Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 24 Number 3, April 2019 ISSN 1531-7714

Causal Inference Methods for selection on observed and unobserved factors: Propensity Score Matching, Heckit Models, and Instrumental Variable Estimation

Paul Wesley Scott, *University of Pittsburgh*

Two approaches to causal inference in the presence of non-random assignment are presented: The Propensity Score approach which pseudo-randomizes by balancing groups on observed propensity to be in treatment, and the Endogenous Treatment Effects approach which utilizes systems of equations to explicitly model selection into treatment. The three methods based on these approaches that are compared in this study are Heckit models, Propensity Score Matching, and Instrumental Variable models. A simulation is presented to demonstrate these models under different specifications of selection observables, selection unobservables, and outcome unobservables in terms of bias in average treatment effect estimates and size of standard errors. Results show that in most cases Heckit models produce the least bias and highest standard errors in average treatment effect estimates. Propensity Score Matching produces the least bias when selection observables are mildly correlated with selection unobservables and outcome unobservables with outcome and selection unobservables being uncorrelated. Instrumental Variable Estimation produces the least bias in two cases: (1) when selection unobservables are correlated with both selection observables and outcome unobservables, while selection observables are unrelated to outcome unobservables; (2) when there are no relations between selection observables, selection unobservables, and outcome unobservables.

J.S. Mill (1843) formulated that the basic criteria to establish a causal relation required that (a) a cause and effect vary in accordance with one another, i.e. a change in a cause corresponds to a change in effect; (b) a cause temporally precedes an effect in a sequence of events; and, (c) that alternate explanations as to how an event came about can be ruled out, i.e. no other thing could have plausibly produced the effect other than the cause.

The first two criteria are easy to satisfy. We can quantify the relationship between two things by calculating their covariance and through design we can measure variables in subsequent occasions. It is with the third criteria that the complications arise. Ideally, the problem of the latter is solved by establishing a counterfactual (e.g., Morgan & Winship, 2014; Murnane & Willet, 2011; Pearl, 2000; Shadish, et al., 2002). That is, if we could establish a condition where we could observe both outcomes under the condition where the posited cause occurred and a condition where it did not, with all else equal, then we would have some validation in ruling out alternate explanations. In consideration of this C.S. Peirce (1883) began emphasizing the importance of randomization in the context of statistical inference, which was later formalized as an official component of experimental design by R.A. Fisher (1935). The idea underlying randomization is that by assuring each individual has a chance of being assigned to any one of the conditions, then we can consider our groups equal in expectation. This equality in expectations bolsters our ability to rule out alternate explanations and strengthens the claim that changes in the outcome are due to changes in the causal variable (Murnane & Willett, 2011).

Causal inference has a central role in educational research as the implementation of new practices, policies, and interventions require evidence that their implementation will lead to the results they claim. Educational settings are characterized by a multitude of aspects, some of which are observed and some of which are not. Furthermore, within the realm of educational research our data often come in the form of observational studies, which means the data are simply collected from intact samples acting in their own unique environments within their own specific contexts (Shadish, et al., 2002). This is problematic in terms of causal inference, because we are unsure of the selection mechanisms that place individuals into their respective groupings. This lack of clarity defies the idea that such groups are equal in expectation, because we have no evidence for random assignment, moreover, we have evidence to the contrary, thus precluding our readily given counterfactuals as offered via random assignment.

Non-random assignment mechanisms can take on various forms. Two of the most pervasive within social research are (1) individuals self-select into one condition or another; and (2) due to one's particular placement in the world they are more prone to be in certain conditions. An example of mechanism (1) could be an individual choosing to participate in a program for the benefits it offers them, while (2) could be a program that is applied to a community due to its characteristics which may be purposive or convenient. A purposive example of (2) may be a curriculum which is intentionally introduced into one school but not another under the belief that the school which receives the curriculum would get the most benefit from the curriculum. A convenience example of (2) may be introducing a curriculum to a school simply because one has affiliation with that school and not to another school because of a lack of affiliation. In cases where non-random selection into treatment is present, treatment assignment can't be assumed as independent of the expected outcomes since both treatment assignment and outcomes share a mutual dependence on sample characteristics. For example, let's say we wanted to evaluate the effectiveness of a program for increasing college enrollments. If this program was targeted at schools with a record of low college enrollments, it wouldn't be appropriate to compare these schools to others which tend to have relatively high college enrollment rates. In this case, selection into the program and college enrollments following program implementation are confounded by pre-existing differences between schools receiving and those not receiving the program. When such mutual dependence exists amongst treatment assignment and the outcome of interest, we refer to treatment selection as endogenous. The presence of endogenous selection into treatment induces bias into our estimation of treatment effects in the outcome because our groups cannot be considered equal in expectation. Reducing this bias is an important activity for educational researchers in their endeavor to establish causal inference.

Some of the factors underlying these selection mechanisms are observed while some are not, which is to say in some cases we can account for differences between groups through measured variables whereas in other cases we cannot. Researchers have described the former as selection on the observables and the latter as selection on the unobservables (e.g., Greene, 2012; Guo & Fraser, 2010; Heckman & Robb, 1985; Morgan & Winship, 2014; Rosenbaum & Rubin, 1983). Selection on the unobservables is particularly problematic because it is hard to gauge what motivates such individual choices. Selection on the observables is less problematic, since we can account for factors leading to selection, we are better positioned to establish a pseudo-randomization that allows us to conceptualize our groups as equal in expectation. This notion brings us to the concept of constructing counterfactuals to estimate a treatment effect.

Defining counterfactuals and treatment effects is simple under the context of randomization where we can assume groups are equal in expectation (Murnane & Willet, 2011). Since randomization induces statistical independence between outcome and treatment assignment, we conceive of the individual level treatment as simply the hypothetical difference between the value they obtain within treatment versus control (Holland, 1986). Without random assignment we cannot assume independence between treatment selection and outcomes. Two broad frameworks (Guo & Fraser, 2010; Morgan & Winship, 2014) for handling this issue of non-random assignment come from the econometric tradition (Angrist, et al., 1996; Heckman, 2005) and the statistical tradition (Rosenbaum & Rubin, 1983).

