HEDG Working Paper 09/09

Econometric Evaluation of Health Policies

Andrew M Jones

Nigel Rice

May 2009

ISSN 1751-1976

http://www.york.ac.uk/res/herc/research/hedg/wp.htm

ECONOMETRIC EVALUATION OF HEALTH POLICIES

ANDREW M. JONES AND NIGEL RICE
University of York

In preparation for *The Oxford Handbook of Health Economics*, Sherry Glied and Peter C Smith (eds), Oxford: Oxford University Press.

Acknowledgments: We gratefully acknowledge funding from the Economic and Social Research Council (ESRC) under the Large Grant Scheme, reference RES-060-25-0045. We are also grateful to Pedro Rosa Dias, Silvana Robone, Ranjeeta Thomas for comments on an earlier version.

Abstract

This chapter provides a concise guide to the econometric methods that are available to evaluate the impact of health policies.

Contents

1. The evaluation problem

1.1 Counterfactuals and treatment effects

The narrow goal of evaluative research is to identify the causal impact of an intervention on outcomes of interest. The broader goal is to understand the mechanisms underlying this impact. In evaluating the cost-effectiveness of medical technologies, randomized controlled trials (RCTs) are often regarded to be the gold standard in identifying internally valid estimates of causal effects. In health policy research, randomized experiments are less prevalent and researchers are more often faced with identifying causal relationships from observational, or non-experimental, sources of data where the assignment of individuals to treatment or control group is beyond the control of the researcher. In such circumstances, the identification of causal effects is often less than straightforward and econometric tools are often called into play.

The evaluation of a treatment, in its broadest sense, refers to the measurement of the impact of an intervention on specified outcomes of interest. In the clinical context, this frequently means identifying the effect of a particular treatment or treatment technology on health outcomes, where the treatment is compared to either no-treatment, a placebo or, more commonly, an existing treatment or technology. For public health and health policy research we might be interested in the evaluation of a particular treatment, but we might also be interested in the evaluation of a broader policy intervention or programme, such as a ban on smoking in public places or a reform of provider reimbursement. Throughout the chapter we use the terms policy intervention, programme or treatment interchangeably, using treatment as shorthand for all three. Similarly, we refer to individuals as the targets of the policy interventions. In practice the unit of analysis may be organisations or groups such as hospitals or other health care providers.

Heckman (2008) argues that the evaluation problem consists of three distinct steps: definition (which relates to policy analysis and relevant theory), identification (which relates to what, in principle, can be learned from the whole population) and inference (which relates to what, in practice, can be estimated from sample data). Angrist and Pischke (2008) pose these steps as questions: what is the causal relationship of interest (which, they argue, can often be clarified by asking what experiment would ideally be used to capture the effect); what is the identification strategy to isolate this effect; and what is the mode of inference to estimate and test the effect with sample data? The analogy between casual effects and experimental design is highly influential in the evaluation literature and shapes much of the language and conceptual framework that is used, although Heckman (2008) strikes a note of caution and stresses the danger of confusing definition of an underlying causal mechanism with a particular identification strategy, such as randomization. In practice, policy effects are typically defined by the agenda of policymakers, coupled with theory from economics or other disciplines that helps to define the relevant outcomes and their relationship to the treatments. This theory may help to shape the identification strategy. Most applied health

economics is not concerned with pure testing of scientific hypotheses, instead it focuses on estimating economically relevant magnitudes of policy effects and on providing evidence to inform decision-making[1].

The search for a convincing identification strategy usually boils down to finding a source of variation in the treatment that is independent of other factors that influence outcomes. Randomizing the assignment of treatment, in the context of a social experiment, is one such source of independent variation. While randomized trials are the norm in the evaluation of new clinical therapies, and despite the precedent set by the RAND Health Insurance Experiment, their use for the evaluation of broader health and social programmes remains relatively rare (e.g., Björkman and Svensson 2009; Gertler 2004; Miguel and Kremer 2004). In non-experimental settings the identification strategy may mean appealing to a rich enough set of observed confounders so that any remaining variation in treatment is effectively randomized. Often it means turning to variation over time or across groups that can be used to mimic the features of a randomized experiment so that the analysis can be interpreted as a quasi-experimental design, or "natural experiment" (see e.g., Meyer 1995). A neat example of the analogy between a designed experiment and a natural experiment is the work of Stillman et al. (2009) who make use of a random ballot of Tongans who applied to emigrate to New Zealand to identify the effects of migration on mental health. The recent health economics literature contains many examples of proposed natural experiments and these are reviewed in more detail in Jones (2009). For example, to try to identify independent variation in individual health Almond (2006) uses the 1918 influenza pandemic while Doyle (2005) uses severe traffic accidents. To identify independent variation in the use of prenatal care Evans and Lien (2005) use the impact of the 1992 Port Authority Transit (PAT) strike in Pennsylvania on access to health care. Lindahl (2005) proposes lottery winnings as a source of exogenous variation in income in a study of the income-health gradient. Broader macroeconomic shocks have been used as natural experiments: for example, Frijters et al. (2005) use the reunification of Germany in 1990; Jensen and Richter (2004) use the 1996 crisis in the public pension system in Russia; Duflo (2000) uses reform of public pension provision for black African women at the end of the Apartheid era in South Africa. Natural experiments may call on historical data: Van den Berg et al. (2006) use the state of the Dutch economy at the end of the 19th Century to study long-term consequences for mortality rates during the 20th Century. Institutional and policy reforms are often the source of natural experiments. Lleras-Muney (2005) uses state-level reforms the US educational system to provide variation in educational attainment that is independent of individual decisions. Health policies and reforms are of particular relevance for health economics: for example, Bleakley (2007) focuses on a programme aimed at the eradication of hookworm in the Southern US, while Pop-Eleches (2006) uses the 1966 ban on abortion and family planning under the Ceausescu regime in Romania.

[1] This explains the prominence of statistical decision analysis in the recent economic literature on health technology assessment (see e.g., Claxton 1999) which has parallels with the work of Manski (2005) in the general program evaluation literature.

To put these identification strategies into practice, natural experiments are often used to support estimation approaches such as instrumental variables (IV), regression discontinuity (RD) and difference-in-differences (DD). The problem of inference focuses on the assumptions required to implement a given identification strategy with a particular sample or samples of data. In recent years there has been a heavy emphasis on keeping parametric assumptions to a minimum. The literature has favoured methods such as matching, nonparametric regression, and control function approaches and on making inferences that are robust to functional form and distributional assumptions. Similarly computation of standard errors needs to take account of sampling methodologies and data structures and make appropriate adjustments for features such as clustering of observations and heteroscedasticity.

In defining the evaluation problem, a fundamental problem arises in attempting to derive inference of a causal relationship between a treatment, denoted *d*, and an outcome, denoted *y*. The treatment effect of interest, Δ, is the change in potential outcome for individual, i, when exposed to the intervention compared to an alternative (referred to as the control) and can be defined as:

$$\Delta_i = y_i^1 - y_i^0 \tag{1}$$

Where superscript 1 denotes treatment and 0 denotes the control. The evaluation problem is that an individual cannot be observed to be under treatment and under the control at the same time. At any particular point in time only one of the potential outcomes can be observed (Roy 1951; Rubin 1974). This framework, which emphasises unobserved potential outcomes, is often referred to as the Rubin Causal Model (Holland 1986). Methods to define and estimate the counterfactual outcome lie at the heart of all attempts to identify causal relationships between treatment assignment and the outcomes of interest[2].

1.2 Average treatment effects

A common approach to addressing the evaluation problem is to focus on average treatment effects (ATE). For example the population average treatment effect (PATE) is the difference in the average potential outcomes in the treated and control groups for the population as a whole[3]:

$$PATE = E\left[y^1 - y^0 \right] = E\left[y^1 \right] - E\left[y^0 \right] \tag{2}$$

[2] In this sense the evaluation problem can be seen as a particular brand of missing data problem and many of the methods used to evaluate treatment effects, such as those based on propensity scores, are used in that more general context as well.

[3] Where appropriate, and for ease of notation, we suppress the *i* subscript that denotes that variables vary across individuals. The same applies to the *t* subscript when variables vary over time, as in the case of panel data.

For a particular sample of data the analogue of the PATE is the sample average treatment effect (SATE):

$$SATE = \frac{1}{n^1}\sum_{d=1} y^1 - \frac{1}{n^0}\sum_{d=0} y^0 \tag{3}$$

Where n^1 and n^0 are the numbers of treated and controls in the sample.

More often, the relevant concept is the treatment effect for the subset of the population who would actually be assigned to treatment (e.g., Heckman, LaLonde and Smith 1999). This is the treatment effect on the treated (ATT, or sometimes ATET or TT). The population average treatment effect on the treated (PATT) is:

$$PATT = E\left[y^1 - y^0 \mid d = 1\right] = E\left[y^1 \mid d = 1\right] - E\left[y^0 \mid d = 1\right] \tag{4}$$

The PATT represents the expected gain from the intervention for an individual randomly selected from the treated population, rather than for any individual in the general population. For example, the PATT would be applicable to a study of an area-based intervention where the area is chosen on the basis of certain characteristics or for any interventions where those who self-select into treatment respond differently from those who do not. The sample analogue is the sample average treatment effect (SATT).

It is worth emphasising that the fundamental treatment effect of interest, Δ_i, is individual specific and that there may be considerable heterogeneity in treatment effects across individuals. While much of the evaluation literature focuses on average treatment effects, the broader impact on the full distribution of treatment effects is also of interest. Heckman, Smith and Clements (1997) note that while randomized experiments can identify average treatment effects, due to the additivity of the expectations operator as seen in equation (2), they may not identify other features of the distribution without further assumptions being imposed. They discuss other measures of the impact of a treatment such as the proportion of people receiving the treatment who benefit from it, $P(\Delta > 0 \mid d = 1)$; the proportion of the total population that benefits, $P(\Delta > 0 \mid d = 1)P(d = 1)$; selected quantiles of the distribution of treatment effects, such as median treatment effects; and the distribution of gains from treatment at selected values of covariates.

1.3 Selection bias

Selection bias arises in situations where assignment to the treatment is correlated with observed or unobserved factors ('confounders') that are also correlated with the outcome. Where evaluative research relies on observational data, obtained from non-experimental study designs, the effects of selection are likely to be more pronounced than those found in experimental settings where the study design aims to remove confounding effects. If

sufficient characteristics that determine the assignment of treatment are observed, then methods based on *selection on observables* can be used to consistently estimate the treatment effect. These methods include regression analysis, matching estimators and inverse probability weighted estimators. If, however, selection into treatment is based on characteristics that are unobservable to the researcher then methods that allow for *selection on unobservables* are appropriate. These approaches are mostly based on finding factors that predict treatment assignment but, crucially, that do not have a direct effect on the outcome of interest. Methods include instrumental variables, control functions and the joint estimation of the outcome and treatment in a structural approach. Longitudinal data may provide a way of dealing with unobservables as in the difference-in-differences approach and panel data regression methods.

