From Micro Data to Causality: Forty Years of Empirical Labor Economics

Bas van der Klaauw*

February 26, 2014

Abstract

This overview describes the development of methods for empirical research in the field of labor economics during the past four decades. This period is characterized by the use of micro data to answer policy relevant research question. Prominent in the literature is the search for exogenous variation in treatment assignment which can be exploited to estimate causal effects. With the increased availability of detailed administrative data empirical labor economics and more generally empirical microeconomics will become an even more prominent field in economics research.

Keywords: treatment effects, endogeneity, selection, experiments, labor market behavior, microeconometrics.

JEL-code: C21, C26, C93, J68.

*VU University Amsterdam, Tinbergen Institute.
Address: Department of Economics, VU University Amsterdam, De Boelelaan 1105, NL–1081 HV Amsterdam, The Netherlands. E-mail: b.vander.klaauw@vu.nl

Thanks to Joshua Angrist for many useful comments during his discussion at the EALE 2013 meeting in Turin, and to Joop Hartog for carefully commenting on the paper.

1 Introduction

Research in labor economics is closely related to policy, and, therefore, labor economists often aim to provide evidence on the causal effect of either a policy intervention (e.g. minimum wage) or an individual choice variable (e.g. education, fertility, child care) on labor market outcomes. During the last decades, labor economists have been very prominent in developing microeconometric methods for estimating such causal effects. This has had substantial spillovers to other fields in economics, now often using similar methods as used in empirical labor economics.

This paper describes the development of methods for empirical research in the field of labor economics during the past few decades. The focus is on microeconometrics used for analyzing labor market behavior, which gained popularity during the early 1970s when labor economists realized that administrative micro data are essential to answer policy relevant research questions (Ashenfelter (1974)). This intensified the collection and use of detailed data at the individual level.

Already in the 1970s it was realized that standard regression methods, such as ordinary linear squares, probit and logit, suffer from endogeneity and selection problems (Heckman (1974)). This causes estimators for policy relevant parameters to become inconsistent, which triggered the development of econometric approaches correcting for these sources of inconsistencies. Heckman (1979) introduced methods for dealing with only observing outcomes for a selective subsample.

LaLonde (1986) showed empirically that endogeneity can be a major problem in microeconometric research. He compared the results from a randomized experiment with a series of non-experimental estimates for the effects of an employment program for disadvantaged workers. The non-experimental results are often different than those from a randomized experiment, implying that controlling for a limited set of observed individual characteristics is not sufficient to deal with the endogeneity problem.

The insight from LaLonde (1986) has been very influential on empirical research in labor economics. In general, researchers started to think more carefully about endogeneity and the identification of their parameters of interest. Since then, it has often been argued that randomized experiments are ideal when studying causal effects. However, in many economic settings randomized experiments are

very difficult to implement. For example, randomly assigning years of education or wages to individuals is often infeasible.[1] Therefore, since the late 1980s researchers started exploiting natural experiments (e.g. Angrist (1990), Angrist and Krueger (1991), Card (1990), and Card and Krueger (1994) for early contributions).

The idea of a natural experiment is to find exogenous variation in some treatment variable when estimating the effect of this treatment on individual outcomes.[2] Often the exogenous variation comes from institutional rules causing that (almost) identical individuals are exposed to different treatment regimes. The use of natural experiments when estimating causal effects induced a change in empirical research, which until the 1980s had mainly focused on developing microeconometric techniques dealing with selectivity and endogeneity. Microeconometricians started using methods such as instrumental variables and difference-in-differences much more frequently. In the late 1990s, economists also adopted regression discontinuity estimation as method for causal inference (e.g. Angrist and Lavy (1999) and Van der Klaauw (2002)). Regression discontinuity estimation was already discussed by Thistlewaite and Campbell (1960) in the educational sciences.

The use of natural experiments changed data requirements for empirical research. These methods require detailed information about both the cause of the exogenous variation as well as a sufficient number of individuals at the margin of the natural experiment. For example, Lalive (2008) who uses a regression-discontinuity design to study the effect of extended benefits entitlement on job finding, requires exact information on the age at the moment of becoming unemployed and a sufficiently large number of individuals who entered unemployment around the age of 50. Surveys often do not satisfy these data requirements. Maybe because age and the start of the unemployment period are imprecisely observed, or because there are only very few people in the survey who entered unemployment around the age of 50. Administrative data do not have such problems, which can explain the increased popularity of using administrative data in microeconomic research. Alternatively, researchers can collect their own data focussing on the relevant population and paying particular attention to the relevant variables. This type of data collection is often done in combination with a field experiment. Card

[1] Gneezy and List (2006) discuss the results of a field experiment with randomly assigned wages.

[2] Treatment is very broadly defined. It can also be an individual choice variable such as years of education or an endogenous variable such as wages.

et al. (2011) document the increased popularity of field experiments in economic research, particularly since the mid 1990s.

Natural experiments as approach to estimating causal effects have also been criticized. Rosenzweig and Wolpin (2000) discuss behavioral responses of individuals to the institutional setting and argue that not all natural experiments generate variation that is truly exogenous. Imbens and Angrist (1994) stress that empirical results using instrumental variables should often be interpreted locally. And Hahn et al. (2001) show that regression discontinuity methods provide a treatment effect at the margin of the discontinuity. Heckman and Urzúa (2010) criticize the focus on these local effects. Heckman and Vytlacil (2001) present policy relevant treatment effects, which link marginal treatment effects to an economic meaningful parameter. Chetty (2009) uses a sufficient statistics approach to establish a link between a welfare analysis and reduced-form treatment evaluation.

The remainder of this paper is organized as follows. In Section 2 we discuss two empirical models traditionally used in labor economics, the Mincerian wage equation and neoclassical labor supply model. These models are used to illustrate the failure of straightforward regression using microeconomic data. Section 3 presents the sample selection model as introduced by Heckman (1979) as an illustration for the use of econometric techniques dealing with selection issues. Section 4 provides a discussion of the potential outcomes model and discusses the use of social and field experiments. Next, an overview of natural experiment methods is presented in Section 5. Section 6 relates the treatment evaluation literature more explicitly to labor market behavior and dynamics. Finally, Section 7 provides some concluding remarks.

2 Two traditional labor market models

To illustrate the development of empirical microeconometric research in labor economics we briefly discuss two traditional models. The first is the Mincerian wage equation and the second the neoclassical labor supply model. Traditionally empirical economists used ordinary least squares (OLS) for estimating such labor market models. In both models most likely the classical assumptions for using OLS will be violated, implying that estimators are not consistent.

Human capital theory describes that workers invest in their productivity by

following education or by obtaining work experience. A higher productivity should be reflected in the wage, which provides a (reduced-form) relation between wages and human capital. The most prominent wage equation is provided by Mincer (1974),

$$\log \text{wage}_i = \beta_0 + \beta_1 \text{schooling}_i + \beta_2 \text{experience}_i + \beta_3 \text{experience}_i^2 + U_i$$

The logarithm of the wage of worker i depends on her years of schooling and work experience. For ease of presentation other observed worker characteristics are not mentioned explicitly, but these are often taken into account in an empirical analysis. The disturbance term U_i contains the effects of unobserved characteristics and shocks on wages. The key parameter of interest is β_1, which describes the *returns to education*. This is an important policy parameter since most governments subsidize schooling and impose other regulations such as minimum school leaving ages.

