MEDIATING INSTRUMENTAL VARIABLES

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

1 INTRODUCTION

In an effort to avert confounding in observational studies, economists and social scientists have devised a method called "instrumental variables" [Reiersol, 1945] which is based on the following basic principle. Assume we are interested in estimating the influence of X on Y as given by the structural equation:

$$Y = bX + U \tag{1}$$

where U represents unobserved (zero-mean) disturbances. It is well known that b, the parameter of interest, cannot be estimated by ordinary least square methods unless U is uncorrelated with X. In other words, if we compute the regression of Y on X, r_{yx}, we have

$$r_{yx} = b + r_{xu} \tag{2}$$

hence, if we suspect that r_{xu} is non-zero, and we do not have other means of estimating r_{xu}, we are unable to estimate b. Now, suppose we find a third variable Z that is correlated with X and can safely be assumed to be uncorrelated with U. Under such conditions we can multiply Eq. (1) by Z, take the expectation and obtain an expression for b:

$$b = r_{yz}/r_{xz} \tag{3}$$

This ratio and its various matrix manifestations came to be known as the Instrumental-Variable (IV) estimator [Bowden & Turkington, 1984].

In spite of its general appeal and wide use, the method of instrumental variables suffers from two major drawbacks.

1. The method does not extend naturally to non-linear models. For example, if instead of the linear model of Eq. (1) we were to write:

$$Y = f(X, U) \tag{4}$$

where f is an arbitrary function, then even if we find an instrument Z that is perfectly independent of U we still cannot estimate the target quantity $E_u[f(X, U)]$, unless we specify precisely the functional form of f and the joint distribution of X and U [1].

Nonlinear extensions are needed for dealing with discrete or truncated variables or with mixtures of normal distributions. Analyses of non-parametric models show that, in general, instrumental variables can only produce bounds, rather than point estimates, for the causal effect of X on Y [Robins, 1989, Manski, 1990, Angrist et al, 1993, Balke & Pearl, 1993].

2. The IV-estimator is highly biased if a slight correlation exists between Z and U [Bartels, 1991] and, since U is unobservable, there is no effective test to reveal such correlation. Undesired correlations between Z and U can emanate from two sources; unspecified factors influencing both Z and Y, and direct influence of Z on Y. Although procedures were developed for minimizing the possibility of selecting inadequate instruments [Wu, 1973] these procedures are only partially successful as they rely on comparisons to other estimators and are uniformative in case of disagreements. All in all, the selection of reliable instrumental variable must be based on subjective judgment about cause-effect relationships in the domain.

This paper does not attempt to correct for shortcomings of the traditional IV method but, rather, to develop a complementary method which can provide unbiased estimates under conditions where the IV method fails. The method relies on finding an auxiliary variable Z' which fulfills radically different conditions than those demanded of Z. In fact, the assumptions underlying our method are almost orthogonal to those of standard instrumental variable, hence, agreement between the estimators produced by the two methods would provide strong evidence for the reliability of the estimation. Another useful feature of the proposed approach is that it extends naturally to nonparametric models, thus requiring only qualitative structural assumptions on the part of the investigator, with no commitment to any functional form.

To illustrate the basic idea behind the proposed method, we will consider again the linear model of Eq. (1) and show how an estimator for b is constructed from correlation coefficients.

We start with $Y = bX + U$ but, instead of seeking a variable Z that is independent of U, i.e., one that has no effect on the interaction between X and Y, we now seek a variable Z' that intercepts, or *mediates* that interaction. Formally, Z' should satisfy the following relationships relative to X and Y

$$Z' = cX + \epsilon_{z'} \tag{5}$$

$$Y = dZ' + U' \tag{6}$$

with $\epsilon_{z'}$ being an exogenous disturbance, independent of X and U'. Since X does not appear in the equation for Y, it has no effect on Y except the one mediated by Z'. The relation of

[1] $E_u[f(X, U)]$ represents the causal effect of X on Y – the nonlinear analog of bX in Eq. (1).

c and d to our original parameter b is clear; eliminating Z' from the equations we get

$$Y = cdX + d\epsilon_{z'} + U' \tag{7}$$

giving $b = cd$. Thus, estimating b amounts to finding estimators for c and d in Eqs. (5) - (6), both of which can be found with relative ease, as shown below.

