Instrumental Variables: An Econometrician's Perspective*

Guido W. Imbens†

Graduate School of Business, Stanford University

March 2014

Abstract

I review recent work in the statistics literature on instrumental variables methods from an econometrics perspective. I discuss some of the older, economic, applications including supply and demand models and relate them to the recent applications in settings of randomized experiments with noncompliance. I discuss the assumptions underlying instrumental variables methods and in what settings these may be plausible. By providing context to the current applications a better understanding of the applicability of these methods may arise.

Keywords: Simultaneous Equations Models, Randomized Experiments, Potential Outcomes, Noncompliance, Selection Models

*Financial support for this research was generously provided through NSF grants 0820361 and 0961707. I am grateful to Joshua Angrist who got me interested in these topics many years ago, and, over the the years, has taught me much about the issues discussed in this manuscript, the editor of *Statistical Science* for suggesting this review, and three anonymous referees who wrote remarkably thoughtful reviews.

†Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu, http://www.gsb.stanford.edu/users/imbens.

1 Introduction

Instrumental Variables (IV) refers to a set of methods developed in econometrics starting in the 1920s to draw causal inferences in settings where the treatment of interest cannot be credibly viewed as randomly assigned, even after conditioning on additional covariates.[1] In the last two decades these methods have attracted considerable attention in the statistics literature. Although this recent statistics literature builds on the earlier econometric literature, there are nevertheless important differences. First, the recent statistics literature primarily focuses on the binary treatment case. Second, the recent literature explicitly allows for treatment effect heterogeneity. Third, the recent instrumental variables literature explicitly uses the potential outcome framework used by Neyman for randomized experiments and generalized to observational studies by Rubin (1974, 1978, 1990). Fourth, in the applications this literature has concentrated on, including randomized experiments with non-compliance, the intention-to-treat or reduced-form estimates are often of greater interest than they are in the traditional econometric simultaneous equations applications.

Partly the recent statistics literature has been motivated by the earlier econometric literature on instrumental variables, starting with Wright (1928) (see the discussion on the origins of instrumental variables in Stock and Trebbi, 2003). However, there are also other antecedents, outside of the traditional econometric instrumental variables literature, notably the work by Zelen on encouragement designs (Zelen, 1979, 1990). Early papers in the recent statistics literature include Angrist, Imbens and Rubin (1996), Robins (1989), and McClellan and Newhouse (1994). Recent reviews include Rosenbaum (2010), Vansteelandt, Bowden, Babanezhad, and Goetghebeur (2011) and Hernán and Robins (2006). Although these reviews include many references to the earlier economics literature, it might still be useful to discuss the econometric literature in more detail to provide some background and perspective on the applicability of instrumental variables methods in other fields. In this discussion I will do so.

Instrumental variables methods have been a central part of the econometrics canon

[1] There is another literature in econometrics using instrumental variables methods to deal with classical measurement error (where explanatory variables are measured with error that is independent of the true values). My remarks in the current paper do not directly reflect on that literature. See Sargan (1958) for a classical paper, and Hillier (1992) and Arellano (2002) for more recent discussions.

since the first half of the twentieth century, and continue to be an integral part of most graduate and undergraduate textbooks (*e.g.,* Angrist and Pischke, 2008; Bowden and Turkington, 1984; Greene, 2011; Hayashi, 2000; Manski, 1995; Stock and Watson, 2010; Wooldridge, 2002, 2008). Like the statisticians Fisher and Neyman (Fisher, 1925; Neyman, 1923), early econometricians such as Wright (1928), Working (1927), Tinbergen (1930) and Haavelmo (1943) were interested in drawing causal inferences, in their case about the effect of economic policies on economic behavior. However, in sharp contrast to the statistical literature on causal inference, the starting point for these econometricians was *not* the randomized experiment. From the outset there was a recognition that in the settings they studied, the causes, or treatments, were not *assigned* to passive units (economic agents in their setting, such as individuals, households, firms, or countries). Instead the economic agents actively influence, or even explicitly choose, the level of the treatment they receive. Choice, rather than chance, was the starting point for thinking about the assignment mechanism in the econometrics literature. In this perspective, units receiving the active treatment are different from those receiving the control treatment not just because of the receipt of the treatment: they (choose to) receive the active treatment because they are different to begin with. This makes the treatment potentially *endogenous*, and creates what is sometimes in the econometrics literature referred to as the *selection problem* (Heckman, 1979).

The early econometrics literature on instrumental variables did not have much of an impact on thinking in the statistics community. Although some of the technical work on large sample properties of various estimators did get published in statistics journals (*e.g.*, the still influential Anderson and Rubin (1948) paper), applications by non-economists were rare. It is not clear exactly what the reasons for this are. One possibility is the fact that the early literature on instrumental variables was closely tied to substantive economic questions (*e.g.*, interventions in markets), using theoretical economic concepts that may have appeared irrelevant or difficult to translate to other fields (e.g., supply and demand). This may have suggested to non-economists that the instrumental variables methods in general had limited applicability outside of economics. The use of economic concepts was not entirely unavoidable, as the critical assumptions underlying instrumental variables methods are substantive and require subtle subject matter knowledge. A second reason may be that although the early work by Tinbergen and Haavelmo used

a notation that is very similar to what Rubin (1974) later called the potential outcome notation, quickly the literature settled on a notation only involving realized or obseved outcomes. See for a historial perspective Hendry and Morgan (1992) and Imbens (1997). This realized-outcome notation that remains common in the econometric textbooks obscures the connections between the Fisher and Neyman work on randomized experiments and the instrumental variables literature. It is only in the 1990s that econometricians returned to the potential outcome notation for causal questions (e.g., Heckman, 1990, Manski, 1990; Imbens and Angrist, 1994), facilitating and initiating a dialogue with statisticians on instrumental variable methods.

The main theme of the current paper is that the early work in econometrics is helpful in understanding the modern instrumental variables literature, and furthermore, is potentially useful in improving applications of these methods and identifying potential instruments. These methods may in fact be useful in many settings statisticians study. Exposure to treatment is rarely solely a matter of chance or solely a matter of choice. Both aspects are important and help to understand when causal inferences are credible and when they are not. In order to make these points I will discuss some of the early work and put it in a modern framework and notation. In doing so I will address some of the concerns that have been raised about the applicability of instrumental variables methods in statistics. I will also discuss some areas where the recent statistics literature has extended and improved our understanding of instrumental variables methods. Finally I will review some of the econometric terminology and relate it to the statistical literature to remove some of the semantic barriers that continue to separate the literatures. I should emphasize that many of the topics discussed in this review continue to be active research areas, about which there is considerable controversy both inside and outside of econometrics.

The remainder of the paper is organized as follows. In Section 2 I will discuss the distinction between the statistics literature on causality with its primary focus on chance, arising from its origins in the experimental literature, and the econometrics or economics literature with its emphasis on choice. The next two sections discuss in detail two classes of examples. In Section 3 I discuss the canonical example of instrumental variables in economics, the estimation of supply and demand functions. In Section 4 I discuss a modern class of examples, randomized experiments with noncompliance. In Section 5 I

discuss the substantive content of the critical assumptions, and in Section 6 I link the current literature to the older textbook discussions. In Section 7 I discuss some of the recent extensions of traditional instrumental variables methods. Section 8 concludes.

2 Choice versus Chance in Treatment Assignment

Although the objectives of causal analyses in statistics and econometrics are very similar, or even identical, namely the estimation of, and inference for, causal effects, traditionally statisticians and economists have approached these questions very differently. A key difference in the approaches taken in the statistical and econometric literatures is the focus on different assignment mechanisms, those with an emphasis on chance versus those with an emphasis on choice. Although in practice in many observational studies assignment mechanisms have elements of both chance and choice, the traditional starting points in the two literatures are very different, and it is only recently that these literatures have discovered how much they have in common.[2]

2.1 The Statistics Literature: The Focus on Chance

The starting point in the statistics literature, going back to Fisher (1925) and Neyman (1923), is the randomized experiment, with both Fisher and Neyman motivated by agricultural applications where the units of analysis are plots of land. To be specific, suppose we are interested in the average causal effect of a binary treatment or intervention, say fertilizer A or fertilizer B, on plot yields. In the modern notation and language originating with Rubin (1974), the unit (plot) level causal effect is a comparison between the two potential outcomes, $Y_i(A)$ and $Y_i(B)$ (*e.g.*, the difference $\tau_i = Y_i(B) - Y_i(A)$), where $Y_i(A)$ is the potential outcome given fertilizer A and $Y_i(B)$ is the potential outcome given fertilizer B, both for plot i. In a completely randomized experiment with N plots we select M (with $M \in \{1, \ldots, N-1\}$) plots at random to receive fertilizer B, with the

[2]In both literatures it is typically assumed that there is no interference between units. In the statistics literature this is often referred to as the *Stable Unit Treatment Value Assumption* (SUTVA, Rubin, 1978). In economics there are many cases where this is not a reasonable assumption because there are *general equilibrium* effects. In an interesting recent experiment Crépon, Duflo, Gurgand, Rathelot, and Zamoray (2012) varied the scale of experimental interventions (job training programs in their case) in different labor markets and found that the scale substantially affected the average effects of the interventions. There is also a growing literature on settings directly modelling interactions. In this discussion I will largely ignore the complications arising from interference between units.

remaining $N-M$ plots assigned to fertilizer A. Thus, the treatment assignment, denoted by $W_i \in \{A, B\}$ for plot i, is by design independent of the potential outcomes. In this specific setting the work by Fisher and Neyman shows how one can draw exact causal inferences. Fisher focused on calculating exact p-values for sharp null hypotheses, typically the null hypothesis of no effect whatsoever, $Y_i(A) = Y_i(B)$ for all plots. Neyman focused on developing unbiased estimators for the average treatment effect $\sum_i (Y_i(A) - Y_i(B))/N$ and the variance of those estimators.

The subsequent literature in statistics, much of it associated with the work by Rubin and coauthors (Cochran, 1968; Cochran and Rubin, 1973; Rubin, 1974, 1990, 2006; Rosenbaum and Rubin, 1983; 1984; Rubin and Thomas 1992; Rosenbaum, 2002, 2010; Holland, 1986) has focused on extending and generalizing the Fisher and Neyman results that were derived explicitly for randomized experiments to the more general setting of observational studies. A large part of this literature focuses on the case where the researcher has additional background information available about the units in the study. The additional information is in the form of pretreatment variables or covariates not affected by the treatment. Let X_i denote these covariates. A key assumption in this literature is that conditional on these pretreatment variables the assignment to treatment is independent of the treatment assignment. Formally,

$$W_i \perp Y_i(A), Y_i(B) \,\Big|\, X_i. \qquad \text{(unconfoundedness)}$$

Following Rubin (1990), I refer to this assumption as *unconfoundedness given* X_i, also known as *no unmeasured confounders*. This assumption, in combination with the auxiliary assumption that for all values of the covariates the probability of being assigned to each level of the treatment is strictly positive is referred to as *strong ignorability* (Rosenbaum and Rubin, 1984). If we assume only that $W_i \perp Y_i(A)|X_i$ and $W_i \perp Y_i(B)|X_i$ rather than jointly, the assumption is referred to as *weak unconfoundedness* (Imbens, 2000), and the combination as *weak ignorability*. Substantively it is not clear that there are cases in the setting with binary treatments where the weak version of plausible but not the strong version, although the difference between the two assumptions has some content in the multivalued treatment case (Imbens, 2000). In the econometric literature closely related assumptions are referred to as *selection-on-observables* (Barnow, Cain and Goldberger (1980) or *exogeneity*.

Under weak ignorability (and thus also under strong ignorability) it is possible to estimate the average effect of the treatment in large samples, in other words, the average effect of the treatment is *identified*. Various specific methods have been proposed, including matching, subclassification, and regression. See Rosenbaum (2010), Rubin (2006), Imbens (2004, 2014), Gelman and Hill (2006), and Angrist and Pischke (2009) for general discussions and surveys. Robins and coauthors (Robins, 1986; Gill and Robins, 2001; Richardson, and Robins, 2013; Van der Laan and Robins, 2003), have extended this approach to settings with sequential treatments.

2.2 The Econometrics Literature: The Focus on Choice

In contrast to the statistics literature whose point of departure was the randomized experiment, the starting point in the economics and econometrics literatures for studying causal effects emphasizes the choices that led to the treatment received. Unlike the original applications in statistics where the units are passive, for example plots of land, with no influence over their treatment exposure, units in economic analyses are typically economic agents, for example, individuals, families, firms, or administrations. These are agents with objectives and the ability to pursue these objectives within constraints. The objectives are typically closely related to the outcomes under the various treatments. The constraints may be legal, financial, or information-based.

The starting point of economic science is to model these agents as behaving optimally. More specifically this implies that economists think of everyone of these agents as choosing the level of the treatment to most efficiently pursue their objectives given the constraints they face.[3] In practice, of course, there is considerable evidence that not all agents behave optimally. Nevertheless, the starting point is the presumption that optimal behavior is a reasonable approximation to actual behavior, and the models economists take to the data often reflect this.

2.3 Some Examples

Let us contrast the statistical and econometric approaches in a highly stylized example. Roy (1951) studies the problem of occupational choice and the implications for the ob-

[3] In principle these objectives may include the effort it takes to find the optimal strategy, although it is rare that these costs are taken into account.

served distribution of earnings. He focuses on an example where individuals can choose between two occupations, hunting and fishing. Each individual has a level of productivity associated with each occupation, say, the total value of the catch per day. For individual i, the two productivity levels are $Y_i(h)$ and $Y_i(f)$, for the productivity level if hunting and fishing respectively.[4] Suppose the researcher is interested in the average difference in productivity in these two occupations, $\tau = \mathbb{E}[Y_i(f) - Y_i(h)]$, where the averaging is over the population of individuals.[5] The researcher observes for all units in the sample the occupation they choose (W_i, equal to h for hunters and f for fishermen) and the productivity in their choosen occupation,

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(h) & \text{if } W_i = h, \\ Y_i(f) & \text{if } W_i = f. \end{cases}$$

In the Fisher-Neyman-Rubin statistics tradition, one might start by estimating τ by comparing productivity levels by occupation:

$$\hat{\tau} = \overline{Y}_f^{\text{obs}} - \overline{Y}_h^{\text{obs}},$$

where

$$\overline{Y}_f^{\text{obs}} = \frac{1}{N_f} \sum_{i:W_i=f} Y_i^{\text{obs}}, \quad \overline{Y}_h^{\text{obs}} = \frac{1}{N_h} \sum_{i:W_i=h} Y_i^{\text{obs}},$$

$$N_f = \sum_{i=1}^{N} \mathbf{1}_{W_i=f}, \quad \text{and} \quad N_h = N - N_f.$$

If there is concern that these unadjusted differences are not credible as estimates of the average causal effect, the next step in this approach would be to adjust for observed individual characteristics such as education levels, or family background. This would be justified if individuals can be thought of as choosing, at least within homogenous groups defined by covariates, randomly which occupation to engage in.

Roy, in the economics tradition, starts from a very different place. Instead of assuming that individuals choose their occupation (possibly after conditioning on covariates)

[4] In this example the no-interference (SUTVA) assumption that there are no effects of other individual's choices, and therefore that the individual level potential outcomes are well-defined is tenuous—if one hunter is successful that will reduce the number of animals available to other hunters—but I will ignore these issues here.

[5] That is not actually the goal of Roy's original study, but that is beside the point here.

randomly, he assumes that each individual chooses her occupation optimally, that is, the occupation that maximizes her productivity:

$$W_i = \begin{cases} f & \text{if } Y_i(f) \geq Y_i(h), \\ h & \text{otherwise.} \end{cases}$$

There need not be a solution in all cases, especially if there is interference and thus there are general equilibrium effects, but I will assume here that such a solution exists. If this assumption were true, it would be difficult to learn much about τ from data on occupations and earnings. In the spirit of research by Manski (1990, 1992, 2003, 2008) and Robins (1989), one can derive bounds on τ, exploiting the fact that if $W_i = f$, then the unobserved $Y_i(h)$ must satisfy $Y_i(h) \leq Y_i(f)$, with $Y_i(f)$ observed. For the Roy model the specific calculations have been reported in Manski (1995, Section 2.6). Without additional information or restrictions these bounds would be fairly wide, and one would not learn much about τ. However, the original version of the Roy model, where individuals know *ex ante* the exact value of the potential outcomes and choose the level of the treatment corresponding to the maximum of those, is ultimately not plausible in practice. It is likely that that individuals face uncertainty regarding their future productivity and thus may not be able to choose the *ex post* optimal occupation. See for bounds under that scenario Manski and Nagin (1998). Alternatively, and this is emphasized in Athey and Stern (1998), individuals may have more complex objective functions taking into account heterogenous costs or non-monetary benefits associated with each occupation. This creates a wedge between the outcomes that the researcher focuses on and the outcomes that the agent optimizes over. What is key here in relation to the statistics literature is that under the Roy model and its generalizations the very fact that two individuals have different occupations is seen as indicative that they have different potential outcomes, thus fundamentally calling into question the unconfoundedness assumption that individuals with similar pretreatment variables but different treatment levels are comparable. This concern about differences between individuals with the same values for pretreatment variables but different treatment levels underlies many econometric analyses of causal effects, specifically in the literature on selection models. See Heckman and Robb (1985) for a general discussion.

Let me discuss two additional examples. There is a large literature in economics concerned with estimating the causal effect of educational achievement (measured as years

of education) on earnings. See for general discussions Griliches (1977) and Card (2001). One starting point, and in fact the basis of a large empirical literature, is to compare earnings for individuals who look similar in terms of background characteristics, but who differ in terms of educational achievement. The concern in an equally large literature is that those individuals who choose to acquire higher levels of education did so precisely because they expected their returns to additional years of education to be higher then individuals who choose not to acquire higher levels of education expected their returns to be. In the terminology of the returns-to-education literature, the individuals choosing higher levels of education may have higher levels of ability, which lead to higher earnings for given levels of education.

Another canonical example is that of voluntary job training programs. One approach to estimate the causal effect of training programs on subsequent earnings would be to compare earnings for those participating in the program with earnings for those who did not. Again the concern would be that those who choose to participate did so because they expected bigger benefits (financial or otherwise) from doing so than individuals who choose not to participate.

These issues also arise in the missing data literature. The statistics literature (Rubin, 1976, 1987; Little and Rubin, 1987) has primarily focused on models that assume that units with item nonresponse are comparable to units with complete response, conditional on covariates that are always observed. The econometrics literature (Heckman, 1976, 1979) has focused more heavily on models that interpret the nonresponse as the result of systematic differences between units. Philipson (1997ab) takes this even further, viewing survey response as a market transaction, where individuals not responding do so because to the potential respondents the costs of responding outweights the benefits. The Heckman-style selection models often assume strong parametric alternatives to the Rubin and Little missing-at-random or ignorability condition. This has often in turn led to estimators that are sensitive to small changes in the data generating process.