Years of schooling is a choice variable. When individuals make schooling decisions, they can take all relevant heterogeneity into account. More able individuals attend schooling for more years, and ability might affect wages as well. If the econometrician does not observe ability or other relevant individual characteristics, these are included in the error term U_i. In that case years of schooling is an endogeneous variable. This causes that OLS will not provide a consistent and unbiased estimate for the returns to education β_1. A possible solution would be to add many other covariates, which should reduce the omitted variable bias. However, such a kitchen sink approach does not guarantee that a consistent estimator for β_1 will be obtained.

The theory of labor supply is based on traditional neoclassical utility models in which workers face the trade-off between leisure and income. The individual choice variable is how many hours to work. Working more hours increases earnings which can be used for consumption, but it reduces leisure. The key element in these models are hourly wages, which indicate how much additional consumption one hour of leisure is worth. Empirical research focuses on how hours of work is affected by the hourly wage. Often the reduced-form labor supply model is used

(e.g. Heckman (1974)),

$$\text{hours of work}_i = \beta_0 + \beta_1 \log \text{wage}_i + \beta_2 \text{other income}_i + U_i$$

Other income includes all income of the individual which is not earned within the labor market, for example, social insurance benefits and subsidies. The most important policy parameter is β_1, which describes the curvature of the labor supply function. Because taxes affect the after-tax hourly wage, the parameter β_1 informs policy makers how labor supply changes when modifying the tax system.

When estimating the labor supply model, there are two major complications. First, wages are likely to be endogenous, i.e. there may be unobserved individual characteristics which affect both the individual's wage and preferences for working included in the error term U_i. Second, there are individuals who do not work, and for those individuals hourly wages remain unobserved. Nonparticipation in the labor market can be selective. For example, the choice to participate in the labor market may be related to both the hourly wage and preferences for working. These complications cause that estimating the labor supply equation using OLS may yield an inconsistent and biased estimate for β_1.

3 Selection models

Before discussing the issue of endogeneity, we first pay attention to sample selection, which implies that outcomes are only observed for a (nonrandom) part of the sample. As discussed above this is likely to be present in the labor supply model, where labor supply and wages are only observed for employed workers. To deal with problems arising due to selectively observing outcomes, Heckman (1979) introduced the sample selection model. Consider a sample of N individuals, and for each individual $i = 1 \ldots, N$ the relevant outcome is described by Y_i^*. The researcher is primarily interested in how outcomes are related to exogenous variables X_i, which is described by a regression equation,

$$Y_i^* = \beta_0 + \beta_1 X_i + U_i$$

The key variable of interest in this model is β_1.

However, the outcomes Y_i^* are latent, because they are not observed for a (possibly selective) subsample. There may be both observed and unobserved individual characteristics affecting both outcomes and whether or not the outcome will be observed. If unobserved characteristics are important, the selection process is related to the dependent variable. For example, years of education (observed) and motivation (unobserved) affect both wages (outcomes) and labor force participation (selection). Therefore, the sample selection model has a separate selection equation, which indicates if the outcome variable is observed

$$I_i^* = \gamma_0 + \gamma_1 Z_i + V_i$$

The indicator I_i takes value 1 if $I_i^* > 0$, and value 0 if $I_i^* \leq 0$. If the indicator I_i equals 1, the outcome is observed in the data. The observed outcome is thus given by

$$Y_i = \begin{cases} Y_i^* & \text{if} \quad I_i = 1 \\ \text{missing} & \text{if} \quad I_i = 0 \end{cases}$$

Within the subsample of individuals for which the outcome is observed ($I_i = 1$), the expected value of Y_i conditional on X_i and Z_i equals

$$\begin{aligned} \mathrm{E}[Y_i | I_i = 1, X_i, Z_i] &= \beta_0 + \beta_1 X_i + \mathrm{E}[U_i | I_i = 1, X_i, Z_i] \\ &= \beta_0 + \beta_1 X_i + \mathrm{E}[U_i | V_i > -\gamma_0 - \gamma_1 Z_i, X_i, Z_i] \end{aligned}$$

This expression provides insight in the causes of selection bias when applying OLS. Let us assume that the covariates X_i and Z_i are exogenous ($\mathrm{E}[U_i | X_i, Z_i] = 0$). OLS only provides consistent estimators either if U_i and V_i are independent (sampling of outcomes is random), or if X_i and Z_i are uncorrelated. The latter refers to the case where the sampling of outcomes is determined by other covariates than those affecting outcomes, which are also uncorrelated to each other. If neither of the two conditions is satisfied, there is so-called *selection bias* when applying OLS.

In general, there are two ways to deal with sample selection. First, one can impose a functional form on the joint distribution of both error terms U_i and V_i. Such an approach is not advisable when X_i and Z_i contain the same variables. The estimation results are not very robust against departures from the imposed distribution. An alternative two-step approach has been proposed by Heckman

(1979).

In the first step, Heckman (1979) uses probit to estimate the parameters γ_0 and γ_1 of the binary choice model. When U_i and V_i follow a bivariate normal distribution function with correlation ρ, then $\mathrm{E}[U_i|V_i > -\gamma_0 - \gamma_1 Z_i, X_i, Z_i]$ equals $\rho\sigma\frac{\phi(\hat{\gamma}_0+\hat{\gamma}_1 Z_i)}{\Phi(\hat{\gamma}_0+\hat{\gamma}_1 Z_i)}$ (where σ^2 is the variance of U_i). This selection correction term is used to specify the regression equation

$$Y_i = \beta_0 + \beta_1 X_i + \rho\sigma\frac{\phi(\hat{\gamma}_0 + \hat{\gamma}_1 Z_i)}{\Phi(\hat{\gamma}_0 + \hat{\gamma}_1 Z_i)} + U_i^*$$

which can be estimated using OLS only on the (selected) sample for which the outcome Y_i is observed. The estimated inverse Mills ratio $\frac{\phi(\hat{\gamma}_0+\hat{\gamma}_1 Z_i)}{\Phi(\hat{\gamma}_0+\hat{\gamma}_1 Z_i)}$ is simply treated as regressor and $\rho\sigma$ as unknown regressor.

This regression equation shows that if the covariates Z_i in the selection equation are the same as the covariates X_i in the regression equation, then the identification hinges on the nonlinearity of the inverse Mills ratio. Locally the inverse Mills ratio is quite linear. So, if there is not much variation in $\gamma_0 + \gamma_1 Z_i$, identification is problematic. This stresses the importance of using an exclusion restriction, i.e. a variable included in Z_i and excluded from X_i.

Blundell et al. (1998) apply the sample selection model to estimating a model for labor supply decisions of women. They use changes in tax rates as exclusion restriction. The idea is that changes in tax rates exogenously change incentives for participating in the labor market. They also generate exogenous variation in the after-tax hourly wage, which is used to take account of endogeneity in wages when estimating the labor supply model. In the actual estimating Blundell et al. (1998) apply the grouping estimator, which uses a large set of exclusion restrictions. Their results do not show significant selectivity in labor force participation. Bosch and Van der Klaauw (2012) provide a similar analysis for female labor supply in the Netherlands. They find evidence in favor of selective labor force participation, i.e. the coefficient of the inverse Mills ratio in the hours of work equation is significant.

4 Counterfactuals

4.1 Potential outcomes model

Labor economists are often interest in the causal effect of a specific (choice) variable on future outcomes. For example, when analyzing how the wage rate affects working hours, and when evaluating how years of education affects earnings. The choice variable can also be whether or not someone participates in a labor market program or it can describe the health status of an individual.