Multiplying (5) by X and taking expectations gives

$$c = r_{xz'} \tag{8}$$

further multiplying (6) by X, then by Y, taking expectations and eliminating r_{xu}, gives

$$d = \frac{(r_{z'y} - r_{xz'} r_{xy})}{1 - r_{xz'^2}} = \beta_{z'y \cdot x} \tag{9}$$

where $\beta_{z'y\cdot x}$ is the standardized regression coefficient (Betta weight). Thus, we obtain the following consistent estimator for b

$$b = cd = r_{xz'} \beta_{z'y\cdot x} \tag{10}$$

To distinguish this from the standard IV-estimator, we call the formula in Eq. (10) the "Mediating Variable" (MV) estimator.

The intuition behind Eq. (10) can best be seen by adding a third equation for X and representing the correlation between X and Y explicitly through a latent (unobserved) variable U.

$$\begin{aligned} X &= aU + \epsilon_x \\ Z' &= cX + \epsilon_{z'} \\ Y &= dZ' + U + \epsilon_y \end{aligned} \tag{11}$$

with $U, \epsilon_x, \epsilon_{z'}$ and ϵ_y being uncorrelated disturbances. The path diagram corresponding to Eqs. (11), is depicted in Figure 1 below, where, for comparison, we also show the position of a standard instrumental variable Z.

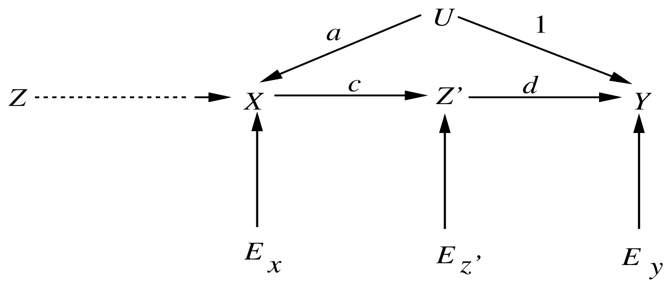

Figure 1

The diagram shows the role of Z' as a mediator between X and Y, while U serves as a confounder which induces spurious correlation between X and Y. By comparison, the standard instrumental variable Z is required to be exogenous relative to the $X \to Y$ interaction.

These diverse roles are reflected in different independence conditions for the two variables, Z and Z'. Whereas Z is required to satisfy the conditions:

$$\begin{array}{ll} 1. & r_{zu} = 0 \\ 2. & r_{zx} \neq 0 \\ 3. & r_{yz \cdot xu} = 0 \end{array} \tag{12}$$

the mediating variable Z' is required to satisfy:

$$\begin{array}{ll} 1'. & r_{uz' \cdot x} = 0 \\ 2'. & r_{z'x} \neq 1 \\ 3'. & r_{yx \cdot uz'} = 0 \end{array} \tag{13}$$

$1'$ requires that the entire correlation between U and Z' be mediated by X, $2'$ requires that Z and X not be perfectly correlated, and $3'$ requires that the influence of X on Y be mediated by Z'. In other words, we now seek an auxiliary variable Z' which transmits, rather than stimulates, the influence we seek to estimate. It should be noted that, since U is unobserved, only assumptions 2 and $2'$ can be tested empirically. All other assumptions, both in the traditional IV approach and the MV approach, rest on judgmental knowledge of the domain which must be considered carefully before instruments are selected.

The structure of the MV formula (10) can be interpreted as a peculiar form of two-steps regression, quite unlike the methods of "two-stage least-square" estimation [Theil, 1954]. In the first stage, a regression of the mediating variable Z' on the explanatory variable X produces the coefficient $r_{z'x}$ and the fitted values $Z' = r_{z'x}X$. In the second stage, instead of regressing the dependent variable Y on the fitted values from the first stage, we now regressed Y on the actual values of Z but only after adjusting for X itself.

In summary, the MV approach possesses two unique features:

1. Instead of basing the estimation on instrumental variables that are exogenous to the relation we wish to estimate, the MV method uses variables that are endogenous to the relation under study. As a result, such variables will not normally be present in the initial phase of the system specification, but would need to be identified and introduced into the analysis by conscience effort. (In our example, it was the inability to estimate b directly which has led us to breakup this structural parameter (b) into finer parameters (c and d) that were not part of the initial specification.)