To illustrate, consider Figure 1 which presents a stylised view of the potential outcomes for the treated (y^1) and controls (y^0) plotted against time. For simplicity, these are shown to follow a common linear trend over time. Treatment is assigned at the point shown by the dotted vertical line and, in this example, sample data may be observed before treatment (t-1) and after treatment (t). Notice that there is a gap between the potential outcomes prior to treatment, which may be due to observed (x) or unobserved (u) confounders. In the example the treated have better outcomes whether of not they receive the treatment, so selection bias is positive, as shown by the vertical distance $\{x,u\}$. The challenge for the analyst is to identify the causal effect of the treatment, allowing for the confounding effect of selection bias.

Insert Figure 1 around here

Now consider the kinds of data that might be available to evaluate the treatment. First we may have cross-section data collected after treatment at time t. Then a simple comparison of the (mean) outcomes for the treated and untreated will equal the distance CD. This overestimates the true treatment effect, that equals CC′, by the amount of selection bias C′D. If the selection bias is fully accounted for by observable confounders x, then methods such as regression or matching may be used to deal with the problem, effectively adjusting the counterfactual outcomes (y^0) so that they are comparable to y^1. Graphically this would eliminate the vertical gap between the two lines prior to treatment so that both potential outcomes were observed to have equivalent pre-treatment trajectories. If the selection bias is due to unobservables u, then other methods, such as instrumental variables, should be considered.

Now consider a before and after comparison of the outcomes for the treated cases. In this case using the vertical difference in (mean) outcomes before and after, CA, will overestimate the true treatment effect CC' by the vertical distance C'A. This bias reflects the underlying trend in outcomes and has nothing to do with the treatment itself. One way to account for the trend, and also for fixed differences due to x and u, is to take the difference in differences. This estimates the treatment effect by taking the difference between the change over time for the treated (CA) and the change over time for the controls (DB). So long as the two groups share the same trend ("parallel trends"), the resulting difference equals the

true treatment effect CC'. The difference-in-differences approach will work as long as the common trend assumption holds. Differences in the trend that are attributable to observables can be captured by controlling for those covariates, through regression or matching. Differences due to unobservables that vary over time are more problematic.

To define selection bias more formally, consider the difference between the population means of the outcome y for the treated and controls, which can be decomposed as follows (see Heackman, Ichimura, Smith and Todd 1998):

$$\begin{aligned} & E\left[y \mid d=1\right]-E\left[y \mid d=0\right] \\ & =E\left[y^{1} \mid d=1\right]-E\left[y^{0} \mid d=0\right] \\ & =E\left[y^{1}-y^{0} \mid d=1\right]+\left\{E\left[y^{0} \mid d=1\right]-E\left[y^{0} \mid d=0\right]\right\} \\ & =PATT+Bias \end{aligned} \tag{5}$$

In this case the simple difference in population means equals the average treatment effect on the treated plus a selection bias that captures the underlying difference in potential outcomes, in the absence of treatment, between those assigned to treatment and those assigned to the control. This selection bias may be attributable to observables or to unobservables. The bias term can be further decomposed (Heckman, Ichimura, Smith and Todd 1998; King and Zeng 2007). Bias can arise due to: failing to control for relevant confounders (*omitted variable bias*); inclusion of covariates that are themselves affected by the treatment (*post-treatment bias*); failure to adequately control for covariates within the observed range of data, for example when applying a linear model to a nonlinear relationship (*interpolation bias*); or failure to adequately control for covariates when extrapolating to areas outside the observed range of the data, for example if a linear approximation holds within the sample but not beyond it (*extrapolation bias*). For a well-designed experiment, where randomization fully determines assignment to treatment, $E\left[y^{0} \mid d=1\right]-E\left[y^{0} \mid d=0\right]=0$ eliminating the bias.

The sources of selection bias can be illustrated further by expressing the potential outcomes as regression functions, using some fairly general notation. Let the potential outcomes be (additive) functions of observable (x) and unobservable (u) confounders (e.g., Heckman, Ichimura and Todd 1997):

$$\begin{aligned} y^{0} &= \mu^{0}(x)+u^{0} \\ y^{1} &= \mu^{1}(x)+u^{1} \end{aligned} \tag{6}$$

Then, by definition:

$$y=(1-d)y^{0}+dy^{1}$$

So:

$$\begin{aligned} y &= y^0 + d(y^1 - y^0) \\ &= \mu^0(x) + d(\mu^1(x) - \mu^0(x)) + u^0 + d(u^1 - u^0) \end{aligned} \tag{7}$$

In this formulation the "treatment regression" (7) consists of an "intercept", $\mu^0(x)$, which reflects the way in which baseline (pre-treatment) outcomes depend on *x*. Then the treatment effect is $\left(\mu^1(x) - \mu^0(x)\right)$ which may vary with observable characteristics (captured by interactions between *d* and *x*). The error term reflects unobservable baseline differences in the outcome, u^0, as well as unobservable idiosyncratic differences in the benefits of the treatment, $\left(u^1 - u^0\right)$. Treatment assignment is likely to be influenced by both of these terms, creating selection bias on unobservables. For example a doctor may take account of their personal assessment of a specific patient's capacity to benefit when deciding which treatment regime to adopt.

The model can be augmented by a model of treatment assignment/participation:

$$d = 1 \quad if \quad d^* = \mu^d(z) + u^d > 0 \tag{8}$$

Where z are covariates that influence assignment to treatment and where u^d may be correlated with $u^0 + d\left(u^1 - u^0\right)$. Using linear functions for the regressions (6) coupled with (8), gives Roy's (1951) model:

$$y = x'\beta_j + u_j, \; j = 0,1 \tag{9}$$

where the regression coefficients can differ by treatment regime.

2. Selection on observables

2.1 Regression analysis

In situations where selection into treatment is based only on observables, we can use outcome data from a set of potential comparison individuals for whom observed characteristics are comparable to those of the treated individuals, so that like is compared with like. This is the idea behind the use of standard regression analysis and the use of matching (Cochran and Rubin 1973; Rubin 1973a; Heckman, Ichimura and Todd 1997; Deheija and Wahba 1999) and inverse probability weights (e.g., Hirano, Imbens and Ridder 2006).

The key assumption of the selection on observables approach is that, conditional on the chosen set of matching variables *x*, selection into treatment is independent of the outcomes of interest (Heckman and Robb 1985). Recall equation (5), which can now be rewritten conditional on the set of observed covariates *x:*

$$\begin{aligned} & E\left[y \mid d=1, x\right]-E\left[y \mid d=0, x\right] \\ & =E\left[y^{1}-y^{0} \mid d=1, x\right]+\left\{E\left[y^{0} \mid d=1, x\right]-E\left[y^{0} \mid d=0, x\right]\right\} \\ & =PATT\mid_{x}+Bias \end{aligned} \tag{10}$$

This implies that a minimal assumption to eliminate the bias term, and therefore for the identification of the PATT, is *unconfoundedness*, *ignorability*, or *conditional independence* (Rosenbaum and Rubin, 1983):

$$y^{0} \perp d \mid x \tag{11}$$

Condition (11) states that, conditional on *x,* the assignment of treatment *d* is independent of the potential outcome y^0 and, hence, would make the bias term in (10) disappear. In estimating (10) a weaker version of this condition can be expressed in terms of expected values, but this is not invariant to transformations of y (Heckman, Ichimura and Todd 1997). Note that assumptions about the distribution of y^1 among the treated are not required for identification of the PATT as this is identified from the observed data.

In practice and, in particular if covariates are well balanced between the treated and controls, analysts may be willing to assume the parametric structure implied by a regression model. Given the unconfoundedness assumption, (11), a simple way to estimate average treatment effects is the standard linear regression framework. To see this consider the definition of the PATT:

$$\begin{aligned} PATT\mid_{x} &= E\left[y^{1}-y^{0} \mid d=1, x\right] \\ &= E\left[y^{1} \mid d=1, x\right]-E\left[y^{0} \mid d=1, x\right] \\ &= \tau(.) \end{aligned} \tag{12}$$

This implies:

$$E\left[y^{1} \mid d=1, x\right]=E\left[y^{0} \mid d=1, x\right]+\tau(.) \tag{13}$$

Then, the basic identifying assumption of the linear regression model is:

$$E\left[y^0 \mid d=1, x \right] = E\left[y^0 \mid d=0, x \right] = x'\beta_0 \tag{14}$$

This combines unconfoundedness (11) with the assumption that the conditional mean of y^0 is a linear function of the covariates. In this context, as noted by Hirano and Imbens (2001, p.263), "linearity is not really restrictive, as we can include functions of the original covariates in the vector x".

Then, by definition:

$$E(y \mid d, x) = d.E(y^1 \mid d, x) + (1-d).E(y^0 \mid d, x) \tag{15}$$

Then, using (13) and (14):

$$\begin{aligned} E(y \mid d, x) &= d\left\{x'\beta_0 + \tau(.)\right\} + (1-d)\left\{x'\beta_0\right\} \\ &= x'\beta_0 + \tau(.)d \end{aligned} \tag{16}$$

So the linear regression estimate of the ATT is the coefficient, $\tau(.)$, on the treatment indicator d. This can be estimated by least squares. The simplest version of this model treats $\tau(.)$ as a fixed coefficient. But the treatment effect need not be constant. To allow for heterogeneity in the treatment effect d could be interacted with the x variables or it could be treated as a random parameter. The model with interaction terms gives Roy's model (9) but, given the unconfoundedness assumption (11), the regression models can be estimated independently for the treated and controls. The ATT can then be estimated as the difference in the average of the predicted values of the two regressions, estimated for the sample of treated individuals:

$$\hat{\tau}_{LR} = \frac{1}{n^1}\sum_{i\in\{d=1\}}\left(x_i'\hat{\beta}_1 - x_i'\hat{\beta}_0\right) = \overline{x}_1'\left(\hat{\beta}_1 - \hat{\beta}_0\right) \tag{17}$$

Notice that the counterfactual outcome $E(y^0 \mid d=1, x)$ is estimated by:

$$\frac{1}{n^1}\sum_{i\in\{d=1\}}\left(x_i'\hat{\beta}_0\right) = \overline{x}_1'\hat{\beta}_0 = \overline{x}_0'\hat{\beta}_0 + \left(\overline{x}_1 - \overline{x}_0\right)'\hat{\beta}_0 \tag{18}$$

While the regression model for the controls may do a good job of estimating β_0 for the observed sample of controls, if $\overline{x}_0$ and $\overline{x}_1$ are far apart, the model may do a poor job of extrapolating to the sample of treated observations. This makes it clear that, for linear models, it is balancing of the means of the covariates that is important, but this is contingent on the linear specification being correct. For nonlinear specifications other facets of the distribution will be important.