Early empirical studies on causal effects focused on training programs (Ashenfelter (1974) and Ashenfelter and Card (1985)). Whereas Ashenfelter (1974) combined administrative micro data on training participation with labor market outcomes, Ashenfelter and Card (1985) focused on the actual research question.[3] They framed the research question in terms of what could be expected from training participants in a counterfactual world in which the participants would not receive training. This strongly relates to the potential outcomes framework developed by Rubin (1974), which relates back to Neyman (1923).

The potential outcomes model provides a general framework for *ex-post* treatment evaluation. Each individual has two potential outcomes, Y_{1i}^* with treatment and Y_{0i}^* without treatment. The difference between the two potential outcomes is the causal effect Δ_i for individual i of participating in the treatment

$$\Delta_i = Y_{1i}^* - Y_{0i}^*$$

Since each individual is either in the treated or the untreated state, only one potential outcome can be observed. The unobserved outcome is the counterfactual outcome. Therefore, Δ_i is always an unobserved random variable. This is what Holland (1986) refers to as *the fundamental problem of causal inference*.

The causal effects Δ_i can differ between individuals. To summarize the individual causal effects, an often considered *parameter of interest* is the *average*

[3]Joshua Angrist stressed during his discussing at the EALE meeting 2013 that both data sets used by Ashenfelter (1974) and Ashenfelter and Card (1985) were later embargoed, which forced researchers to rely on surveys for which large amounts of research funding were made available. Nowadays, there is worldwide a trend of making detailed administrative data available for researchers.

treatment effect (ATE)

$$ATE = \mathrm{E}[\Delta] = \mathrm{E}[Y_1^* - Y_0^*] = \mathrm{E}[Y_1^*] - \mathrm{E}[Y_0^*]$$

where $\mathrm{E}[Y_1^*]$ and $\mathrm{E}[Y_0^*]$ refer to the average expected potential outcomes in the population of interest. This average treatment effect describes how average outcomes in the population change if the full population receives treatment compared to not being treated. Defining the average treatment effect requires making the assumption that the potential outcomes of each individual are not affected by the actual assignment of treatment within the population. This is what Cox (1958) refers to as the stable unit treatment value assumption (SUTVA).

If treatment is usually imposed on a selective group of individuals, the average treatment effect may not be the most informative parameter of interest. It may be more useful to evaluate the effects of treatment only for those individuals who are exposed to the treatment. Define D_i as an indicator for receiving treatment ($D_i = 1$) or not ($D_i = 0$). The alternative parameter of interest is the *average treatment effect on the treated (ATET)*

$$ATET = \mathrm{E}[\Delta|D = 1] = \mathrm{E}[Y_1^* - Y_0^*|D = 1] = \mathrm{E}[Y_1^*|D = 1] - \mathrm{E}[Y_0^*|D = 1]$$

They key empirical problem is that treatment participation is often not independent of the potential outcomes, because individuals *self-select* into treatment. This is the case when individuals make schooling decisions or when casemanagers assign unemployed worker to a job search assistance program. Individuals with a large treatment effect Δ_i may be more likely to receive treatment.

The potential outcomes model remained relatively unnoticed in economics until Heckman (1990) and Manski (1990). Instead microeconometrics focused directly on the observed outcome $Y_i = D_i Y_{1i}^* + (1 - D_i)Y_{0i}^*$ and developed methods for dealing with endogeneity of treatment assignment D_i (e.g. Heckman and Robb Jr. (1985)). This differed from the statistical literature, which often imposes a conditional independence assumption to deal with selective treatment participation (Rosenbaum and Rubin (1983)). The idea of this conditional independence assumption is that observing a set of individual characteristics X_i is sufficient to ensure that treatment participation is independent of potential outcomes. Dehejia

and Wahba (1999) and Heckman et al. (1997) show that if this conditional independence assumption holds, propensity score methods can relatively easily deal with the selection of treatment participants.

Heckman and Honoré (1990) considered an alternative selection rule into treatment, which is related to the Roy (1951) model. This model assumes that individuals self select in the treatment status which gives the most favorable outcome. If a higher value of the outcome variable relates to a better outcome, individual i decides to participate in treatment if $Y_{1i}^* > Y_{0i}^*$. This selection rule is consistent with the behavior of a utility maximizing economic agent, and, therefore, links the potential outcomes model to microeconomic theory and structural econometrics. However, Heckman and Honoré (1990) show that imposing this selection rule is not sufficient to nonparametrically identify the potential outcomes model using nonexperimental data. Identification requires either distributional assumptions about the joint distribution of potential outcomes or sufficient exogenous variation in potential outcomes due to differences in observed individual characteristics.

4.2 Social experiments

The key problem in analyzing the potential outcomes model is that individuals self select into treatment. Therefore, a comparison of outcomes between individuals in the treatment and control group does not provide a proper estimate of the average treatment effect. In many applications this is also likely to be the case after controlling for differences in observed individual characteristics. Often, it is argued that randomly assigning treatment within the population of interest is the golden standard of treatment evaluation. Such social experiments ensure that treatment assignment is independent of potential outcomes. Because the compositions of the treatment and control group are similar, differences in outcomes between both groups can only be due to the treatment.

LaLonde (1986) proved the value of a social experiment. He evaluated an employment program for disadvantaged workers using data from a social experiment. Next, he showed that various microeconometric methods using nonexperimental data could not replicate the results from the social experiment. However, for a long time social experiments have been much less common in economics than randomized experiments in other sciences. Maybe the first social experiment is a negative

tax experiment in the late 1960s in several cities in New Jersey and Pennsylvania described in Ross (1970).

Four other large scale social experiments have been extensively studied in the economics literature. These are the Rand health insurance experiment and the Star class size experiment in the US, the Canadian self-sufficiency project and Progresa in Mexico. The Rand health insurance experiment randomly assigned various health insurance plans to previously uninsured individuals. This experiment is used to show the presence of moral hazard, i.e. providing more extensive insurance coverage increased health care use (Manning et al. (1987)). The Star class size experiment randomized pupils in classes with different sizes. Krueger (1999) showed that pupils in smaller classes perform better on standardized tests. Card and Robins (1998) studied the Canadian self-sufficiency project, which provides earnings subsidies for long-term welfare recipients who find work for at least thirty hours earning at least the minimum wage. They find that this program significantly increased labor market attachment and reduced welfare participation. However, Card and Hyslop (2005) find that these effects disappeared at the moment entitlement to the subsidies ended. The Mexican Progresa program offered grants to women from poor household when their children attended school and preventive health measures were taken. Schultz (2004) provided an early evaluation of Progresa and showed the effectiveness of this grant scheme.

Randomized experiments have the advantage that they allow for making causal inference without making functional-form or behavioral assumptions. Therefore, the results are convincing and due to the straightforward design also easy to understand for, for example, policymakers. The latter is illustrated by Progresa. After the initial evaluation by the International Food Policy Institute, the grant program was scaled up in Mexico and similar programs have been implemented in other countries such as Honduras, Nicaragua, Ecuador, Brazil. Evaluations in these countries have shown similar results on the effectiveness.