2. Unlike most IV estimation, the MV approach is generalizable to non-parametric models, that is, it yields a non-parametric unbiased estimation of causal effects in cases where we do not wish to commit to any particular functional form.

These features will be demonstrated in the next sections.

2 An Example: Smoking and the Genotype Theory

Consider the century old debate on the effect of smoking (X) on lung cancer (Y). According to legend, the tobacco industry has managed to stay anti-smoking legislation on the theory that the observed correlation between smoking and cancer could be explained by some sort of carcinogenic genotype (U) which also induces inborn craving for nicotine[2]. The instrumental variable approach to estimating the causal effect of smoking would be to seek a variable z such as "cigarette price" or "smoking advertisement" which is exogenous to the *smoking* $\rightarrow$ *cancer* linkage, then use the IV-ratio (Eq. (3)) to estimate the strength of the linkage. (For non-parametric bounds on this structure, see [Balke & Pearl, 1993]).

The MV-approach would be to seek an endogenous variable, like the amount of tar deposits in a person's lung, which is believed to mediate the linkage *smoking* $\rightarrow$ *cancer*. The adequacy of this variable as an MV instrument rests on whether it is reasonable for us to assume that it meets the conditions listed in Eq. (13). Translated to our settings, one of these conditions ($2'$) states that high levels of tar deposits could be realized not only through cigarette smoking but also through other means, e.g., exposure to environmental pollutants; and may be absent in some smokers. We must also assume that whatever genotype might be acting to aggravate cancer production, it has no effect on the amount of tar-deposits in the lungs, except indirectly, through cigarette smoking ($1'$). Similarly, we must assume that cigarette smoking has no other effect on cancer production except the one mediated through tar deposits ($3'$). These assumptions are represented in the causal network of Figure 2 (adopted from [Spirtes, et al, 1993], page 231). It is identical to the path diagram of Figure 1, but does not show the disturbances and parameters explicitly. In this diagram, as well as in the rest of the paper we have removed the notational distinction between Z and Z'.

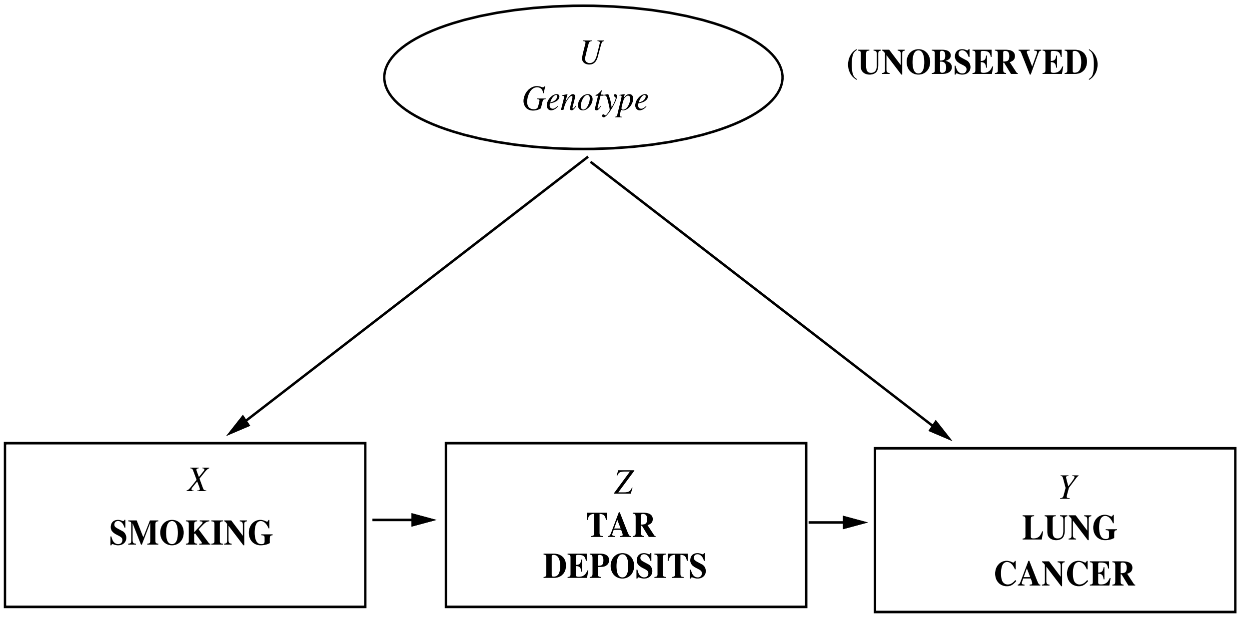

Figure 2

To demonstrate how the MV method assesses the degree to which cigarette smoking increases (or decreases) lung cancer risk, we will assume a hypothetical study in which the following factors were measured simultaneously on a large, randomly selected sample from the population.