These issues of non-random selection are of course not special to economics. Outside of randomized experiments the exposure to treatment is typically also chosen to achieve some objectives, rather than randomly within homogenous populations. For example, physicians presumably choose treatments for their patients optimally, given their knowledge and given other constraints (*e.g.*, financial). Similarly, in economics and other

social sciences one may view individuals as making optimal decisions, but these are typically made given incomplete information, leading to errors that may make the ultimate decisions appear as good as random within homogenous subpopulations. What is important is that the starting point is different in the two disciplines, and this has led to the development of substantially different methods for causal inference.

2.4 Instrumental Variables

How do instrumental variables methods address the type of selection issues the Roy model raises? At the core instrumental variables change the incentives for agents to choose a particular level of the treatment, without affecting the potential outcomes associated with these treatment levels. Consider a job training program example where the researcher is interested in the average effect of the training program on earnings. Each individual is characterized by two potential earnings outcomes, earnings given the training and earnings in the absence of the training. Each individual chooses to participate or not based on their perceived net benefits from doing so. As pointed out in Athey and Stern (1998), it is important that these net benefits differ from the earnings. Thye do so by the costs associated with participating in that regime. Suppose that there is variation in the costs individuals incur with participation in the training program. The costs are broadly defined, and may include travel time to the program facilities, or the effort required to become informed about the program. Furthermore suppose that these costs are independent of the potential outcomes. This is a strong assumption, often made more plausible by conditioning on covariates. Measures of the participation cost may then serve as instrument variables and aid in the identification of the causal effects of the program. Ultimately we compare earnings for individuals with low costs of participation in the program with those for individuals with high costs of participation and attribute the difference in average earnings to the increased rate of participation in the program among the two groups.

In almost all cases the assumption that there is no direct effect of the change in incentives on the potential outcomes is controversial, and it needs to be assessed at a case-by-case level. The second part of the assumption, that the costs are independent of the potential outcomes, possibly after conditioning on covariates, is qualitatively very

different. In some cases it is satisfied by design, *e.g.,* if the incentives are randomized. In observational studies it is a substantive, unconfoundedness-type, assumption, that may be more plausible or at least approximately hold after conditioning on covariates. For example, in a number of studies researchers have used physical distance to facilities as instruments for exposure to treatments available at such facilities. Such studies include McClellan and Newhouse (1994) and Baiocchi, Small, Lorch and Rosenbaum (2010) who use distance to hospitals with particular capabilities as an instrument for treatments associated with those capabilities, after conditioning on distance to the nearest medical facility, and Card (1995), who uses distance to colleges as an instrument for attending college.

3 The Classic Example: Supply and Demand

In this section I will discuss the classic example of instrumental variables methods in econometrics, that is, simultaneous equations. Simultaneous equations models are both at the core of the econometrics canon and at the core of the confusion concerning instrumental variables methods in the statistics literature. More precisely, in this section I will look at supply and demand models that motivated the original research into instrumental variables. Here the *endogeneity*, that is, the violation of unconfoundedness, arises from an equilibrium condition. I will discuss the model in a very specific example to make the issues clear, as I think that perhaps the level of abstraction used in the older econometric text books has hampered communication with researchers in other fields.

3.1 Discussions in the Statistics Literature

To show the level of frustration and confusion in the statistics literature with these models, let me present some quotes. In a comment on Pratt and Schlaifer (1984), Dawid (1984) writes "I despair of ever understanding the logic of simultaneous equations well enough to tackle them,"(1984, page 24). Cox (1992) writes in a discussion on causality "it seems reasonable that models should be specified in a way that would allow direct computer simulation of the data ..... This for example precludes the use of y_2 as an explanatory variable for y_1 if at the same time y_1 is an explanatory variable for y_2"(page 294). This seems to directly contradict the first model Haavelmo considers, e.g., equations

(1.1) and (1.2) (Haavelmo, 1943, p. 2):

$$Y = aX + \epsilon_1 \qquad X = bY + \epsilon_2.$$

In fact, the comment by Cox appears to rule out simultaneous equations models of the type studied by economists. Holland (1988), in comment on structural equation methods in econometrics, writes "why should [this disturbance] be independent of [the instrument] ... when the very definition of [this disturbance] involves [the instrument],"(page 460). Freedman writes "Additionally, some variables are taken to be exogenous (independent of the disturbance terms) and some endogenous (dependent on the disturbance terms). The rationale is seldom clear, because–among other things–there is seldom any very clear description of what the disturbance terms mean, or where they come from," (Freedman, 2006, p. 699).

3.2 The Market for Fish

The specific example I will use in this section is the market for whiting (a particular white fish, often used in fish sticks) traded at the Fulton fish market in New York City. Whiting was sold at the Fulton fish market at the time by a small number of dealers to a large number of buyers. Graddy collected data on quantities and prices of whiting sold by a particular trader at the Fulton fish market on 111 days between December 2nd 1991 and May 8th 1992 (Graddy, 1995, 1996; Angrist, Graddy and Imbens, 2000). I will take as the unit of analysis a day, and interchangeably refer to this as a market. Each day, or market, during the period covered in this data set, indexed by $t = 1, \ldots, 111$, a number of pounds of whiting are sold by this particular trader, denoted by Q_t^{obs}. Not every transaction on the same day involves the same price, but to focus on the essentials I will aggregate the total amount of whiting sold and the total amount of money it was sold for, and calculate a price per pound (in cents) for each of the 111 days, denoted by P_t^{obs}. Figure 1 presents a scatterplot of the observed log price and log quantity data. The average quantity sold was 6,335 pounds, with a standard deviation of 4,040 pounds, for an average price of 88 cts per pound and a standard deviation of 34 cts. For example, on the first day of this period 8,058 pounds were sold for an average of 65 cents, and the next day 2,224 pounds were sold for an average of 100 cents. Table 1 presents averages of log prices and log quantities for the fish data.

Now suppose we are interested in predicting the effect of a tax in this market. To be specific, suppose the government is considering imposing a $100 \times r$ % tax (*e.g.*, a 10% tax) on all whiting sold, but before doing so it wishes to predict the average percentage change in the quantity sold as a result of the tax. We may formalize that by looking at the average effect on the logarithm of the quantity, $\tau = \mathbb{E}[\ln Q_t(r) - \ln Q_t(0)]$, where $Q_t(r)$ is the quantity traded in market/day t if the tax rate were set at r. The problem, substantially worse than in the standard causal inference setting where for some units we observe one of the two potential outcomes and for other units we observe the other potential outcome, is that in all 111 markets we observe the quantity traded at tax rate 0, $Q_t^{\text{obs}} = Q_t(0)$, and we never see the quantity traded at the tax rate contemplated by the government, $Q_t(r)$. Because only $\mathbb{E}[\ln Q_t(0)]$ is directly estimable from data on the quantities we observe, the question is how to draw inferences about $\mathbb{E}[\ln Q_t(r)]$.

A naive approach would be to assume that a tax increase by 10% would simply raise prices by 10%. If one additionally is willing to make the unconfoundedness assumption that prices can be viewed as set independently of market conditions on a particular day, it follows that those markets after the introduction of the tax where the price net of taxes is \$1.00 would on average be like those markets prior to the introduction of the 10% tax where the price was \$1.10. Formally, this approach assumes that

$$\mathbb{E}[\ln Q_t(r)|P_t^{\text{obs}} = p] = \mathbb{E}[\ln Q_t(0)|P_t^{\text{obs}} = (1+r) \times p], \tag{3.1}$$

implying that

$$\begin{aligned}\mathbb{E}[\ln Q_t(r) - \ln Q_t(0)|P_t^{\text{obs}} = p] &= \mathbb{E}[\ln Q_t^{\text{obs}}|P_t^{\text{obs}} = (1+r) \times p] - \mathbb{E}[\ln Q_t^{\text{obs}}|P_t^{\text{obs}} = p]. \\ &\approx \mathbb{E}[\ln Q_t^{\text{obs}}| \ln P_t^{\text{obs}} = r + \ln p] - \mathbb{E}[\ln Q_t^{\text{obs}}| \ln P_t^{\text{obs}} = \ln p].\end{aligned}$$

The last quantity is often estimated using linear regression methods. Typically the regression function is assumed to be linear in logarithms with constant coefficients,

$$\ln Q_t^{\text{obs}} = \alpha^{\text{ls}} + \beta^{\text{ls}} \times \ \ln P_t^{\text{obs}} + \varepsilon_t. \tag{3.2}$$

Ordinary least squares estimation with the Fulton fish market data collected by Graddy leads to

$$\widehat{\ln Q_t^{\text{obs}}} = \underset{(0.08)}{8.42} - \underset{(0.18)}{0.54} \times \ \ln P_t^{\text{obs}}.$$

The estimated regression line is also plotted in Figure 1. Interestingly this is what Working (1927) calls the "statistical 'demand curve'," as opposed to the concept of a demand curve in economic theory. This simple regression, in combination with the assumption embodied in (3.1), suggests that the quantity traded would go down, on average, by 5.4% in response to a 10% tax.

$$\hat{\tau} = -0.054 \qquad \text{(s.e. } 0.018\text{)}.$$

Why does this answer, or at least the method in which it was derived, not make any sense to an economist? The answer assumes that prices can be viewed as independent of the potential quantities traded, or, in other words, unconfounded. This assignment mechanism is unrealistic. In reality, it is likely the markets/days, prior to the introduction of the tax, when the price was \$1.10 were systematically different from those where the price was \$1.00. From an economists' perspective the fact that the price was \$1.10 rather than \$1.00 implies that market conditions *must* have been different, and it is likely that these differences are directly related to the quantities traded. For example, on days where the price was high there may have been more buyers, or buyers may have been interested in buying larger quantities, or there may have been less fish brought ashore. In order to predict the effect of the tax we need to think about the responses of both buyers and sellers to changes in prices, and about the determination of prices. This is where economic theory comes in.

3.3 The Supply of and Demand for Fish

So, how do economists go about analyzing questions such as this one if not by regressing quantities on prices? The starting point for economists is to think of an economic model for the determination of prices (the treatment assignment mechanism in Rubin's potential outcome terminology). The first part of the most simple model an economist would consider for this type of setting is a pair of functions, the demand and supply functions. Think of the buyers coming to the Fulton fishmarket on a given market/day (say, day t) with a demand function $Q_t^d(p)$. This function tells us, for that particular morning, how much fish all buyers combined would be willing to buy if the price on that day were p, for any value of p. This function is conceptually exactly like the potential outcomes set up commonly used in causal inference in the modern literature. It is more complicated

than the binary treatment case with two potential outcomes, because there is a potential outcome for each value of the price, with more or less a continuum of possible price values, but it is in line with continuous treatment extensions such as those in Gill and Robins (2001). Common sense, and economic theory, suggests that this demand function is a downward sloping function: buyers would likely be willing to buy more pounds of whiting if it were cheaper. Traditionally the demand function is specified parametrically, for example linear in logarithms:

$$\ln Q_t^d(p) = \alpha^d + \beta^d \times \ln p + \varepsilon_t^d, \tag{3.3}$$

where β^d is the price elasticity of demand. This equation is *not* a regression function like (3.2). It is interpreted as a *structural equation* or behavioral equation, and, in the treatment effect literature terminology, it is a model for the potential outcomes. Part of the confusion between the model for the potential outcomes in (3.3) and the regression function in (3.2) may stem from the traditional notation in the econometrics literature where the same symbol (*e.g.,* Q_t) would be used for the observed outcomes (Q_t^{obs} in our notation) and the potential outcome function ($Q_t^d(p)$ in our notation), and the same symbol (*e.g.,* P_t) would be used for the observed value of the treatment (P_t^{obs} in our notation) and the argument in the potential outcome function (p in our notation). Interestingly the pioneers in this literature, Tinbergen (1928) and Haavelmo (1943), *did* distinguish between these concepts in their notation, but the subsequent literature on simultaneous equations dropped that distinction and adopted a notation that did not distinguish between observed and potential outcomes. My view is that dropping this distinction was merely incidental, and that implicitly the interpretation of the simultaneous equations models remained that in terms of potential outcomes.[6]

Implicit (by the lack of a subscript on the coefficients) in the specification of the demand function in (3.3) is the strong assumption that the effect of a unit change in the logarithm of the price (equal to β^d) is the same for all values of the price, and that the effect is the same in all markets. This is clearly a very strong assumption, and the

[6]As a reviewer pointed out, once one views simultaneous equations in terms of potential outcomes, there is a natural normalization of the equations. This suggests that perhaps the discussions of issues concerning normalizations of equations in simultaneous equations models ((*e.g.,* Basmann (1963ab, 1965, Hillier, 1990) implicitly rely on a different interpretation, for example thinking of the endogeneity arising from measurement error. Throughout this discussion I will interpret simultaneous equations in terms of potential outcomes, viewing the realized outcome notation simply as obscuring that.

modern literature on simultaneous equations (see Matzkin (2007) for an overview) has developed less restrictive specifications allowing for nonlinear and nonadditive effects while maintaining identification. The unobserved component in the demand function, denoted by ε_t^d, represents unobserved determinants of the demand on any given day/market: a particular buyer may be sick on a particular day and not go to the market, or may be expecting a client wanting to purchase a large quantity of whiting. We can normalize this unobserved component to have expectation zero, where the expectation is taken over all markets or days:

$$\mathbb{E}\left[\ln Q_t^d(p)\right] = \alpha^d + \beta^d \times \ln p.$$

The interpretation of this expectation is subtle, and again it is part of the confusion that sometimes arises. Consider the expected demand at $p = 1$, $\mathbb{E}\left[\ln Q_t^d(1)\right]$, under the linear specification in (3.3) equal to $\alpha^d + \beta^d \cdot \ln(1) = \alpha^d$. This α^d is the average of all demand functions, evaluated at price equal to \$1.00, irrespective of what the actual price in the market is, where the expectation is taken over *all* markets. It is *not*, and this is key, the conditional expectation of the observed quantity in markets where the price is equal to \$1.00 (or which is the same the demand function at 1 in those markets), which is $\mathbb{E}[\ln Q_t^{\text{obs}}|P_t^{\text{obs}} = 1] = \mathbb{E}[\ln Q_t^d(1)|P_t^{\text{obs}} = 1]$, for example, under a linear specification as in (3.2), equal to $\alpha^{\text{ls}} + \beta^{\text{ls}} \cdot \ln(1) = \alpha^{\text{ls}}$. Here the original Tinbergen and Haavelmo notation and the modern potential outcome version is much clearer than the sixties econometrics textbook notation.[7]

Similar to the demand function, the supply function $Q_t^s(p)$ represents the quantity of whiting the sellers collectively are willing to sell at any given price p, on day t. Here common sense would suggest that this function is sloping upward: the higher the price, the more the sellers are willing to sell. As with the demand function the supply function is typically specified parametrically with constant coefficients:

$$\ln Q_t^s(p) = \alpha^s + \beta^s \times \ln p + \varepsilon_t^s, \tag{3.4}$$

[7]Other notations have been recently been proposed to stress the difference between the conditional expectation of the observed outcome and the expectation of the potential outcome. Pearl (2000) writes the expected demand when the price is *set* to \$1.00 as $\mathbb{E}\left[\ln Q_t^d|\text{do}(P_t = 1)\right]$, rather than conditional on the price being observed to be \$1.00. Hernán and Robins (2006) write this average potential outcome as $\mathbb{E}\left[\ln Q_t^d(P_t = 1)\right]$, whereas Lauritzen and Richardson (2002) write it as $\mathbb{E}[\ln Q_t^{\text{obs}}||P_t^{\text{obs}} = 1]$ where the double $||$ implies *conditioning by intervention*.

where β^s is the price elasticity of supply. Again we can normalize the expectation of ε_t^s to zero (where the expectation is taken over markets), and write

$$\mathbb{E}\left[\ln Q_t^s(p)\right] = \alpha^s + \beta^s \times \ln p.$$

Note that the ε_t^d and ε_t^s are not assumed to be independent in general, although in some applications that may be a reasonable assumption.

3.4 Market Equilibrium

Now comes the second part of the simple economic model, the determination of the price, or, in the terminology of the treatment effect literature, the assignment mechanism. The conventional assumption in this type of market is that the price that is observed, that is the price at which the fish is traded in market/day t, is the (unique) market clearing price at which demand and supply are equal. In other words, this is the price at which the market is in *equilibrium*, denoted by P_t^{obs}. This equilibrium price solves:

$$Q_t^d(P_t^{\text{obs}}) = Q_t^s(P_t^{\text{obs}}). \tag{3.5}$$

The observed quantity on that day, that is the quantity actually traded, denoted by Q_t^{obs}, is then equal to the demand function at the equilibrium price (or, equivalently, because of the equilibrium assumption, the supply function at that price):

$$Q_t^{\text{obs}} = Q_t^d(P_t^{\text{obs}}) = Q_t^s(P_t^{\text{obs}}). \tag{3.6}$$

Assuming that the demand function does slope downward and the supply function does slope upward, and both are linear in logarithms, the equilibrium price exists and is unique, and we can solve for the observed price and quantities in terms of the parameters of the model and the unobserved components:

$$\ln P_t^{\text{obs}} = \frac{\alpha^d - \alpha^s}{\beta^s - \beta^d} + \frac{\varepsilon_t^d - \varepsilon_t^s}{\beta^s - \beta^d}, \qquad \text{and} \quad \ln Q_t^{\text{obs}} = \frac{\beta^s \cdot \alpha^d - \beta^d \cdot \alpha^s}{\beta^s - \beta^d} + \frac{\beta^s \cdot \varepsilon_t^d - \beta^d \cdot \varepsilon_t^s}{\beta^s - \beta^d}.$$

For economists this is a more plausible model for the determination of realized prices and quantities than the model that assumes prices are independent of market conditions. It is not without its problems though. Chief among these from our perspective is the complication that, just as in the Roy model, we cannot necessarily infer the values of the unknown parameters in this model even if we have data on many markets.

Another issue is how buyers and sellers arrive at the equilibrium price. There is a theoretical economic literature addressing this question. Often the idea is that there is a sequential process of buyers making bids, and suppliers responding with offers of quantities at those prices, with this process repeating itself until it arrives at a price at which supply and demand are equal. In practice economists often refrain from specifying the details of this process and simply assume that the market is in equilibrium. If the process is fast enough, it may be reasonable to ignore the fact the specifics of the process and analyze the data as if equilibrium was instantaneous.[8] A related issue is whether this model with an equilibrium prices that equates supply and demand is a reasonable approximation to the actual process that determines prices and quantities. In fact, Graddy's data contains information showing that the seller would trade at different prices on the same day, so strictly speaking this model does not hold. There is a long tradition in economics, however, of using such models as approximations to price determination and we will do so here.