2.2 Matching

The method of matching avoids the need to make parametric assumptions such as those implied by the linear regression model (16). The idea behind matching estimators is that if a suitably matched group of individuals can be identified, then the average outcome across these individuals provides the counterfactual for the mean outcome for treated individuals in the absence of treatment (Cochran and Rubin 1973; Rubin 1973a; Heckman, Ichimura and Todd 1997; Deheija and Wahba 1999). For matching to provide a nonparametric estimate the unconfoundedness assumption (11) has to be coupled with the requirement of weak overlap:

$$\mathrm{P}(d = 1 \mid x) < 1 \tag{19}$$

so that it is possible to find controls who share the same x values as each treated case. Otherwise regression models are required to extrapolate counterfactual outcomes to areas outside the range of common support[4], meaning that estimates of the counterfactual are *model dependent*.

A number of alternative methods for defining the matched group of individuals have been proposed. Exact matching consists of finding a match to a treated individual, from the pool of controls, based on the set of observed characteristics of the individual. Individuals can then be compared to their matched counterparts. The choice of observed characteristics is important and should be based on all factors that affect both treatment assignment and outcomes of interest but are not affected by the treatment itself (to avoid *post treatment bias*). Where this is not the case, matching fails to control for treatment selection. Accordingly, matching is based on pre-treatment characteristics of individuals. It is assumed that for each combination of values of x among the treated, an untreated individual can be found. This ensures common support over x. Individuals with characteristics only observed within those treated, or indeed, only within comparison cases, may be ignored as appropriate matches are not available. But in doing so, the population of interest and, hence, the relevant treatment effect is redefined to the area of common support.

The method of exact matching is most practicable when the number of observed characteristics is reasonably small and where characteristics are measured as discrete variables. In such circumstances, the ability to locate matched cases is enhanced. However unconfoundedness is unlikely to hold if the list of observables is short. For the majority of empirical problems, where matches are sought over a larger number of covariates, or for continuous variables, finding exact matches is more difficult and alternative methods are usually required.

[4] Common support requires that for each level of $p(x)$, the probability of observing a non-treated individual is positive. Accordingly, the PATT is identified by restricting attention to comparative non-treated individuals that fall within the support of the propensity score distribution of the treatment group.

2.3 Propensity scores

When there are many covariates exact matching is often impracticable and inexact methods of matching are required. The leading method is propensity score matching (Rosenbaum and Rubin, 1983). Rosenbaum and Rubin (1983) show that the curse of dimensionality can be overcome by using a balancing score[5]. They show that unconfoundness in terms of the full set of covariates implies unconfoundedness in terms of a balancing score *b(x)*:

$$y^0 \perp d \mid x \quad \Rightarrow \quad y^0 \perp d \mid b(x) \tag{20}$$

There are many balancing scores, including x itself, but the one that is most commonly used is the propensity score:

$$p(x) = E(d = 1 \mid x) = \mathrm{P}(d = 1 \mid x) \tag{21}$$

which is the population conditional expectation of d given x. Then, rather than conditioning on the full set of the covariates, the propensity score approach conditions only on the propensity score[6]. For example the PATT may be redefined as:

$$PATT_{PS} = E_{p(x),d=1}\left\{E(y^1 \mid d = 1, p(x)) - E(y^0 \mid d = 1, p(x))\right\} \tag{22}$$

Matching is just one way in which the propensity score can be used to estimate (22). Other methods include blocking (Rosenbaum and Rubin 1983), inverse probability weighting (Hirano, Imbens and Ridder 2006) and regression on the propensity score using semiparametric regression models such as series approximations or local linear regression (Hahn 1998; Heckman, Ichimura and Todd 1998).

Blocking (stratification or *interval matching*) divides the region of common support into 'blocks' of pre-defined width (Rosenbaum and Rubin 1983, 1984). Each block contains both treated individuals and their corresponding controls based on the propensity scores and the criterion used to define the common support (for example, a radius criterion). The treatment effect is then computed separately for each block and the overall ATT as the weighted average of the individual block effects, with weights defined by the number of treated individuals within each block. This can be seen as a simple form of nonparametric regression using a step function (Imbens 2004). For example, Dehejia and Wahba (1999) define the intervals on the basis of treatment and controls failing to exhibit a statistically

[5] Hahn (1998) established that the value of conditioning on the propensity score, rather than on the elements of x directly, stems from the reduction in dimensionality rather than a gain in efficiency for semiparametric estimators.

[6] In practice matching on the propensity score may be combined with exact matching on specific covariates or matching within subgroups of the sample defined by the covariates. The Mahalanobis metric, which scales differences in x by the inverse of their covariance matrix, is often used for exact matching of covariates.

significant difference in estimated propensity scores (this is implemented in Stata by Becker and Ichino (2002)).

A general expression for the propensity score matching estimator of the PATT is:

$$\hat{\tau}_{PS} = \frac{1}{n^1} \sum_{i \in \{d=1\}} \left(y_i - \sum_{j \in \{C(p_i(x))\}} w_{ij} y_j \right) \tag{23}$$

where w_{ij} is a weight applied to the contribution of comparison individual, j, as an appropriate counterfactual for treated individual, i, and is determined by the distance between their respective propensity scores. $C\left(p_i\left(x\right)\right)$ represents the set of comparable neighbours for each treated individual, i, where closeness between the propensity scores of the multiple js for each i is defined by some appropriate criterion. Notice, that the final outcome of the matching process is to compare the sample mean for the treated with sample mean for an appropriately selected and weighted set of control observations. In this sense the fact that particular treated cases are paired with particular controls during the matching process is irrelevant and matching is simply a way of restricting the sample that is used for comparisons. This leads Ho et al. (2007) to suggest that the use of term matching is unfortunate and could be a source of misunderstanding. They suggest that the approach would be better labelled as "pruning" rather than matching[7].

Empirical implementation of the propensity score approach involves estimating the individual's propensity to receive treatment, most commonly using either a probit or logit regression when the treatment is binary, and matching treated individuals to controls with a similar propensity for treatment. Again, it is important that there is common support over the estimated propensities for both treatment and control groups, even if their respective densities are different.

Two concerns are relevant in the application of propensity score matching techniques. Firstly, whether matching is performed with or without replacement, and secondly, the number of matches to use in the control group. Matching without replacement means that an individual in the control group is matched to only the closest individual in the treatment group. This can result in poor matches between control and treated individuals unless there are sufficient individual observations in the potential control group to ensure that close matches are obtained for all individuals. Poor matches result where propensity scores are not close resulting in bias in the matching estimator of the treatment effect. Matching with replacement avoids this problem by allowing each control individual to be matched to multiple treated individuals, again selected on the basis of closest propensity score match. Accordingly, matching with replacement is helpful in reducing bias since comparison individuals can be used more than once if they happen to be the closest match to more than

[7] Note that this argument applies for the estimation of average treatment effects but does not apply to other measures such as quantile treatment effects.

one treated individual. However, the variance of the propensity score matching estimator is likely to be greater than the corresponding estimator without replacement due to greater reliance on a smaller number of comparison individuals. Allowing multiple matches for each treated individual tends to reduce the variance of the propensity score matching estimator of the treatment effect due to the increased information used to obtain the counterfactual. The choice of the number of controls to match to each treated individual, however, involves a trade-off between bias and variance. By including only the closest match bias is reduced, however, the variance of the estimator can be reduced by including more matches, but at the expense of increasing the bias if the additional matched controls are not close matches to the treated individual.

An often neglected point is that matching only requires information on the treatment and on the covariates, not on the outcomes. Good practice in empirical analysis is to conduct the matching exercise without access to the outcome data to avoid them influencing the results (Rubin 2006, p.3). Repeated specification searches, aimed at improving the balance of treated and controls, only involves the covariates and will not bias subsequent analysis of the treatment effects.

The simplest form of propensity score matching is *nearest-neighbour* (NN) matching, which seeks to match each treated individual to the closest untreated individual, based on their estimated propensity scores. Matching may be performed with or without replacement. This form of matching can lead to substantial bias if the nearest-neighbours turn out to be poor matches, in the sense that the propensity scores are not close. In addition, given that, when sampling without replacement, each control individual is matched to at most one treated individual, the order in which matches are sought may influence the resulting estimator of the PATT..

The methods of *caliper* and *radius matching* offer a compromise between the bias and variance of the propensity score matching estimator. The methods define a tolerated radius or neighbourhood around the propensity score for treated individuals beyond which matched controls are excluded. The caliper method only uses the nearest neighbour if it lies within the radius, while the radius method uses all of the matches that lie within the radius (Cochran and Ruben 1973; Deheija and Wahba 2002). The choice of the radius is subjective and involves a compromise between the desire to obtain a small bias by having a small radius and to exclude as few treated individuals as possible (excluded due to failing to find an appropriate match). Common support is often invoked by the requirement for at least one comparison individual to be within the defined propensity score radius criterion.

Kernel density matching provides a further way to construct the counterfactual for treated individuals (Heckman, Ichimura and Todd 1997). For each treated individual the counterfactual is the kernel weighted average of the multiple comparison individuals. The contribution of a potential match to the counterfactual will depend on the distance of the propensity score to that of the treated individual, together with the chosen bandwidth for the kernel function. Silverman (1986) provides guidance on the choice of bandwidth. The closer the match, in terms of propensity scores, the greater the weight placed on the match.

Empirical research has often used bootstrap methods for conducting inference with matching estimators (for example, Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998)). Abadie and Imbens (2008), however, describe situations where the validity of the bootstrap is not justified and in the context of nearest neighbour matching suggest the use of analytical estimators of the asymptotic variance as an alternative (Abadie and Imbens, 2006).

To assess the degree of common support kernel density estimators can be used to obtain smoothed estimates of the densities of the propensity score for the treated and constructed control groups which can then be plotted against each other to identify areas where common support breaks down. Also it is important to check whether the covariates are successfully balanced between the treated and controls.

The success of matching methods relies on the ability of the chosen set of covariates to induce unconfoundedness in terms of the balancing score (20). This can be appraised using various methods that assess the degree to which matching has improved covariate balance between the treated and controls. One commonly used measure to assess the difference in the means of the covariates, used before and after matching, is the normalised difference (Rosenbaum and Rubin 1983; Lalonde 1986):

$$\frac{\overline{x}^1 - \overline{x}^0}{\sqrt{(Var(x^1) + Var(x^0)}} \tag{24}$$

Also t-tests for the difference in means are often proposed as a way of checking for balancing. This approach is criticised by Ho et al. (2007) and Imbens and Wooldridge (2008): for example, "the critical misunderstood point is that balance is a characteristic of the observed sample, not some hypothetical population. The idea that hypothesis tests are useful for checking balance is therefore incorrect." (Ho et al. 2007). They argue that this is compounded by the fact that pruning affects the statistical power of the hypothesis tests and that it is therefore misleading to use tests, such as t-ratios for the difference in means, as a guide to the quality of matching.