Heckman's Scientific Model of Causality and the Neyman-Rubin Statistical Model of Causality

James Heckman (2005) names the econometric approach to causal modeling as scientific to clarify that scientific theory is being invoked in modeling causal

effects. Specifically, he makes an appeal to what in economic theory is known as ex ante and ex post evaluations (Harsanyi,1955; Vickrey,1960), and emphasizes that both types of valuation are incorporated into the theory of his model (Carneiro et al., 2001; Heckman, 2005; Heckman & Navarro, 2004). To clarify, ex ante refers to anticipated returns for an individual due to entering treatment, while ex post pertains to the actual outcomes that follow from being in treatment.

This approach intends to account for the influence of the individual's choice on treatment and outcome by separately modeling ex ante expectations and ex post realizations, then allowing them to relate to one another through their respective unobserved factors. In this way, the econometric model aims to utilize scientific theory (particularly, rational choice theory) to make the selection mechanisms explicit. In order to accomplish this, we incorporate parallel models for selection and outcomes. The account of selection is given in ex ante expectations of returns to participation in treatment, expressed as:

$$E[V(Y(s,\omega),P(s,\omega),C_s(\omega),\omega)|I_\omega], s \in S, \qquad (1)$$

In short, this states that given some set of information, I_ω, a person ω, will evaluate (V) their potential outcomes of taking treatment s, $Y(s,\omega)$, the potential cost of doing so $P(s,\omega)$, and the characteristics of the treatment as known to the subject $C_s(\omega)$. Actual realizations (ex post) are not necessarily known at the time of treatment and we assume that selection is made under some uncertainty in the subjective evaluations of returns and cost of being in treatment. What this allows is for treatment selection to be conditioned on an information set. By allowing for subjective information sets to influence treatment selection, we anticipate that unobserved characteristics are apt to be involved. The factors that enter into the *ex ante* evaluation of treatment selection and those that enter into the ex post realizations can be shared, but by using parallel models we allow for the influence of these factors to be separated out between the selection and outcome models. The modeling of the *ex ante* selection process and the ex post treatment effect in parallel resides in the covariation of the errors from the outcome and selection models. It is through this residual covariance that unobserved selection factors and unobserved influences on the outcome are taken into account.

The primary distinction between the statistical approach and the econometric approach is due to the assumption that given a set of observable features determining selection, the outcome is rendered independent of treatment (Guo & Fraser, 2010), denoted $(Y \perp\!\!\!\perp D)|W$, where Y is the outcome, D the treatment, and W a set of observable variables. This assumption refers to the notion of selection on the observables, which is a core assumption implicit in the Neyman-Rubin model (Guo & Fraser, 2010; Heckman, 2005; Holland, 1986; Neyman, 1923; Rosenbaum & Rubin, 1983). This model comes out of the statistical literature and follows from treatment effect assumptions under randomization, but assumes that it corrects for selection bias by conditioning on observable features amongst groups thus allowing for a pseudo-randomization by covariate balance. This presumed correction allows the statistical model to maintain the assumptions we can make under randomization into treatment. Namely, a set of counterfactuals can be defined in *ex post* outcomes as an averaged treatment effect between treatment and control which is invariant to assignment mechanism, i.e. regardless of how an individual comes to receive treatment the same outcome will result. Also implicit in the statistical model is that social interactions don't influence an individual's outcomes, i.e. their outcome will not be altered due to the composition or size of the group receiving treatment. This point concerning the ignorability of social interactions relates to the stable unit treatment value assumption (SUTVA), which fundamentally expresses that an individual's treatment effect will not depend on that of another (Rubin, 1986).

Because the scientific model is arising directly from social sciences it aims to incorporate the subjective valuations of selection and outcomes thus doesn't maintain SUTVA. The econometric model of causal inference incorporates a choice of treatment model into its identification strategy which derives from characteristics of the treatment for an individual, observed and unobserved costs/preferences in taking up treatment given some information set. In contrast to the statistical approach, which establishes its identification strategy on the pseudo-randomization of subjects to treatment via balancing observed covariates amongst groups.

In the following, we will present three related methods: propensity score methods for covariate balance, which are based in the statistical model; endogenous treatment effect models, which are based in the econometric model; and, instrumental variable estimation, which is also based in the econometric model, but to a weaker extent than the endogenous treatment model. Based on the conceptualizations of the models, a simulation study will be presented to demonstrate their expected performance under varying conditions of observable and unobservable influences on selection and outcomes.

Analytic Models

To start off let's set out some basic terms that will be used throughout. First we have X and Z which are respectively observable determinants of outcome and selection, where they are not distinguished we will use W= (Z, X) to simply denote observable characteristics of a sample. Y is used to represent an outcome variable, D denotes selected/assigned treatment, U denotes unobservable influences on the outcome, and V unobservable influences in the treatment selection mechanism. To illustrate, if we observe that a student's socio-economic status influences their academic achievement, then we would refer to this as an observed influence [X] on an outcome [Y]. Further, if we observe that a student's socio-economic status influences the likelihood that they will participate in a program for improving their academic achievement, then we would refer to this as an observed influence on selection [Z] into treatment [D]. In the case where we make no distinction between the influence of socio-economic status on selection into a program vs. academic achievement, then we refer to this as the joint influence of socio-economic status on both selection into a program and academic achievement [W]. If unobserved parental motivation was increasing the likelihood that a student would participate in a program, then we would refer to this as unobserved influences on selection [V] into treatment [D]. If the unobserved engagement of a student with their studies was influencing their academic achievement, then we would refer to this as an unobserved influence [U] on the outcome [Y]. Graphically, observed variables are represented by boxes, and circles represent latent/unobserved variables, single headed arrows imply a directed regression path, and double headed arrows represent an undirected correlation. Figure 1 and 2 represent selection on the observables and selection on the unobservables respectively.

Figure 1. Graphical Representation of Selection on the observables

Figure 2. Graphical Representation of Selection on the unobservables

These visually represent that in the covariance matrix for the model we have freely estimated covariances amongst these features. Since our unobservables represent the errors in our treatment and outcome, one can gather that when selection on observables is present it implies that such error is attributed to observed variables, namely the observed selection determinants.

