participants of the IWH Workshop on "Evaluating Active Labor Market Policies and Welfare Programs" in Halle, specifically to Michael Lechner and Jeffrey Smith, to seminar participants in Heidelberg, specifically to Hans Gersbach, and to participants of the first IZA European Summer School in Labor Economics. I would like to express my thanks to David Card for discussions and his hospitality at the Center for Labor Economics, UC Berkeley, and to Rainer Dahlhaus, Hartmut Kogelschatz, Enno Mammen, and Günther Sawitzki for teaching me statistics. Finally, financial support by the Friedrich-Ebert-Stiftung is gratefully acknowledged.