Analysts are often tempted to stop after checking for balance in the sample means of the covariates but the balancing condition relates to the full empirical distribution and it is wise to check higher moments (variance, skewness, kurtosis) and cross-moments (such as the covariance). Ho et al. (2007) suggest that nonparametric density plots and quantile-quantile (QQ) plots for each covariate and their interactions should be compared for the treated and controls.

Perfect balancing is unlikely to be achieved in practice and that, rather than simply comparing means after matching, running parametric regression models on the matched sample is likely to improve causal inferences (Rubin 1973b, 1979, 2006; Heckman, Ichimura

and Todd 1998; Imbens 2004; Abadie and Imbens 2006; Ho et al 2007). In this sense, matching can be used as a nonparametric preprocessing of the data to select observations prior to parametric modelling. Alternatively, Indurkhya et al. (2006) suggest conditioning regression models on the propensity score as a means of achieving balance and lessening reliance on the parametric assumptions underlying the specification of a model conditioned on all the confounding variables. Similarly, propensity score weights may be combined with regression models in a weighted least squares approach (see e.g., Robins and Rotnitzky 1995). Robins and Ritov (1997) argue that this provides a *doubly robust* estimator as it is consistent so long as either the propensity score or the regression model is correctly specified.

3. Selection on unobservables

3.1 Structural models and control functions

The use of standard regression models, such as equation (16), to estimate treatment effects relies on the unconfoundedness assumption (11). This implies that, conditional on the observed covariates, treatment assignment is uncorrelated with any unobservable factors that are captured by the error term in the simple treatment regression model (ε):

$$y = x'\beta + \tau d + \varepsilon \tag{25}$$

Now consider estimating this regression model when the treatment variable, *d*, is correlated with the error term, even after conditioning on the set of observed covariates, *x*. This reflects selection on unobservables and, due to omitted variable bias, OLS estimation would be a biased and inconsistent estimator of the treatment effect parameter τ. One approach to identifying the treatment effect in this case involves specifying a model to determine treatment assignment and estimating this jointly with the outcome equation of interest. Often this approach involves strong distributional assumptions and is fully parametric[8]. For example, consider the following specification for the treatment equation:

$$d^* = x'\gamma + z'\theta + \eta_i \tag{26}$$

Where d^* is a latent variable with the following observation mechanism:

[8] The strong assumptions that are often required to achieve identification of point estimates for treatment effects, in the presence of selection on unobservables, lead Manski to focus on partial identification and to propose estimates of the bounds for treatment effects (e.g., Manski 1990).

$$d = \begin{cases} 1 & iff \quad d^* > 0 \\ 0 & otherwise \end{cases} \tag{27}$$

The set of variables, z, provide a source of independent variation in d (independent from y) and, due to the fact that it does not enter the outcome equation directly, is referred to as an exclusion restriction. While these variables should be correlated with treatment assignment, they should be uncorrelated with the outcome, y, except through their effect on d.

By assuming a joint distribution for the two error terms ε and η the model can be estimated by full information maximum likelihood (FIML) estimation. For example it may be assumed that they are bivariate normally distributed with zero mean and covariance matrix given by:

$$\sum = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\eta} \\ \sigma_{\varepsilon\eta} & 1 \end{bmatrix} \tag{28}$$

In (28) the covariance terms, $\sigma_{\varepsilon\eta}$ reflect the endogeneity of the treatment variable. The restriction that the variance of the error for treatment assignment is unity is a requirement for the probit model for the binary treatment. In practice identification may be fragile due to the assumption of joint normality which may not be appropriate. Furthermore, finding appropriate instruments to include in z might prove difficult, or they might only be weakly related to treatment assignment or correlated with the error in the outcome equation of interest.

An alternative to estimating the model as a system is to take a control function approach. This focuses on estimating the outcome equation (25) on its own and captures selection bias by including estimates for the terms $E(\varepsilon \mid x, d=1) = E(\varepsilon \mid x, \eta > -x'\gamma - z'\theta)$ and $E(\varepsilon \mid x, d=0) = E(\varepsilon \mid x, \eta \leq -x'\gamma - z'\theta)$ as controls. In the case of bivariate normal error terms this corresponds to the Heckman (1979) two-step estimator, which uses the inverse Mills ratio from a probit model for the control function. For more general versions of the index models the control functions can be written as functions of the propensity score $P(d=1 \mid x, z)$ (Heckman and Robb 1985).

The example of a structural model used here is very simple, involving a single outcome and a single binary treatment with a linear equation for the outcome. In practice health economics applications may involve multiple outcomes and treatments and these may be binary, count data or multinomial variables. The FIML approach can be extended to specify a complete system of equations for the outcomes and treatments and estimates them jointly, allowing for common unobservable factors and identifying the model through exclusion restrictions. Estimation can be done by maximum likelihood (MLE), maximum simulated likelihood (MSL), Bayesian MCMC, discrete factor models (DFM) or using copulas (see e.g.,

Aakvik et al., 2003; Aakvik et al., 2005; Deb and Trivedi, 2006; Deb et al., 2006a; Deb et al., 2006b; Geweke et al., 2003; Zimmer and Trivedi, 2006). See Jones (2009) for a full discussion of these methods and applications.

3.2 Instrumental variables

The idea of instrumental variables estimators is to find one or more variables that are predictive of treatment assignment but which are not directly correlated with the outcome. For a linear regression equation like (25) this implies finding a set of variables, z, that are correlated with d, but uncorrelated with ε. Formally, these conditions can be written $E(z'\varepsilon) = 0$ and $E(z'd) \neq 0$. The variables, z, are referred to as instruments and can be used to derive consistent estimators of the treatment effect. Where the two conditions are met, the set of instruments form exclusion restrictions on the model.

The IV approach is often motivated by the analogy with randomized experiments and the notion that randomization is a perfect instrument. If subjects were randomized, say, by the toss of a fair coin then when z takes a value 1 (heads) the subject would be allocated to treatment and when it takes a value 0 (tails) they would be allocated to the control. The value of z perfectly predicts the assignment of treatment and, ordinarily, there would be no reason to expect that its value would be associated with outcomes. The challenge for researchers is to find natural instruments that mimic the process of randomization. In the absence of randomization it is unlikely that an instrument will be found that perfectly predicts treatment assignment for all individuals. Coupled with the fact that there are likely to be unobservable differences in the magnitude of the treatment effect across individuals and that these differences will often influence the assignment of treatment (recall equation (7)) this creates a fundamental difficulty with the IV approach that the treatment effect identified in sample data by the IV estimator will be contingent on the instruments used. This has been addressed through the concept of local average treatment effects (LATE) which is introduced below.

Conceptually, the mechanics of the IV estimator are best explained as a two-part process reflected in two-stage least squares (2SLS). In the first part, the treatment variable, d, is regressed on the set of instruments, z, and the set of included regressors in the outcome equation, x.[9] From this regression a prediction of treatment assignment can be obtained which replaces the treatment variable in the second stage regression of the outcome of interest. The estimated coefficient on this variable is the IV estimator of the treatment effect. Note, however, that since the treatment variable in the second stage is replaced by its predicted value from the first stage regression (with the corresponding uncertainty surrounding these estimates), OLS standard errors in the second stage will be biased and

[9] If any of the set of exogenous explanatory variables, x, are omitted from the first stage, then this might induce correlation between the omitted variables and the second stage residuals, potentially leading to an inconsistent estimator of the treatment effect.

require correction. This correction is automatically implemented in 2SLS routines in standard econometric software packages.

Standard IV estimators such as 2SLS are consistent in the presence of heteroscedastic errors. Standard errors, however, will be biased leading to invalid inference. Further corresponding diagnostic tests, and importantly, tests of overidentifying restrictions will also be invalid if the errors are heteroscedastic. A potential solution is to use "robust" estimators of the covariance matrix. More commonly, however, is the use of generalised method of moments (GMM) estimators (Hansen, 1982). The GMM IV estimator exploits the set of *l* (number of instruments) moment conditions to derive a consistent and efficient estimator of the treatment effect in the presence of arbitrary heteroscedasdicity (Baum et al., 2003).

While IV estimators offer a potential solution to the problem of selection on unobservables, they are biased in finite samples. Accordingly, they are best employed when dealing with large samples where one can appeal to the consistency properties of the estimators. The credibility of the IV approach and the extent of bias relies on the assumptions outlined above. These assumptions are ultimately untestable. But, to gauge their credibility, a number of tests should be performed prior to reporting the estimated treatment effect and it is considered good practice to report tests for the: endogeneity of the treatment variable; instrument validity; model identification (relevance); the problem of weak instruments; and model specification. We cover each of these briefly.

There is no direct test of the validity of an instrument set, $E(z'\varepsilon) = 0$ since ε is unknown. However, if there are more instruments than endogenous regressors a test of over-identifying restrictions is available which should be routinely reported for any practical application of instrumental variables. For IV where the errors are homoscedastic, the appropriate statistic is provided through the Sargan test (Sargan 1958). This can be viewed as nR^2 of a regression of the estimated IV residuals on the set of instruments, where n is the sample size and R^2 is the uncentered r-squared (see Baum et al., 2003). For heteroscedastic errors, the GMM alternative is the *J* statistic of Hansen (1982). This is essentially a test of a function of the corresponding moment conditions derived from minimising the GMM criterion function. Under the assumption of homoscedasticity the *J* statistic and Sargan statistic coincide. Both provide tests of the null hypothesis that the instruments are uncorrelated with the error term (that they are *valid*) and that they are therefore legitimately excluded from the outcome equation. The statistics are distributed as chi-squared with degrees of freedom equal to the number of overidentifying restrictions (number of instruments, *l*, minus the number of endogenous regressors, *k*). Rejection of the null hypothesis implies that the orthogonality conditions do not hold and the approach to defining instruments should be reviewed. The test is only applicable to situations where there are more instruments than endogenous regressors and hence the model is said to be overidentified. Where this is not the case, and further instruments are not available, higher order terms of the instruments or interactions with exogenous regressors, *x*, might be used to obtain appropriate over-identifying restrictions.