Propensity Score Balancing

Within Propensity Score methods no distinction is made between observed features bearing upon treatment versus outcome, hence propensity score methods assume that once observed differences between treatment groups are accounted for then the outcome can be rendered independent of the treatment. For example, once we account for differences in socio-economic status amongst individuals in a program vs. those not in that program, we also account for the influence of socio-economic status on the academic achievement outcomes when evaluating the effects of the program on academic achievement. This is referred to as the conditional independence assumption, i.e. $(Y_1, Y_0) \perp\!\!\!\perp D|P(W)$. This expresses that the outcome within each treatment group is independent of treatment when conditioned on the observables (W). The propensity score P(W) is given as the probability of

being in D=1 given the observed W, it is this conditioning on the propensity score which supports the conditional independence assumptions in the propensity balancing approach. Further, the unobserved factors of selection are independent of the unobserved factors in outcomes within each group when conditioned on the observed W, $U_v \perp\!\!\!\perp (U_1, U_0)|W$. This can be stated as an assumption on the ignorability of unobserved factors. From the above example, this assumption would imply that once we account for socio-economic status all other potential confounding factors can be ignored. An implicit requirement in such an assumption is that the unobservables take on an independent, identical random distribution across treatment groups. One can see that within this framework the observable factors are doing all the work in correcting for selection bias and strong (and often inaccurate) assumptions are being made concerning the unobservables. This strong reliance on the observed variables inspired the phrase "selection on the observables" (Heckman and Robb, 1985). The figure below represents selection on observables as is done within the propensity score framework, within a model covariance matrix representation we find that $Cov(W,U) \neq 0$ and $Cov(W,V) \neq 0$, while $Cov(U,V)=0$.

In application, the steps taken to implement a propensity score method begin with estimating the propensity scores from a selection model, e.g. based on socio-economic status what is the probability that an individual will enter the program or not. Generally, a logistic or probit model will be used to estimate this probability. The propensity score can then be implemented in various ways (Guo & Fraser, 2010). The primary considerations about propensity score methods most useful for our purposes is that (1) propensity score methods function as selection on the observables thus neglect the unobserved characteristics of selection into treatment; and, (2) for propensity methods to succeed we require that adequate and appropriate overlap exist in propensity scores between the two groups. Consideration (2) is most problematic under situations where there is particularly high selectivity into treatment[1] such that groups are highly dissimilar in terms of factors determining selection, and (1) proves problematic in the case where one fails to account for the proper selection factors, i.e. when unobservables are driving selection.

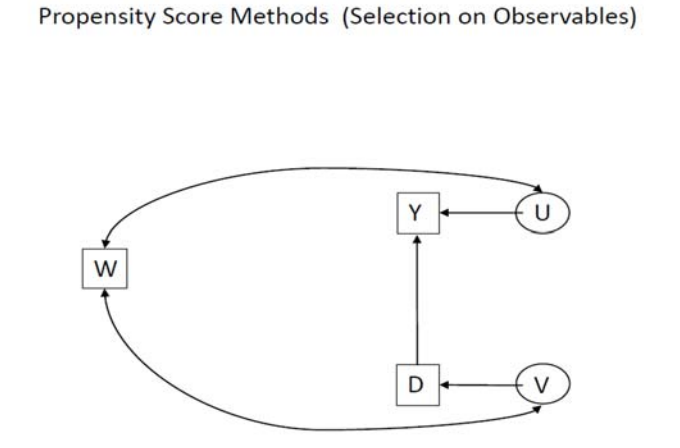

Figure 3 Propensity Score Balancing Methods

Endogenous treatment effects (Heckit Models)

The endogenous treatment effects (Cameron & Trivedi, 2005; Greene, 2000; Woolridge, 2010) model has been dubbed Heckit models to reflect their creation by James Heckman (1976) and as one may have guessed these models relate to the Heckman scientific model of causal inference. These models were originally developed by Heckman to correct for sample selection bias where an observation would be missing unless an individual had selected into a situation. The classic example being the analysis of women's wages in the labor force during the 1970's (Heckman,1976); namely, wages would only be observed for women participating in the labor force, which would be further based on other factors determining whether a woman would enter the workforce or not.

Similar to propensity score methods such as matching, the Heckit method conditions expected outcomes on observable variables and treatment to derive a probability of selection into treatment. The differences, however, being that Heckit methods separate out determinants of selection (Z) from determinants of outcomes (X) allowing any or all predictor variables to be in involved with both X and Z. For example, we would first determine the probability that an individual would participate in a program given their socio-economic status then further consider what is the influence of socio-economic status on academic achievement. Heckit methods invoke a requirement that

[1]This issue was demonstrated in a prior simulation study presented by Scott et al. at AERA 2016; details are available upon request (pws5@pitt.edu)

a model should relate outcome unobservables to choice of treatment to correct sample selection bias. For example, if unobserved student engagement differentially influences those who participate in a program vs. those who do not, then it makes sense to let unmeasured variance vary depending on whether an individual enters the program or not. Conceptually, the outcome expectations within this framework are expressed as:

$$E(Y_1|X,Z,D=1) = \mu_1(X) + E(U_1|X,Z,D=1) \quad (2)$$

$$E(Y_0|X,Z,D=0) = \mu_0(X) + E(U_0|X,Z,D=0) \quad (3)$$

It can be seen here that variation in the expectations of the outcome are due to variations of group specific unobserved factors which are conditioned on the X, Z, and D. This simply presents the unobserved influence on the outcome as the remaining unexplained variance once we account for the observed influences on the outcome [X], the observed determinants of selection [Z], and treatment assignment [D].To represent this as a function of the propensity score we must adopt an assumption that the unobservables are independent of the propensity scores, $(U_1, U_0, U_v) \perp\!\!\!\perp P(X,Z)$. From the example above, this would mean that the student's unobserved engagement and the parent's unobserved motivation influence program selection and academic achievement above and beyond the observed influence of socio-economic status on both selection into the program as well as academic achievement. This assumption allows the formulation of our expectations in terms of propensity scores. First we construct our expectations on the outcome unobservables using the selection unobservables $U_v = D - P(X,Z)$:

$$E(U_1|X,Z,D=1) = E(U_1|U_v \geq -\mu_v(X,Z)) = K_1(P(X,Z)) \quad (4)$$

$$E(U_0|X,Z,D=0) = E(U_0|U_v - \mu_v(X,Z)) = K_0(P(X,Z)) \quad (5)$$

The selection unobservables expresses the unexplained variance after accounting for selection into treatment, this is key as it allows the unexplained variance in selection into treatment to be carried over into the outcome model. The variables denoted by K are what we call control functions, these play a key role in the Heckit methods (Maddala,1983). Control functions operate by modeling endogeneity into the residual terms of the outcome model to control for bias. From the above we recast our expectations to incorporate control functions which depend only on the propensity of selection:

$$E(Y_1|X,Z,D=1) = \mu_1(X) + K_1(P(X,Z)) \quad (6)$$

$$E(Y_0|X,Z,D=0) = \mu_0(X) + K_0(P(X,Z)) \quad (7)$$

These functions imply a more general approach to selection bias where selection is due to both observable and unobservable factors. Because the unobservables are functionally modeled on the basis of the propensity scores, the stronger our prediction of selection is, the more optimally the control functions can perform as they are better informed about the selection mechanisms. From our example above, this implies that once we account for the influence of socio-economic status on selecting into a program and academic achievement, any remaining unobserved factors, such as student engagement and parental motivation, are incorporated into the evaluation of the program's effect on academic achievement.