Finally, let me connect this to the textbook discussion of supply and demand models. In many textbooks the demand and supply equations would be written directly in terms of the observed (equilibrium) quantities and prices as

$$Q_t^{\text{obs}} = \alpha^s + \beta^s \times \ln P_t^{\text{obs}} + \varepsilon_t^s, \tag{3.7}$$

$$Q_t^{\text{obs}} = \alpha^d + \beta^d \times \ln P_t^{\text{obs}} + \varepsilon_t^d. \tag{3.8}$$

This representation leaves out much of the structure that gives the demand and supply function their meaning, that is, the demand equation (3.3), the supply equation (3.4), and the equilibrium condition (3.5). As Strotz and Wold (1960) write "Those who write such systems [(3.8) and (3.8)] do not, however, really mean what they write, but introduce an ellipsis which is familiar to economists" (p. 425), with the ellipsis referring to the market equilibrium condition that is left out.

3.5 The Statistical Demand Curve

Given this set up, let me discuss two issues. First, let us explore, under this model, the interpretation of what Working (1927) called the "statistical demand curve." The covari-

[8] See Shapley and Shubik (1977) and Giraud (2003), and for some experimental evidence, Plott and Smith, (1978) and Smith (1982).

ance between observed (equilibrium) log quantities and log prices is $\text{Cov}\left(\ln Q_t^{\text{obs}}, \ln P_t^{\text{obs}}\right) = (\beta^s \cdot \sigma_d^2 + \beta^d \cdot \sigma_s^2 - \rho \cdot \sigma_d \cdot \sigma_s \cdot (\beta^d + \beta^s))/(\left(\beta^s - \beta^d\right)^2)$, where σ_d and σ_s are the standard deviations of ε_t^d and ε_t^s respectively, and ρ is their correlation. Because the variance of $\ln P_t^{\text{obs}}$ is $(\sigma_s^2 + \sigma_d^2 - 2 \cdot \rho \cdot \sigma_d \cdot \sigma_s)/(\beta^s - \beta^d)^2$, it follows that the regression coefficient in the regression of log quantities on log prices is

$$\frac{\text{cov}\left(\ln Q_t^{\text{obs}}, \ln P_t^{\text{obs}}\right)}{\text{var}\left(\ln P_t^{\text{obs}}\right)} = \frac{\beta^s \cdot \sigma_d^2 + \beta^d \cdot \sigma_s^2 - \rho \cdot \sigma_d \cdot \sigma_s \cdot (\beta^d + \beta^s)}{\sigma_s^2 + \sigma_d^2 - 2 \cdot \rho \cdot \sigma_d \cdot \sigma_s}.$$

Working focuses on the interpretation of this relation between equilibrium quantities and prices. Suppose that the correlation between ε_t^d and ε_t^s, denoted by ρ, is zero. Then the regression coefficient is a weighted average of the two slope coefficients of the supply and demand function, weighted by the variances of the residuals:

$$\frac{\text{cov}\left(\ln Q_t^{\text{obs}}, \ln P_t^{\text{obs}}\right)}{\text{var}\left(\ln P_t^{\text{obs}}\right)} = \beta^s \cdot \frac{\sigma_d^2}{\sigma_s^2 + \sigma_d^2} + \beta^d \cdot \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2}.$$

If σ_d^2 is small relative to σ_s^2, then we estimate something close to the slope of the demand function, and if σ_s^2 is small relative to σ_d^2, then we estimate something close to the slope of the supply function. In general, however, as Working stresses, the "statistical demand curve" is not informative about the demand function (or about the supply function). See also Leamer (1981).

3.6 The Effect of a Tax Increase

The second question is how this model with supply and demand functions and a market clearing price helps us answer the substantive question of interest. The specific question considered is the effect of the tax increase on the average quantity traded. In a given market, let p be the price sellers receive per pound of whiting, and let $\tilde{p} = p \times (1 + r)$ the price buyers pay after the tax has been imposed. The key assumption is that the only way buyers and sellers respond to the tax is through the effect of the tax on prices: they do not change how much they would be willing to buy or sell at any given price, and the process that determines the equilibrium price does not change. The technical econometric term for this is that the demand and supply functions are *structural* or *invariant* in the sense that they are not affected by changes in the treatment, taxes in this case. This may not be a perfect assumption, but certainly in many cases it is reasonable: if I have

to pay \$1.10 per pound of whiting, I probably do not care whether 10cts of that goes to the government and \$1 to the seller, or all of it goes to the seller. If we are willing to make that assumption, we can solve for the new equilibrium price and quantity. Let $P_t(r)$ be the new equilibrium price (net of taxes, that is, the price sellers receive), given a tax rate r, with in our example $r = 0.1$. This price solves

$$Q_t^d(P_t(r) \times (1+r)) = Q_t^s(P_t(r)).$$

Given the linear specification for the demand and supply functions, this leads to

$$\ln Q_t(r) = \frac{\alpha^d - \alpha^s}{\beta^s - (1+r) \times \beta^d} + \frac{\varepsilon_t^d - \varepsilon_t^s}{\beta^s - (1+r) \times \beta^d}.$$

The result of the tax is that the average price that sellers receive with a positive tax rate r is less than what they would have received in the absence of the tax rate:

$$\mathbb{E}\left[\ln P_t(r)\right] = \frac{\alpha^d - \alpha^s}{\beta^s - (1+r) \times \beta^d} \leq \frac{\alpha^d - \alpha^s}{\beta^s - \beta^d} = \mathbb{E}\left[\ln P_t(0)\right].$$

(Note that $\beta^d < 0$.) On the other hand, the buyers will pay more on average:

$$\mathbb{E}\left[\ln P_t(r)\right] = \frac{\alpha^d - \alpha^s}{\beta^s - \beta^d} \leq (1+r) \times \frac{\alpha^d - \alpha^s}{\beta^s - (1+r) \times \beta^d} = (1+r) \times \mathbb{E}\left[\ln P_t(0)\right].$$

The quantity traded after the tax increase is

$$\ln Q_t(r) = \frac{\beta^s \cdot \alpha^d - (1+r) \cdot \beta^d \cdot \alpha^s}{\beta^s - (1+r) \cdot \beta^d} + \frac{\beta^s \cdot \varepsilon_t^d - (1+r) \cdot \beta^d \cdot \varepsilon_t^s}{\beta^s - (1+r) \cdot \beta^d},$$

which is less than the quantity that would be traded in the absence of the tax increase. The market-level causal effect is

$$\ln Q_t(r) - \ln Q_t(0) = \frac{\beta^s \cdot \alpha^d - (1+r) \cdot \beta^d \cdot \alpha^s}{\beta^s - (1+r) \cdot \beta^d} + \frac{\beta^s \cdot \varepsilon_t^d - 1.1 \cdot \beta^d \cdot \varepsilon_t^s}{\beta^s - (1+r) \cdot \beta^d}$$
$$- \frac{\beta^s \cdot \alpha^d - \beta^d \cdot \alpha^s}{\beta^s - \beta^d} - \frac{\beta^s \cdot \varepsilon_t^d - \beta^d \cdot \varepsilon_t^s}{\beta^s - \beta^d},$$

with the average causal effect of the $100 \times r\%$ tax increase equal to

$$\tau = \mathbb{E}[\ln Q_t(r) - \ln Q_t(0)] = \frac{r \cdot \beta^d \cdot \beta^s \cdot (\alpha^d - \alpha^s)}{(\beta^s - \beta^d) \times (\beta^s - (1+r) \cdot \beta^d)} < 0.$$

What should we take away from this discussion? There are three points. First, the regression coefficient in the regression of log quantity on log prices does not tell us much

about the effect of new tax. The sign of this regression coefficient is ambiguous, depending on the variances and covariance of the unobserved determinants of supply and demand. Second, in order to predict the magnitude of the effect of a new tax we need to learn about the demand and supply functions separately, or in the econometrics terminology, *identify* the supply and demand function. Third, observations on equilibrium prices and quantities by themselves do not identify these functions.

3.7 Identification with Instrumental Variables

Given this identification problem, how *do* we identify the demand and supply functions? This is where instrumental variables enter the discussion. To identify the demand function we look for determinants of the supply of whiting that do not affect the demand for whiting, and, similarly, to identify the supply function we look for determinants of the demand for whiting that do not affect the supply. In this specific case Graddy (1995, 1996) assumes that weather conditions at sea on the days prior to market t, denoted by Z_t, affect supply but do not affect demand. Certainly it appears reasonable to think that weather is a direct determinant of supply: having high waves and strong winds makes it harder to catch fish. On the other hand, there does not seem to be any reason why demand on day t, at a given price p, would be correlated with wave height or wind speed on previous days. This assumption may be made more plausible by conditioning on covariates. For example, if one is concerned that weather conditions on land affect demand, one may wish to condition on those, and only look at variation in weather conditions at sea given similar weather conditions on land as an instrument. Formally, the key assumptions are that

$$Q_t^d(p) \perp Z_t, \qquad \text{and} \quad Q_t^s(p) \not\perp Z_t,$$

possibly conditional on covariates. If both these conditions hold we can use weather conditions as an instrument.

How do we exploit these assumptions? The traditional approach is to generalize the functional form of the supply function to explicitly incorporate the effect of the instrument on the supply of whiting. In our notation,

$$\ln Q_t^s(p, z) = \alpha^s + \beta^s \times \ln p + \gamma^s \times z + \varepsilon_t^s.$$

The demand function remains unchanged, capturing the fact that demand is not affected by the instrument:

$$\ln Q_t^d(p, z) = \alpha^d + \beta^d \times \ln p + \varepsilon_t^d.$$

We assume that the unobserved components of supply and demand are independent of (or at least uncorrelated with) the weather conditions:

$$\left(\varepsilon_t^d, \varepsilon_t^s\right) \perp Z_t.$$

The equilibrium price P_t^{obs} is the solution for p in the equation

$$Q^d(p, Z_t) = Q_t^s(p, Z_t),$$

leading to:

$$\ln P_t^{\text{obs}} = \frac{\alpha^d - \alpha^s}{\beta^s - \beta^d} + \frac{\varepsilon_t^d - \varepsilon_t^s}{\beta^s - \beta^d} - \frac{\gamma^s \cdot Z_t}{\beta^s - \beta^d},$$

and

$$\ln Q_t^{\text{obs}} = \frac{\beta^s \cdot \alpha^d - \beta^d \cdot \alpha^s}{\beta^s - \beta^d} + \frac{\beta^s \cdot \varepsilon_t^d - \beta^d \cdot \varepsilon_t^s}{\beta^s - \beta^d} - \frac{\gamma^s \cdot \beta^d \cdot Z_t}{\beta^s - \beta^d}.$$

Now consider the expected value of the equilibrium price and quantity given the weather conditions:

$$\mathbb{E}\left[\ln Q_t^{\text{obs}} \middle| Z_t = z\right] = \frac{\beta^s \cdot \alpha^d - \beta^d \cdot \alpha^s}{\beta^s - \beta^d} - \frac{\gamma^s \cdot \beta^d}{\beta^s - \beta^d} \cdot z, \tag{3.9}$$

and

$$\mathbb{E}\left[\ln P_t^{\text{obs}} \middle| Z_t = z\right] = \frac{\alpha^d - \alpha^s}{\beta^s - \beta^d} - \frac{\gamma^s}{\beta^s - \beta^d} \cdot z. \tag{3.10}$$

Equations (3.9) and (3.10) are what is called in econometrics the *reduced form* of the simultaneous equations model. It expresses the *endogenous* variables (those whose values are determined inside the model, price and quantity in this example) in terms of the *exogenous* variables (those whose values are not determined within the model, weather conditions in this example). The slope coefficients on the instrument in these reduced form equations are what in randomized experiments with noncompliance would be called the *intention-to-treat* effects. One can estimate the coefficients in the reduced form by

least squares methods. The key insight is that the ratio of the coefficients on the weather conditions in the two regression functions, $\gamma^s \cdot \beta^d/(\beta^s - \beta^d)$ in the quantity regression and $\gamma^s/(\beta^s - \beta^d)$ in the price regression, is equal to the slope coefficient in the demand function.

For some purposes the reduced-form or intention-to-treat effects may be of substantive interest. In the Fulton fish market example people attempting to predict prices and quantities under the current regime may find these estimates of interest. They are of less interest to policy makers contemplating the introduction of a new tax. In simultaneous equations settings the demand and supply functions are viewed as *structural* in the sense that they are not affected by interventions in the market such as new taxes. As such they, and not the reduced-form regression functions, are the key components of predictions of market outcomes under new regimes. This is somewhat different in many of the recent applications of instrumental variables methods in the statistics literature in the context of randomized experiments with noncompliance where the intention-to-treat effects are traditionally of primary interest.

Let me illustrate this with the Fulton Fish Market data collected by Graddy. For ease of illustration let me simplify the instrument to a binary one: the weather conditions are good for catching fish ($Z_t = 0$, fair weather, corresponding to low wind speed and low wave height) or stormy ($Z_t = 1$, corresponding to relatively strong winds and high waves).[9] The price is the average daily price in cents for one dealer, and the quantity is the daily quantity in pounds. The two estimated reduced forms are

$$\widehat{\ln Q_t}^{\text{obs}} = \underset{(0.08)}{8.63} - \underset{(0.15)}{0.36} \times Z_t.$$

and

$$\widehat{\ln P_t}^{\text{obs}} = - \underset{(0.04)}{0.29} + \underset{(0.07)}{0.34} \times Z_t.$$

Hence the instrumental variables estimate of the slope of the demand function is

$$\hat{\beta}^d = \frac{-0.36}{0.34} = -1.08 \quad \text{(s.e. 0.46)}.$$

[9]The formal definition I use, following Angrist, Graddy and Imbens (2000) is that stormy is defined as wind speed greater than 18knots and wave height more than 4.5ft, and fair weather is anything else.

Another, perhaps more intuitive way of looking at these estimates is to consider the location of the average log quantity and average log price separately by weather conditions. Figure 2 presents the scatter plot of log quantity and log prices, with the stars indicating stormy days and the plus signs indicating calm days. On fair weather days the average log price is -0.29, and the average log quantity is 8.6. On stormy days the average log price is 0.04, and the average log quantity is 8.3. These two loci are marked by by circles in Figure 2. On stormy days the price is higher and the quantity traded is lower than on fair weather days. This is used to estimate the slope of the demand function. The figure also includes the estimated demand function based on using the indicator for stormy days as an instrument for the price.

With the data collected by Graddy it is more difficult to point identify the supply curve. The traditional route towards identifying the supply curve would rely on finding an instrument that shifts demand without directly affecting supply. Without such an instrument we cannot point identify the effect of the introduction of the tax on quantity and prices. It is possible under weaker assumptions to find bounds on these estimands (*e.g.,* Leamer, 1981; Manski 2003), but we do not pursue this here.

3.8 Recent Research on Simultaneous Equations Models

The traditional econometric literature on simultaneous equations models is surveyed in Hausman (1983). Compared to the discussion in the preceeding sections, this literature focuses on a more general case, allowing for multiple endogenous variables and multiple instruments. The modern econometric literature, starting in the 1980s, has relaxed the linearity and additivity assumptions in specification (3.3) substantially. Key references to this literature are Brown (1983), Roehrig (1988), Newey and Powell (2001), Benkard and Berry (2006), Matzkin (2003, 2007), Altonji and Matzkin (2005), Imbens and Newey (2009), Hoderlein and Mammen (2007), Horowitz (2011), and Horowitz and Lee (2007). Matzkin (2007) provides a recent survey of this technically demanding literature. This literature has continued to use the observed outcome notation, making it more difficult to connect to the statistical literature. Here I briefly review some of this literature. The starting point is a structural equation, in the potential outcome notation,

$$Y_i(x) = \alpha + \beta \cdot x + \varepsilon_i,$$

and an instrument Z_i that satisfies

$$Z_i \perp \varepsilon_i, \qquad \text{and} \quad Z_i \not\perp X_i.$$

The traditional econometric literature would formulate this in the observed outcome notation as

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i, \qquad Z_i \perp \varepsilon_i, \qquad \text{and} \quad Z_i \not\perp X_i.$$

There are a number of generalizations considered in the modern literature. First, instead of assuming independence of the unobserved component and the instrument, part of the current literature assumes only that the conditional mean of the unobserved component given the instrument is free of dependence on the instrument, allowing the variance and other distributional aspects to depend on the value of the instrument. See Horowitz (2011). Another generalization of the linear model allows for general nonlinear function forms of the type

$$Y_i = g(X_i) + \varepsilon_i, \qquad Z_i \perp \varepsilon_i, \qquad \text{and} \quad Z_i \not\perp X_i,$$

where the focus is on nonparametric identification and estimation of $g(x)$. See Brown (1983), Roehrig (1988), Benkard and Berry (2006). Allowing for even more generality researchers have studied non-additive versions of these models with

$$Y_i = g(X_i, \varepsilon_i), \qquad Z_i \perp \varepsilon_i, \qquad \text{and} \quad Z_i \not\perp X_i,$$

with $g(x, \varepsilon)$ strictly monotone in a scalar unobserved component ε. In these settings point identification often requires strong assumptions on the support of the instrument and its relation to the endogenous regressor, and therefore researchers have also explored bounds. See Matzkin (2003, 2007, 2008).

4 A Modern Example: Randomized Experiments with Noncompliance and Heterogenous Treatment Effects

In this section I will discuss part of the modern literature on instrumental variables methods that has evolved simultaneously in the statistics and econometrics literature. I

will do so in the context of a second example. On the one hand concern arose in the econometric literature about the restrictiveness of the functional form assumptions in the traditional instrumental variables methods and in particular with the constant treatment effect assumption that were commonly used in the so-called selection models (Heckman, 1979; Heckman and Robb, 1985). The initial results in this literature demonstrated the difficulties in establishing point identification (Heckman, 1990; Manski, 1990), leading to the bounds approach developed by Manski (1995; 2003). At the same time statisticians analyzed the complications arising from noncompliance in randomized experiments (Robins, 1989) and the merits of encouragement designs (Zelen, 1979; 1990). By adopting a common framework and notation, these literatures have become closely connected and influenced each other substantially.