Identification and weak instruments refer to related concepts, although the implication of each is different. Identification refers to the strength of the relationship between the instruments, z, and the endogenous regressor, d, that is $E(z'd) \neq 0$ and is often referred to as the *relevance* of the instruments. Tests of identification are often tests of the rank of $E(z'd)$. Where there is more than one endogenous regressor, the rank test refers to the ability of the set of instruments to induce independent variation in the endogenous regressors. Where this is not the case, then the model is not be identified through the chosen instrument set. Test statistics, such as that suggested by Anderson (1984) and Cragg and Donald (1993) provide tests of whether at least one endogenous regressor has no correlation with the set of instruments.

A further problem arises, however, even where we are satisfied that both the conditions, $E(z'\varepsilon) = 0$ and $E(z'd) \neq 0$ hold. If some elements of $E(z'd)$ are close to zero, that is if one or more of the endogenous regressors are poorly correlated with the set of instruments, then we face a problem of *weak instruments*. A consequence of weak instruments is an increase in IV bias (Hahn and Hausman, 2002). To illustrate consider the simple case where there is a single treatment d and instrument z. Then the probability limit of the IV (2SLS) estimator is:

$$\hat{\tau}_{IV} \underset{p}{\rightarrow} \tau + \frac{Cov(z,\varepsilon)}{Cov(z,d)} \tag{29}$$

So, if $Cov(z,\varepsilon) = 0$ exactly, the IV estimator is asymptotically unbiased. But so long as there is some residual correlation between the instrument and the outcome, as is likely to be the case in practice, then weak instruments, implying a small value for $Cov(z,d)$, may lead to explosive bias. In the extreme, weak instruments would result in a bias in the IV estimate equivalent to the bias observed in simple OLS in the presence of an endogenous regressor. Accordingly, IV regression would offer no improvement over OLS. The test statistic of Cragg and Donald (1993) has been suggested as providing a test for weak instruments. For a single endogenous regressor this statistic is equivalent to an F-test of the joint significance of the set of instruments in the first-stage regression of 2SLS. In these circumstances, Staiger and Stock (1997), suggest as a rule of thumb that a first-stage F statistic of less than 10 is course for concern. Moreover, Stock and Yogo (2002) provide critical values for the Cragg and Donald statistic that limit the IV bias to a given percentage of the bias of OLS. If the chosen instruments are suspected to be weak, then a search for alternative instruments should be undertaken. This might include consideration of transformations of the existing instruments. A further concern is that IV bias is an increasing function of the number of instruments and hence, caution should be used in selecting overidentifying instruments.

The fundamental problem we are attempting to resolve through the use of IV is that the treatment variable, d, is endogenous, that is it is correlated with the error, ε. A simple test of whether treatment is endogenous follows the two-stage approach to estimation outlined above. As before, the first stage consists of the regression of d, on the set of instruments, z,

and the set of included regressors in the outcome equation, *x*. From this regression compute the predicted error, $\hat{\eta} = d - \hat{d}$. This can then be included in the second stage regression of *y* on *d*, *x* and $\hat{\eta}_i$. A test of significance of the estimated coefficient on $\hat{\eta}_i$ is a test of endogeneity of the treatment variable. Further, for this model, the estimated coefficient on *d* is the IV estimator of the treatment effect.

Finally, it is good practice to apply tests of model specification. A general test of specification is provided through the use of Ramsey's (1969) reset test adapted for instrumental variables (see, Pesaran and Taylor 1999; Pagan and Hall 1983). The test is more properly thought of as a test of a linearity assumption in the mean function or a test of functional form restrictions and omitted variables (see for example, Wooldridge, 2002) and can be useful as a general check of model specification. For OLS the test simply consists of computing the predicted values from a regression, $\hat{y}$, and inserting powers (for example, $\hat{y}^2$, $\hat{y}^3$ and $\hat{y}^4$) of the predictions into the regression and re-estimating the model. A Wald test of the joint significance of these terms provides a test of specification under the null hypothesis of no neglected nonlinearities. The test can be adapted for IV regression by forming the predictions based on x and $\hat{d}$, the predicted value of the endogenous regressor from the first-stage regression (Pesaran and Taylor 1999). Alternatively, the Pagan and Hall (1983) version of the test forms the predictions from a reduced form regression of *y* on the set of instruments, *z*, and exogenous regressors, *x*.

As mentioned above, if there is heterogeneity in the response to treatment the IV estimator identifies a local average treatment effect, or LATE (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996). This is the average treatment effect over the sub-group of the population that are induced to participate in the treatment due to variation in the instrument[10]. The fact that IV estimates only identify the LATE and that the results are therefore contingent on the set of instruments explains why different empirical studies can often produce quite different estimates, even though they examine the same outcomes and treatments. The fact that the definition of the LATE involves the values of the instrument is criticized by Heckman (2008) who argues that the definition of treatment effects should be kept distinct from the particular strategy used to identify and make inferences about the effect. Heckman and Vytlacil (1999, 2007) have extended the analysis of local treatment effects by specifying a latent index model for the assignment of treatment and using it to identify those individuals who are indifferent between treatments. The method is applied by Basu et al, (2007) in a study of treatments for breast cancer. This approach defines the marginal treatment effect (MTE): the treatment effect among those individuals at the margin of treatment. The MTE is the building block for the LATE, average treatment effect on the treated (ATT) and average treatment effect (ATE). It can be identified using Local-IV methods or by specifying multiple equation models with a common factor structure (e.g., Basu et al., 2007; Aakvik et al., 2005).

[10] The definition of the LATE relates to instruments that are monotonically related to treatment assignment.

3.3 Regression discontinuity

The regression discontinuity (RD) design exploits situations where the assignment to treatment changes discontinuously with respect to a threshold value of one or more exogenous variables (the *forcing variables*). For example, eligibility for free prescription drugs may be a deterministic function of an individual's age. The contrast between individuals on either side of the discontinuity is used to identify the treatment effect. In a sense it can be seen as an IV strategy, in that the forcing variable predicts assignment of the treatment. But rather than excluding a direct effect of the forcing variable on the outcome, identification relies on the discrete jump in the outcomes at the point of discontinuity. In a sharp regression discontinuity design, passing the threshold completely determines the allocation of treatment. In a fuzzy design, which is more likely in practice, the allocation of treatment is stochastic and the threshold creates a discontinuity in the probability of treatment. The discontinuity design relies on a comparison of observations "before and after" the threshold and does not have a separate control group. For this reason, applications typically use a narrowly defined neighbourhood around the discontinuity, to try and ensure that the treated and untreated observations are comparable in other respects, and often exploit graphical and nonparametric methods to identify any discontinuity. This approach is yet to be used widely in health economics. Studies that implicitly use a discontinuity design include Almond (2006), Lleras-Muney (2005) and Pop-Eleches (2006).

3.4 Difference-in-differences

In applied economics one of the most commonly used methods of evaluating the effect of a policy intervention is the difference-in-differences, or double differences, (DD) approach (Ashenfelter 1978; Ashenfelter and Card 1985; Heckman and Robb 1985). DD is essentially a before & after design with controls and there are many recent examples of its use in the health economics literature (see Jones 2009). The basic method requires data in both a pre-treatment and post-treatment period, $t = 0, 1$, on the treated (g=1) and control (g=0) groups. These may be constructed from longitudinal data on the same individuals or from repeated cross sections (Heckman, Ichimura and Todd 1997). DD is often used to estimate treatment effects in the context of randomized social experiments, especially if blocked randomization is used and regression analysis is still required to adjust for imbalance in covariates post-randomization. As an illustration, for the stylised example depicted in Figure 1, the DD treatment effect is the quantity (CA) – (DB). Given a "parallel trends" assumption (DB) = (C'A) and therefore the treatment effect is the quantity: CC'.

Consider again the definition of the PATT, now expressed in terms of the indicators *t* and *g*:

$$
\begin{aligned}
PATT\,|_x &= E\left[y^1 - y^0 \mid d=1, x \right] \\
&= E\left[y^1 - y^0 \mid t=1, g=1, x \right] \\
&= E\left[y^1 \mid t=1, g=1, x \right] - E\left[y^0 \mid t=1, g=1, x \right] \\
&= \tau(.)
\end{aligned}
\tag{30}
$$

This implies:

$$
E\left[y^1 \mid t=1, g=1, x \right] = \tau(.) + E\left[y^0 \mid t=1, g=1, x \right] \tag{31}
$$

Then, the basic identifying assumption of linear DD is:

$$
E\left[y^0 \mid t=1, g=1, x \right] = x'\beta + \alpha t + \gamma g \tag{32}
$$

In (32) the time effect (α) is common to both treated and controls ("parallel trends"). The group effect γ captures any fixed (time-invariant) difference between the treated and controls, whether this is due to observable or unobservable confounders. The covariates x are included to account for any time-varying confounders.

By definition:

$$
E(y \mid t, g, x) = dE(y^1 \mid t, g, x) + (1-d)E(y^0 \mid t, g, x) \tag{33}
$$

Then, using (32) gives:

$$
\begin{aligned}
&E(y \mid t, g, x) = d\left\{ \tau(.) + x'\beta + \alpha t + \gamma g \right\} + (1-d)\left\{ x'\beta + \alpha t + \gamma g \right\} \\
&= \tau(.)d + x'\beta + \alpha t + \gamma g
\end{aligned}
\tag{34}
$$

So the DD estimate of the PATT is the coefficient on the interaction term $d=t.g$ in a model that also includes main effects for time (t) and group (g). As with the standard linear regression estimator, $\tau(.)$ need not be constant. To allow for heterogeneity in the treatment effect d could be interacted with the x variables or it could be treated as a random parameter.

The DD approach is applicable to repeated cross-sections of data as long as it can be assumed that the composition of the observations in the two cross-sections has not changed over time. It further assumes that it is possible to identify individuals in the first period who are eligible for treatment. This would be straightforward if, for example, a policy was implemented in a given area only, or directed at individuals with certain characteristics (for example, young smokers).

An important assumption of the DD method is the common time trend for both the treated and control group. This assumes that in the absence of treatment, the average change in the outcomes would be the same for treated individuals as for untreated individuals. Failure of this assumption would confound the estimated treatment effect with a natural time trend producing biased inference. An example of the assumption failing to hold is when there is a pre-treatment dip in outcomes for the treated cases (known in the labour economics literature as an *Ashenfelter dip*).

For identification, the closer the control and treatment group in terms of both observable and unobservable characteristics, the greater the credibility of the DD approach in recovering the treatment effect. To enhance comparability between treatment and control group, the approach can be combined with matching. Accordingly, pre-treatment controls can be matched with treated individuals to ensure that the characteristics of treated and controls are close and hence are more comparable.

An advantage of combining the DD approach with matching, for example using propensity scores, is that the method of matching relies on the assumption that selection into treatment is based on observable measured characteristics (Heckman, Ichimura, Smith and Todd 1998). Where this assumption is untenable, inference will be contaminated with omitted variable bias through a failure to control for important but unobserved or unobservable characteristics. Combining matching with the DD approach further allows control for unobserved time-invariant components and accordingly, increasing the credibility of the identification of the treatment effect.