In this way, the Heckit methods incorporate both selection on the unobservables and observables. Thus we state the selection and error covariance terms in the following ways: *Cov (U, V) ≠ 0*, *Cov (Z, U) ≠ 0*, and *Cov(Z,V) ≠ 0*. This allows, from our example, that parental motivation and student engagement be related [*Cov (U, V)*], as well as socio-economic status being related to both student engagement [*Cov (Z, U)*] and parental motivation [*Cov(Z,V)*]. Figure 4 gives a visualization of such a model.

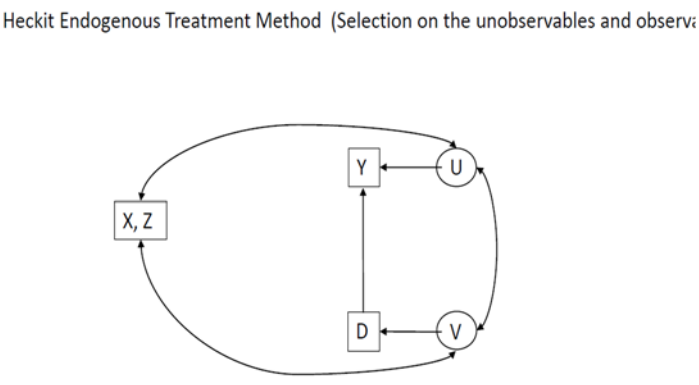

Figure 4. Heckit Endogenous Treatment Method (aka method of Control Functions)

The models used in analysis are formulated as such: we have an outcome model where y is a function of the observed covariates, the odds of selection (δ) into treatment t, and the unobserved residuals,

$$y_j = \boldsymbol{x_j\beta} + \delta t_j + U_{y_j}, \quad (8)$$

where the t_j is conceptualized as resulting from a latent variable which accounts for both observed and unobserved influences on selection into treatment

$$t_j^* = \boldsymbol{z_j\Upsilon} + U_{v_j} \quad (9)$$

in a manner such that:

$$t_j = \begin{cases} 1, & if \quad t_j^* > 0 \\ 0, & otherwise \end{cases} \quad (10)$$

Which expresses, if there are factors influencing selection into treatment, these will be accounted for when evaluating treatment effects on the outcome, and not otherwise.

We further assume that $U_{y_j} \sim N(0, \sigma^2)$ and $U_{v_j} \sim N(0,1)$, with $\rho = Cov(U_{y_j}, U_{v_j})$. ρ serves to indicate the extent to which sample selectivity is of concern and as such establishes the value reflected in δ. In the presence of a ρ=0 we have no evidence for sample selection and our results will reduce to the OLS estimate of the treatment effect. Thus when fitting such models it is essential to check the hypothesis test for whether ρ=0 or not. The Heckit method, as mentioned, is founded in the concept of control functions. Another familiar econometric approach is the instrumental variable method given in the following section. The major distinction between the control function and instrumental variable methods is that the Heckit method handles endogeneity by directly imposing a model onto the error structure such that selection and outcome errors are related to one another while instrumental variables account for error in the treatment selection due to a presumable exogenous source.

Instrumental Variable Estimation (IVE)

The instrumental variable approach (Angrist et al., 1996; Heckman & Vytlacil, 1999; Imbens & Angrist, 1994) also separates out observables except it only concerns the influence of observables on increasing the likelihood of selection into treatment. Note, one can incorporate observables into the outcome model but these should be independent of observable influences on the outcome. In other words, Z is excluded from the outcome model. More specifically, the assumption underlying IVE is that the instrument Z influences the outcome only through the causal variable (i.e., treatment). To exemplify, let's consider the situation where schools will only be invited to participate in a program for increasing college enrollment if they are within so many miles of the organizations main office. In this case, it is fair to assume that miles from an organization would then predict selection into treatment, but not necessarily increases in college enrollments. However, given that the program does indeed boost college enrollments, then by extension miles from the organization would also predict higher enrollments but only by way of increased likelihood of program participation. In the propensity score balancing method and the Heckit control function method, P(X, Z) arises directly from setting D as a function of X & Z (with X,Z=W in the case of propensity balancing) such that P(X,Z)= Pr(D=1|X,Z). Instrumental variable approaches differ in that the P(X, Z) come about as a function of the instrument Z alone, so P(X,Z)= Pr(D=1|Z). The X are held separately as covariates which bear upon outcomes but not on selection into treatment once the instrument is construed in the form of the propensity In terms of the unobservables, we now require an equivalence in the treatment and control group unobservables on the outcomes, U1 = U0, and a correlation between treatment and U0, here we can state that the selection unobservables covary with the outcome unobservables which are unconditioned on Z, i.e. there is selection on the unobservables but the selection only influences the outcome via the relationship between Z and D. The propensity score is now seen as a valid instrument in so far as the instrument explains treatment and, given selection, the outcome unobservables are not conditioned on the propensity instrument:

$$E(U_0|P(X,Z),X) = E(U_0|X) \quad (11)$$

and

$$U_v \perp\!\!\!\perp (U_1 - U_0)|X \quad (12)$$

The proposition given in (12) is simply stating that given X, treatment unobservables are independent of the difference between the outcome unobservables in treatment and control, it is from this assumption that we base the assumption that $U_1 = U_0$. In other words, once we account for miles from an organization, any other confounding influences on increasing college enrollment

will be the same regardless of whether a schools participates in the organizations program or not. The idea is that the instrument is an adequate means of explaining out any unobserved selection bias bearing upon the treatment outcome, hence in the outcome model it is only the control unobservables that are present.