4.3 Nonparametric bounds

Without data from a randomized experiment, additional assumptions to identify the treatment effect are necessary. The credibility of the inference decreases with the strength of the assumptions, which Manski (2003) refers to as *the law of de-*

creasing credibility. Manski (2003) discusses partial identification of treatment effects. If the support of the potential outcomes is bounded, an identification region for the treatment effect can be constructed. For example, when the outcome variable is binary, without making any assumptions the identification region is given by

$$\begin{aligned}\Pr(Y=1|D=1)\Pr(D=1)-\Pr(Y=1|D=0)\Pr(D=0)-\Pr(D=1)\\ \leq \mathrm{E}[Y_1^*]-\mathrm{E}[Y_0^*]\leq\\ \Pr(Y=1|D=1)\Pr(D=1)+\Pr(D=0)-\Pr(Y=1|D=0)\Pr(D=0)\end{aligned}$$

Since $\Pr(D=1)+\Pr(D=0)=1$, the wideness of the no-assumption bounds is always one. Without any data the identification region is given by $-1 \leq \mathrm{E}[Y_1^*]-\mathrm{E}[Y_0^*] \leq 1$. So data reduce the identification region to half its logical range.

Also when outcome variables are not binary, without making any assumptions the identification region is often wide. The identification region can be narrowed by imposing additional assumptions which are often weaker and maybe more credible than those to achieve point identification. For example, imposing the Roy assumption that individuals choose the treatment status with the most favorable outcome reduces the identification region for binary outcomes to

$$-\Pr(Y=1|D=0)\Pr(D=0)\leq \mathrm{E}[Y_1^*]-\mathrm{E}[Y_0^*]\leq \Pr(Y=1|D=1)\Pr(D=1)$$

Because the lower bound is non-positive and the upper bound is non-negative, a zero average treatment effect can never be ruled out. Furthermore, this identification region confirms that without any further assumptions, imposing the Roy assumption on treatment selection is not sufficient to achieve point identification.

5 Natural experiments

Experiments have become more popular in economics in the past decade (Card et al. (2011)). These are often field experiments which have a much smaller scale than the social experiments discussed above. But randomized experiments remain relatively rare in economics. As an alternative in the past 25 years economists have used natural experiment in empirical research. Many of the associated methods

have been introduced by labor economists in the economics literature (e.g. Angrist (1990), Angrist and Krueger (1991), Ashenfelter and Card (1985), Van der Klaauw (2002)). The success factor of any of these methods is finding exogenous variation in the assignment of treatment. However, economists were not the first to exploit natural experiments, Campbell (1969) discusses nonexperiment methods in treatment evaluation using institutional reforms in education.

What the appropriate method is depends on the treatment assignment mechanism and the data availability. DiNardo and Lee (2011) summarize this by providing three criteria for judging the quality of methods for empirical treatment evaluation. First, the method should provide an appropriate description of the treatment assignment mechanism. The researcher should have detailed insight in how in practice treatment is assigned to individuals. Second, the method should be consistent with a wide class of behavioral models, which implies that parametric or functional-form assumptions should be limited as much as possible. And third, the method should yield testable implications, such that using the available data the validity of the method can be tested.

5.1 Difference-in-difference

Ashenfelter and Card (1985) used difference-in-difference estimation to estimate the effects of a training program on earnings. Difference-in-difference estimation requires that for the treatment group and the control group outcomes are observed both before and after the moment of treatment. Difference-in-difference is thus a panel data method and can be written as a regression model for the outcome of individual i at time t

$$Y_{it} = \alpha_i + \delta D_{it} + \eta_t + U_{it}$$

The α-parameters describe individual fixed effects and the η-parameters time fixed effects. Since the treatment indicator D_{it} denotes if individual i was treated at time period t, the parameter of interest δ describes the average treatment effect on the treated (ATET). The key identifying assumption is that the treatment and control group are exposed to the same time trend.

A famous example of violation of this common trend assumption is the Ashenfelter dip, which describes a drop in average outcomes of the treatment group prior to the start of treatment. Ashenfelter (1978) noted this pre-programme dip

when showing that the earnings of participants in a training programme reduced substantially just prior to entering the program. This dip in outcomes has also been found in subsequent studies on active labor market programs (e.g. Heckman and Smith (1999)).

The most prominent study using difference-in-difference estimation is the minimum wage study by Card and Krueger (1994). In April 1992 the minimum wage in New Jersey was increased from \$4.25 to \$5.05, while it remained \$4.25 in Pennsylvania. Card and Krueger (1994) collected data on employment in fast-food restaurants in both states both before and after the minimum wage increase. The fast-food sector is characterized by having many minimum-wage jobs. Whereas in New Jersey employment slightly increased around the minimum wage increase, it reduced in Pennsylvania. From this Card and Krueger (1994) concluded that increasing the minimum wage increases employment. This finding contradicts the conventional wisdom from competitive models for the labor market.

Eissa and Liebman (1996) used a difference-in-difference model to study the labor market effects of providing earned income tax credits to low-income workers. They exploit a fiscal reform in the US in 1986, which substantially increased this earned income tax credit for workers with children. The treatment group are single women with children and the control group single women without children. Eissa and Liebman (1996) report that the labor force participation rate of women with children increased significantly, while for women without children it remained constant. From which they conclude that introducing the earned income tax credit increased female labor force participation.

The outcomes of individuals within groups may be correlated to each other, which may affect the precision of estimators. Moulton (1986) provides formulas for computing standard errors when observations are clustered. Donald and Lang (2007) consider inference based on difference-in-difference models, where there are only a small number of groups with clustered observations. They argue that standard asymptotics cannot be applied in such cases. In the presence of group specific shocks, standard errors around the policy parameters will be estimated too low even if the number of individuals within groups goes to infinity. Bertrand et al. (2004) also consider computing standard errors when the number of groups is limited. They focus on the consequences of autocorrelation in the error terms.

The difference-in-difference approach hinges on the common trend assump-

tion. In some applications it is clear that this assumption is violated, for example, from considering trends in outcomes in the treatment and control group in the pre-treatment period. Abadie et al. (2010) provide a solution to violation of the common trend assumption. They construct synthetic control groups as weighted averages of available control groups. The weights are chosen to minimize the difference in outcomes between the treatment and control group in the pre-treatment periods.

An alternative approach to relax the common trend assumption is taken by Athey and Imbens (2006). Their changes-in-changes approach allows time trends to systematically differ between groups. The key assumption is that regardless of the intervention individuals stay in the same quantile. They use this to reweight individuals in the control group such that in the pre-treatment period the outcome distribution in the control group is identical to that of the treatment group. This approach allows to estimate entire counterfactual distributions and is robust against rescaling outcome variables.

5.2 Instrumental variables

In Subsection 4.1 we saw that observed outcomes can be written in terms of potential outcomes as $Y_i = D_i Y_{1i}^* + (1 - D_i) Y_{0i}^*$. If for the moment we assume a homogeneous treatment effect $\delta = Y_{1i}^* - Y_{0i}^*$ and we rewrite the potential untreated outcome as $Y_{0i}^* = \alpha + U_i$, then we obtain the regression equation

$$Y_i = \alpha + \delta D_i + U_i$$

The treatment assignment indicator D_i is endogenous when treatment is assigned selectively. Instrumental variable methods can account for endogeneity of a regressor. Early and important contributions to the use of instrumental variable methods in empirical (labor) economics were made by Angrist (1990) and Angrist and Krueger (1991).

The idea is to find an instrumental variable Z_i which is correlated to treatment assignment D_i, but uncorrelated to the error term U_i. In that case the empirical framework can be extended by the first-stage equation

$$D_i = \gamma_0 + \gamma_1 Z_i + V_i$$

Instrumental variables or two-stage least squares estimation are applied to estimate the parameter of interest δ. Angrist (1990) used instrumental variable methods to estimate the effect of military service during the Vietnam war on future earnings. He exploited that drafts for military service were based on lottery numbers. A low lottery number increased the risk of military service, but since lottery numbers were randomly assigned these were supposed to be uncorrelated to the error terms.