Given sufficient information on both controls and treated individuals prior to treatment, it is possible to test the credibility of the parallel trend assumption. An example of this is Galiani et al.'s (2005) evaluation of the impact of the privatisation of local water services on child mortality in Argentina. They estimate a *placebo regression* using only data from the pre-treatment period, but including an indicator of those cases that would go on to be treated. Also they include measures of deaths from infectious and parasitic diseases and from causes unrelated to water quality. The fact that they detect a reduction for the former but not for the latter creates confidence in their identification strategy.

Equation (32) shows that linearity is central to the standard DD approach. Athey and Imbens (2006) propose a more general approach, labelled changes-in-changes (CC), that relaxes the additivity assumption required for DD. Their approach allows pre-treatment outcomes to be general function of unobservables, with the restrictions that outcomes are strictly monotonic in the unobservables and that the unobservables are time invariant within groups. Let F_{gt} denote the distribution function for y for group g (0,1) at time t (0,1), The CC approach uses the distribution F_{01}, for the controls after treatment, to construct the counterfactual. The counterfactual value of an observed outcome for a treated individual, y, is given by $F_{01}^{-1}(F_{00}(y))$, where $F_{00}(y)$ is the probability corresponding to y in the distribution of outcomes for the controls before treatment and $F_{01}^{-1}(.)$ gives the quantile of that probability in the post-treatment distribution for the controls.

3.5 Panel data models

Where repeated observations are available on individuals we can extend the linear DD model, (34), to a panel data regression model that includes error components:

$$y_{it} = \tau d_{it} + x'_{it}\beta + \alpha_t + \gamma_i + \varepsilon_{it} \tag{35}$$

where $i\ (i = 1,\ldots n)$ indexes individuals and t time $(t = 1,\ldots,T)$. This is a two-way fixed effects specification (2FE) where the time effects (α_t) capture the common trend, usually measured by including a dummy variable for each wave of the panel, and each individual has their own fixed effect (γ_i) to capture their time-invariant characteristics.

With individual panel data an individual's own history, prior to treatment is used to construct the counterfactual. In effect, they are used as their own control group. But there may be other "natural control groups" that can be used to form the counterfactual. For example, variation within families can be used to control for shared environmental and genetic traits: this can involve using information on parents or grandparents (Auld and Sidhu 2005); information on siblings (Currie and Stabile 2006); or on twins (Almond et al. 2005; Behrman and Rosenzweig 2004; Black et al. 2007).

In this context, the estimate of τ is a difference-in-differences estimator. The consistency of this estimate relies on a particular variant of the unconfoundedness or conditional independence assumption (11):

$$E(y^0_{it} \mid d_{it}, x_{it}, t, \gamma_i) = E(y^0_{it} \mid x_{it}, t, \gamma_i) \tag{36}$$

The individual effect in (35) can be removed from the equation by transforming the variables to represent deviations from their within individual means:

$$y_{it} - \overline{y}_i = \tau\left(d_{it} - \overline{d}_i\right) + \left(x'_{it} - \overline{x}'_i\right)\beta_1 + (\alpha_t - \overline{\alpha}) + \left(\varepsilon_{it} - \overline{\varepsilon}_i\right) \tag{37}$$

with $\overline{y}_i = 1/T\sum_{t=1}^{T} y_{it}$ etc. Alternatively, first differencing the data has the same desired effect:

$$y_{it} - y_{it-1} = \tau\left(d_{it} - d_{it-1}\right) + \left(x'_{it} - x'_{it-1}\right)\beta_1 + (\alpha_t - \alpha_{t-1}) + \left(\varepsilon_{it} - \varepsilon_{it-1}\right) \tag{38}$$

Note, however, that in both approaches not only is the unobserved individual effect removed but also all time-invariant regressors. A further approach is to specify fixed effects

by including dummy variables for each individual[11]. This approach becomes burdensome where the number of individuals is large.[12]

The fixed effects model identifies the treatment effect through variation within individuals and over time but does not exploit variation across individuals (the individual effect is removed). Accordingly, the approach is less efficient than the random effects specification which exploits variation both within and between individuals. Further, unlike random effects, fixed effects approach also does not estimate the impact of time-invariant regressors on outcomes. However, fixed effects deals explicitly with correlation between the treatment variable and the error disturbance, as long as this correlation is confined to the unobserved individual-specific effect. Hausman (1978) provides a test of fixed versus random effects. It is based on the idea that the fixed effects estimator is unbiased and consistent if inefficient, whereas the random effects estimator is biased and inconsistent in the presence of correlation between the regressors and the time-invariant individual error component. In the absence of such correlation, it is, however, more efficient than the fixed effects estimator due to it exploiting variation both within and between individuals. Parameter estimates from the random effects estimator can be compared to those from fixed effects estimation and the degree of discrepancy between the sets of estimates provides the basis of a test for evidence that the regressors are correlated with the unobserved individual error component.

The credibility of the conditional independence assumption used above in equation (36) may be strengthened when the panel data model conditions on lagged values of the regressors and, more importantly on lagged outcomes. For example with a one-period lag of the outcome this gives the dynamic specification:

$$y_{it} = \tau d_{it} + \delta y_{it-1} + x'_{it}\beta + \alpha_t + \gamma_i + \varepsilon_{it} \tag{39}$$

Which implies a modified version of the conditional independence assumption:

$$E(y^0_{it} \mid d_{it}, y_{it-1}, x_{it}, t, \gamma_i) = E(y^0_{it} \mid y_{it-1}, x_{it}, t, \gamma_i) \tag{40}$$

For models like (39), that include dynamics as well as an individual effect, the usual fixed effects estimators such as within-groups or first differences break-down. This is because the transformations induce correlation with the error term ε_{it}. Arellano and Bond (1991) proposed generalised method of moments (GMM) estimators for these kinds of dynamic panel data models. Instruments are created within the model by first taking differences of the equation to sweep out the individual effect and then using lagged levels or differences of

[11] Excluding one individual to represent the baseline case against which the effects of others can be contrasted.

[12] Inferences concerning the estimate of τ depend on the assumptions made about the error term ε_{it} and it may not be reasonable to assume serial independence. Recent work by Bertrand, Duflo and Mullainthan (2004) have suggested that standard errors should be adjusted to allow for clustering within individuals in applications of DD.

the regressors as instruments. The validity of this approach rests on the degree of autocorrelation in the error term. Bover and Arellano (1997) extended the use of GMM to dynamic specifications for categorical and limited dependent variable models, where it is not possible to take first differences or deviations as the latent variable y^* is unobserved.

Contoyannis et al. (2003) consider the determinants of a binary indicator for functional limitations using seven waves of the British Household Panel Survey (BHPS). Their models allow for persistence in the observed outcomes due to state dependence, unobservable individual effects and persistence in the transitory error component. These are estimated by Maximum Simulated Likelihood using the GHK simulator. In related work Contoyannis et al. (2004) explore the dynamics of self-assessed health (SAH) in the BHPS. In this kind of application it is quite likely that the unobserved individual effect will be correlated with the observed regressors including indicators of treatment. Also, it is well known that in dynamic specifications the individual effect will be correlated with the lagged dependent variable, which gives rise to what is known as the *initial conditions problem*. To deal with the problem of initial conditions an attractively simple approach suggested by Wooldridge (2005) can be used. This specifies the distribution of the individual effects as a linear function of the outcome at the first wave of the panel and of the within-individual means of endogenous regressors.

4. Ex ante evaluation and microsimulation

The techniques outlined above offer approaches to the ex-post evaluation of treatment effects by comparing outcomes across suitably constructed treatment and control groups. Clearly, ex-post evaluation is crucial in understanding the true impact of a treatment or policy. There may be situations, however, where the likely impact of the introduction of a policy or treatment is required, but where experimental approaches to ex-post evaluation are not feasible, for example due to cost, or perhaps ethical or political concerns. In such circumstances, the ability to simulate the effects of a policy that currently does not exist but potentially could be implemented has become the focus of recent research. Ex-ante approaches to evaluation have been used in the field of labour economics, where simulation of the labour supply response to changes in aspects of the tax-benefit system is of interest (for example, Creedy and Duncan 2002; Blundell and MaCurdy 1999). In development economics ex-ante evaluation has been used to investigate the effects of conditional cash transfer programmes for stimulating improved social outcomes (Todd and Wolpin 2009).

While ex-post evaluation uses data on both the treated and non-treated, for ex-ante evaluation the treatment group is simulated to represent the population characteristics of interest as they would appear under the hypothetical policy change. Ex ante evaluation can be based on either structural estimation or reduced form estimation, built on behavioural models. For example, this might involve changing household budget constraints to reflect the impact of wage subsidies, income support or conditional cash transfers (CCTs), as in Todd and Wolpin (2009), who evaluate Mexico's PROGRESA experiment. They adopt a

reduced form approach that avoids the use of structural estimation and specific functional forms. They show that, if the impact of the policy reform can be captured wholly through changes in the budget constraint, a fully nonparametric approach can be used and they derive a matching estimator to identify the treatment effects. This places ex ante evaluation within the potential outcomes framework and constructs counterfactuals by matching untreated individuals with other untreated individuals. The matching is based on functions of observable variables that are generated by the model, reflecting the impact of the policies on shadow prices and full income in the budget constraint. For example, if the impact of a proposed policy would be to increase a household's income by $500, that household would be matched with another that has an income $500 greater than theirs' but that is comparable in terms of other relevant characteristics.

Microsimulation has been suggested as an instrument for ex-ante evaluative analysis (Bourguignon and Spadaro 2006). Microsimulation models consider the changes, at a micro level, in the economic circumstances brought about by a policy and the corresponding imputed behavioural responses and outcomes. Typically microsimulation models consider representative samples of agents such as individuals or households and assess the impact of policy on these samples. An advantage of the microsimulation approach is that the imputed consequences of a policy change can be made over various time horizons. Further, since imputed changes can be made at the individual level, evaluation of the distributional impact of policies can be undertaken (Spadaro 2005).

Microsimulation is an established and widely used tool for analyzing the impact of policies and has been developed in areas outside the health sector, most commonly focusing on the impact of fiscal policies on population income and welfare distributions. The approach, however, has yet to be used widely in health economics. Microsimulation models can be arithmetical or behavioural and static or dynamic (Bourguignon and Spadaro 2006). Arithmetical models are simply concerned with the gainers and losers from a policy and ignore any behavioural responses that might be brought about by the reform. In this sense, arithmetic models estimate the immediate impact of a policy. Behavioural microsimulation, in contrast, accounts for the behavioural response of individuals to the policy intervention often achieved through the use of structural behavioural models based on a utility maximizing framework subject to a budget constraint. The approach has the benefit of allowing policy evaluation to be assessed on the basis of corresponding social welfare functions (Spadaro, 2005). The various parameters of the utility function are estimated using appropriate data, for example, a household survey. Once estimates of the parameters are known, simulating outcomes of interest under alternative policy regimes can be undertaken.