4.1 The McDonald and Tierney (1992) Data

The canonical example in this literature is that of a randomized experiment with noncompliance. I will use here the application in Hirano, Imbens, Rubin and Zhou (2000) to illustrate the issues. Hirano, Imbens, Rubin and Zhou re-analyze data previously analyzed by McDonald and Tierney (1992). McDonald and Tierney (1992) carried out a randomized experiment to evaluate the effect of an influenza vaccination on flu-related hospital visits. Instead of randomly assigning individuals to receive the vaccination, the researchers randomly assigned physicians to receive letters reminding them of the upcoming flu season and encouraging them to vaccinate their patients. This is what Zelen (1979, 1990) refers to as an *encouragement design*. I discuss this using the potential outcome notation used for this particular set up in Angrist, Imbens and Rubin (1996), and in general sometimes referred to as the Rubin Causal Model (Holland, 1986), although there are important antecedents in Neyman (1923, 1990). I consider two distinct treatments: the first the receipt of the letter, and second the receipt of the influenza vaccination. Let $Z_i \in \{0, 1\}$ be the indicator for the receipt of the letter, and let $X_i \in \{0, 1\}$ be the indicator for the receipt of the vaccination. We start by postulating the existence of four potential outcomes. Let $Y_i(z, x)$ be the potential outcome corresponding to the receipt of letter equal to $Z_i = z$, and the receipt of vaccination equal to $X_i = x$, for $z = 0, 1$ and $x = 0, 1$. In addition we postulate the existence of two potential outcomes corresponding

to the receipt of the vaccination as a function of the receipt of the letter, $X_i(z)$, for $z = 0, 1$. We observe for each unit in a population of size $N = 2861$ the value of the assignment, Z_i, the treatment actually received, $X_i^{\text{obs}} = X_i(Z_i)$ and the potential outcome corresponding to the assignment and treatment received, $Y_i^{\text{obs}} = Y_i(Z_i, X_i(Z_i))$. Table 2 presents the number of individuals for each of the eight values of the triple $(Z_i, X_i^{\text{obs}}, Y_i^{\text{obs}})$ in the McDonald and Tierney data set. It should be noted that the randomization in this experiment is at the physician level. I do not have physician indicators, and therefore ignore the clustering. This will tend to lead to under-estimation of the standard errors.

4.2 Instrumental Variables Assumptions

There are four key of assumptions underlying instrumental variables methods beyond the no-interference assumption or SUTVA, with different versions for some of them. I will introduce these assumptions in this section, and in Section 5 discuss their substantive content in the context of some examples. The first assumption concerns the assignment to the instrument Z_i, in the flu example the receipt of the letter by the physician. The assumption requires that the instrument is as good as randomly assigned:

$$Z_i \perp \Big(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), X_i(0), X_i(1)\Big) \qquad \text{(random assignment).} \quad (4.1)$$

This assumption is often satisfied by design: if the assignment is physically randomized, as the letter in the flu example and as in many of the applications in the statistics literature (*e.g.,* see the discussion in Robins, 1989), it is automatically satisfied. In other applications with observational data, common in the econometrics literature, this assumption is more controversial. It can in those cases be relaxed by requiring it to hold only within subpopulations defined by covariates, assuming the assignment of the instrument is unconfounded:

$$Z_i \perp \Big(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), X_i(0), X_i(1)\Big) \;\Big|\; X_i \qquad \text{(unconfounded assignment given } X_i\text{).}$$

This is identical to the generalization from random assignment to unconfounded assignment in observational studies. Either version of this assumption justifies the causal interpretation of *Intention-To-Treat* (ITT) effects, the comparison of outcomes by assignment to the treatment. In many cases these ITT effects are only of limited interest, however, and this motivates the consideration of additional assumptions that do allow

the researcher to make statements about the causal effects of the treatment of interest. It should be stressed however, that in order to draw inferences beyond ITT effects, additional assumptions will be used. Whether the resulting inferences are credible will depend on the credibility of these assumptions.

The second class of assumptions limits or rules out completely direct effects of the assignment (the receipt of the letter in the flu example) on the outcome, other than through the effect of the assignment on the receipt of the treatment of interest (the receipt of the vaccine). This is the most critical, and typically most controversial assumption underlying instrumental variables methods, sometimes viewed as the defining characteristic of instruments. One way of formulating this assumption is as

$$Y_i(0,x) = Y_i(1,x) \quad \text{for } x = 0,1, \text{ for all } i. \qquad \text{(exclusion restriction)}$$

Robins (1989) formulates this assumption as the requirement that the instrument is "not an independent causal risk factor," (Robins, 1989, p. 119). Under this assumption we can drop the z argument of the potential outcomes and write the potential outcomes without ambiguity as $Y_i(x)$. This assumption is typically a substantive one. In the flu example, one might be concerned that the physician, in response to the receipt of the letter, takes actions that affect the likelihood of the patient getting infected with the flu other than simply administering the flu vaccine. In randomized experiments with noncompliance the exclusion restriction is sometimes made implicitly by indexing the potential outcomes only by the treatment x and not the instrument z (*e.g.*, Zelen, 1990).

There are other, weaker versions of this assumption. Hirano, Imbens, Rubin and Zhou (2000) use a stochastic version of the exclusion restriction that only requires that the distribution of $Y_i(0,x)$ is the same as the distribution of $Y_i(1,x)$. Manski (1990) uses a weaker restriction that he calls a *level set restriction*, which requires that the average value of $Y_i(0,x)$ is equal to the average value of $Y_i(1,x)$. In another approach Manski and Pepper (2004) consider monotonicity assumptions that restrict the sign of $Y_i(1,x) - Y_i(0,x)$ across individuals without requiring that the effects are completely absent.

Imbens and Angrist (1994) combine the random assignment assumption and the exclusion restriction by postulating the existence of a pair of potential outcomes $Y_i(x)$, for

$x = 0, 1$, and directly assuming that

$$Z_i \perp \Big(Y_i(0), Y_i(1)\Big).$$

A disadvantage of this formulation is that it becomes less clear exactly what role randomization of the instrument plays. Another version of this combination of the exclusion restriction and random assignment assumption does not require full independence, but assumes that the conditional mean of $Y_i(0)$ and $Y_i(1)$ given the instrument is free of dependence on the instrument. A concern with such assumptions is that they are functional form dependent: if they hold in levels, they do not hold in logarithms unless full independence holds.

A third assumption that is often used, labelled *monotonicity* by Imbens and Angrist (1994), requires that

$$X_i(1) \geq X_i(0), \quad \text{for all } i, \qquad \text{(monotonicity)},$$

for all units. This assumption rules out the presence of units who always do the opposite of their assignment (units with $X_i(0) = 1$ and $X_i(1) = 0$), and is therefore also referred to as the *no-defiance* assumption (Balke and Pearl, 1995). It is implicit in the latent index models often used in econometric evaluation models (*e.g.,* Heckman and Robb, 1984). In the randomized experiments such as the flu example this assumption is often plausible. There it requires that in response to the receipt of the letter by their physician, no patient is less likely to get the vaccine. Robins (1989) makes this assumption in the context of a randomized trial for the effect of AZT on Aids, and describes the assumption as "often, but not always, reasonable," (Robins, 1989, p. 122).

Finally, we need the instrument to be correlated with the treatment, or the instrument to be *relevant* in the terminology of Staiger and Stock (1997):

$$X_i \not\perp Z_i.$$

In practice we need the correlation to be substantial in order to draw precise inferences. A recent literature on *weak instruments* is concerned with credible inference in settings where this correlation between the instrument and the treatment is weak. See Staiger and Stock (1997).

The random assignment assumption and the exclusion restriction are conveniently captured by the graphical model below, although the monotonicity assumption does not fit in as easily. The unobserved component U has a direct effect on both the treatment X and the outcome Y (captured by arrows from U to X and to Y). The instrument Z is not related to the unobserved component U (captured by the absence of a link between U and Z), and is only related to the outcome Y through the treatment X (as captured by the arrow from Z to X and an arrow from X to Y, and the absence of an arrow between Z and Y).

$$\begin{tikzcd} & U \arrow[dl] \arrow[dr] & \\ Z \arrow[r] & X \arrow[r] & Y \end{tikzcd}$$

I will primarily focus on the case with all four assumptions maintained, random assignment, the exclusion restriction, monotonicity, and instrument relevance, without additional covariates, because this case has been the focus of, or a special case of the focus of, many studies, allowing me to compare different approaches. Methodological studies considering essentially this set of assumptions, sometimes without explicitly stating instrument relevance, and sometimes adding additional assumptions, include Robins (1989), Heckman (1990), Manski (1990), Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996), Robins and Greenland (1996), Balke and Pearl (1995, 1997), Greenland (2000), Hernán and Robins (2006), Robins (1994), Robins and Rotnitzky (2004), Vansteelandt and Goetghebeur (2003), Vansteelandt, Bowden, Babanezhad, and Ghoetghebeur (2011), Hirano, Imbens, Rubin and Zhou (2000), Tan (2006, 2010) and others. Many more studies make the same assumptions in combination with a constant treatment effect assumption.

The modern literature analyzed this setting from a number of different approaches. Initially the literature focused on the inability, under these four assumptions, to identify the average effect of the treatment. Some researchers, including prominently Manski (1990), Balke and Pearl (1995), and Robins (1989), showed that although one could not point-identify the average effect under these assumptions, there was information about the average effect in the data under these assumptions and they derived bounds for it. Another strand of the literature, starting with Imbens and Angrist (1994) and Angrist,

Imbens and Rubin (1996) abandoned the effort to do inference for the overall average effect, and focused on subpopulations for which the average effect could be identified, the so-called compliers. We discuss the bounds approach in the next section (Section 4.3) and the local average treatment effect approach in Sections 4.4-4.6.

4.3 Point Identification versus Bounds

In a number of studies the primary estimand is the average effect of the treatment, or the average effect for the treated:

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right], \qquad \text{and} \quad \tau_t = \mathbb{E}\left[Y_i(1) - Y_i(0)|X_i = 1\right] \tag{4.2}$$

With only the four assumptions, random assignment, the exclusion restriction, monotonicity, and instrument relevance Robins (1989), Manski (1990) and Balke and Pearl (1995) established that the average treatment effect can often not be consistently estimated even in large samples, in other words, that it is often *not point-identified*.

Following this result a number of different approaches have been taken. Heckman (1990) showed that if the instrument takes on values such that the probability of treatment given the instrument can be arbitrarily close to zero and one, then the average effect is identified. This is sometimes referred to as *identification at infinity*. Robins (1989) also formulates assumptions that allow for point identification, focusing on the average effect for the treated, τ_t. These assumptions restrict the average value of the potential outcomes when not observed in terms of average outcomes that are observed. For example, Robins formulates the condition that

$$\mathbb{E}\left[Y_i(1) - Y_i(0)|Z_i = 1, X_i = 1\right] = \mathbb{E}\left[Y_i(1) - Y_i(0)|Z_i = 0, X_i = 1\right],$$

which, in combination with the random assigment and the exclusion restriction, this allows for point identification of the average effect for the treated. Robins also formulates two other assumptions, including one where the effects are proportional to survival rates $\mathbb{E}[Y_i(1)|Z_i = 1, X_i = 1]$ and $\mathbb{E}[Y_i(1)|Z_i = 0, X_i = 1]$ respectively, that also point-identifies the average effect for the treated. However, Robsin questions the applicability of these results by commenting that "it would be hard to imagine that there is sufficient understanding of the biological mechanism ... to have strong beliefs that any of the three conditions ... is more likely to hold than either of the other two" (Robins, 1989, p. 122).

As an alternative to adding assumptions, Robins (1989), Manski (1990), and Balke and Pearl (1995), focused on the question what can be learned about τ or τ_t given these four assumptions that do not allow for point identification. Here I focus on the case where the three assumptions, random assignment, the exclusion restriction, and monotonicity, are maintained (without necessarily instrument relevance holding), although Robins (1989) and Manski (1990) also consider other combinations of assumptions. For ease of exposition I focus on the bounds for the average treatment effect τ under these assumptions, in the case where $Y_i(0)$ and $Y_i(1)$ are binary. Then:

$$\begin{aligned}
&\mathbb{E}[Y_i(1) - Y_i(0)] \in \\
&\quad \Big[-(1 - \mathbb{E}[X_i|Z_i = 1]) \cdot \mathbb{E}[Y_i|Z_i = 1, X_i = 0] + \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \\
&\qquad\qquad + \mathbb{E}[X_i|Z_i = 0] \cdot (\mathbb{E}[Y_i|Z_i = 0, X_i = 1] - 1), \\
&\qquad (1 - \mathbb{E}[X_i|Z_i = 1]) \cdot (1 - \mathbb{E}[Y_i|Z_i = 1, X_i = 0]) + \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \\
&\qquad\qquad + \mathbb{E}[X_i|Z_i = 0] \cdot \mathbb{E}[Y_i|Z_i = 0, X_i = 1]\Big],
\end{aligned}$$

which are known at the *natural bounds*. In this simple setting this is a straightforward calculation. Work by Manski (1995, 2003, 2005, 2008), Robins (1989) and Hernán and Robins (2006) extends the partial identification approach to substantially more complex settings.

For the MacDonald-Tierney flu data the estimated identified set for the population average treatment effect is

$$\mathbb{E}[Y_i(1) - Y_i(0)] \in \Big[-0.24, 0.64,\Big].$$

There is obviously also uncertainty associated with these bounds. There is a growing literature developing methods for establishing confidence intervals for parameters in settings with partial identification. See Imbens and Manski (2004) and Chernozhukov, Hong and Tamer (2007).

4.4 Compliance Types

Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) take a different approach. Rather than focusing on the average effect for the population that is not identified

under the three assumptions given in Section 4.2, they focus on different average causal effects. A first key step in the Angrist-Imbens-Rubin set up is that we can think of four different compliance types defined by the pair of values of $(X_i(0), X_i(1))$, that is, defined by how individuals would respond to different assignments in terms of receipt of the treatment:[10]

$$T_i = \begin{cases} n \text{ (never} - \text{taker)} & \text{if } \; X_i(0) = X_i(1) = 0 \\ c \text{ (complier)} & \text{if } \; X_i(0) = 0, X_i(1) = 1, \\ d \text{ (defier)} & \text{if } \; X_i(0) = 1, X_i(1) = 0 \\ a \text{ (always} - \text{taker)} & \text{if } \; X_i(0) = X_i(1) = 1 \end{cases}$$

Given the existence of deterministic potential outcomes this partitioning of the population into four subpopulations is simply a definition.[11] It clarifies immediately that it will be difficult to identify the average effect of the primary treatment (the receipt of the vaccine) for the entire population: never-takers and always-takers can only be observed exposed to a single level of the treatment of interest, and thus for these groups any point estimates of the causal effect of the treatment must be based on extrapolation.

We cannot infer without additional assumptions the compliance type of any unit: for each unit we observe $X_i(Z_i)$, but the data contain no information about the value of $X_i(1 - Z_i)$. For each unit there are therefore two compliance types consistent with the observed behavior. We can also not identify the proportion of individuals of each compliance type without additional restrictions. The monotonicity assumption implies that there are no defiers. This, in combination with random assignment, implies that we can identify the population shares of the remaining three compliance types. The proportion of always-takers and never-takers are

$$\pi_a = \text{pr}(T_i = a) = \text{pr}(X_i = 1 | Z_i = 0), \qquad \text{and } \; \pi_n = \text{pr}(T_i = n) = \text{pr}(X_i = 0 | Z_i = 1),$$

respectively, and the proportion of compliers is the remainder:

$$\pi_c = \text{pr}(T_i = c) = 1 - \pi_a - \pi_n.$$

For the McDonald-Tierney data these shares are estimated to be

$$\hat{\pi}_a = 0.189, \qquad \hat{\pi}_n = 0.692, \qquad \hat{\pi}_c = 0.119,$$

[10]Frangakis and Rubin (2002) generalize this notion of subpopulations whose membership is not completely observed into their *principal stratification* approach. See also Section 7.2.

[11]Outside of this framework the existence of these four subpopulations would be an assumption.

although, as I discuss in section 5.2, these shares may not be consistent with the exclusion restriction.

4.5 Local Average Treatment Effects

If we also assume that the exclusion restriction holds, Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) show that the *local average treatment effect* or *complier average causal effect* is identified:

$$\tau_{\text{late}} = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| T_i = \text{complier}\right] = \frac{\mathbb{E}\left[Y_i|Z_i = 1\right] - \mathbb{E}\left[Y_i|Z_i = 0\right]}{\mathbb{E}\left[X_i|Z_i = 1\right] - \mathbb{E}\left[X_i|Z_i = 0\right]}. \tag{4.3}$$

The components of the righthand side of this expression can be estimated consistently from a random sample $(Z_i, X_i, Y_i)_{i=1}^N$. For the McDonald-Tierney data this leads to

$$\hat{\tau}_{\text{late}} = -0.125 \quad (s.e\ 0.090)$$

Note that just as in the supply and demand example, the causal estimand is the ratio of the intention-to-treat effects of the letter on hospitalization and of the letter on the receipt of the vaccine. These intention-to-treat effects are

$$\widehat{\text{ITT}}_Y = -0.015 \quad (s.e.\ 0.011) \qquad\qquad \widehat{\text{ITT}}_X = \hat{\pi}_c = 0.119 \quad (s.e.\ 0.016),$$

with the latter equal to the estimated proportion of compliers in the population.

Without the monotonicity assumption, but maintaining the random assignment assumption and the exclusion restriction, the ratio of ITT effects still has a clear interpretation. In that case it is equal to a linear combination average of the effect of the treatment for compliers and defiers:

$$\begin{aligned} &\frac{\mathbb{E}\left[Y_i|Z_i = 1\right] - \mathbb{E}\left[Y_i|Z_i = 0\right]}{\mathbb{E}\left[X_i|Z_i = 1\right] - \mathbb{E}\left[X_i|Z_i = 0\right]} \\ &\quad = \frac{\Pr(T_i = \text{complier})}{\Pr(T_i = \text{complier}) - \Pr(T_i = \text{defier})}\mathbb{E}\left[Y_i(1) - Y_i(0) \middle| T_i = \text{complier}\right] \\ &\qquad - \frac{\Pr(T_i = \text{defier})}{\Pr(T_i = \text{complier}) - \Pr(T_i = \text{defier})}\mathbb{E}\left[Y_i(1) - Y_i(0) \middle| T_i = \text{defier}\right]. \end{aligned} \tag{4.4}$$

This estimand has a clear interpretation if the treatment effect is constant across all units, but if there is heterogeneity in the treatment effects it is a weighted average with some weights negative. This representation shows that if the monotonicity assumption is violated, but the proportion of defiers is small relative to that of compliers, the interpretation of the instrumental variables estimand is not severely impacted.

4.6 Do We Care About the Local Average Treatment Effect?

The local average treatment effect is an unusual estimand. It is an average effect of the treatment for a subpopulation that cannot be identified in the sense that there are no units whom we know to belong to this subpopulation, although there are some units whom we know do not belong to it. A more typical approach is to start an analysis by clearly articulating the object of interest, say the average effect of a treatment for a well-defined population. There may be challenges in obtaining credible estimates of this object of interest, and along the way one may make more or less credible assumptions, but typically the focus remains squarely on the originally specified object of interest.

Here the approach appears to be quite different. We started off by defining unit-level treatment effects for all units. We did not articulate explicitly what the target estimand was. In the MacDonald-Tierney influenza-vaccine application a natural estimand might be the population average effect of the vaccine. Then, apparently more or less by accident, the definition of the compliance types led us to focus on the average effects for compliers. In this example the compliers were defined by the response in terms of the receipt of the vaccine to the receipt of the letter. It appears difficult to argue that this is a substantially interesting group, and in fact no attempt was made to do so.