1. amount of smoking (X)

[2] For an excellent historical account of this debate, see [Spirtes, et al, 1993, pp. 291-302].

2. amount of tar deposits in the lungs (Z)

3. whether lung cancer has been found (Y)

To simplify the exposition and to demonstrate the application of the method to non-linear models, we will further assume that all three variables are binary, taking on true or false values. A hypothetical data from such study is presented in Table 1. It shows that

	Group Type	$P(x,z)$ Group Size (% of Population)	$P(Y=1\|x,z)$ % of Cancer Cases in Group
$X=0,\ Z=0$	Non-Smokers, No-tar	47.5	10
$X=1,\ Z=0$	Smokers, No-tar	2.5	90
$X=0,\ Z=1$	Non-Smokers, Tar	2.5	5
$X=1,\ Z=1$	Smokers, Tar	47.5	85

Table 1

95% of smokers and only 5% of non-smokers have developed high levels of tar deposits. Moreover, 81% of subjects with tar deposits have developed lung cancer, compared to only 9% among those with no tar deposits. Finally, within each of these two groups (tar and no-tar), smokers show a much higher percentage of cancer than non-smokers. These results, taken at face value, seem to prove conclusively that smoking is a major contributor to cancer. However, the table was especially crafted to tell a different story – smoking would actually decrease, not increase, one's risk of lung cancer.

To convince the reader that this conclusion is inevitable from the table, we first derive a general, non-parametric formula for the causal effect of X on Y, then we apply the formula to the table, and finally we provide an intuitive (though biologically unfounded) explanation of the finding. The point of this exercise is to show that, given the assumptions stated above, non-experimental data of the type shown in Table 1 allow the precise calculation of the actual effect of smoking on cancer.

3 Non-parametric MV-formula

Let X, Y, Z, and U be discrete [3] variables, structured in accordance with the diagram of Figure 2. Formally, the assumptions embedded in the diagram are equivalent to a set of three structural equations,

$$\begin{aligned} X &= f_x(U) \\ Z &= f_z(X, \epsilon_z) \\ Y &= f_y(Z, U) \end{aligned} \tag{14}$$

in which U and ϵ_z are mutually independent disturbances, and f_x, f_y, and f_z arbitrary deterministic functions. These equations are the non-parametric analogue of Eq.(11))

[3] Analogous formulae can be derived in case X, Y, Z, and U represent continuous variables as well as vectors consisting of continuous and discrete variables.

Our task is to compute the causal effect of X on Y, which we denote as [4]

$$P^*(y|x) = \sum_u P(y|x,u)P(u) \tag{15}$$

which stands for the probability of the event $Y = y$ under the condition that X is held constant at x, say by external force. This can be seen by simulating the process of intervention (i.e., setting X to x) on the structural equations in (14), that is, replacing the equation for X by $X = x$, and treating x as a constant. Eq. (15) is clearly different from the familiar conditional probability

$$P(y|x) = \sum_u P(y|x,u)P(u|x) \tag{16}$$

which stands for the probability of $Y = y$ given that event $X = x$ was observed under no intervention. The reason that Eq. (15) invokes the term $P(u)$, and not $P(u|x)$ is clear – holding X fixed (unlike observing $X = x$) provides no information about U (or any other cause of X).

Our task now is to express $P^*(y|x)$ in terms of probabilities of observed variables, namely, eliminate u from the rhs of (15). This can be done by resorting to the two conditional independence assumptions embodied in the structure of Eq. (14) (or Figure 2) which are generalizations of conditions (1') and (3') in Eq. (13):

$$\begin{aligned} P(z|u,x) &= P(z|x) && (17) \\ P(y|x,z,u) &= P(y|z,u) && (18) \end{aligned}$$