Static micro-simulation models offer a snap-shot of the impact of a policy reform at a particular point in time. The simulated population is the same as the reference population with the exception that relevant characteristics (income, consumption, health etc.) are updated. This approach is appropriate if there are few long-run behavioral responses to the policy. Dynamic models, in contrast, project samples of individuals over time. Often this is achieved by creating a synthetic panel that simulates individual or household trajectories. The approach incorporates relevant life-course events such as changes in demographics,

household composition, educational attainment and labour market transitions using what are termed dynamic ageing techniques. Once the synthetic panel is constructed simulations of the impact of an intervention or policy reform can be computed More ambitious approaches incorporate dynamics with behavioural responses to policy reforms. This adds an additional layer of complexity to the simulation whereby transition probabilities are assumed to be endogenous to the budget constraint requiring further estimation of more sophisticated structural models (Bourguignon and Spadaro, 2006).

5. Recommended reading

This chapter has skimmed the surface of a large and ever-growing literature. Comprehensive recent reviews are provided by: Angrist and Pischke (2008); Blundell and Costa-Dias (2008); Frölich (2004), who pays particular attention to multiple treatments; Heckman and Vytlacil (2007); Heckman (2008); Imbens and Wooldridge (2008), who give a general review and history of evaluation methods in econometrics and statistics and who provide a formal treatment of some of the most recent technical developments; Lee (.2005); and Todd (2008), who provides examples from development economics. Imbens (2004) concentrates on methods that apply under the assumption of unconfoundedness. Rubin (2006) collects together his contributions on the subject of matching. Caliendo and Kopeining (2008) focus on the practical application of propensity score matching. Auld (2006) discusses the pros and cons of the instrumental variable approach in the context of health economics. Heckman and Smith (1995), Banerjee and Duflo (2008) and Deaton (2008) debate the pros and cons of randomized social experiments. Bourguignon and Spadaro (2006) discuss microsimulation techniques as a tool for the evaluation of public policies.

Jones (2009) provides a detailed review of the health econometrics literature since 2000 and focuses on the identification strategies that have been adopted in recent work with health data.

Cameron and Trivedi (2009) provide a comprehensive guide to the use of Stata for microeconometrics including many of the approaches discussed in this chapter. Jones, Rice, Bago d'Uva and Balia (2007) illustrate the use of Stata for applied health economics, with a particular emphasis on panel data regression. Becker and Ichino (2002) and Nichols (2007) provide guidance on the application of estimators for treatment effects, especially matching approaches, in Stata.

REFERENCES

Aakvik, A., J.J. Heckman and E.J. Vytlacil (2005). 'Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs'. *Journal of Econometrics,* 125: 15-51.

Aakvik, A., T.H. Holmas and E. Kjerstad (2003). 'A low-key social insurance reform-effects of multidisciplinary outpatient treatment for back pain patients in Norway'. *Journal of Health Economics,* 22: 747-62.

Abadie, A. and G. Imbens (2006). 'Large sample properties of matching estimators for average treatment effects'. *Econometrica,* 74: 235-67.

Abadie, A. and G. Imbens (2008). 'On the Failure of the Bootstrap for Matching Estimators.' *Econometrica*, 76: 1537-57.

Almond, D. (2006). 'Is the 1918 influenza pandemic over? Long term effects of in utero influenza exposure in the post 1940 US'. *Journal of Political Economy,* 114: 672-712.

Almond, D., K.Y. Chay and D.S. Lee (2005). 'The costs of low birth weight'. *Quarterly Journal of Economics,* 120: 1031-83.

Angrist, J., G. Imbens and D. Rubin (1996). 'Identification of causal effects using instrumental variables'. *Journal of the American Statistical Association,* 91: 444-72.

Anderson, T. W. (1984). *Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley & Sons

Angrist, J. and S. Pischke (2008). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: University of Princeton Press.

Arellano, M. and S. Bond (1991). 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations'. *Review of Economic Studies,* 58: 277 - 97.

Ashenfelter, O. (1978). 'Estimating the effect of training programs on earnings'. *Review of Economics and Statistics,* 60: 47-57.

Ashenfelter, O. and D. Card (1985). 'Using the longitudinal structure of earnings to estimate the effect of training programs'. *Review of Economics and Statistics*, 67: 648-60.

Athey, S. and G. W. Imbens (2006). 'Identification and inference in nonlinear difference-in-differences models'. *Econometrica,* 74: 431-97.

Auld, M.C. (2006). 'Using observational data to identify the causal effects of health-related behaviour'. In A. M. Jones (ed.) *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Auld, M.C. and N. Sidhu (2005). 'Schooling, cognitive ability and health'. *Health Economics,* 14: 1019-34.

Banerjee, A. V. and E. Duflo (2008). 'The experimental approach to development economics'. NBER Working Paper w14467.

Baum, C.F., Schaffer, M.E., Stillman, S. (2003). 'Instrumental variables and GMM: estimation and testing'. *Stata Journal,* 3: 1-31..

Basu, A., J. Heckman, S. Navarro and S. Urzua (2007). 'Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients'. *Health Economics* 16: 1133-57.

Becker, S. and A. Ichino (2002). 'Estimation of average treatment effects based on propensity scores'. *The Stata Journal*, 2: 358-77.

Behrman, J.R. and M.R. Rosenzweig (2004). 'Returns to birthweight'. *Review of Economics and Statistics,* 86: 586-601.

Bertrand, M., E. Duflo and S. Mullainathan (2004). 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics*, 119: 249-75.

Björkman, M., J. Svensson (2009). 'Power to the people: evidence from a natural randomized field experiment on community-based monitoring in Uganda.' *The Quarterly Journal of Economics*, 124: 735-69.

Black, S., P. Devereux and K. Salvanes (2007). 'From the cradle to the labour market ? The effect of birth weight on adult outcomes'. *The Quarterly Journal of Economics*, 409 - 39.

Bleakley, H. (2007). 'Disease and development evidence from hookwork eradication in the Americal South'. *The Quarterly Journal of Economics,* 122: 73-117.

Blundell, R. and M. Costa-Dias (2008). 'Alternative approaches to evaluation in empirical microeconomics'. CEMMAP working paper CWP26/08.

Blundell, R and T. MaCurdy, (1999). 'Labour Supply: A Review of Alternative Approaches.' In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Volume 3A, Amsterdam: Elsevier.

Bover, O. and M. Arellano (1997). 'Estimating dynamic limited-dependent variable models from panel data'. *Investigaciones Economicas,* 21: 141 - 65.

Bourguignon, F. and A. Spadaro, (2006). 'Microsimulation as a tool for evaluating redistrbution policies.' *Journal of Economic Inequality*, 4: 77-106.

Caliendo, M. and S. Kopeneinig (2008). 'Some practical guidance for the implementation of propensity score matching'. *Journal of Economic Surveys,* 22: 31-72.

Cameron, A. C. and P. K. Trivedi (2009). *Microeconometrics Using Stata. College Station Texas: Stata Press.*

Claxton, K. P. (1999). 'The irrelevance of inference: a decision making approach to the stochastic evaluation of health care technologies'. *Journal of Health Economics,* 17: 341-64.

Cochran, W. and D. Rubin (1973). 'Controlling bias in observational studies: a review'. *Sankhya,* 35: 417-46.

Contoyannis, P., A.M. Jones and N. Rice (2003). 'Simulation-based inference in dynamic panel probit models: an application to health'. *Empirical Economics,* 28: 1-29.

Contoyannis, P., A.M. Jones and N. Rice (2004). 'The dynamics of health in the British household panel survey'. *Journal of Applied Econometrics,* 19: 473-503.

Cragg, J. G. and S. G. Donald (1993). 'Testing indentifiability and specification in instrumental variables models'. *Econometric Theory,* 9: 222-40.

Creedy, J., and Duncan, A. (2002) 'Behavioral micro-simulation with labor supply responses'. *Journal of Economic Surveys*, 16: 1-39.

Deaton, A.S. (2008). 'Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development'. NBER Working Paper 14690.

Deb, P., C. Li, P.K. Trivedi and D.M. Zimmer (2006a). 'The effect of managed care on use of health care services: results from two contemporaneous household surveys'. *Health Economics,* 15: 743-60.

Deb, P., M.K. Munkin and P.K. Trivedi (2006b). 'Bayesian analysis of the two-part model with endogeneity: application to health care expenditure'. *Journal of Applied Econometrics,* 21: 1081-99.

Deb, P. and P.K. Trivedi (2006). 'Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization'. *Econometrics Journal,* 9: 307-31.

Deheija, R. and S. Wahba (1999). 'Causal effects in nonexperimental studies: reevaluating the evaluation of training programs'. *Journal of the American Statistical Association,* 94: 1053-62.

Deheija, R. and S. Wahba (2002). 'Propensity score matching methods for nonexperimental causal studies'. *Review of Economic Studies,* 84: 151-61.

Doyle, J.J. (2005). 'Health insurance, treatment and outcomes: using auto accidents as health shocks'. *Review of Economics and Statistics,* 87: 256-70.

Duflo, E. (2000). 'Child health and household resources in South Africa: evidence from the old age pension program'. *American Economic Review,* 90: 393 - 98.

Evans, W.N. and D.S. Lien (2005). 'The benefits of prenatal care: evidence from the PAT bus strike'. *Journal of Econometrics,* 125: 207-39.

Frijters, P., J.P. Haisken-Denew and M.A. Shields (2005). 'The causal effect of income on health: evidence from German reunification'. *Journal of Health Economics,* 24: 997-1017.

Frölich, M. (2004). 'Programme evaluation with multiple treatments'. *Journal of Economic Surveys,* 18: 181-224.

Galiani, S., P. Gertler and S. E. Schargrodsky (2005). 'Water for life: the impact of the privatization of water services on child mortality'. *Journal of Political Economy,* 113: 83-120.

Gertler, P. (2004). 'Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment'. *American Economic Review,* 94: 336-41.

Geweke, J., G. Gowrisankaran and R.J. Town (2003). 'Bayesian inference for hospital quality in a selection model'. *Econometrica,* 71: 1215-38.

Hahn, J. (1998). 'On the role of the propensity score in efficient semiparametric estimation of average treatment effects'. *Econometrica,* 66: 315-31.

Hahn, J., and J. Hausman (2002). 'Notes on bias in estimators for simultaneous equation models'. *Economics Letters*, 75: 237-41.