This type of example has led distinguised researchers both in economics and in statistics to wonder whether and why one should care about the local average treatment effect. Deaton writes "I find it hard to make any sense of the LATE [local average treatment effect]," (Deaton, 2010, p 430). Pearl similarly wonders "Realizing that the population averaged treatment effect (ATE) is not identifiable in experiments marred by noncompliance, they have shifted attention to a specific response type (i.e., compliers) for which the causal effect was identifiable, and presented the latter [the local average treatment effect] as an approximation for ATE. ... However, most authors in this category do not state explicitly whether their focus on a specific stratum is motivated by mathematical convenience, mathematical necessity (to achieve identification) or a genuine interest in the stratum under analysis," (Pearl, 2011, p 3). Freedman writes "In many circumstances, the instrumental-variables estimator turns out to be estimating some data-dependent average of structural parameters, whose meaning would have to be elucidated," (Freedman, 2006, p 700-701). Let me attempt to clear up this confusion. An instrumental

variables analysis is an analysis in a second-best setting. It would have been preferable if one had been able to carry out a well-designed randomized experiment. However, such an experiment was not carried out, and we have noncompliance. As a result we cannot answer all the questions we might have wanted to ask. Specifically, if the noncompliance is substantial, we are limited in the questions we can answer credibly and precisely. More precisely, there is only one subpopulation we can credibly (point-)identify the average effect of the treatment for, namely the compliers.

It may be useful to draw an analogy. Suppose a researcher is interested in evaluating a medical treatment and suppose a randomized experiment had been carried out to estimate the average effect of this new treatment. However, the population of the randomized experiment included only men, and the researcher is interested in the average effect for the entire population, including both men and women. What should the researcher do? I would argue that the researcher should report the results for the men, and acknowledge the limitation of the results for the original question of interest. Similarly in the instrumental variables I see the limitation of the results to the compliers as one that was unintended, but driven by the lack of identification for other subpopulations given the design of the study. This limitation should be acknowledged, but one should not drop the analysis simply because the original estimand cannot be identified. Note that our case with instrumental variables is slightly worse than in the gender example, because we cannot actually identify all individuals with certainty as compliers.

There are alternatives to this view. One approach is to focus solely or primarily on intention-to-treat effects. The strongest argument for that is in the context of randomized experiments with noncompliance. The causal interpretation of intention-to-treat effects is justified by the randomization. As Freedman writes, "Experimental data should therefore be analyzed first by comparing rates or averages, following the intention-to-treat principle. Such comparisons are justified because the treatment and control groups are balanced, within the limits of chance variation, by randomization," (Freedman, 2006, p.701). Even in that case one may wish to also report estimates of the local average treatment effects because they may correspond more closely to the object of ultimate interest. The argument for focusing on intention-to-treat or reduced-form estimates is weaker in other settings. For example, in the Fulton Fish Market demand and supply application the intention-to-treat effects are the effects of weather conditions on prices

and quantities. These effects may be of little substantive interest to policy makers interested in tax policy. The substantive interest for these policy makers is almost exclusively in the *structural* effects of price changes on demand and supply, and reduced form effects are only of interest in sofar as they are informative about those structural effects. Of course one should bear in mind that the reduced form or intention-to-treat effects rely on fewer assumptions.

A second alternative is associated with the partial identification approach by Manski (1990, 1992, 2003, 2008). See also Robins (1989) and Leamer (1981) for antecedents. In this setting that suggests maintaining the focus on the original estimand, say the overall average effect. We cannot estimate that accurately because we cannot estimate the average value of $Y_i(0)$ for always-takers or the average value of $Y_i(1)$ for nevertakers, but we can *bound* the average effect of interest because we know *a priori* that the average value of $Y_i(0)$ for always-takers and the average value of $Y_i(0)$ for nevertakers is restricted to lie in the unit interval. Manski's is a principled and coherent approach. One concern with the approach is that it has often focused on reporting solely these bounds, leading researchers to miss relevant information that is available given the maintained assumptions. Two different data sets may lead to the same bounds even though in one case we may know that the average effect for one subpopulation (the compliers) is positive and statistically significantly different from zero whereas in the other case there need not be any evidence of a non-zero effect for any subpopulation. It would appear to be useful to distinguish between such cases by reporting both the local average treatment effect and the bounds.

5 The Substantive Content of the Instrumental Variables Assumptions

In this section I will discuss the substantive content of the three key assumptions, random assignment, the exclusion restriction, and the monotonicity assumption. I will not discuss here the fourth assumption, instrument relevance. In practice the main issue with that assumption concerns the quality of inferences when the assumption is close to being violated. See Section 7.5 for more discussion, and Staiger and Stock (1997) for a detailed study.

5.1 Unconfoundedness of the Instrument

First, consider the random assignment or unconfoundedness assumption. In a slightly different setting this is a very familiar assumption. Matching methods often rely on random assignment, either unconditionally or conditionally, for their justification.

In some of the leading applications of instrumental variables methods this assumption is satisfied by design, when the instrument is physically randomized. For example, in the draft lottery example (Angrist, 1989), draft priority is used as an instrument for veteran status in an evaluation of the causal effect of veteran status on mortality and earnings. In that case the instrument, the draft priority number was assigned by randomization. Similarly, in the flu example (Hirano, Imbens, Rubin and Zhou, 2001), the instrument for influenza vaccinations, the letter to the physician, was randomly assigned.

In other cases the conditional version of this assumption is more plausible. In the McClellan and Newhouse (1994) study proximity of an individual to a hospital with particular facilities is used as an instrument for the receipt of intensive treatment of acute myocardial infarction. This proximity measure is not randomly assigned, and McClellan and Newhouse use covariates to make the unconfoundedness assumption more plausible. For example, they worry about differences between individuals living in rural versus urban areas. To adjust for such differences they use as one of the covariates the distance to the nearest hospital (regardless of the facilities at the nearest hospital).

A key issue is that although on its own this random assignment or unconfoundedness assumption justifies a causal interpretation of the intention-to-treat effects, it is *not* sufficient for a causal interpretation of the instrumental variables estimand, the ratio of the ITT effects for outcome and treatment.

5.2 The Exclusion Restriction

Second, consider the exclusion restriction. This is the most critical and typically most controversial assumption underlying instrumental variables methods.

First of all, it has some testable implications. See Balke and Pearl (1997), and the recent discussions in Kitagawa (2009) and Ramsahai and Lauritzen (2011). This can be seen most easily in a binary outcome setting. Under the three assumptions, random assignment, the exclusion restriction, and monotonicity, the intention-to-treatment effect

of the assignment on the outcome is the product of two causal effects. First, the average effect of the assignment on the outcome for compliers, and second, the intention-to-treat effect of the assignment on receipt of the treatment, which is equal to the population proportion of compliers. If the outcome is binary, the first factor is between -1 and 1. Hence the intention-to-treat effect of the assignment on the outcome has to be bounded in absolute value by the intention-to-treat effect of the assignment on the receipt of the treatment. This is a testable restriction. If the outcomes are multivalued, there is in fact a range of restrictions implied by the assumptions. However, there exist no consistent tests that will reject the null hypothesis with probability going to one as the sample size increases in all scenarios where the null hypothesis is wrong.

Let us assess these restrictions in the flu example. Because

$$\text{pr}(Y_i = 1, X_i = 0|Z_i = 1) = \text{pr}(Y_i = 1|\text{nevertaker}) \cdot \text{pr}(\text{nevertaker}),$$

and

$$\begin{aligned}\text{pr}(Y_i = 1, X_i = 0|Z_i = 0) &= \text{pr}(Y_i = 1|\text{nevertaker or complier})\cdot\text{pr}(\text{nevertaker or complier})\\ &= \text{pr}(Y_i = 1|\text{nevertaker}) \cdot \frac{\text{pr}(\text{nevertaker})}{\text{pr}(\text{nevertaker or complier})}\\ &\quad +\text{pr}(Y_i = 1|\text{complier}) \cdot \frac{\text{pr}(\text{complier})}{\text{pr}(\text{nevertaker or complier})}.\end{aligned}$$

it follows that

$$\text{pr}(Y_i = 1, X_i = 0|Z_i = 1) \leq \text{pr}(Y_i = 1, X_i = 0|Z_i = 0). \tag{5.1}$$

There are three more restrictions in this setting with a binary outcome, binary treatment and binary instrument. See Balke and Pearl (1997) and Richardson, Evans and Robins (2011) for details. For the flu data, the simple frequency estimator for the left hand side of (5.1) is $30/1389 = 0.0216$, and the right hand side is $31/72 = 0.0211$, leading to a slight violation as pointed out in Richardson, Evans and Robins (2011) and Imbens and Rubin (2014). Although not statistically significant, it shows that these restrictions can be important in practice.

To assess the plausibility of the exclusion restriction it is often helpful to do so separately in subpopulations defined by compliance status. Let us first consider the exclusion

restriction for always-takers, who would receive the influenza vaccine irrespective of the receipt of the letter by their physician. Presumably such patients are generally at higher risk for the flu. Why would such patients be affected by a letter warning their physicians about the upcoming flu season when they will get innoculated irrespective of this warning? It may be that the letter led the physician to take other actions beyond giving the flu vaccine, such as encouraging the patient to avoid exposure. These other actions may affect health outcomes, in which case the exclusion restriction would be violated. The exclusion restriction for never-takers has different content. These patients would not receive the vaccine in any case. If their physicians did not regard the risk of flu as sufficiently high to encourage their patients to have the vaccination, presumably the physician would not take other actions either. For these patients the exclusion restriction may therefore be reasonable.

Consider the draft lottery example. In that case the always-takers are individuals who volunteer for military service irrespective of their draft priority number. It seems plausible that the draft priority number has no causal effect on their outcomes. never-takers are individuals who do not serve in the military irrespective of their draft priority number. If this is for medical reason, or more generally reasons that make them ineligible to servce this seems plausible. If, on the other hand these are individuals fit but unwilling to serve they may have had to take actions to stay out of the military that could have affected their subsequent civilian labor market careers. Such actions may include extending their educational career, or temporarily leaving the country. Note that these issues are not addressed by the random assignment of the instrument.

In general, the concern is that the instrument creates incentives not only to receive the treatment, but also to take additional actions that may affect the outcome of interest. The nature of these actions may well differ by compliance type. Most important is to keep in mind that this assumption is typically a substantive assumption, not satisfied by design outside of double-blind, single-dose placebo control randomized experiments with non-compliance.

5.3 Monotonicity

Finally consider the monotonicity or no-defiers assumption. Even though this assumption is often the least controversial of the three instrumental variables assumptions, it is still sometimes viewed with suspicion. For example, whereas Robins views the assumption as "often, but not always reasonable," (Robins, 1989, p. 122), Freedman (2006) wonders: "The identifying restriction for the instrumental-variables estimator is troublesome: just why are there no defiers?"(Freedman, 2006, p. 700). In many applications it is perfectly clear why there should be no or at most few defiers. The instrument plays the role of an *incentive* for the individual to choose the active treatment by either making it more attractive to take the active treatment or less attractive to take the control treatment. As long as individuals do not respond perversely to this incentive, monotonicity is plausible with either no or a neglible proportion of defiers in the population. The term incentive is used broadly here: it may be a financial incentive, or the provision of information, or an imperfectly monitored legal requirement, but in all cases something that makes it more likely, at the individual level, that the individual participates in the treatment.

Let us consider some examples. If non-compliance is one-sided, and those assigned to the control group are effectively embargoed from receiving the treatment, monotonicity is automatically satisified. In that case $X_i(0) = 0$, and there are no always-takers or defiers. The example discussed in Sommer and Zeger (1991), Imbens and Rubin (1997), and Greenland (2000) fits this set up.

In the flu application introduced in Section 4, the letter to the physician creates an additional incentive for the physician to provide the flu vaccine to a patient, something beyond any incentives the physician may have had already to provide the vaccine. Some individuals may already be committed to the vaccine, irrespective of the letter (the always-takers), and some may not be swayed by the receipt of the letter (the never-takers), and that is consistent with this assumption. Monotonicity only requires that there is no patient, who, if their physician receives the letter, would not take the vaccine, whereas they would have taken the vaccine in the absence of the letter.

Consider a second example, the influential draft lottery application by Angrist (1990) (see also Hearst, Newman, and Hulley, 1986). Angrist is interested in evaluating the effect of military service on subsequent civilian earnings, using the draft priority established by

the draft lottery as an instrument. Monotonicity requires that assigning an individual priority for the draft rather than not, may induce them to serve in the military, or may not affect them, but cannot induce them to switch from serving to not serving in the military. Again that seems plausible. Having high priority for the draft increases the cost of staying out of the military: that may not be enough to change behavior, but it would be unusual if the increased cost of staying out of the military induced an individual to switch from serving in the military to not serving.

As a third example, consider the Permutt and Hebel (1989) study of the effect of smoking on birthweight. Permutt and Hebel use the random assignment to a smoking-cessation program as an instrument for the amount of smoking. In this case the monotonicity assumption requires that there are no individuals who as a causal effect of the assignment to the smoking-cessation program end up smoking more. There may be individuals who continue to smoke as much under either assignment and individuals who reduce smoking as a result of the assignment, but the assumption is that there is nobody who increases their smoking as a result of the smoking-cessation program. In all these examples monotonicity requires individuals not to respond perversely to changes in incentives. Systematic and major violations in such settings seem unlikely.

In other settings the assumption is less attractive. Suppose a program has eligibility criteria that are checked by two administrators. Individuals applying for entry to the program are assigned randomly to one of the two administrators. The eligibility criteria may be interpreted slightly differently by the two administrators, with on average administrator A being more strict than administrator B. Monotonicity requires that anyone admitted by administrator A would also be admitted by administrator B, or *vice-versa*. In this type of setting monotonicity does not appear to be as plausible as it is in the settings where the instrument can be viewed as creating an incentive to participate in the treatment.

The discussion in this section focuses primarily on the case with a binary treatment and a binary instrument. In cases with multivalued treatments the monotonicity can be generalized in two different ways. In both cases it may be less plausible than in the binary case. Let $X_i(z)$ be the potential treatment level associated with the assignment z. One can generalize the monotonicity assumption for the binary instrument case to

this case as

$$X_i(z) \text{ is non} - \text{decreasing in } z, \qquad \text{for all } i \qquad \text{(monotonicity in instrument).}$$

This generalization is used in Angrist and Imbens (1995). It is consistent with the view of the instrument as changing the incentive to participate in the treatment: increasing the incentive cannot decrease the level of the treatment received. Angrist and Imbens show that this assumption has testable implications.

An alternative generalization is

$$\text{if } X_i(z) > X_j(z), \text{ then } X_i(z') \geq X_j(z') \qquad \text{for all } z, z', i, j \qquad \text{(monotonicity in unobservables).}$$

This assumption, referred to as *rank preservation* in Robins (1986), implicitly ranks all units in terms of some unobservables (Imbens, 2006). It assumes this ranking is invariant to the level of the instrument. It implies that if $X_i(z) > X_j(z)$, then it cannot be that $X_j(z') > X_i(z')$. It is equivalent to the "continuous prescribing preference" in Hernán and Robins (2006).

In both cases the special case with a binary treatment is identical to the previously stated monotonicity. In settings with multivalued treatments these assumptions are more restrictive than in the binary treatment case. In the demand and supply example in Section 3with linear supply and demand functions, both the monotonicity in the instrument and monotonicity in the unobservables conditions are satisfied.

6 The Link to the Textbook Discussions of Instrumental Variables

Most textbook discussions of instrumental variables use a framework that is quite different at first sight from the potential outcome set up used in Sections 4 and 5. These textbook discussions (graduate texts include Wooldridge (2002), Angrist and Pischke (2009), Greene (2011), and Hayashi (2000), and introductory undergraduate textbooks include Wooldridge (2008) and Stock and Watson (2010)) are often closer to the simultaneous equations example from Section 3. An exception is Manski (2007) who uses the potential outcome set up used in this discussion. In this section I will discuss the standard textbook set up and relate it to the potential outcome framework and the simultaneous equations set up.

The textbook version of instrumental variables does not explicitly define the potential outcomes. Instead the starting point is a linear regression function describing the relation between the realized (observed) outcome Y_i, the endogenous regressor of interest X_i and other regressors V_i:

$$Y_i^{\text{obs}} = \beta_0 + \beta_1 X_i + \beta_2' V_i + \varepsilon_i. \tag{6.1}$$

These other regressors a well as the instruments are often referred to in the econometric literature as *exogenous* variables. Although this term does not have a well-defined meaning, informally it includes variables that Cox (1992) called *attributes*, as well as potential causes that whose assignment is unconfounded. This set up covers both the demand function setting and the randomized experiment example. Although this equation looks like a standard regression function, that similarity is misleading. Equation (6.1) is not an ordinary regression function in the sense that the first part does *not* represent the conditional expectation of the outcome Y_i given the right hand side variables X_i and V_i. Instead it is what is sometimes called a *structural equation* representing the causal response to changes in the input X_i.

The key assumption in this formulation is that the unobserved component ε_i in this regression function is independent of the exogenous regressors V_i and the instruments Z_i, or, formally

$$\varepsilon_i \perp \Big(Z_i, V_i\Big). \tag{6.2}$$

The unobserved component is *not* independent of the endogenous regressor X_i though. The value of the regressor X_i may be partly chosen by individual i to optimize some objection function as in the noncompliance example, or the result of an equilibrium condition as in the supply and demand model. The precise relation between X_i and ε_i is often not fully specified.

How does this set up relate to the earlier discussion involving potential outcomes? Implicitly there is in the background of this set up a causal, unit-level response function. In the potential outcome notation, let $Y_i(x)$ denote this causal response function for unit i, describing for each value of x the potential outcome corresponding to that level of the treatment for that unit. Suppose the conditional expectation of this causal response

function is linear in x and some exogenous covariates:

$$\mathbb{E}[Y_i(x)|V_i] = \beta_0 + \beta_1 \cdot x + \beta_x' V_i. \tag{6.3}$$

Moreover let us make the (strong) assumption that the difference between the response function $Y_i(x)$ and its conditional expectation does not depend on x, so we can define the residual unambiguously as

$$\varepsilon_i = Y_i(x) - (\beta_0 + \beta_1 \cdot x + \beta_x' V_i)\,,$$

with the equality holding for all x. The residual ε_i is now uncorrelated with V_i by definition. We will assume that it is in fact independent of V_i. Now suppose we have an instrument Z_i such that

$$Y_i(x) \perp Z_i \;\Big|\; V_i.$$

This assumption is, given the linear representation for $Y_i(x)$, equivalent to

$$\varepsilon_i \perp Z_i \;\Big|\; V_i.$$

In combination with the assumption that $\varepsilon_i \perp V_i$, this gives us the textbook version of the assumption given in (6.2). We observe V_i, X_i, the instrument Z_i, and the realized outcome

$$Y_i^{\text{obs}} = Y_i(X_i) = \beta_0 + \beta_1 X_i + \beta_2' V_i + \varepsilon_i,$$

which is the starting point in the econometric textbook discussion (6.1).