These yield the equality

$$P(y|x,z) = \sum_u P(y|x,z,u)P(u|x,z) = \sum_u P(y|z,u)P(u|x) \tag{19}$$

and allow the reduction of Eq. (15) to the desired form:

$$\begin{aligned} P^*(y|x) &= \sum_u \sum_z P(y|x,z,u)P(u)P(z|x) \\ &= \sum_z P(z|x) \sum_u P(y|z,u)P(u) \\ &= \sum_z P(z|x) \sum_u P(y|z,u) \sum_{x'} P(u|x')P(x') \\ &= \sum_z P(z|x) \sum_{x'} [\sum_u P(y|z,u)P(u|x')]P(x') \\ &= \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x') && (20) \end{aligned}$$

Since all factors on the r.h.s of (20) are consistently estimatable from non-experimental data, it follows that $P^*(y|x)$ is estimable as well. Thus, we are in the possession of a general non-parametric estimator for the causal effect of a potential cause X on a potential response

[4] In [Pearl, 1993] I have used the notation $P(y|set(X = x))$ which evoked objections from a number of traditionalists. The lack of mathematical notation for representing interventions is and will continue to be a glaring deficiency in the statistical literature. Still, while I do urge readers to help correct this deficiency, I wish not to offend the guardians of inadequate traditions.

Y, assuming of course that we find a mediating instrument Z that meets the structural conditions set forth in Eqs. (14). Eq. (20) is the non-parametric analog of the correlational estimator shown in Eq. (13). The interpretation of Eq. (20) is transparent once we realize that X is an exogenous instrument relative to Z, and Z is exogenous relative to Y, if only we adjust for X. This allows us to write Eq. (20) as

$$P^*(y|x) = \sum_z P^*(y|z)P^*(z|x) \tag{21}$$

which is aesthetically more appealing. However, the similarity to standard probabilistic formulas should be approached with caution, since the conditioning operator in P^* obeys a different set of syntactic rules than ordinary conditioning [Pearl, 1993].

We summarize this result by a theorem, following a formal definition of the assumptions.

Definition 3.1 *A variable Z is said to be a* **mediating instrument** *relative to an ordered pair of variables (X, Y) if the relation between X, Y, and Z is governed by the structural equations*

$$\begin{aligned} X &= f_x(U) \\ Z &= f_z(X, \epsilon_z) \\ Y &= f_y(Z, U) \end{aligned} \tag{22}$$

in which U and ϵ_z are exogeneous, mutually independent disturbances, such that $P(Z, X) > 0$. (No restrictions are placed on the functions f_x, f_y and f_z.)

Definition 3.2 *The* **causal effect** *of X on Y, written $P^*(y|x)$, is the probability of $Y = y$ that obtains by setting $f_x(U) = x$ in (22) i.e., holding X constant (at x) while keeping the probability of U unaltered.*

Theorem 3.3 *If Z is a mediating instrument relative to (X, Y), then the causal effect of X on Y is given by the formula*

$$P^*(y|x) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{23}$$

Generalizations to more intricate structures of mediating instruments, including multiple Z variables, adjustments for observed covariates, and multi-stage estimations are standard extensions which we do not elaborate in this paper.

4 Smoking Revisited

We wish to return now to the example of Section 2 and apply our formula to the data in Table 1. In particular, we wish to calculate the probability that a randomly selected person will develop cancer under each of the following two actions: choosing to smoke (setting $X = 1$) or choosing to refrain from smoking (setting $X = 0$).

Substituting in Eq. (20) the appropriate values of $P(y|x)$, $P(y|x, z)$ and $P(x)$, gives

$$\begin{aligned} P^*(Y=1|X=1) &= .05(.10 \times .50 + .90 \times .50) + .95(.05 \times .50 + .85 \times .50) \\ &= .05 \times .50 + .95 \times .45 = .4525 \\ P^*(Y=1|X=0) &= .95(.10 \times .50 + .90 \times .50) + .05(.05 \times .50 + .85 \times .50) \\ &= .95 \times .50 + .05 \times .45 = .4975 \end{aligned} \tag{24}$$

Thus, contrary to naive expectations, the data proves smoking to be beneficial to one's health.

This conclusion stands out clearly in the table: If I choose to smoke, then my chances of building up tar deposits are 95%, compared to 5% if I choose not to smoke. To evaluate the effect of tar deposits, we look separately at two groups, smokers and non-smokers, and we take the average, weighted by the proportion of smokers in the population (in our case, 50%). As strange as it might sound, the table shows that tar deposits have beneficial affects on either one of the two groups; in smokers it lowers cancer rates from 90% to 85% and in non-smokers it lowers cancer rates from 10% to 5%. Thus, regardless of whether I do have a natural crave to smoke, I should be seeking the remedial effects of tar deposits, and smoking is my only means of achieving that.