Hansen, L. (1982). 'Large sample properties of generalized method of moments estimators'. *Econometrica*, 50: 1029-1054.

Hausman, J.A. (1978). 'Specification tests in econometrics'. *Econometrica*, 46: 1251-71.

Heckman, J. J. (1979). 'Sample selection bias as a specification error'. *Econometrica,* 47: 153-62.

Heckman, J. J. (2008). 'Econometric causality'. CEMMAP working paper CWP1/08.

Heckman, J. J., H. Ichimura and P. E. Todd (1997). 'Matching as an econometric evaluation estimator: evidence from evaluating a job training programme'. *Review of Economic Studies,* 64: 605-54.

Heckman, J. J. , H. Ichimura and P. E. Todd (1998). 'Matching as an econometric evaluation estimator'. *Review of Economic Studies, 65: 261-94.*

Heckman, J. J., H. Ichimura, J. A. Smith and P.E. Todd (1998). 'Characterizing selection bias using experimental data'. *Econometrica,* 66: 1017-98.

Heckman, J. J., R. J. LaLonde and J. A. Smith (1999). 'The economics and econometrics of active labor market programs'. In O. Ashenfelter and D.Card (eds.), *Handbook of Labor Economics Volume III.* Amsterdam: Elsevier.

Heckman, J. J. and R. Robb (1985). 'Alternative models for evaluating the impact of interventions'. In J. J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data.* Cambridge: Cambridge University Press.

Heckman, J. J. and J. A. Smith (1995). ' Assessing the case for social experiments'. *Journal of Economic Perspectives,* 9: 85-115.

Heckman, J. J., J. A. Smith and N. Clements (1997). 'Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts'. *Review of Economic Studies,* 64: 487-535.

Heckman, J. J. and E. Vitlacyl (1999). 'Local instrumental variables and latent variable models for identifying and bounding treatment effects'. *Proceedings of the National Academy of Sciences,* 96: 4730-34.

Heckman, J. J. and E. Vitlacyl (2007). 'Econometric evaluation of social programs'. In J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics Volume 6B.* Amsterdam: Elsevier.

Hirano, K. and G. W. Imbens (2001). 'Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization'. *Health Services & Outcomes Research Methodology,* 2: 259-78.

Hirano, K., G. W. Imbens and G. Ridder (2003). 'Efficient estimations of average treatment effects using the estimated propensity score'. *Econometrica,* 71: 1161-89.

Ho, D. E., K. Imai, G. King and E. A. Stuart (2007). 'Matching as nonparametric preprocessing for reduced model dependence in parametric causal inference'. *Political Analysis*, 15: 199-236.

Holland, P. W. (1986). 'Statistics and causal inference'. *Journal of the American Statistical Association,* 81: 945-60.

Imbens, G.W. (2004). 'Nonparametric estimation of average treatment effects under exogeneity: a review'. *Review of Economics and Statistics,* 86: 4-29.

Imbens, G.W. and J. Angrist (1994). 'Identification and estimation of local average treatment effects'. *Econometrica,* 62: 467-75.

Imbens, G. W. and J. M. Wooldridge (2008). 'Recent developments in the econometrics of program evaluation'. NBER Working Paper 14251.

Indurkhya, A., N. Mitra and D. Schrag (2006). 'Using propensity scores to estimate the cost-effectiveness of medical therapies.' *Statistics in Medicine*, 25: 1561-1576.

Jensen, R.T. and K. Richter (2004). 'The health implications of social security failure: evidence from the Russian pension crisis'. *Journal of Public Economics,* 88: 209-36.

Jones, A.M. (2009). 'Panel data methods and applications to health economics'. In T. C. Mills and K. Patterson (eds.) *Palgrave Handbook of Econometrics. Volume II Applied Econometrics*. Basingstoke: Palgrave MacMillan.

Jones, A.M., N. Rice, T. Bago D'uva and S. Balia (2007). *Applied Health Economics*. London: Routledge.

King, G. and L. Zeng (2007). 'When can history be our guide? The pitfalls of counterfactual inference'. *International Studies Quarterly,* 51: 183-210.

Lalonde, R. J. (1986). 'Evaluating the Econometric Evaluations of Training Programs with Experimental Data'. *American Economic Review,* 76: 604-20.

Lee, M-J. (2005). *Micro-econometrics for policy, program, and treatment effects*. Oxford: Oxford University Press.

Lindahl, M. (2005). 'Estimating the effect of income on health and mortality using lottery prizes as exogenous source of variation in income'. *Journal of Human Resources,* 40: 144-68.

Lleras-Muney, A. (2005). 'The relationship between education and adult mortality in the United States'. *Review of Economic Studies,* 72: 189-221.

Manski, C. F. (1990). 'Nonparametric bounds on treatment effects'. *American Economic Review Papers and Proceedings,* 80: 319-23.

Manski, C. F. (2005). *Social Choice with Partial Knowledge of Treatment Response*. Princeton: Princeton University Press.

Meyer, B. (1995). 'Natural and quasi-experiments in economics'. *Journal of Business and Economic Statistics,* 12: 151-61.

Miguel, E. and M. Kremer (2004). 'Worms: Identifying impacts on education and health in the presence of treatment externalities'. *Econometrica,* 72: 159-217.

Nichols, A. (2007). 'Causal inference with observational data'. *The Stata Journal*, 7: 507-41.

Pagan, A. R. and D. Hall (1983). 'Diagnostic tests as residual analysis'. *Econometric Reviews*, 2: 159-218.

Pesaran, M. H. and L. W. Taylor (1999). 'Diagnostics for IV regressions'. *Oxford Bulletin of Economics and Statistics*, 61: 255-81.

Pop-Eleches, C. (2006). 'The impact of an abortion ban on socioeconomic outcomes of children: evidence from Romania'. *Journal of Political Economy,* 114: 744-73.

Ramsey, J. B. (1969). 'Tests for specification errors in classical linear least squares regression analysis'. *Journal of the Royal Statistical Society Series B,* 31: 350-71.

Robins, J. M. and Y. Ritov (1997). 'Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models'. *Statistics in Medicine,* 16: 285-319.

Robins, J. M. and A. Rotnitzky (1995). 'Semiparametric efficiency in multivariate regression models with missing data'. *Journal of the American Statistical Association,* 90: 122-29.

Rosenbaum, P. R. and D. B. Rubin (1983). 'The central role of the propensity score in observational studies for causal effects'. *Biometrika,* 70: 41-55.

Rosenbaum, P. R. and D. B. Rubin (1984). 'Reducing the bias in observational studies using subclassification on the propensity score'. *Journal of the American Statistical Association,* 79: 516-24.

Roy, A. (1951). 'Some thoughts on the distribution of earnings'. *Oxford Economic Papers,* 3: 135-46.

Rubin, D. B. (1973a). 'Matching to remove bias in observational studies'. *Biometrics,* 29: 159-83.

Rubin, D. B. (1973b). 'The use of matched sampling and regression adjustments to remove bias in observational studies'. *Biometrics,* 29: 185-203.

Rubin, D. B. (1974). 'Estimating causal effects of treatments in randomized and non-randomized studies'. *Journal of Eductational Psychology,* 66: 688-701.

Rubin, D. B. (1979). 'Using multivariate matche sampling and regression adjustment to control bias in observational studies'. *Journal of the American Statistical Association,* 74: 318-28.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Sargan, J. D. (1958). 'The estimation of economic relationships using instrumental variables'. *Econometrica*, 26: 393-415.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Spadaro, A. (2005). 'Micro-simulation and normative policy evaluation: an application to some EU tax benefit systems'. *Journal of Public Economic Theory*, 7: 593-622.

Staiger, D. and J.H. Stock (1997). 'Instrumental variables regression with weak instruments'. *Econometrica*, 65: 557-86.

Stillman, S., D. McKenzie and J. Gibson (2009). 'Migration and mental health: Evidence from a natural experiment'. *Journal of Health Economics*, 28: 677-87.

Stock, J. H. and M. Yogo (2002). *Testing for weak instruments in linear IV regression*. NBER Technical Working Paper No 284. http: www.nber.org/papers/T0284.

Todd, P. E. (2008). 'Evaluating social programs with endogenous program placement and selection of the treated'. In T. P. Schultz and J. A. Strauss (eds.) *Handbook of Development Economics Volume 4*. Amsterdam: Elsevier.

Todd, P.E. and K.I. Wolpin (2009) 'Ex ante evaluation of social programs'. *Annales d'Economie et de Statistiques*, in press.

Van Den Berg, G. J., M. Lindeboom and F. Portrait (2006). 'Economic conditions early in life and individual mortality'. *The American Economic Review,* 96: 290-302.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2005). 'Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity'. *Journal of Applied Econometrics,* 20: 39-54.

Zimmer, D. M. and P. K. Trivedi (2006). 'Using trivariate copulas to model sample selection and treatment effects: application to family health care demand'. *Journal of Business & Economic Statistics,* 24: 63-76.

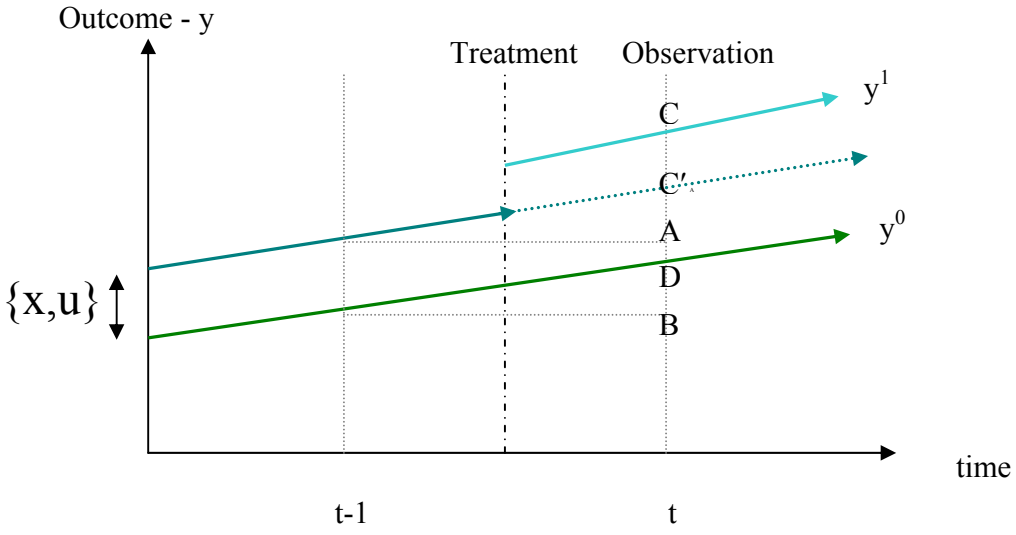

Figure 1: A simple example of potential outcomes.