Lottery numbers are usually considered to be ideal instrumental variables (see for an example also Ketel et al. (2013) who study admission lotteries for medical schools). More often the institutional setting is exploited to generate instrumental variables. Angrist and Krueger (1991) used quarter of birth as instrumental variable for years of schooling when estimating the returns to schooling. The US educational system caused that individuals born in the fourth quarter of the year have more compulsory schooling than individuals born in the first quarter.

Krueger (1999) uses instrumental variable methods to deal with noncompliance in the Star class size experiment. In this experiment compliance to the initial random assignment of pupils in classes was not perfect. If the noncompliance is related to potential outcomes, the actual assignment is endogenous. Krueger (1999) used the initial assignment as instrumental variable for the actual assignment. Because both the instrumental variable and the endogenous regressor are binary, the Wald estimand can be used for the parameter of interest

$$\delta = \frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}$$

If treatment effects are homogenous, it is obvious that the estimate equals the average treatment effect as well as the average treatment effect on the treated. In case of heterogenous treatment effects, the interpretation of the estimated treatment effect is more complicated. Imbens and Angrist (1994) show that under a monotonicity assumption instrumental variables methods estimate a local average treatment effect (LATE). Let $D_i(z)$ denote the treatment assignment of individual i if $Z_i = z$. The monotonicity assumption implies that changing the value of the instrumental variable can only affect the treatment assignment in one direction,

e.g. $D_i(1) \geq D_i(0)$. In this binary case the local average treatment effect equals

$$LATE = \mathrm{E}[Y_1^* - Y_0^* | D(1) - D(0) = 1] = \frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}$$

The local average treatment effect describes the average treatment effect for only those individuals who switch from untreated ($D = 0$) to treated ($D = 1$) when the value of the instrumental variable switches from $Z = 0$ to $Z = 1$. These individuals are the so-called compliers, but these cannot be identified directly. For each individual we only observe the treatment status given the observed value of the instrumental variable. If we observe for an individual $D_i = 0$ and $Z_i = 0$, this individual is either a complier or never taker.[4] Otherwise, if we observe $D_i = 1$ and $Z_i = 1$, the individual can both be a complier or always taker. However, the fraction of compliers in the population can be determined by $1 - \Pr(D = 1|Z = 0) - \Pr(D = 0|Z = 1)$. A second criticism is that different instrumental variables have different groups of compliers and generate different local average treatment effects.[5]

Bound et al. (1995) reanalyzed the data used by Angrist and Krueger (1991), but they replaced the observed quarter of birth with a randomly drawn quarter of birth. Obviously, the random quarter of birth is not correlated to the observed years of schooling. It is, therefore, a valid but also irrelevant instrumental variable. Bound et al. (1995) show that this causes that the IV estimate for the returns to education converges to the OLS estimate. The finding that when an instrumental variable does not explain much variation in the endogenous regressor the bias of the IV estimator might be very substantial even in very large samples, started the weak instruments literature. Hahn and Hausman (2005) characterize the bias of the IV estimator by

$$\text{Bias IV} \approx \frac{\{\#\text{ instruments}\} \times \rho(U,V) \times (1 - R^2_{\text{partial}})}{\{\#\text{ observations}\} \times R^2_{\text{partial}}}$$

[4]Never takers are individuals who never participate in the treatment regardless of the value of the instrument. On the other hand, always takers will always participate in the treatment. The fraction never takers and always takers equal $\Pr(D = 0|Z = 1)$ and $\Pr(D = 0|Z = 1)$, respectively.

[5]Heckman and Vytlacil (2001) show that the local average treatment effect as well as other population treatment parameters can be written in terms of weighted averages of marginal treatment effects.

where R^2_{partial} denotes the contribution of the instruments to the R^2 in the first-stage, and $\rho(U, V)$ the correlation between the error terms in the first-stage and the second-stage equation. Comparing IV estimation to OLS estimation (with one endogenous regressor) gives

$$\frac{\text{Bias IV}}{\text{Bias OLS}} \approx \frac{\{\#\text{ instruments}\}}{\{\#\text{ observations}\} \times R^2_{\text{partial}}}$$

Staiger and Stock (1997) show in simple model (one endogenous regressor, one instrument, no other regressors), that $\frac{1}{F}$ provides a good approximation for the bias of IV compared to OLS. Where F is the test statistics from an F-test for significance of the instrumental variable in the first-stage regression. Staiger and Yogo (2005) argue that instruments are weak if the IV Bias is more than 10% of the OLS Bias. This implies the nowadays used rule of thump for weak instruments that the F-statistic for significance of the instrumental variables in the first-stage regression should exceed 10. When there a multiple instrumental variables it might be better to use the Gragg and Donald test. Limited information maximum likelihood (LIML) can provide a more robust estimator than IV when having multiple (weak) instruments.

5.3 Regression discontinuity

Regression-discontinuity estimation has been discussed in the educational literature already a number of decades ago by Thistlewaite and Campbell (1960). This method remained for a long time unexplored by economists until Van der Klaauw (2002) exploited a GRE-threshold to estimate the effects of offering scholarships on college enrollment. Angrist and Lavy (1999) used the Israeli Maimonidis' maximum class size rule to show that reducing class sizes improves academic achievement of fourth and fifth graders. Hahn et al. (2001) provide a theoretical discussion of regression-discontinuity estimation in which they introduce the sharp and the fuzzy design.

The idea of regression discontinuity is that when a continuous running or assignment variable S_i crosses a threshold $\bar{S}$, this substantially increases the likelihood that this individual receives treatment. In case of a sharp regression-discontinuity design the likelihood of treatment assignment jumps from zero to one when cross-

ing the threshold, $D_i = I(S_i \geq \bar{S})$. The treatment effect at the threshold $\bar{S}$ is then the marginal treatment effect

$$MTE(\bar{S}) = \lim_{s \downarrow \bar{S}} \mathrm{E}[Y|S = s] - \lim_{s \uparrow \bar{S}} \mathrm{E}[Y|S = s]$$

In this sharp design there is locally (for S close to $\bar{S}$) an experiment even though the assignment variable S can also directly affect outcomes Y (e.g. DiNardo and Lee (2011)). The key identifying assumption is that the potential outcomes Y_{0i}^* and Y_{1i}^* are continuous in the assignment variable S_i at $\bar{S}$ (Hahn et al. (2001) and Imbens and Lemieux (2008)).

This sharp regression-discontinuity design was adopted by Lalive (2008), when estimating the effect of extending the unemployment insurance entitlement period on the unemployment duration. He studied Austria, where in June 1988 for individuals who were above age 50 when entering unemployment the benefits entitlement period was extended from 30 to 209 weeks. A similar design is exploited by Schmieder et al. (2012). They focus on an extension of the entitlement period to unemployment insurance benefits for workers above age 42 in Germany. Both studies find that the longer benefits entitlement period reduces job finding rates.