In practice we can use a 2 stage least squares estimation for instrumental variables which follows as such:

$$\text{Stage 1} \quad D = Y_1 Z_i + U_v \quad yielding \quad P(D = 1|Z) \tag{13}$$

$$\text{Stage 2} \quad Y_i = \beta_1 P(D = 1|Z) + \beta_2 X_i + U_i \quad where \; U_0 = U_1 \; as \; U_v \perp\!\!\!\perp (U_1 - U_0)|X \tag{14}$$

Below is a figure conceptually presenting the instrumental variable approach, we have removed the X covariates influence on the outcome as they are not a part of the mechanism for controlling out selection bias.

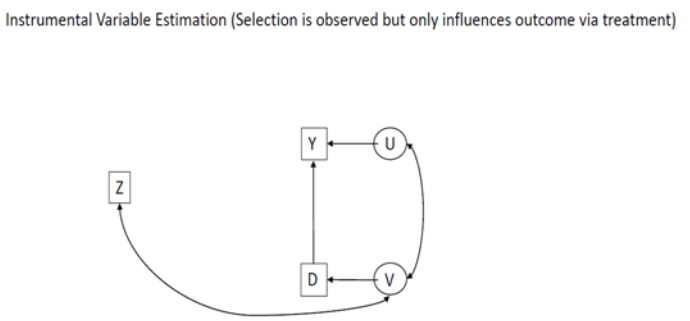

Figure 5. Instrumental Variable Estimate (conventional method for LATE)

Since the treatment effect we receive from this only pertains to those whose selection into treatment is influenced by the instrument, this is called a local average treatment effect (LATE) (Imbens & Angrist, 1994). The above figure illustrates that within the instrumental variable approach we have $Cov(Z, V) \neq 0$, $Cov(U,V) \neq 0$, and $Cov(Z,U) = 0$. The covariance between U and V arises from the entry of P(Z) into the outcome equation. In this way we see that there is an alignment of instrumental variables and control functions within the econometric approach as they are accounting for selection of unobervables. Because of the system of equations approach, as used in both control functions and instrumental variables, we will find that there is higher estimation error due to the fact that we carry error over from one estimating equation into another. The propensity balance approach does better at minimizing error because in using the propensity for balancing covariate between groups it reduces the group differences prior to estimation.

Hypotheses

Following from these conceptualization, we forward some hypotheses about the expected performance of the different methods under varying conditions of selection mechanism (observable and unobservable) and unobserved influences on the outcome. The different conditions will be reflected in the covariance matrices generating the data for demonstrating expected performance of the different analytic models. As a general point we hypothesize that Heckit models will, for the most part, do best at recovering treatment effects, because the Heckit models are the most generalized methods amongst the different models under consideration. It will also be noted that because this method will do well at reducing bias in most cases it will come at the expense of increases in error of estimation, this relates to the bias and variance tradeoff (e.g. Friedman, Hastie, & Tibshirani, 2001).

Our hypotheses of bias in estimation derive directly from the structure of the covariance amongst observed selection mechanisms (Z), unobserved influences on the outcome (U), and unobserved selection mechanisms (V). Table 1 below gives a concise presentation of our hypotheses.

In sum, when there are no covariances present we anticipate ordinary least squares (OLS) to do better at recovering treatment effects, because OLS assumes independence of errors as reflected in the independent covariance structure. In the presence of an unstructured covariance (i.e. all covariances are non-zero), we expect Heckit methods to perform best regardless of correlation magnitude, due to the explicit modeling of error covariance. When the observed selection variable (Z) is unrelated to the outcome unobservables (U) but related to the selection unobservables (V), we will expect IVE to perform best regardless of correlation magnitude, this hypothesis is due to the IVE assumption that the outcome will be rendered independent of the selection observables (instrument Z) once the treatment assignment is conditioned on the instruments. When the unobserved selection (V) and outcome (U) features are uncorrelated but the selection observables (Z) relate to both selection (V) and outcome unobservables (U) we

Table 1. Hypothesized preferred methods under varying conditions of covariance structure

Corr (U,V)	Corr (Z,V)	Corr (Z,U)	Method Hypothesized to yield Least Biased Treatment Effect Estimate
ρ = 0	**ρ = 0**	**ρ = 0**	Ordinary Least Squares Regression (OLS)
ρ = 0.30	**ρ = 0.30**	**ρ = 0.30**	Endogenous Treatment Effect Model (Heckit)
ρ = 0.60	**ρ = 0.60**	**ρ = 0.60**	Endogenous Treatment Effect Model (Heckit)
ρ = 0.30	**ρ = 0.30**	**ρ = 0**	Instrumental Variable Estimation (IVE)
ρ = 0.60	**ρ = 0.60**	**ρ = 0**	Instrumental Variable Estimation (IVE)
ρ = 0	**ρ = 0.30**	**ρ = 0.30**	Propensity Score Matching (PSM)
ρ = 0	**ρ = 0.60**	**ρ = 0.60**	Endogenous Treatment Effect Model (Heckit)
ρ = 0.30	**ρ = 0**	**ρ = 0.30**	Endogenous Treatment Effect Model (Heckit)
ρ = 0.60	**ρ = 0**	**ρ = 0.60**	Endogenous Treatment Effect Model (Heckit)

Note:
U= Unobserved Influences on the Outcome
V= Unobserved Selection Mechanisms
Z = Observed Selection Mechanisms

expect propensity score balancing to perform better, except when the correlations are high, in which case we expect Heckit models to perform better. The reasons for this are discussed above, but in general, this is based in the requirement of overlap in propensity scores amongst groups required by propensity balancing approaches, such non-overlap is anticipated by the Heckit methods. A final example shows a correlation between selection (V) and outcome unobservables (U), a zero correlation between selection observables (Z) and unobservables (V), and a correlation between selection observables (Z) and outcome unobservables (U). This latter example would reflect a situation where selection bias is present, however, our selection observables are doing poorly at explaining selection. Because the Heckit models are doing more in way of handling selection on the unobservables than the other models we anticipate its superior performance regardless of magnitude in correlation.