This set up is more restrictive than it needs to be. For example, the assumption that the difference between the response function $Y_i(x)$ and its conditional expectation does not depend on x can be relaxed to allow for variation in the slope coefficient,

$$Y_i(x) - Y_i(0) = \beta_1 \cdot x + \eta_i \cdot x,$$

as long as the η_i satisfies conditions similar to those on ε_i. The modern literature (*e.g.,* Matzkin, 2007) discusses such models in more detail.

One key feature of the textbook version is that there is no separate role for the monotonicity assumption. Because the linear model implicitly assumes that the per-unit causal effect is constant across units and levels of the treatment, violations of the

monotonicity assumption do not affect the interpretation of the estimand. A second feature of the textbook version is that the exclusion restriction and the random assigment assumption are combined in (6.2). Implicitly, the exclusion restriction is captured by the absence of Z_i in the equation (6.1), and the (conditional) random assignment is captured by (6.2).

7 Extensions and Generalizations

In this section I will briefly review some of other approaches taken in the instrumental variables literature. Some of these originate in the statistics literature, some in the econometrics literature. They reflect different concerns with the traditional instrumental variables methods, sometimes because of different applications, sometimes because of different traditions in econometrics and statistics. This discussion is not exhaustive. I will focus on highlighting the most interesting developments and provide some references to the relevant literature.

7.1 Model-based Approaches to Estimation and Inference

Traditionally instrumental variables analyses relied on linear regression methods. Additional explanatory variables are incorporated linearly in the regression function. The recent work in the statistics literature has explored more flexible approaches to including covariates. These approaches often involve modelling the conditional distribution of the endogenous regressor given the instruments and the exogenous variables. This is in contrast to the traditional econometric literature which has focused on settings and methods that do not rely on such models.

Robins (1989, 1994), Hernán and Robins (2006), Greenland (2000), Robins and Rotznitzky (2004), and Tan (2010) developed an approach that allow for identification of average treatment effect by adding parametric modelling assumptions. This approach starts with the specification of the *structural mean*, the expectation of $Y_i(x)$. Structural is used here in the same meaning as in the econometric literature. This structural mean can be the conditional mean given covariates, or the marginal mean, labeled the *marginal structural mean*. The specification for this expectation is typically parametric. Then estimating equations for the parameters of these models are developed. In the simple setting

considered here this would typically lead to the same estimators considered already. An important virtue of the method is that it has been extended to much more general settings, in particular with time-varying covariates and dynamic treatment regimes in a series of papers. In other settings it has also led to the development of doubly robust estimators (Robins and Rotnitzky, 2004). A key feature of the models is that the models are robust in a particular sense. Specifically, the estimators for the average treatment effects are consistent irrespective of the misspecification of the model, in the absence of intention-to-treat effects (what they call the conditional ITT null).

Imbens and Rubin (1997a) and Hirano, Imbens, Rubin and Zhou (2000) propose building a parametric model for the compliance status in terms of additional covariates, combined with models for the potential outcomes conditional on compliance status and covariates. Given the monotonicity assumption there are three compliance types, never-takers, always-takers and compliers. A natural model for compliance status given individual characteristics V_i is therefore a trinomial logit model:

$$\text{pr}(T_i = n|V_i = v) = \frac{\exp(v'\gamma_n)}{1 + \exp(v'\gamma_n) + \exp(v'\gamma_n)},$$

$$\text{pr}(T_i = a|V_i = v) = \frac{\exp(v'\gamma_a)}{1 + \exp(v'\gamma_n) + \exp(v'\gamma_n)},$$

and

$$\text{pr}(T_i = c|V_i = v) = \frac{1}{1 + \exp(v'\gamma_n) + \exp(v'\gamma_n)}.$$

With continuous outcomes the conditional outcome distributions given compliance status and covariates may be normal:

$$Y_i(x)|T_i = t, V_i = v \sim \mathcal{N}(\beta'_{tx}v, \sigma^2_{tx}),$$

for $(t, x) = (n, 0), (a, 1), (c, 0), (c, 1)$. With binary outcomes one may wish to use logistic regression models here. This specification defines the likelihood function. Hirano, Imbens, Rubin and Zhou (2000) apply this to the flu data discussed before. Simulations in Richard, Evans and Robins (2011) suggest that the modelling of the compliance status here is key. Specifically they point out that even in the absence of ITT effects there can be biases if the model of the compliance status is misspecified.

Like Hirano, Imbens, Rubin and Zhou (2000), Richardson, Evans and Robins (2011) build parametric model only for the identified distributions. They use them to estimate the bounds, so that the parametric assumptions do not contain identifying information.

Little and Yau (2001) similarly model the conditional expectation of the outcome given compliance status and covariates. In their application there are no always-takers, only never-takers and compliers. Their specification specifies parametric forms for the conditional means given the compliance types and the treatment status:

$$\mathbb{E}[Y_i(0)|T_i = n, V_i = v] = \beta_{n0} + \beta'_{n1}v,$$

$$\mathbb{E}[Y_i(0)|T_i = c, V_i = v] = \beta_{c00} + \beta'_{c01}v,$$

and

$$\mathbb{E}[Y_i(1)|T_i = c, V_i = v] = \beta_{c00} + \beta'_{c11}v.$$

7.2 Principal Stratification

Frangakis and Rubin (2002) generalize the latent compliance type approach to instrumental variables in an important way. Their focus is on the causal effect of a binary treatment on some outcome. However, it is not the average effect of the treatment they are interested in, but the average within a subpopulation. It is the way this subpopulation is defined that creates the complications as well as the connection to instrumental variables. There is a post-treatment variable that may be affected by the treatment. Frangakis and Rubin postulate the existence of a pair of potential outcomes for this post-treatment variable. The subpopulation of interest is then defined by the values for the pair of potential outcomes for this post-treatment variables.

Let us consider two examples. First the randomized experiment with non-compliance. The treatment here is the random assignment. The post-treatment variable is the actual receipt of the treatment. The pair of potential outcomes for this post-treatment variable capture the compliance status. The subpopulation of interest is the subpopulation of compliers.

The second example shows how principal stratification generalizes the instrumental variables set up to other cases. Examples of this type are considered in Zhang, Rubin, and Mealli (2009), Frumento, Mealli, Pacini, and Rubin (2011), and Robins (1986). Suppose

we have a randomized experiment with perfect compliance. The primary outcome is survival after one year. For patients who survive a quality of life measure is observed. We may be interested in the effect of the treatment on quality of life. This is only defined for patients who survive up to one year. The principal stratification approach suggests focusing on the subpopulation or *principal stratum* of patients who survive irrespective of the treatment assignment. Membership in this stratum is not observed, and so we cannot directly estimate the average effect of the treatment on quality of life for individuals in this stratum, but the data are generally still informative about such effects, particularly under monotonicity assumptions.

7.3 Randomization Inference with Instrumental Variables

Most of the work on inference in instrumental variables settings is model-based. After specifying a model relating the treatment to the outcome, the conditional distribution or conditional mean of outcomes given instruments is derived. The resulting inferences are conditional on the values of the instruments. A very different approach is taken in Rosenbaum (1996) and Imbens and Rosenbaum (2004).

Rosenbaum focuses on the distribution for statistics generated by the random assignment of the instruments. In the spirit of the work by Fisher (1925) confidence intervals for the parameter of interest, β_1 in equation (6.3) based on this randomization distribution. Similar to confidence intervals for treatment effects based on inverting conventional Fisher p-values these intervals have exact coverage under the stated assumptions. However, these results rely on arguably restrictive constant treatment effect assumptions.

7.4 Matching and Instrumental Variables

In many observational studies using instrumental variables approaches the instruments are not randomly assigned. In that case adjustment for additional pretreatment variables can sometimes make causal inferences more credible. Even if the instrument is randomly assigned, such adjustments can make the inferences more precise. Traditionally in econometrics these adjustments are based on regression methods. Recently in the statistics literature matching methods have been proposed as a way to do the adjustment for pretreatment variables (Baiocchi, Small, Lorch, and Rosenbaum, 2010).

7.5 Weak Instruments

One concern that has arisen in the econometrics literature is about *weak* instruments. For an instrument to be helpful in estimating the effect of the treatment it not only needs to have no direct effect on the outcome, it also needs to be correlated with the treatment. Suppose this correlation is very close to zero. In the simple case the IV estimator is the ratio of covariances,

$$\hat{\beta}_{1,\text{iv}} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(X_i, Z_i)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Z_i - \overline{Z})}.$$

The distribution of this ratio can be approximated by a normal distribution in large samples, as long as the covariance in the denominator is non-zero in the population. If the population value of the covariance in the denominator is exactly zero the distribution of the ratio $\hat{\beta}_{1,\text{iv}}$ is Cauchy in large samples, rather than normal. The weak instrument literature is concerned with the construction of confidence intervals in the case the covariance is close to zero. Interest in this problem rose sharply after a study by Angrist and Krueger (1991), which remains the primary empirical motivation for this literature. Angrist and Krueger were interested in estimating the causal effect of years of education on earnings. They exploited variation in educational achievement by quarter of birth attributed to differences in compulsory schooling laws. These differences in average years of education by quarter of birth were small, and they attempted to improve precision of their estimators by including interactions of the basic instruments, the three quarter of birth dummies, with indicators for year and state of birth. Bound, Jaeger and Baker (1995) showed that the estimates using the interactions as additional instruments were potentially severely affected by the weakness of the instruments. In one striking analysis they re-estimated the Angrist-Krueger regressions using randomly generated quarter of birth data (uncorrelated with earnings or years of education). One might have expected, and hoped, that in that case one would find an imprecisely estimated effect. Surprisingly, Bound, Jaeger and Baker (1995) found that the confidence intervals constructed by Angrist and Krueger suggested precisely estimated effects for the effect of years of education on earnings. It was subsequently found that with weak instruments the TSLS estimator, especially with many instruments, was biased, and that the standard variance estimator led to confidence intervals with substantial undercoverage (Bound, Jaeger and

Baker, 1995; Staiger and Stock, 1997; Chamberlain and Imbens, 2004).

Motivated by the Bound-Jaeger-Baker findings the weak and many instruments literature focused on point and interval estimators with better properties in settings with weak instruments. Starting with Staiger and Stock (1997) a literature developed to construct confidence intervals for the instrumental variables estimand that remained valid irrespective of the strength of the instruments. A key insight was that confidence intervals based on the inversion of Anderson-Rubin (1948) statistics have good properties in settings with weak instruments. See also Moreira (2003), Andrews and Stock (2007), Kleibergen (2002), and Andrews, Moreira and Stock (2006).

Let us look at the simplest case case with a single endogenous regressor, a single instrument, and no additional regressors, and normally distributed residuals:

$$Y_i(x) = \beta_0 + \beta_1 \cdot x + \varepsilon_i, \qquad \text{with } \varepsilon_i | Z_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right).$$

The Anderson-Rubin statistic is, for a given value of b

$$AR(b) = \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (Z_i - \overline{Z}) \cdot (Y_i - b \cdot X_i) \right)^2 \Bigg/ \left(\frac{1}{N} \sum_{i=1}^{N} (Z_i - \overline{Z})^2 \cdot \hat{\sigma}_\varepsilon^2 \right),$$

where $\overline{Z} = \sum_{i=1}^{N} Z_i / N$, and for some estimate of the residual variance σ_ε^2. At the true value $b = \beta_1$ the AR statistic has in large samples a chi-squared distribution with one degree of freedom. Staiger and Stock (1997) propose constructing a confidence interval by inverting this test statistic:

$$CI^{0.95}(\beta_1) = \{b | AR(b) \leq 3.84\}.$$

The subsequent literature has extended this by allowing for multiple instruments and developed various alternatives, all with the focus on methods that remain valid irrespective of the strength of the instruments. See Andrews and Stock (2007) for an overview of this literature.

7.6 Many Instruments

Another strand of the literature motivated by the Angrist-Krueger study focused on settings with many weak instruments. The concern centered on the Bound-Jaeger-Baker (1995) finding that in a setting similar to the Angrist-Krueger setting using TSLS with

many randomly generated instruments led to confidence intervals that had very low coverage rates.

To analyze this setting Bekker (1995) considered the behavior of various estimators under an asymptotic sequence where the number of instruments increases with the sample size. Asymptotic approximations to sampling distributions based on this sequence turned out to be much more accurate than those based on conventional asymptotic approximations. A key finding in Bekker (1995) is that under such sequences one of the leading estimators, Two-Stage-Least-Squares (TSLS, See the appendix for details) estimator is no longer consistent, whereas another estimator, Limited Information Maximum Likelihood (LIML, again see the appendix for details) estimator remains consistent although the variance under this asymptotic sequence differs from that under the standard sequence. See also Kunitomo (1980), Morimune (1983), Bekker and VanderPloeg (2005) Chamberlain and Imbens (2004), Chao and Swanson (2005), Hahn (2002), Hansen, Hausman and Newey (2008) Kolesár, Chetty, Friedman, Glaeser and Imbens (2011), Van Hasselt (2010).

7.7 Proxies for Instruments

Hernán and Robins (2006) and Chalak (2011) explores settings where the instrument is not directly observed. Instead a proxy variable Z_i^* is observed. This proxy variable is correlated with the underlying instrument Z_i, but not perfectly so. The potential outcomes $Y_i(z, x)$ are still defined in terms of the underlying, unobserved instrument Z_i. The unobserved instrument Z_i satisfies the instrumental variables assumptions, random assignment, the exclusion restriction, and the monotonicity assumption. In addition, the observed proxy Z_i^* satisfies

$$Z_i^* \perp Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), X_i(0), X_i(1) | Z_i.$$

Chalak shows that the ratio of covariances (now no longer the ratio of intention-to-treat effects) still has an interpretion of an average causal effect.

7.8 Regression Discontinuity Designs

Regression Discontinuity (RD) designs attempt to estimate causal effects of a binary treatment in settings where the assignment mechanism is a deterministic function of a

pretreatment variable. In the sharp version of the RD design the assigment mechanism takes the form

$$X_i = \mathbf{1}_{V_i \geq c},$$

for some fixed threshold c: all units with a value for the covariate V_i exceeding c receive the treatment and all units with a value for V_i less than c are in the control group. Under smoothness assumptions it is possible in such settings to estimate the average effect of the treatment for units with a value for the pretreatment variable equal to $V_i \approx c$:

$$\mathbb{E}[Y_i(1) - Y_i(0)|V_i = c] = \lim_{w \uparrow c} \mathbb{E}[Y_i|V_i = w] - \lim_{w \downarrow c} \mathbb{E}[Y_i|V_i = w].$$

These designs were introduced by Thistlewaite and Campbell (1960), and have been used in psychology, sociology, political science, and economics. For example, many educational programs have eligibility criteria that allow for the application of RD methods. See Cook (2008) for a recent historical perspective and Imbens and Wooldridge (2009) for a recent review.

A generalization of the sharp RD design is the *Fuzzy Regression Discontinuity* or FRD design. In this case the probability of receipt of the treatment increases discontinuously at the threshold, but not necessarily from zero to one:

$$\lim_{w \downarrow c} \text{pr}(X_i = 1|V_i = w) \neq \lim_{w \uparrow c} \text{pr}(X_i = 1|V_i = w).$$

In that case it is no longer possible to consistently estimate the average effect of the treatment for all units at the threshold. Hahn, Todd, and Van der Klaauw (2000) demonstrate that there is a close link to the instrumental variables set up. Specifically Hahn, Todd and VanderKlaauw show that one can estimate a local average treatment effect at the threshold. To be precise, one can identify the average effect of the treatment for those who are on the margin of getting the treatment:

$$\mathbb{E}\left[Y_i(1) - Y_i(0) \middle| V_i = c, \lim_{w \uparrow c} X_i(w) = 0, \lim_{w \downarrow c} X_i(w) = 1\right] =$$

$$\frac{\lim_{w \uparrow c} \mathbb{E}[Y_i|V_i = w] - \lim_{w \downarrow c} \mathbb{E}[Y_i|V_i = w]}{\lim_{w \uparrow c} \mathbb{E}[X_i|V_i = w] - \lim_{w \downarrow c} \mathbb{E}[X_i|V_i = w]}.$$

This estimand can be estimated as the ratio of an estimator for the discontinuity in the regression function for the outcome and an estimator for the discontinuity in the regression function for the treatment of interest.

8 Conclusion

In this paper I review the connection between the recent statistics literature on instrumental variables and the older econometrics literature. Although the econometric literature on instrumental variables goes back to the 1920s, until recently it had not made much of an impact on the statistics literature. The recent statistics literature has combined some of the older insights from the econometrics instrumental variables literature with the separate literature on causality, enriching both in the process.

Appendix: Estimation and Inference, Two-Stage-Least-Squares and other Traditional Methods

A.1 Set Up

In this section I will discuss the traditional econometric approaches to estimation and inference in instrumental variables settings. Part of the aim of this section is to provide easier access to the econometric literature and terminology on instrumental variables, and to provide a perspective and context for the recent advances.

The textbook setting is the one discussed in the previous section, where a scalar outcome Y_i is linearly related to a scalar covariate of interest X_i. In addition there may be additional exogenous covariates V_i. The traditional model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2' V_i + \varepsilon_i. \tag{8.1}$$

In addition we have a vector of instrumental variables Z_i, with dimension K.

An important distinction in the traditional econometric literature is between the case with a single instrument ($K = 1$), and the case with more than one instrument $K > 1$). More generally, with more than one endogenous regressor, the distinction is between the case with the number of instruments equal to the number of endogenous regressors and the case with the number of instruments larger than the number of endogenous regressors. In the empirical literature there are few credible examples with more than one endogenous regressor, so I focus here on the case with a single endogenous regressor. The first case, with a single instrument, is referred to as the *just-identified* case, and the second, with multiple instruments and a single endogenous regressor, as the *over-identified* case. In the textbook setting with a linear model and constant coefficients this distinction has motivated different estimators and specification tests. In the modern literature, with its explicit allowance for heterogeneity in the treatment effects, these tests, and the distinction between the various estimators, are of less interest. In the recent statistics literature little attention has been paid to the over-identified case with multiple instruments. An exception is Small (2007).

Obviously it is often difficult in applications to find even a single variable that satisfies the conditions for it to be a valid instrument. This raises the question how relevant the literature focusing on methods to deal with multiple instruments is for empirical practice. There are two classes of applications where multiple instruments could credible arise. First, suppose one has a single continous (or multivalued) instrument that satisfies the instrumental variables assumptions, monotonicity, random assignment and the exclusion restriction. Then any monotone function of the instruments also satisfies these assumptions, and one can use multiple monotone functions of the original instrument as instruments. Second, if one has a single instrument in combination with exogenous covariates, then one can use interactions of the instrument and the covariates to generate additional instruments.