Institutions not always provide such a sharp discontinuity. Angrist and Lavy (1999) and Van der Klaauw (2002) report noncompliance to the threshold. Campbell (1969) refers to this case as fuzzy regression discontinuity. This design implies that the probability of receiving treatment is discontinuous at $\bar{S}$,

$$\lim_{s \downarrow \bar{S}} \Pr(D = 1|S = s) \neq \lim_{s \uparrow \bar{S}} \Pr(D = 1|S = s)$$

Van der Klaauw (2002) suggested to use an indicator for crossing the threshold ($I(S_i \geq \bar{S})$) as locally valid instrumental variable for treatment assignment in the fuzzy regression discontinuity design. This approach is similar to dealing within randomized experiments with noncompliance (e.g. Krueger (1999)) and implies the estimand for the marginal treatment effect

$$MTE(\bar{S}) = \frac{\lim_{s \downarrow \bar{S}} \mathrm{E}[Y|S = s] - \lim_{s \uparrow \bar{S}} \mathrm{E}[Y|S = s]}{\lim_{s \downarrow \bar{S}} \Pr(D = 1|S = s) - \lim_{s \uparrow \bar{S}} \Pr(D = 1|S = s)}$$

Because this is an instrumental variable method, the interpretation is similar to

that of the local average treatment effect. The fuzzy regression discontinuity estimation estimates the average treatment effect for the subpopulation of individuals who change from non-treated to treated when crossing the threshold $\bar{S}$ at this threshold.

Marginal treatment effects are defined in the neighborhood of the threshold $\bar{S}$. In many application, there are not enough observations close to this threshold to estimate the marginal treatment effects nonparametrically. Hahn et al. (2001) and Porter (2003) discuss flexible parametric specifications, which allow to use also data further away from the threshold.

The identifying assumption of regression discontinuity estimation is that individuals cannot manipulate their value of the assignment variable, i.e. they cannot sort themselves on the preferred side of the threshold. If the threshold triggers behavioral responses, there will be an excess number of individuals just on one side of the threshold. McCrary (2007) formalized this in testing for continuity of the density of the assignment variable S around the threshold $\bar{S}$. Lee (2008) provides some alternative tests for continuity of variables not affected by the treatment at the threshold.

6 Labor market behavior

The discussion above provides a framework for ex-post treatment evaluation in a static environment.[6] This literature is sometimes criticized for the absence of a direct link to labor market behavior. Reduced-form treatment evaluation may provide the causal effect of an intervention on a number of outcomes, but these different causal effects are not integrated in a welfare analysis. Furthermore, within the labor market there may be interactions between workers causing equilibrium and spillover effects of interventions. And finally, interventions may be applied at different moments in time affecting different groups of workers, and the causal effect may depend on the moment of the intervention. In this section we briefly elaborate on these issues.

[6] Difference-in-difference estimation may be an exception because due to the panel data nature it allows for more dynamics, such as distinguishing between short-run and long-run treatment effects.

6.1 Sufficient statistics

In the economics literature there is an ongoing debate about the (reduced-form) treatment evaluation approach and more structural approaches, which estimate models of individual behavior. The structural approach is more in line with ideas of the Cowles foundation which aim at identifying the economic model generating the data. Obviously a structural approach makes many assumption about functional forms and individual behavior. However, Keane (2010) states that even though at first sight the treatment evaluation literature seems to provide causal effects that are not too sensitive to assumptions, also many behavioral assumptions are made. A similar argument is made by Rosenzweig and Wolpin (2000), who mention that not all variation is truly exogenous. Therefore, Keane (2010) argues that empirical work cannot exist independently of economic theory.

Heckman and Vytlacil (2001) stress that many of the conventional treatment parameters lack a direct link to an interpretable economic framework or, for example, a cost-benefit analysis. Therefore, such treatment parameters may have a limited relevance for a policy analysis. They argue that by using the appropriate weights for their marginal treatment effects, it is possible to construct treatment effects that capture the effects of policy changes. Chetty (2009) argues that a *sufficient statistics* approach combines the advantages from reduced-form treatment evaluation and structural econometrics. This approach has the potential to provide a welfare analysis relying on credible identification.

The sufficient statistics approach derives from economic theory a number of sufficient statistics for a welfare analysis or policy purpose, which can be estimated using the treatment effect methods. For example, Chetty (2006), Chetty (2008), Gruber (1997) and Shimer and Werning (2007) specify formulas for welfare as function of the level of unemployment insurance benefits. Using envelope conditions for optimal behavior, this welfare equation can be expressed in terms of policy variables. Next, the marginal welfare from a change in the policy variable such as increasing the generosity of unemployment insurance benefits can be determined. For different structural models these marginal welfare functions are often similar and functions of only a few statistics, which Chetty (2008) refers to as sufficient statistics. Ideally, these sufficient statistics can be estimated using the treatment evaluation methods discussed above.

According to Chetty (2009) the sufficient statistics approach has three advantages compared to a structural analysis. First, less data and variation in the data are required to estimate the sufficient statistics than to identify all structural parameters. Second, fewer functional form assumption are required to identify the sufficient statistics than for identifying the structural model. And third, the same sufficient statistics may apply to multiple behavioral models, so the sufficient statistics approach might be robust against misspecification of the behaviorial model. However, there are also some disadvantages associated to the sufficient statistics approach. First, each policy question requires different sufficient statistics, which may not be easy to derive. Policy simulations may, therefore, also be easier with a structural model. And second, testing the validity of the underlying behavioral model is not as straightforward in case of a sufficient statistics approach as in case of a structural model.

6.2 Spillover and peer effects

When presenting the potential outcomes framework, the stable unit treatment value assumption was discussed. This assumes that the potential outcomes of individuals are not affected by their treatment status nor the treatment status of other individuals. Imposing this assumption in economics settings is not always straightforward. There may be spillover effects between individuals or equilibrium effects. For example, Miguel and Kremer (2004) find large positive spillover effects of de-worming drugs on schools in Kenya. And Blundell et al. (2003) and Gautier et al. (2012) estimate using difference-in-difference models that equilibrium effects may be substantial when studying the large-scale role-out of active labor market programs. Lise et al. (2004) use a structural approach to show that taking account of equilibrium effects can cause that the results from a cost-benefit analysis reverse. Heckman et al. (1998) find a similar result when studying the effects of tuition fees on college enrollment.

Estimating spillover effects is difficult. In the ideal setting different fully separated local labor market (or schools) are randomly assigned different treatment intensities. Crepon et al. (2013) implement such an experimental design when studying equilibrium effects of an active labor market program in France. The results do not show evidence in favor of equilibrium effects. However, the take-up

in their encouragement design is low and their target population consists of highly educated long-term young unemployed workers, who are only a very small fraction of all job seekers.

Spillover effects are also studied in terms of peer effects. Manski (1993) shows that peer effects are difficult to identify. There can be various reasons why individuals in the same group have similar outcomes. Individuals in the same group can affect each other (endogenous effects), individuals with similar characteristics may sort in the same group (exogenous effects), and individuals in the same group may experience the same group specific shocks (correlated effects).

Angrist (2013) discusses the estimation of peer effects. He shows the problems arising when regressing an individual's outcome on the mean outcome within the group of peers. If the individual's outcome is also included in the mean outcome of peers, Angrist (2013) shows that the coefficient should equal unity by construction. However, leaving out the individual's outcome when computing the group mean does not solve the problem. The individual might be exposed to the same shocks as its peers and the regression coefficient can thus capture correlated shocks as well as true peer effects. Angrist (2013) states that experiments that can manipulate the characteristics of the peers without affecting the individual's characteristic provide the strongest evidence for the existence of spillovers within peer groups.

6.3 Dynamic treatment evaluation

Economic processes happen most often in real time and this is particularly true in labor economics. Also policy interventions happen in real time. For example, not all workers enter a job search assistance program after exactly the same elapsed unemployment duration. Even though there is only a single treatment, the effects may differ between individuals.[7] Not only because individuals are heterogeneous, but also because different individuals may be exposed to the treatment at different time periods. Furthermore, the effects of the program may depend on the moment at which outcomes are observed. Abbring and Heckman (2007) provide a very general potential outcomes framework for treatment evaluation in dynamic settings dealing with these issues. Below, we discuss a slightly simplified framework, which follows Kastoryano and Van der Klaauw (2011).