Methods: Simulation

Purpose behind simulation

A simulation study is presented here to demonstrate the performance of the aforementioned analytical models under consideration. Namely, the aim is to demonstrate how the different models perform in terms of how well they recover the treatment effect under different conditions of selection bias as reflected in different specifications of the correlation matrix for the observed selection (Z), unobserved selection (V), and unobserved outcome (U) components. The different correlation specifications refer to different mechanisms of selection on observables and unobservables. It is worth noting that in a true to life situation one would likely have several observed and unobserved variables influencing both selection and variation in the outcome. For the sake of simplicity in demonstrating these models, only a single variable is incorporated in generating data to represent each aspect under consideration. One can consider Z, V, & U as theoretical factors underlying observed selection and confounding in the outcome, similar in nature to latent variables. Of course, the correct and complete specification of selection mechanisms is unlikely be known with absolute certainty in real life, thus it is imperative that one apply some combination of conceptual understanding of their research situation and an empirical exploration of sample data to determine the most plausible selection mechanisms underlying one's research context.

Procedures[2]

All data generation and analyses were conducted via Stata 14 (StataCorp, 2015). Data were generated using the corr2data command, and the Monte Carlo simulation study was conducted with the simulate command. A total of N=1,000 cases were produced in each simulated dataset and the simulations were conducted over R=1,000 replications for each model under each correlation matrix specification. Data were

[2] Codes used in simulation are available upon request

generated from a multivariate normal distribution using a correlation matrix with error added into the outcome using a random normal distribution with mean 0 and variance of 1 to avoid perfect prediction. The treatment effect was modeled as the mean difference in the outcome between treatment and control. Following Cohen's (1988) criteria, the correlation value of r = 0.3 is used to represent a moderate relationship, corresponding to r^2 = 9% variance explained, while a correlation value of r= 0.6 is used to capture a strong association, corresponding to four-times as much variance explained as the moderate relation, i.e. r^2= 36%.

From the correlation matrix we produced three terms, Z, V, and U, referring to the observables in the selection model (Z), the unobservables in the selection model (V), and the unobservables in the outcome model (U). Non-zero correlations between Z, V imply that observed selection mechanisms are confounded with unobserved selection mechanisms. For example, consider the situation where parents of high achieving students are given the option to enroll their children in an enrichment program. When unmeasured parental motivation (V) leads both to a child being enrolled in an enrichment program (i.e., parental selecting child into treatment) as well as higher measured academic performance (Z) of the child (i.e., criteria making child eligible for treatment) we have a confounding of observed with unobserved selection mechanisms. Non-zero correlations between (U, V) pertain to the situation where unobserved selection mechanisms induce cofounding in the outcome variable regardless of treatment assignment. For example, students are offered an after school program to help them secure financial aid; If students' unmeasured ambition made them more likely to enter the program (i.e. self-select into treatment (V)) as well as securing more financial aid (i.e., the ambitious students would naturally secure more financial aid (U)), then we'd suggest that some unobserved characteristic is creating a confound between both selection into treatment as well as expectation on the outcome. Non-zero correlations between (Z, U) indicate a situation where some observed differences between groups is making them unequal in expectation on the outcome. For example, various schools have implemented programs to increase college enrollments, but the implementation of these programs is more likely to occur in low-income schools (Z) (i.e., low income rates are measured). If lower income students are less compelled to enroll in college (U) (i.e., with or without the program they will be less likely to enroll in college), then we would say that observable selection mechanisms are confounding outcomes with selection into treatment. The nine correlation conditions are as follows

	Corr (U,V)	Corr (Z,V)	Corr (Z,U)
A	0	0	0
B1	0.30	0.30	0.30
B2	0.60	0.60	0.60
C1	0.30	0.30	0
C2	0.60	0.60	0
D1	0	0.30	0.30
D2	0	0.60	0.60
E1	0.30	0	0.30
E2	0.60	0	0.60

From this we generated the treatment D of individuals by first creating a probability scale using

$$p = \frac{\exp(Z+V)}{(\exp(Z+V))+1} \tag{15}$$

then assigning an individual to $D = 1$ if $p > 0.50$ and $D=0$ if $p <= 0.50$. From here we constructed the outcome as

$$Y_i = 2D_i + U_i + \varepsilon_i \ , \quad U_i \sim N(0,1) \ \& \ \varepsilon_i \sim N(0,1) \tag{16}$$

Analytic Plan

The treatment effect to recover is (Y_1 - Y_0 = 2), parameter recovery will be gauged in the form of mean bias over *R* replications, $Mean\ Bias = [(2-\hat{\beta})/R$. We also present the distribution (Mean and Standard Deviation) of the standard errors across the 1,000 replications for each model by correlation matrix condition to give researchers a sense of how much error in estimation they may anticipate when using the given models under the different correlation conditions. Along with this we also give an 90% capture interval that shows at its upper bound the 95th percentile for treatment effect estimates and at its lower bound the 5th percentile treatment effect estimate for each of the models per correlation configuration over the 1,000 replications.

Results

Table 2 gives the results from the simulation for each of the nine correlation conditions per each of the four models. The results yielded here align with the hypotheses we presented at the end of our introductory

section, except in the case of no covariances where we assumed OLS would exhibit the best performance. What we find instead is that IVE is producing the least bias. These results are sensible when we consider that the instrument Z is determining treatment D but not influencing the outcome Y, which is a fundamental assumption of IVE. Consequentially, we may observe increased error in the IVE produced from the fact that error is being carried over from stage 1 to stage 2. If we reference again the notion of the bias-variance tradeoff we may anticipate that this increased variance may be resulting in a decrease in the bias of the estimation of the treatment effect. In table 2, one can observe that while OLS exhibits an average bias that is about 7.7 times greater than that exhibited by IVE, it is also yielding an average standard error that is only about 2/3 of that as produced by IVE. One can see this reflected in the interval as well where OLS is producing a much tighter bound than IVE, a span of about .25 of a unit as opposed to .38, i.e. IVE produces a 90% capture interval that is about 1.5 times wider than that produced by OLS under an independent covariance structure. In the table,