Consider for example the Fulton fish market study by Graddy (1995, 1996). Graddy uses weather conditions as an instrument that affects supply but not demand. Specifically she measures wind speed and wave height, giving her two basic instruments. She also constructs functions of these basic instruments, such as indicators that the wind speed or wave height exceeds some threshold.

A.2 The Just-Identified Case with no Additional Covariates

The traditional approach to estimation in this case is to use what is known in the econometrics literature as the instrumental variables estimator. In the case without additional exogenous covariates the most widely used estimator is simply the ratio of two covariances:

$$\hat{\beta}_{1,\text{iv}} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(X_i, Z_i)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Z_i - \overline{Z})},$$

where $\overline{Y}$, $\overline{Z}$, and $\overline{X}$ denote sample averages. If the instrument Z_i is binary, this is also known as the Wald estimator:

$$\hat{\beta}_{1,\text{iv}} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{X}_1 - \overline{X}_0},$$

where for $z = 0, 1$

$$\overline{Y}_z = \frac{1}{N_z}\sum_{i:Z_i=z} Y_i, \quad \overline{X}_z = \frac{1}{N_z}\sum_{i:Z_i=z} X_i,$$

and $N_1 = \sum_{i=1}^{N} Z_i$ and $N_0 = \sum_{i=1}^{N}(1 - Z_i)$.

One can interpret this estimator in two different ways. These interpretations are useful for motivating extensions to settings with multiple instruments and additional exogenous regressors. First, the *indirect least squares* interpretation. This relies on first estimating separately the two *reduced form regressions*, the regressions of the outcome on the instrument:

$$Y_i = \pi_{10} + \pi_{11} \cdot Z_i + \varepsilon_{1i},$$

and the regression of the endogenous regressor on the instrument:

$$X_i = \pi_{20} + \pi_{21} \cdot Z_i + \varepsilon_{2i}.$$

The indirect least squares estimator is the ratio of the least squares estimates of π_{11} and π_{21}, or $\hat{\beta}_{1,\text{ils}} = \hat{\pi}_{11}/\hat{\pi}_{21}$. Note that in the randomized experiment example where X_i and Z_i are binary, the π_{11} and π_{12} are the *intention-to-treat* effects, with $\hat{\pi}_{11} = \overline{Y}_1 - \overline{Y}_0$ and $\hat{\pi}_{12} = \overline{X}_1 - \overline{X}_0$.

Second, I discuss the two-stage-least-squares interpretation of the instrumental variables estimator. First estimate the reduced form regression of the treatment on the instruments and the exogenous covariates. Calculate the predicted value for the endogenous regressor from this regression:

$$\hat{X}_i = \hat{\pi}_{20} + \hat{\pi}_{21} \cdot Z_i.$$

The estimate the regression of the outcome on the predicted endogenous regressor and the additional covariates,

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \eta_i,$$

by least squares to get the TSLS estimator $\hat{\beta}_{\text{tsls}}$. In this just-identified setting the three estimators for β_1 are numerically identical: $\hat{\beta}_{1,\text{iv}} = \hat{\beta}_{1,\text{ils}} = \hat{\beta}_{1,\text{tsls}}$.

A.3 The Just-Identified Case with Additional Covariates

In most econometric applications the instrument is not physically randomized. There is in those cases no guarantee that the instrument is independent of the potential outcomes.

Often researchers use covariates to weaken the requirement on the instrument to conditional independence given the exogenous covariates. In addition the additional exogenous covariates can serve to increase precision. In that case with additional covariates the estimation strategy changes slightly. The two reduced form regressions now take the form

$$Y_i = \pi_{10} + \pi_{11} \cdot Z_i + \pi_{12}' V_i + \varepsilon_{1i},$$

and the regression of the endogenous regressor on the instrument:

$$X_i = \pi_{20} + \pi_{21} \cdot Z_i + \pi_{22}' V_i + \varepsilon_{2i}.$$

The indirect least squares estimator is again the ratio of the least squares estimates of π_{11} and π_{21}, or $\hat{\beta}_{1,\text{ils}} = \hat{\pi}_{11}/\hat{\pi}_{21}$.

For the two-stage-least-squares estimator we again first estimate the regression of the endogenous regressor on the instrument, now also including the exogenous regressors. The next step is to predict the endogenous covariate:

$$\hat{X}_i = \hat{\pi}_{20} + \hat{\pi}_{21} \cdot Z_i + \hat{\pi}_{22}' V_i.$$

Finally the outcome is regressed on the predicted value of the endogenous regressor and the actual values of the exogenous variables:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2' V_i + \eta_i.$$

The TSLS estimator is again identical to the ILS estimator.

For inference the traditional approach is to assume homoskedasticity of the residuals $Y_i - \beta_0 - \beta_1 X_i - \beta_2' V_i$ with variance σ_ε^2. In large samples the distribution of the estimator $\hat{\beta}_{\text{iv}}$ is approximately normal, centered around the true value β_1. Typically the variance is estimated as

$$\widehat{\mathbb{V}} = \hat{\sigma}_\varepsilon^2 \cdot \left(\begin{pmatrix} 1 \\ \hat{X}_i \\ V_i \end{pmatrix} \begin{pmatrix} 1 \\ \hat{X}_i \\ V_i \end{pmatrix}' \right)^{-1}.$$

See the textbook discussion in Wooldridge (2002).

A.4 The Over-Identified Case

The second case of interest is the overidentified case. The main equation remains

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2' V_i + \varepsilon_i,$$

but now the instrument Z_i has dimension $K > 1$. We continue to assume that the residuals ε_i are independent of the instruments with mean zero and variance σ_ε^2. This case is the subject of a large literature, and many estimators have been proposed. I will briefly discuss two. For a more detailed discussion see Wooldridge (2002).

A.5 Two-Stage-Least-Squares

The TSLS approach extends naturally to the setting with multiple instruments. First estimate the reduced form regression of the endogenous variable X_i on the instruments Z_i and the exogenous variables V_i,

$$X_i = \pi_{20} + \pi_{21}' Z_i + \pi_{22}' V_i + \varepsilon_{2i},$$

by least squares. Next calculate the predicted value,

$$\hat{X}_i = \hat{\pi}_{20} + \hat{\pi}'_{21} Z_i + \hat{\pi}'_{22} V_i.$$

Finally, regress the outcome on the predicted value from this regression:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta'_2 V_i + \eta_i.$$

The fact that the dimension of the instrument Z_i is greater than one does not affect the mechanics of the procedure.

To illustrate this, consider the Graddy Fulton Fish Market data. Instead of simply using the binary indicator stormy/not-stormy as the instrument, we can use the tri-valued weather indicator, stormy/mixed/fair to generate two instruments. This leads to TSLS estimates equal to

$$\hat{\beta}_{1,\text{tsls}} = -1.014 \ (s.e.\ 0.384).$$

A.6 Limited-Information-Maximum-Likelihood

The second most popular estimator in this over-identified setting is the limited-information-maximum-likelihood (LIML) estimator, originally proposed by Anderson and Rubin (1948) in the statistics literature. The likelihood is based on joint normality of the joint endogenous variables, $(Y_i, X_i)'$, given the instruments and exogenous variables (Z_i, V_i):

$$\left(\begin{array}{c} Y_i \\ X_i \end{array} \right) \Bigg| \, Z_i, V_i \sim \mathcal{N} \left(\left(\begin{array}{c} \pi_{10} + \beta_1 \pi'_{21} Z_i + \pi'_{12} V_i \\ \pi_{20} + \pi'_{21} Z_i + \pi'_{22} V_i \end{array} \right), \Omega \right).$$

The LIML estimator can be expressed in terms of some eigen value calculations, so that it is computationally fairly simple, though more complicated than the TSLS estimator which only requires matrix inversion. Although motivated by a normal-distribution-based likelihood function, the LIML estimator is consistent under much weaker conditions, as long as $(\varepsilon_{1i}, \varepsilon_{2i})'$ are independent of (Z_i, V_i) and the model (8.1) is correct with ε_i independent of (Z_i, V_i).

Both the TSLS and LIML estimators are consistent and asymptotically normally distributed with the same variance. In the just-identified case the two estimators are numerically identical. The variance can be estimated as in the just-identified case as

$$\widehat{\mathbb{V}} = \hat{\sigma}^2_\varepsilon \cdot \left(\left(\begin{array}{c} 1 \\ \hat{X}_i \\ V_i \end{array} \right) \left(\begin{array}{c} 1 \\ \hat{X}_i \\ V_i \end{array} \right)' \right)^{-1}.$$

In practice there can be substantial differences between the TSLS and LIML estimators when the instruments are weak (see Section 7.5) or when there are many instruments (see Section 7.6), that is, when the degree of overidentification is high.

For the fish data the LIML estimates are

$$\hat{\beta}_{1,\text{liml}} = -1.016 \ (s.e.\ 0.384).$$

A.6 Testing the Over-Indentifying Restrictions

The indirect least squares procedure does not work well in the case with multiple instruments. The two reduced form regressions are

$$X_i = \pi_{20} + \pi'_{21} Z_i + \pi'_{22} V_i + \varepsilon_{2i},$$

and

$$Y_i = \pi_{10} + \pi_{11}' Z_i + \pi_{12}' V_i + \varepsilon_{1i}.$$

If the model is correctly specified, the K-component vector π_{11} should be equal to $\beta_1 \cdot \pi_{21}$. However, there is nothing in the reduced form estimates that imposes proportionality of the estimates. In principle we can use any element of the K-component vector or ratios $\hat{\pi}_{21}/\pi_{11}$ as an estimator for β_1. If the assumption that ε_{1i} is independent of Z_i is true for each component of the instrument, all estimators will estimate the same object, and differences between them should be due to sampling variation. Comparisons of these K estimators can therefore be used to test the assumptions that all instruments are valid.

Although such tests have been popular in the econometrics literature, they are also sensitive to the other maintained assumptions in the model, notably linearity in the endogenous regressor and the constant effect assumption. In the local-average-treatment-effect set up from Section 4.5, differences in estimators based on different instruments can simply be due to the fact that the different instruments correspond to different populations of compliers.

References

Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *Journal of the American Statistical Association*, 97, 284-292, 231-263.

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231-263.

Andrews, D., M. Moreira, and J. Stock (2006), "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, Vol. 74(3), 715-752.

Andrews, D., and J. Stock, (2007), "Inference with Weak Instruments," *Advances in Economics and Econometrics*, Vol III, Blundel,, Newey and Persson (eds.), 122-173.

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J., K. Graddy and G. Imbens, (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economics Studies* 67(3):499-527.

Angrist, J., and G. Imbens, (1995), "Two–Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, Vol 90, No. 430, 431-442.

Angrist, J., G. Imbens and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, (with discussion) Vol. 91, 444-472.

Angrist, J., and A. Krueger, (1991), "Does Compulsory School Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics*, 106, 979-1014.

Angrist, J., and S. Pischke, (2009), *Mostly Harmless Econometrics*, Princeton University Press, Princeton, NJ.

Arellano, M., (2002), "Sargan's Instrumental Variables Estimation and the Generalized Method of Moments," *Journal of Business and Economic Statistics*, Vol. 20(4): 450-459.

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design," NBER working paper 6600.

Baiocchi, M., D. Small, S. Lorch and P. Rosenbaum, (2010),"Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants," *Journal of the American Statistical Association*, Vol. 105(492): 1285-1296.

Balke, A. and Pearl, J. (1995), "Counterfactuals and policy analysis in struc- tural models," in *Uncertainty in Artificial Intelligence 11*, P. Besnard and S. Hanks (eds.), Morgan Kaufmann, San Francisco, 1118.

Balke, A. and J. Pearl, (1997),"Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, Vol. 92(439): 1171-1176.

Barnow, B.S., G.G. Cain and A.S. Goldberger (1980), "Issues in the Analysis of Selectivity Bias,"in Evaluation Studies , vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Basmann, R. (1963a): "The Causal Interpretation of Non-Triangular Systems of Economic Relations," *Econometrica*, 31(3): 439-448.

Basmann, R. (1963b): "On the Causal Interpretation of Non-Triangular Systems of Economic Relations: A Rejoinder," *Econometrica*, 31(3): 451-453.

Basmann, R. (1965): "Causal Systems and Stability: Reply to R. W. Clower," *Econometrica*, 33(1): 242-243.

Bekker, P. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62 (3), 657–681.

Benkard, L., and S. Berry (2006): "On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on Brown (1983) and Roehrig (1988)," *Econometrica*, 74(5).

Bound, J., D. Jaeger, and R. Baker, (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak,"*Journal of the American Statistical Association*, 90, 443–450.

Bowden, R., and D. Turkington, (1984), *Instrumental Variables* Cambridge University Press.

Brookhart, M., P. Wang, D. Solomon, and S. Schneeweiss (2006), "Evaluating Short-Term Drug Effects Using a Physician- Specific Prescribing Preference as an Instrumental Variable,"*Epidemiology*, Vol. 17(4): 268-275

Brown, B. (1983): "The Identification Problem in Systems Nonlinear in the Variables,"*Econometrica*, 51(1), 175-196.

Card, D., (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, Toronto: University of Toronto Press.

Card, D., (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69, 1127-1160.

Chalak, K., (2011), "Identification of Local Treatment Effects Using a Proxy for an Instrument," unpublished manuscript, Department of Economics, Boston College.

Chamberlain, G., and G. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables." *Econometrica*, 72 (1), 295–306.

Chao, J., and N. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73 (5), 1673–1692.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Christ, C., (1994), "The Cowles Commission's Contributions to Econometrics at Chicago, 1939-1955," *Journal of Economic Literature*, 32(1): 30-59.

Clark, P., and F. Windmeier, (2012), "Instrumental Variables Estimators for Binary Outcomes," *Journal of the American Statistical Association*, Vol. 107(500): 1638-1652

Cochran, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.

Cochran, W., and D. Rubin (1973) "Controlling Bias in Observational Studies: A Review," *Sankhya*, 35, 417-46.

Cook, T., (2008), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, Journal of Econometrics. Vol 142(2):636-654

Cox, D., (1992), "Causality: Some Statistical Aspects," *Journal of the Royal Statistical Society*, series A, 155, Part 2, 291–301.

Crépon, B., E. Duflo, M. Gurgand, M. Rathelot, and P. Zamoray (2012): "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment," Unpublished Manuscript.

Dawid, P., (1984), "Causal Inference from Messy Data, Comment on 'On the Nature and Discovery of Structure'" *Journal of the American Statistical Assocation*, Vol. 79 (385), 22–24.

Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, Vol 48(2): 424455.

Duflo, E., R. Glennester, and M. Kremer, (2007), "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, forthcoming.

Fisher, R. A., (1925), *The Design of Experiments*, 1st ed, Oliver and Boyd, London.

Freedman, D., (2006), "Statistical models for causation: what inferential leverage do they provide?" *Evaluation Review*, 30(6): 691713.

Frangakis, C., and D. Rubin, (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58(1), 21-29.

Frumento P., Mealli F., Pacini B., and Rubin D. (2011), "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data," *Journal of the American Statistical Association*, forthcoming.

Gelman, A. (2009), "A statistician's perspective on Mostly Harmless Econometrics: An Empiricists Companion, by Joshua D. Angrist and Jorn-Steffen Pischke," *The Stata Journal*, .Vol. 9(2): 315-320.

Gelman, A., and J. Hill, (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press, Cambridge.

Gill, R., and J. Robins, J., (2001), "Causal Inference for Complex Longitudinal Data: The Continuous Case," *Annals of Statistics*, 29(6): 1785-1811.

Giraud, G., (2003), "Strategic market games: an introduction,"

Journal of Mathematical Economics, Vol. 39(5-6): 355-375.

Graddy, K., (1995), "Who Pays More? Essays on Bargaining and Price Discrimination," PhD thesis, Department of Economics, Princeton University.

Graddy, K., (1996), "Testing for Imperfect Competition at the Fulton Fish Market," *RAND Journal of Economics*, Vol. 26, No. 1, 75–92.

Greene, W., (2011), *Econometric Analyses*, 7th Edition, Prentice Hall.

Greenland, S., (2000), "An Introducation to Instrumental Variables for Epidemiologists," *international Journal of Epidemiology* Vol. 29: 722-729.

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

Haavelmo, T., (1943): "The Statistical Implications of a System of Simultaneous Equations," *Econometrica* 11(1): 1-12.

Haavelmo, T., (1944): "The Probability Approach in Econometrics," *Econometrica* 12.

Hahn, J., P. Todd, and W. van der Klaauw, (2000), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1): 201-209.

Hausman, J., (1983), "Specification and Estimation of Simultaneous Equations Models," in Grilliches and Intrilligator, (editors) *Handbook of Econometrics*, Vol. 1, North Holland.

Hayashi, F., (2000), *Econometrics* , Princeton University Press.

Hearst, N., Newman, T., and S. Hulley, (1986), "Delayed Effects of the Military Draft on Mortality: A Randomized Natural Experiment," *New England Journal of Medicine*, 314 (March 6), 620-624.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, Vol. 5(4): 475-492.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1): 153-161.

Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review*, Papers and Proceedings, 80, 313-318.

Heckman, J., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs,"(with discussion), *Journal of the American Statistical Association.*, Vol. 84, No. 804, 862-874.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Hernán, M., and J. Robins (2006), "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology*, Vol. 17(4): 360-372

Hendry, D., and M. Morgan, (1992), *The Foundations of Econometric Analysis* Cambridge University Press.

Hillier, G., (1990), "On the Normalization of Structural Equations: Properties of Direction Estimators," *Econometrica*, Vol. 58(5): 1181-1194.

Hirano, K., G. Imbens, D. Rubin, and X. Zhou (2000), "Identification and Estimation of Local Average Treatment Effects," *Biostatistics*, Vol. 1(1), 69-88.

Hoderlein, S., and E. Mammen, (2007), "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, Vol. 75(5): 1513-1518.

Holland, P., (1986), "Statistics and Causal Inference," with discussion, *Journal of the American Statistical Association*, 81, 945-970.

Holland, P., (1988). "Causal Inference, Path Analysis, and Recursive Structural Equations Models,"Chapter 13 in: *Sociological Methodology*, Washington: American Sociological Association.

Holland, P., and D. Rubin, (1983), "On Lords Paradox," *Principles of Modern Psychological Measurement: A Festschrift for Frederick Lord*, Wainer and Messick (eds.). Erlbaum.

Horowitz, J. (2011), "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, Vol. 79(2): 347-394.

Horowitz, J., and S. Lee (2007), "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, Vol. 75(4): 1191-1208.

Imbens, G., (1997): "Book Review of 'The Foundations of Econometric Analysis', by David Hendry and Mary Morgan," *Journal of Applied Econometrics*, Vol. 12, 91-94.

