Labour Economics 30 (2014) 88–97

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

# From micro data to causality: Forty years of empirical labor economics ☆

Bas van der Klaauw

*Department of Economics, VU University Amsterdam, De Boelelaan 1105, NL-1081 HV Amsterdam, The Netherlands, and Tinbergen Institute, The Netherlands*

ARTICLE INFO

Available online 19 June 2014

*JEL classification:*
C21
C26
C93
J68

*Keywords:*
Treatment effects
Endogeneity
Selection
Experiments
Labor market behavior
Microeconometrics

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 has become a prominent field in economics research.


## 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 from those obtained 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 for 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; 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

☆ 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.
*E-mail address:* b.vander.klaauw@vu.nl.

[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.

http://dx.doi.org/10.1016/j.labeco.2014.06.009
0927-5371/

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-difference much more frequently. In the late 1990s, economists also adopted regression discontinuity estimation as method for causal inference (e.g. Angrist and Lavy, 1999; 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 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),

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

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 $\beta_1$, 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 $\beta_1$. 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 $\beta_1$ 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 h 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 $\beta_1$, which describes the curvature of the labor supply function. Because taxes affect the after-tax hourly wage, the parameter $\beta_1$ 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 these 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 $\beta_1$.

## 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