Table 2. Detailed Results from the Simulation

	OLS			Heckit			PSM			IVE		
	Bias M (SD)	SE M (SD)	90% interval	Bias M (SD)	SE M (SD)	90% interval	Bias M (SD)	SE M (SD)	90% interval	Bias M (SD)	SE M (SD)	90% interval
Matrix A												
ρ (U, V) = 0	-.023	.108	*1.85,*	.228	.618*	*1.25,*	-.079	.208	*1.71,*	**-.003**	.159	*1.81,*
ρ (Z, V) = 0	(.076)	(.002)	*2.10*	(.593)	(.185)	*3.21*	(.130)	(.045)	*2.24*	**(.115)**	(.003)	*2.19*
ρ (Z,U) = 0												
Matrix B1												
ρ (U, V) = 0.3	.361	.110	2.23,	**-.063**	.466*	*1.37,*	.460	.192	2.20,	.972	.143	2.80,
ρ (Z, V) = 0.3	(.078)	(.002)	2.49	**(.406)**	(.106)	*2.69*	(.155)	(.052)	2.71	(.106)	(.003)	3.15
ρ (Z,U) = 0.3												
Matrix B2												
ρ (U, V) = 0.6	.444	.113	2.30,	**-.118**	.272*	*1.53,*	.858	.265	2.40,	1.697	.120	3.55,
ρ (Z, V) = 0.6	(.087)	(.002)	2.58	**(.224)**	(.028)	*2.27*	(.271)	(.120)	3.31	(.092)	(.003)	3.85
ρ (Z,U) = 0.6												
Matrix C1												
ρ (U, V) = 0.3	.517	.112	2.38,	-.134	.436*	*1.34,*	.635	.199	2.37,	**-.003**	.145	*1.82,*
ρ (Z, V) = 0.3	(.078)	(.002)	2.65	(.366)	(.093)	*2.51*	(.155)	(.053)	2.89	**(.106)**	(.003)	*2.17*
ρ (Z,U) = 0												
Matrix C2												
ρ (U, V) = 0.6	1.093	.121	2.95,	-.195	.215	*1.54,*	1.727	.304*	3.27,	**-.002**	.127	*1.85,*
ρ (Z, V) = 0.6	(.087)	(.002)	3.23	(.165)	(.015)	*2.08*	(.271)	(.140)	4.18	**(.092)**	(.003)	*2.15*
ρ (Z,U) = 0												
Matrix D1												
ρ (U, V) = 0	-.158	.111	1.71,	.240	.479*	*1.59,*	**-.154**	.189	*1.58,*	.972	.149	2.80,
ρ (Z, V) = 0.3	(.078)	(.002)	1.97	(.411)	(.115)	*2.92*	**(.155)**	(.050)	*2.10*	(.106)	(.003)	3.15
ρ (Z,U) = 0.3												
Matrix D2												
ρ (U, V) = 0	-.641	.113	1.22,	**.197**	.234	*1.85,*	-.789	.254*	0.76,	1.697	.135	3.55,
ρ (Z, V) = 0.6	(.087)	(.002)	1.50	**(.205)**	(.020)	*2.53*	(.271)	(.111)	1.67	(.092)	(.003)	3.85
ρ (Z,U) = 0.6												
Matrix E1												
ρ (U, V) = 0.3	.483	.105	2.36,	**-.085**	.515*	*1.24,*	.487	.203	2.28,	1.064	.156	2.87,
ρ (Z, V) = 0	(.076)	(.002)	2.61	**(.526)**	(.144)	*2.91*	(.130)	(.046)	2.71	(.115)	(.003)	3.26
ρ (Z,U) = 0.3												
Matrix E2												
ρ (U, V) = 0.6	.995	.093	2.87,	**-.111**	.302*	*1.48,*	1.075	.166	2.86,	2.130	.146	3.94,
ρ (Z, V) = 0	(.076)	(.002)	3.12	**(.288)**	(.058)	*2.35*	(.130)	(.042)	3.30	(.115)	(.004)	4.32
ρ (Z,U) = 0.6												

least bias models are put in bold, highest error models are asterisked, and every interval containing the treatment effect value is italicized.

As indicated before we do see that despite the general superiority in performance of the Heckit model across conditions, it should also be noted that the Heckit models often produce the highest standard errors. The only case where this is not the case is when we have moderately high correlations between the selection observables and unobservables, moderately high correlations between the outcome and selection unobservables, but no correlation between the selection observables and the outcome unobservables. In this case propensity score balancing exhibits not only the highest error but also the greatest bias. This relates to propensity score balance relying on a reduction of selection bias by controlling selection observables out of the outcome. The fact that Heckit models are allotting more error to decrease bias is also reflected when considering that under all conditions these models are capable of capturing the treatment effect in their 90% capture interval. Under the specification of the independent correlation matrix every model captures the treatment effect within their 90% capture intervals, this isn't entirely surprising given that we already accept under such conditions of independence of selection and error covariance that bias doesn't pose a threat to our analyses. Across conditions we also observe that OLS is yielding the lowest standard errors, and in the case of relatively modest selection on the observables has a comparatively similar bias as that produced by propensity score balance, thus in the condition when propensity score balance was slated to perform best it appears to only be performing slightly better than OLS. Of course, it should be noted here that we have one factor for observable selection which is being included as a covariate in the OLS, so essentially the OLS is serving the same covariate balance correction as a propensity score would be. This result is not particularly useful when considering a multitude of observed selection characteristics, where it might be presumed that a propensity score would likely be preferable.

Discussion

As indicated by simulation results, in many cases the Heckit model will be acceptable, but it is important to regard this in light of the amount of estimation error permitted. Note that in the case of the independent correlation structure (i.e., no issue with selection bias) that it performs poorly producing both high bias and error. Thus, when implementing the Heckit models one will want to insure that there is indeed an issue with sample selectivity. Stata will output a Wald test for the independence of the selection and outcome equations (i.e. a test on whether ρ as discussed in the section on Heckit models is equal to zero), if this test is non-significant then Heckit models will yield poor results. Generally speaking, there is no model which is the best in all conditions, rather some become preferable under different conditions and as mentioned before one will need to evaluate the most plausible situation pertaining to their given sample.

One further point is that all of these models rely on correctly specifying the selection variables. It is not to be mistaken that models which incorporate unobservables do so in a vacuum, the unobservable factors are accounted for by conditioning on the observables, and the better we account for our selection factors the more information we will have to work with. As with any analysis one must always begin by closely exploring their research context to guide them in their choice of analytic methods.