As bizzare as the conclusions are, one must still explain why cancer rates are so much higher among persons with tar deposits (81%) as compared with no tar deposit (9%). This, of course, is perfectly consistent with the old genotype theory; persons with tar deposits have higher rates of cancer simply because those persons tend to have an inborn crave for nicotine which, in turns, is highly indicative of that deadly genotype,....

5 Discussion

The data in Table 1 is obviously unrealistic, as it was purposely crafted to support the genotype theory. In reality, we would expect observational studies involving mediating variables to refute the genotype theory by showing, for example, that mediating consequences of smoking (Z) tend to increase, not decrease, the risk of cancer in both smokers and non-smokers alike. The MV-estimator could then be used for quantifying the causal effect of smoking on cancer.

This example illustrates the difficulty of finding good mediating instruments. To measure such an instrument, we need to penetrate a stable mechanism that is fairly isolated from the rest of the system. In the smoking story, for example, we had to penetrate the anatomy of the lungs and identify a measurable quantity (tar-deposits) that is unaffected by any other disturbance (e.g., U), yet fluctuates in response to its own, internal disturbance mechanism (e.g., pollutants). Such instruments are common in ensembles of loosely coupled subsystems, such as the genetic and anatomical subsystems in our medical example, but might be scarce in socio-economic environment.

The MV-estimator is not very efficient. Its power depends on the number of samples obtained in the exceptional classes ($X = 1$, $Z = 0$) and ($X = 0$, $Z = 1$). The same problem plagues other techniques of minimizing bias by adjusting for observed covariates, and seems to be a universal price we must pay for relying on natural rather than controlled experiments. The contribution proposed in this paper is to add another estimator, based on mediating variables, to the library of available techniques for combating confounding bias. It is hoped that by making this library richer we could improve the capacity to produce accurate assessments of causal effects.

References

[Angrist et al, 1993] Angrist, J.D., Imbens, G.W., and Rubin, D.B., "Identification of causal effects using instrumental variables," Technical Report No. 136, Department of Economics, Harvard University, Cambridge, MA, June 1993.

[Balke & Pearl, 1993] Balke, A. and Pearl, J., 1993. "Nonparametric Bounds on Causal Effects from Partial Compliance Data," UCLA Computer Science Department, Technical Report (R-199), September 1993. Submitted to the *Journal of the American Statistical Association.*

[Bartels, 1991] Bartels, L.M., "Instrumental and 'Quasi-Instrumental' Variables," *American Journal of Political Science*, 35, 777-800, 1991.

[Bowden & Turkington, 1984] Bowden, R.J. and Turkington, D.A., *Instrumental Variables*, Cambridge University Press, Cambridge , MA, 1984.

[Manski, 1990] Manski, C.F., "Nonparametric bounds on treatment effects," *American Economic Review, Papers and Proceedings,* 80, 319-323, May 1990.

[Pearl, 1993] Pearl, J., "Aspects of Graphical Models Connected With Causality," in *Proceedings of the 49th Session of the International Statistical Institute*, Tome LV, Book1, Florence, Italy, 391-401, 1993.

[Reiersol, 1945] Reiersol, O., "Confluence Analysis by Means of Instrumental Sets of Variables," *Arkiv for Mathematik, Astronomi, Och Fysik*, 32A, 1-119, 1945.

[Robins, 1989] Robins, J.M., "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies," in L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A Focus on AIDS*, NCHSR, U.S. Public Health Service, 113-159, 1989.

[Spirtes, et al, 1993] Spirtes, P., Glymour, C., and Schienes, R., *Causation, Prediction, and Search*, Springer-Verlag, New York, 1993.

[Theil, 1954] Theil, H., "Repeated Least Squares Applied to Complete Equation Systems," *Mimeographed*, The Hague: Central Planning Bureau, 1954.

[Wu, 1973] Wu, D-M., "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica*, 41, 733-750, 1973.

ϵ_x $\epsilon_{z'}$ ϵ_y

ϵ_x $\epsilon_{z'}$ ϵ_y