[7]Ketel et al. (2013) discuss local average treatment effects in a dynamic setting.

The potential outcomes framework can be extended to a dynamic setting. For ease of exposition, we assume that there is only a single treatment, which is provided at most once but at different time periods. Let $Y^*_{1,t}(s;p)$ describe the potential outcome after t periods if an individual received treatment after s periods under policy regime p. We make the explicit distinction between the policy regime p and the moment of the intervention s. This allows to distinguish between the effects of an actual treatment intervention within a particular policy regime, and the effects of changing the policy regime. Most microeconometric evaluations focus on the first effect. For example, Abbring et al. (2005) and Van den Berg et al. (2004) estimate the effect of actually imposing a benefits sanction on the job finding rate of unemployed workers. This is the ex-post effect which is different from the effect which changing the sanction regime may have on the job search behavior of benefits recipients.

Ideally, the potential untreated outcome would be defined as the outcomes in which the individual will never receive treatment $Y^*_{0,t}(g)$ as $\lim_{s\to\infty} Y^*_{1,t}(s;g)$. Since observation periods are limited, this potential outcome can generally not be observed. In addition when studying, for example, training programs for unemployed workers, it remains generally unobserved when the unemployed workers would have enrolled in the training, if the worker left unemployment before having receiving the training (e.g. Ham and LaLonde (1996)). To define counterfactuals Abbring and Van den Berg (2003) make a no-anticipation assumption

$$Y^*_{1,t}(s;g) = Y^*_{1,t}(s';g) \qquad \text{if } s \neq s' \qquad \forall t < s, s'$$

Abbring and Heckman (2007) refer to this assumption as no causal dependence of outcomes on future treatments. Due to this assumption the potential untreated outcomes equal

$$Y^*_{0,t}(g) = Y^*_{1,t}(s;g) \qquad \forall t < s$$

Abbring and Van den Berg (2003) study the case in which the outcome Y_t describes still being in the initial state after t time periods. They build a bivariate hazard rate framework in which the process until leaving the initial state is jointly modeled with the process until entering treatment. Abbring and Van den Berg (2003) show that if both hazard rates follow a mixed proportional structure and the no-anticipation assumption holds, then the ex-post effect of the intervention

on the exit rate until leaving the initial state can be identified. Their framework allows this ex-post effect to depend on the moment of the intervention, the elapsed duration since the intervention and both observed and unobserved characteristics. Using this no-anticipation assumption, the ex-post effect of the intervention can be defined as the effect of the intervention on the treated survivors

$$ATETS(t, s; g) = \mathrm{E}[Y^*_{1,t}(s; g) - Y^*_{0,t}(g)|Y_s = 0, S = s, g]$$

The framework discussed by Abbring and Van den Berg (2003) allows for selection on unobservables, but it makes some strong assumptions. First, the mixed proportional hazard rate specification restricts the functional form, and second, the no-anticipation assumption makes an assumption about behavior or information provided to individuals.[8]

The question arises if both assumptions can be relaxed when assuming that selection is only on observables. Assume that potential outcomes depend on the observables X and some other individual characteristics U, i.e. $Y^*_{1,t}(s, X, U; g)$. The timing of entry into treatment only depends on X, i.e. $S(X; g)$, so that the conditional independence assumption is satisfied. If individuals anticipate treatment participation and change their behavior already prior to the actual intervention, then within the stock of individuals who survive in the initial state for s periods, those who enter treatment at s will have different characteristics than those who enter later. Therefore, in the presence of anticipation, the actual treatment assignment at time s among the survivors at that moment depends on both observed characteristics X and unobserved characteristics U. The standard conditional independence assumption can thus not replace the no-anticipation assumption when studying ex-post effects of treatment. This relates to the biostatistical literature, which often relies on sequential randomization (e.g. Gill and Robins (2001)).

The no-anticipation assumption seems crucial when the interest is in the treatment effect on the treated survivors. Intuitively, if individuals anticipate treatment, then the survivors in the treatment group may be different than survivors in the control group at the moment the treatment group receives treatment. Only when the treatment moment is assigned to all individuals at time period $t = 0$ and

[8] Abbring and Van den Berg (2005) discuss the role of instrumental variables in duration models.

this is also registered for all individuals in the data, the no-anticipation assumption could be avoided. But in that case not only the ex-post effect of the treatment is estimated, but an effect also including anticipation effects such as, for example, threat effects (Black et al. (2003)).

7 Conclusions

Since the early 1970s research in labor economics has changed dramatically. Labor economists started using micro data and empirical research on evaluating public policies became more popular (e.g. Ashenfelter (1974) and Ashenfelter (1978)). Labor economists quickly acknowledged that simple regressions might not produce the causal effect of interest, due to concerns about endogeneity of policy variables or selection issues (e.g. Heckman (1974) and Heckman (1979)). Whereas during the 1970s and 1980s economists focused on econometric models for the observed outcomes, Ashenfelter and Card (1985) framed the research question in terms of counterfactuals relating more to the statistical literature Rubin (1974)).

LaLonde (1986) showed that econometric methods could not always deal with nonrandom assignment of treatment, which stressed the importance of exogenous variation in treatment assignment when evaluating a treatment. Angrist (1990) and Angrist and Krueger (1991) started using exogenous variation induced by institutions to estimate causal effect. This initiated the natural experiments literature, which is nowadays widely spread in labor economics and also has spillovers to other fields.

The major advantage of social, natural and field experiments is that the methodology is relatively simple and also easily understood by policy makers. Furthermore, the focus of these studies is often on topics which have a direct impact on policy. Therefore, labor economics research became more prominent in the public debate. An example is the minimum wage study by Card and Krueger (1994) which not only initiated a debate in the economics literature on the effects of minimum wages, but was also very influential on public policy both in the US and in other countries.

In the past two decades, research using natural and field experiments has gained popularity (Card et al. (2011) show the increased number of studies using field experiments). This literature now produces many treatment effects. The

main debate is about the interpretation of the estimated treatment effects. Natural experiments often provide causal effects which should be interpreted locally. Heckman and Vytlacil (2001) criticize these estimates for the lack of economic interpretation. The recent literature on sufficient statistics aims at bridging the gap between a welfare analysis and the reduced-form treatment evaluation literature (Chetty (2009)).

Surprisingly, two of the seminal papers in the literature on empirical policy evaluation in labor economics were not published in the traditional top-5 journals (Ashenfelter (1978) and Ashenfelter and Card (1985)). This shows that empirical work was not yet taken as serious in the economics literature as nowadays. LaLonde (1986) made a next step in this respect. Since then empirical micro research became a serious field in economics. And this trend is likely to continue because more and more detailed administrative data become available for research in many countries.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: estimating the effect of Californias tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abbring, J. and J. Heckman (2007). Econometric evaluation of social programs, part iii: distributional treatment effects, dynamic treatment effects and dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6B*, Chapter 72, pp. 5145–5303. Elsevier Science, Amsterdam.

Abbring, J. and G. Van den Berg (2003). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–1517.

Abbring, J. and G. Van den Berg (2005). Social experiments and instrumental variables with duration outcomes. *Tinbergen Institute Discussion Paper 2005-047/3.*.

Abbring, J., G. Van den Berg, and J. Van Ours (2005). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *Economic Journal 115*(505), 602–630.

Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administration record. *American Economic Review 80*(3), 313–336.

Angrist, J. (2013). The perils of peer effects. *NBER Working paper 19774*.

Angrist, J. and A. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics 114*(2), 533–575.

Ashenfelter, O. (1974). The effect of manpower training on earnings: preliminary results. In *Proceedings of the 27th Annual Meeting of the Industrial Relations Research Association.*

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics 60*(1), 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*(4), 648–660.

Athey, S. and G. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–497.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust difference-in-difference estimates. *Quarterly Journal of Economics 119*(1), 249–275.

Black, D., J. Smith, M. Berger, and B. Noel (2003). Is the threat of reemployment services more effective than the services themselves? evidence from random assignment in the ui system. *American Economic Review 93*(4), 1313–1327.

Blundell, R., M. Costa Dias, and C. Meghir (2003). The impact of wage subsidies: a general equilibrium approach. *Working Paper*.

Blundell, R., A. Duncan, and C. Meghir (1998). Estimating labor supply responses using tax reforms. *Econometrica 66*(4), 827–861.

Bosch, N. and B. Van der Klaauw (2012). Analyzing female labor supply – evidence from a Dutch tax reform. *Labour Economics 19*(3), 271–280.

Bound, J., D. Jaeger, and R. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association 90*(430), 443–450.

Campbell, D. (1969). Reforms as experiments. *American Psychologist 24*(4), 409–429.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review 43*(2), 245–257.

Card, D., S. DellaVigna, and U. Malmendier (2011). The role of theory in field experiments. *Journal of Economic Perspectives 25*(3), 39–62.

Card, D. and D. Hyslop (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica 73*(6), 1723–1770.

Card, D. and A. Krueger (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review 84*(4), 772–793.

Card, D. and P. Robins (1998). Do financial incentives encourage welfare recipients to work? evidence from a randomized evaluation of the self-sufficiency project. *Research in Labor Economics 17*, 1–56.

Chetty, R. (2006). A general formula for the optimal level of social insurance. *Journal of Public Economics 90*(10–11), 1879–1901.

Chetty, R. (2008). Moral hazard vs. liquidity and optimal unemployment insurance. *Journal of Political Economy 116*(2), 173–234.

Chetty, R. (2009). Sufficient statistics for welfare analysis: a bridge between structural and reduced-form methods. *Annual Review of Economics 1*, 451–488.

Cox, D. (1958). *Planning of experiments*. New York: John Wiley and Sons.

Crepon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics 128*(2), 531–580.

Dehejia, R. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

DiNardo, J. and D. Lee (2011). Program evaluation and research designs. In O. Ashenfelter and D. Card (Eds.), *Handbook in Labor Economics 4A*, Chapter 5, pp. 463–536. Elsevier Science, Amsterdam.

Donald, S. and K. Lang (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics 89*(2), 221–233.

Eissa, N. and J. Liebman (1996). Labor supply responses to the earned income tax credit. *Quarterly Journal of Economics 111*(2), 605–637.

Gautier, P., P. Muller, B. Van der Klaauw, M. Rosholm, and M. Svarer (2012). Estimating equilibrium effects of job search assistance. *IZA Discussion Paper no. 6748*.

Gill, R. and J. Robins (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics 29*(6), 1785–1811.

Gneezy, U. and J. List (2006). Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. *Econometrica 74*(5), 1365–1384.

Gruber, J. (1997). The consumption smooting benefits of unemployment insurance. *American Economic Review 87*(1), 192–205.

Hahn, J. and J. Hausman (2005). Estimation with valid and invalid instruments. *Annales dEconomie et de Statistique 79/80*, 25–57.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Ham, J. and R. LaLonde (1996). The effect of sample selection and initial conditions in duration models: evidence from experimental data on training. *Econometrica 64*(1), 175–205.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heckman, J. (1990). Varieties of selection bias. *American Economic Reviwe, Papers and Proceedings 80*(2), 313–318.

Heckman, J. and B. Honoré (1990). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J., H. Ichimura, and P. Todd (1997). Matching as an econometric estimator: evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654.

Heckman, J., L. Lochner, and C. Taber (1998, May). General-equilibrium treatment effects: A study of tuition policy. *American Economic Review 88*(2), 381–86.

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

Heckman, J. and J. Smith (1999). The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies. *Economic Journal 109*(457), 313–348.

Heckman, J. and S. Urzúa (2010). Comparing IV with structural models: what simple IV can and cannot identify. *Journal of Econometrics 156*(1), 27–37.

Heckman, J. and E. Vytlacil (2001). Policy-relevant treatment effects. *American Economic Review Papers and Proceedings 91*(2), 107–111.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615–635.

Kastoryano, S. and B. Van der Klaauw (2011). Dynamic evaluation of job search assistance'. *Discussion Paper 5424, IZA Bonn*.

Keane, M. (2010). A structural perspective on the experimentalist school. *Journal of Economic Perspectives 24*(2), 47–58.

Ketel, N., E. Leuven, H. Oosterbeek, and B. Van der Klaauw (2013). The returns to medical school in a regulated labor market: evidence from admission lotteries. *Mimeo, Tinbergen Institute Amsterdam*.

Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics 114*(2), 497–532.

Lalive, R. (2008). How do extended benefits affect unemployment duration? a regression discontinuity approach. *Journal of Econometrics 142*(2), 785–806.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*(4), 604–620.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics 142*(2), 675–697.

Lise, J., S. Seitz, and J. A. Smith (2004). Equilibrium policy experiments and the evaluation of social programs. *NBER Working paper 10283*.

Manning, W., J. Newhouse, N. Duan, E. Keeler, and A. Leibowitz (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review 77*(3), 251–277.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings 80*(2), 319–323.

Manski, C. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies 60*(3), 531–542.

Manski, C. (2003). *Partial identification of probability distributions*. Springer, New York.

McCrary, J. (2007). Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics 142*(2), 698–714.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Mincer, J. (1974). *Schooling, Experience and Earnings*. Columbia University Press for National Bureau of Economic Research, New York.

Moulton, B. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics 32*(3), 385–397.

Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society II Supplement 2*, 107–180.

Porter, J. (2003). Estimation in the regression discontinuity model. *University of Wisconsin, Unpublished Manuscript*.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenzweig, M. and K. Wolpin (2000). Natural "natural experiments" in economics. *Journal of Economic Literature 38*(4), 827–874.

Ross, H. (1970). An experimental study of the negative income tax. *PhD thesis MIT*.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Schmieder, J., T. Von Wachter, and S. Bender (2012). The effects of extended unemployment insurance over the business cycle: Evidence from regression discontinuity estimates over 20 years. *The Quarterly Journal of Economics 127*(2), 701–752.

Schultz, T. (2004). School subsidies for the poor: evaluating the mexican progresa poverty program. *Journal of Development Economics 74*(1), 199–250.

Shimer, R. and I. Werning (2007). Reservation wages and unemployment insurance. *Quarterly Journal of Economics 122*(3), 1145–1185.

Staiger, D. and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Staiger, J. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In D. Andrews and J. Stock (Eds.), *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Chapter 5, pp. 80–108. New York: Cambridge University Press.

Thistlewaite, D. and D. Campbell (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology 51*(6), 309–317.

Van den Berg, G., B. Van der Klaauw, and J. Van Ours (2004). Punitive sanctions and the transition rate from welfare to work. *Journal of Labor Economics 22*(1), 211–241.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. *International Economic Review 43*(4), 1249–1287.