In general, when self-selection is a problem such that an unmeasured characteristic, such as motivation, is likely to lead to differences in outcomes between groups regardless of treatment one would opt for a Heckit model. Also, when there are marked observed differences between treatment groups Heckit models are also preferable. If, however, there are only moderate differences between treatment groups, such that decent overlap can be found between the treatment groups, then a Propensity Score Balancing Method would be preferable. Instrumental Variable Estimation (IVE) is preferable when observed differences in groups are influencing selection into treatment but are not confounded with the outcome (i.e., performance on the outcome depends on selection into treatment, though individuals would not be considered equal in expectation nor equally likely to uptake treatment). Though in the context of this paper we focus on treatment groups in the traditional experimental design context, it is worth noting that IVE could also be used when the treatment of interest is not categorical in nature. For example, if one wished to evaluate how variation in the length of a school day affected college enrollment rates for recent high school graduates, they could instrument school day lengths from different high schools on some observed factors believed to predict variation in school day

lengths. Additionally, it is worth noting that with IVE There is a theoretical burden that the research must bear when justifying their choice of instrument. One must be able to argue that an instrument only influences an outcome through its relationship to the treatment of interest.

Upon deciding amongst the above discussed approaches for causal inference with observational studies for evaluating a program's effects, Stata v.14 (2015) & onward provides ready-made commands for applying these methods. Table 3 summarizes the usage of these commands.

This paper sought to discuss various topics underlying our way of conceptualizing causality in such a way that would lend itself to quantitative modeling. On the basis of these conceptualizations we presented some models for causal inference and demonstrated them under different conditions through simulation. Namely, if one has adequate observed variables to achieve good balance between two groups, then a propensity score method would be a good option that is well aligned to the notion of random assignment in experimental design. If, however, one is concerned that unobserved factors influencing selection into treatment may be confounding the analysis of treatment effects, then they may wish to incorporate Heckit Models. As a caveat, Heckit Models have new layers of complexity, and one should understand that though these models will reduce bias in the treatment effect estimate they will do so at the expense of inducing more error into estimation, i.e. though treatment effect estimate will be more valid, standard errors are apt to be inflated. In the case where one can identify a good instrument such that it predicts selection into treatment but has no direct influence on the treatment effect, then IVE will give one the opportunity to estimate a treatment effect with low bias and less error in estimation. It is, however, important to note that this estimate should be interpreted as a treatment effect which pertains specifically to those influenced by the instrumental variable. Hopefully, this paper offers the reader new insights into the nature of causality and gives some guidance in model selection for causal inference that is most plausible given one's specific research context.

Table 3. Summary of Stata commands for the different methods

Method	Stata Command	Example	Notes
Propensity Score Balancing *	teffects psmatch (Y) (D W) teffects ipw (Y) (D W)	Predict Program Selection (D) from Socio-Economic Status (W), then balance groups on this propensity to analyze Program effects on Academic Achievement (Y)	No distinction between predictors of outcome vs. predictors of selection, both are entered into the treatment selection model
Heckit Models	etregress (Y X) (D Z)	Predict Program Selection (D) from Socio-Economic Status (Z), then incorporate this into a model analyzing the Program effects on Academic Achievement (Y), while controlling for the unique influence of Socio-Economic Status (X) on Academic Achievement above and beyond its influence on Program Selection (D).	Heckit models allows predicting variables to serve as predictors for both selection & outcomes
Instrumental Variable Estimation (IVE)	ivregress 2sls Y X (D=Z)	Predict Program Participation (D) from School's Proximity to an Organization's Main Office (Z), then analyze Program effects on College Enrollment (Y) while controlling for Prior College Enrollment (X).**	With IVE, selection predictors correlate with outcome predictors, and should not be entered into the outcome equation

Notes:
*psmatch & ipw are just two options, the reader is encouraged to explore other options as given in the Stata documentation

** this assumes that Proximity (Z) influences Program participation (D), but that Proximity (Z) only influences Enrollments (Y) by influencing Program Participation (D)

References

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-55.

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. New York: Cambridge University Press.

Carneiro, P., Hansen, K. T., & Heckman, J. J. (2002). Removing the veil of ignorance in assessing the distributional impacts of social policies (No. w8840). National Bureau of Economic Research.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Fisher, R. A. (1935). The design of experiments. Edinburgh, UK: Oliver and Boyd.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

Greene, W.H. (2000). Econometric Analysis. Upper Saddle River: Prentice Hall.

Guo, S. & Fraser, M.W. (2010). Propensity Score Analysis: Statistical Methods and Applications. Thousand Oaks, CA: Sage Publications.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4), 309-321

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, Volume 5, number 4 (pp. 475-492). NBER.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, 30(1-2), 239-267.

Heckman, J., & Navarro, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and statistics*, 86(1), 30-57.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, 35(1), 1-97.

Heckman, J.J. & Vytlacil,E. (1999). Local Instrumental Variables and latent variable models for identifying and bounding treatment effects. Proceedings of the *National Academy of Sciences*, 96: 4730-34.

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

Imbens, G.W. & Angrist, J.D. (1994). Idenitification and estimation of Local Average Treatment Effects. *Econometrica*, 62(2): 467-75.

Maddala, G.S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

Mill, J.S. (1843). A System of Logic. (can be accessed at Project Gutenberg [www.gutenberg.org])

Morgan, S. L., & Winship, C. (2014). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press.

Murnane, R.J. and Willett, J.B. (2011). Methods Matter: improving causal inference in educational and social science. New York: Oxford University Press, Inc.

Neyman, J. (1923). Statistical Problems in Agricultural Experiments. *Journal of the Royal Statistical Society Series B* (suppl.) (2): 107-80

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge, UK: Cambridge University Press

Peirce, C. S. (1883). A Theory of Probable Inference. In C. S. Peirce (Ed.), Studies in Logic by Members of the Johns Hopkins University (pp. 126-181). Boston, MA: Little, Brown, and Company.

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

Rubin, D.B. (1986). Statistics and causal inference: Comment: Which Ifs have causal answers. *Journal of the American Statistical Association*, 81(396): 961-62.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Belmont, CA: Wadsworth (CENGAGE Learning).

StataCorp. (2015). Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Vickrey, W. (1960). Utility, strategy, and social decision rules. *The Quarterly Journal of Economics*, 74(4), 507-535.

Woolridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data, 2nd ed. Cambridge, MA: MIT Press

Citation:

Scott, Paul Wesley. (2019). Causal Inference Methods for selection on observed and unobserved factors: Propensity Score Matching, Heckit Models, and Instrumental Variable Estimation. *Practical Assessment, Research & Evaluation*, 24(3). Available online: http://pareonline.net/getvn.asp?v=24&n=3

Corresponding Author

Paul Wesley Scott
Department of Health & Community Systems
Center for Research & Evaluation
University of Pittsburgh, School of Nursing
Pittsburgh, PA, USA

email: pws5 [at] pitt.edu