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, Vol. 87, No. 3, 706-710.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., (2006), "Nonadditive Models with Endogenous Regressors," *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Blundell, Newey and Persson (eds.), Cambridge University Press.

Imbens, G., (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, Vol. 48(2): 399-423.

Imbens, G., (forthcoming), "Matching in Practice," *Journal of Human Rescources.*

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., and W. Newey (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, Vol. 77(5): 1481-1512.

Imbens, G., and P. Rosenbaum, (2005), "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education," *Journal of the Royal Statistical Society, Series A - Statistics in Society*, Vol. 168, 109-126.

Imbens, G., and D. Rubin, (1997a), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *Annals of Statistics*, Vol. 25, No. 1, 305–327.

Imbens, G., and D. Rubin, (1997b): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555-574.

Imbens, G., and D. Rubin, (forthcoming), *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*, Cambridge University Press.

Imbens, G., and J. Wooldridge, (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, Vol 47(1): 5-86.

Kitagawa, T., (2009), "Identification Region of the Potential Outcome Distributions under Instrument Independence," manuscript, Department of Economics, UCL.

Kleibergen, F., (2002), "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica* 70(5), 1781-1803.

Lauritzen, S. and T. Richardson, (2002) "Chain graph models and their causal interpretation" (with discussion) *Journal of the Royal Statistical Society, Series B,* Vol. 64: 321-361.

Leamer, E., (1981), "Is it a Demand Curve, or is it a Supply Curve? Partial Identification through inequality constraints," *Review of Economisc and Statistics*, Vol. 63(3): 319-327.

Leamer, E., (1988), "Discussion on Marini, Singer, Glymour, Scheines, Spirtes, and Holland," *Sociological Methodology*, Washington: American Sociological Association, Vol. 18, 485-493.

Little, R., and Rubin, D., (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Little, R., and Yau, L., (1998), "Statistical Techniques for Analyzing Data from Prevention Trials: Treatment of No–Shows Using Rubin's Causal Model," *Psychological Methods*, Vol. 3, No. 2, 147-159.

McClellan, M., and J. P. Newhouse, (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality," *Journal of the American Medical Association*, Vol 272, No 11, 859–866.

McDonald, C., Hiu, S., and Tierney, W., (1992), "Effects of Computer Reminders for Influenza Vaccination on Morbidity during Influenza Epidemics," *MD Computing*, 9, 304–312.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. and D. Nagin, 1998. "Bounding Disagreements about Treatment Effects: a Case Study of Sentencing and Recidivism," *Sociological Methodology,* Vol. 28: 99137.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, Vol 87, no. 417, 25–37.

Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge, Harvard University Press.

Manski, C., (2000a), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, 14(3), 115-136.

Manski, C., (2000b), "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treaetment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415-442.

Manski, C., (2001), "Designing Programs for Heterogenous Populations: The Value of Covariate Information," *American Economic Review Papers and Proceedings*, 91, 103-1-6.

Manski, C., (2002), "Treatment Choice Under Ambiguity Induced by Inferential Problems," *Journal of Statistical Planning and Inference*, 105, 67-82.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., (2004), "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72(4), 1221-1246.

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

Manski, C. (2007), *Identification for Prediction and Decision*, Princeton, Princeton University Press.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Manski, C., and J. Pepper, (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997-1010.

Martens, E. , W. Pestman, A. de Boer, S. Belitser, and O. Klungel (2006), "Instrumental Variables: Application and Limitations," *Epidemiology*, Vol. 17(4): 260-267

Matzkin, R., (2007), "Nonparametric Identification," in Heckman and Leamer (editors) *Handbook of Econometrics*, Vol. 6B, North Holland.

Matzkin, R. (2008): "Identification in Nonparametric Simultaneous Equations Models," *Econometrica*, 76(5), 945-978.

Matzkin, R. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71(5), 1339-1375.

Maztkin, R., and J. Altonji, (2005)): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053-1102.

Moreira, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027-1048.

Morgan, S. and C. Winship, (2007), *Counterfactuals and Causal Inference*, Cambridge University Press, Cambridge.

Neyman, J., (1923, 1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, (with discussion), Vol 5, No 4, 465–480, 1990.

Pearl, J. (1995), "Causal Diagrams for Empirical Research", *Biometrika*, 82, 669–688.

Pearl, J., (2000), *Causality*, Cambridge University Press

Pearl, J., (2011), "Principal Stratification – a Goal or a Tool?," *International Journal of Biostatistics*, Vol. 7, No. 1, p. 1-13.

Permutt, T., and Hebel, J., (1989), "Simultaneous–Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight," *Biometrics*, 45, 619–622.

Philipson, T., (1997a), "The Evaluation of New Health Care Technology: The Labor Economics of Statistics," *Journal of Econometrics* 76(1-2): 375-396.

Philipson, T., (1997b), "Data Markets and the Production of Surveys," *Review of Economic Studies* 64(1): 47-73.

Philipson, T., and L. Hedges, (1998), "Subject Evaluation in Social Experiments," *Econometrica* 66(2): 381-408.

Philipson, T., and J. DeSimone, (1997), "Experiments and Subject Sampling." *Biometrika* 84(3): 618-632.

Plott, C., and V. Smith, (1987), "An Experimental Examination of Two Exchange Institutions," *The Review of Economic Studies*, Vol. 45(1): 133-153.

Pratt, J., and Shlaifer, (1984), "On the Nature and Discovery of Structure," *Journal of the American Statistical Association* vol. 79(385): 9-21.

Poirier, D., (1994) "The Methodology of Econometrics," *The International Library of Critical Writings in Econometrics* Vol 6, Elgar Publishing Limited, Aldershot.

Ramsahai, R. and S. Lauritzen (2011): "Likelihood Analysis of the Binary Instrumental Variables Model," *Biometrika*, Vol 98(4): 987-994.

Richardson, T., R. Evans, and J. Robins, (2011) "Transparent parametrizations of models for potential outcomes,"in Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith and West (Eds.) *Bayesian Statistics,* Vol 9.

Richardson, T., and J. Robins, (2013) "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality," Working Paper 128, Center for Statistics and the Social Sciences, Universit of Washington.

Robins, James M. (1986), "A new approach to causal Inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7:1393-1512.

Robins, James M. (1989), "The Analysis of Randomized and Non–Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," *Health Service Research Methodology: A Focus on AIDS*, edited by L. Sechrest, H. Freeman, and A. Bailey, NCHSR, U.S. Public Health Service.

Robins, J. (1994), "Correcting for Non-compliance in Randomized Trials using Structural Nested Mean Models," *Communications in Statistics*, 23: 2379-2412.

Robins, J., and S. Greenland, (1996), "Comment on: Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91: 456-468.

Robins, J., and A. Rotnitzky, (2004), "Estimation of Treatment Effects in Randomized Trials with non-compliance and Observational Studies," *Biometrika*, 91: 763-783.

Roehrig, R. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56(2), 433-447.

Rosenbaum, P., (1996), "Comment on: 'Identification of Causal Effects Using Instrumental Variables'," *Journal of the American Statistical Association*, Vol. 91, 465-468.

Rosenbaum, P., (2009), *Design of Observational Studies*, Springer Series in Statistics, Springer.

Rosenbaum, P., (2010), *Observational Studies*, Second Edition, Springer Series in Statistics, Springer.

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

Roy, A., (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economics Papers*, 3, 135-146.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-592.

Rubin, D. , (1978), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics*, 6:34–58.

Rubin, D. B., (1990), "Formal Modes of Statistical Inference for Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279-292.

Rubin, D. B., (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473-489-292.

Rubin, D., (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

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

Rubin, D., and N. Thomas (1992), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics*, 20(2): 1079-1093.

Sargan, D., (1958) "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26(3): 393-415.

Shapley, L., and M. Shubik, (1977), "Trade Using One Commodity as a Means of Payment," *Journal of Political Economy*, Vol. 85(5): 937-968.

Small, D. (2007) "Sensitivity analysis for instrumental variables regression with overidentifying restrictions," *Journal of the American Statistical Association*, 102, 10491058.

Smith, V. (1982), "Markets as Economizers of Information: Experimental Examination of the Hayek Hypothesis," *Economic Inquiry*.

Sommer, A., and S. Zeger, (1991), "On Estimating Efficacy from Clinical Trials", *Statistics in Medicine*, Vol. 10, 45–52.

Staiger, D., and J. Stock, (1997),"Instrumental Variables Regression with Weak Instruments," *Econometrica*, Vol 65, 557-586.

Strotz, R. (1963): "Interdependence As a Specification Error (Part II of a Triptych on Causal Chain Systems)," *Econometrica*, Vol. 28(2): 428-442.

Strotz, R., and H. Wold, (1963): "Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems)," *Econometrica*, Vol. 28(2): 417-427.

Strotz, R., and H. Wold, (1965): "The Causal Interpretability of Structural Parameters: A Reply," *Econometrica*, Vol. 31(3): 449-450.

Stock, J., and F. Trebbi, (2003),"Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives*, Vol 17, 177194.

Stock, J., and M. Watson, (2010), *Introduction to Econometrics* , 3rd edition, Addison-Wesley.

Tan, Z. (2006), "Regression and Weighint Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association* 101: 1607-1618.

Tan, Z. (2010), "Marginal and Nested Structural Models Using Instrumental Variables," *Journal of the American Statistical Association* 105: 157-169.

Tinbergen, J. (1930), "Bestimmung und Deuting von Angebotskurven. Ein Beispiel." *Zeitschrift fur Nationalokonomie*, Vol. 1 669-679, translated as "Determination and Interpretation of Supply Curves. An Example,"in *The Foundations of Econometric Analysis*, ed David Hendry and Mary Morgan, p. 233:245.

Thistlewaite, D., and D. Campbell, (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," *Journal of Educational Psychology* 51, 309-317.

Van der Laan, M., and J. Robins, (2003) *Unified Methods for Censored Longitudinal Data and Causality*, Springer, Berlin.

Vansteelandt, S. J. Bowden, M. Babanezhad, and E. Goetghebeur, (2011), "On Instrumental Variables Estimation of Causal Odds Ratios," *Statistical Science*, Vol. 26, No. 3, 403422.

Vansteelandt, S. and E. Goetghebeur, (2003), "Causal Inference with Generalized Structural Mean Models," *Journal of the Royal Statistical Socienty, Series B*, Vol. 65: 817-835.

Wold, H. (1960): "A Generalization of Causal Chain Models (Part III of a Triptych on Causal Chain Systems)," *Econometrica*, Vol 28(2): 443-463.

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

Wooldridge, J., (2008), *Introductory Econometrics* , South-Western College Pub.

Working, E., (1927), "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics*, 41(1) 212–35.

Wright, P., (1928) *The Tariff on Animal and Vegetable Oils*, New York, MacMillan.

Yau, L., and R. Little, (2001), "Inference for the Complier-Average Causal Effect from Longitudinal Data Subject to Noncompliance and Missing Data, with Application to a Job Training Assessment for the Unemployed" *Journal of the American Statistical Association*, Vol. 96, No. 456, 1232-1244.

Zelen, M., (1979), "A New Design for Randomized Clinical Trials," *New England Journal of Medicine*, 300, 1242–1245.

Zelen, M., (1990), "Randomized Consent Designs for Clinical Trials: An Update," *Statistics in Medicine*, Vol. 9, 645–656.

Zhang, J., D. Rubin, and F. Mealli, (2009), "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification" *Journal of the American Statistical Association*, Vol. 104, No. 485, 166-176.

Table 1: Fulton Fish Market Data ($N = 111$)

	number of observations	logarithm of price average	logarithm of price standard deviation	logarithm of quantity average	logarithm of quantity standard deviation
all	111	-0.19	(0.38)	8.52	(0.74)
stormy	32	0.04	(0.35)	8.27	(0.71)
not-stormy	79	-0.29	(0.35)	8.63	(0.73)
stormy	32	0.04	(0.35)	8.27	(0.71)
mixed	34	-0.16	(0.35)	8.51	(0.77)
fair	45	-0.39	(0.37)	8.71	(0.69)

Table 2: Influenza Data ($N = 2861$)

Hospitalized for Flu-Related Reasons Y_i^{obs}	Influenza Vaccine X_i^{obs}	Letter Z_i	Number of Individuals
No	No	No	1027
No	No	Yes	935
No	Yes	No	233
No	Yes	Yes	422
Yes	No	No	99
Yes	No	Yes	84
Yes	Yes	No	30
Yes	Yes	Yes	31

Figure 1: Scatterplot of log prices and log quantities

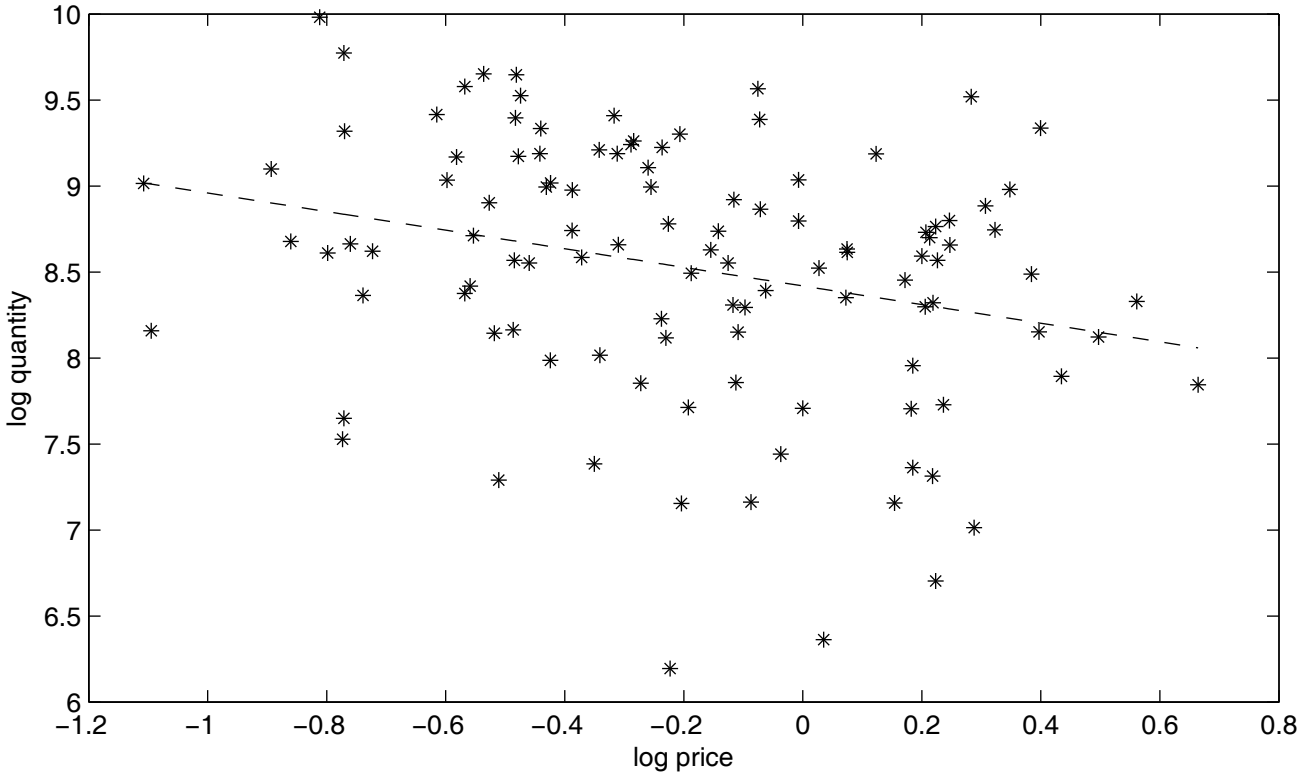
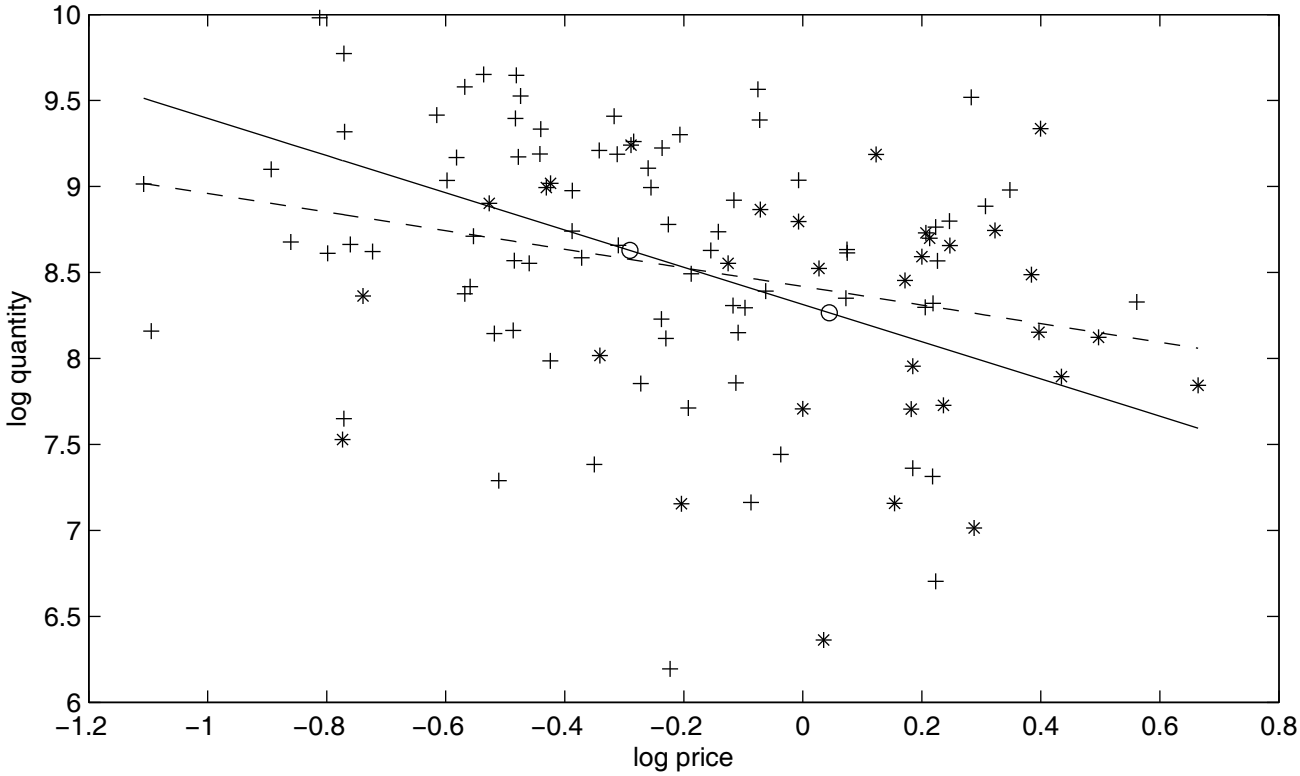
Figure 2: Scatterplot of log prices and log quantities by weather conditions