TECHNICAL REPROT
R-405
December 2012

A solution to a class of selection-bias problems

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
(310) 825-3243
judea@cs.ucla.edu

1 Introduction

In a recent article (Pearl, 2012), I posed a series of questions to econometricians, to demonstrate how long-standing problems in econometric research can be resolved using modern techniques of causal analysis. Among the questions were three pertaining to sample selection bias, to which this paper provides a general solution. The original questions are presented in Section 2.1, their solutions in Section 2.2, and further generalizations are then developed in Section 3.

2 Sampling Selection Bias

2.1 Three questions concerning bias removal

Consider a nonparametric structural model defined over a set of endogenous variables $\{Y, X, Z_1, Z_2, Z_3, W_1, W_2, W_3\}$, and unobserved exogenous variables $\{U, U', U_1, U_2, U_3, U'_1, U'_2\}$. The equations are structured as follows:

Model 1

$$\begin{array}{rclcrcl} Y &=& f(W_3, Z_3, W_2, U) & \quad & X &=& g(W_1, Z_3, U') \\ W_3 &=& g_3(X, U'_3) & & W_1 &=& g_1(Z_1, U'_1) \\ Z_3 &=& f_3(Z_1, Z_2, U_3) & & Z_1 &=& f_1(U_1) \\ W_2 &=& g_2(Z_2, U'_2) & & Z_2 &=& f_2(U_2) \end{array}$$

where $f, g, f_1, f_2, f_3, g_1, g_2, g_3$ are arbitrary, unknown functions, and all exogenous variables are mutually independent but otherwise arbitrarily distributed.

Due to its non-parametric nature, the algebraic representation of the model is superfluous and can be replaced, without loss of information, with the diagram depicted in Fig. 1.[1]

[1] This is entirely optional; readers comfortable with algebraic representations are invited to stay in their comfort zone.

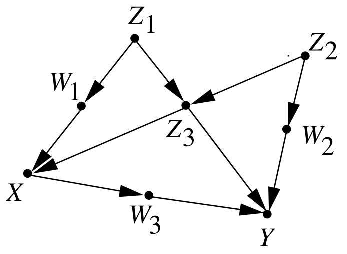

Figure 1: A graphical representation of Model 1. Error terms are assumed mutually independent and not shown explicitly.

Suppose our aim is to estimate the conditional probability $P(Y|X = x)$, yet samples are preferentially selected to the dataset depending on a set V_S of variables.

(a) Let $V_S = \{W_1, W_2\}$, what variables need be measured to correct for selection bias?

(b) In general, for what sets, V_S, would selection bias be correctable, and by what measurements.

(c) Repeat (a) and (b) assuming that our aim is to estimate the causal effect of X on Y.

2.2 Solutions

To students of graphical models, the solution to the three questions above is rather trivial, and can best be expressed in terms of the augmented graph G_S in Fig. 2. Here the node named S indicates actual selection to the dataset (i.e., $S = 1$ indicating selection and $S = 0$ exclusion.)

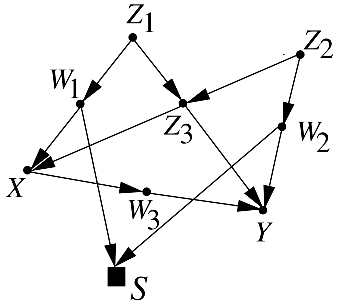

Figure 2: An augmented graph G_S, in which conditioning on S indicate actual selection into the dataset.

2.2.1 Solution to Question (a)

Since the selection set $V_S = \{W_1, W_2\}$ separates S from all other nodes in the graph, a trivial solution to problem is that measurement of the set $M = \{X, Y, W_1, W_2\}$

would be sufficient to correct for selection bias, provided the marginal probability $P(x, w_1, w_2)$ is available from external sources. This can be seen immediately by conditioning on W_1 and W_2:

$$\begin{aligned} P(y|x) &= \sum_{w_1,w_2} P(y|x, w_1, w_2)P(w_1, w_2|x) \\ &= \sum_{w_1,w_2} P(y|x, w_1, w_2, S = 1)P(w_1, w_2|x) \end{aligned} \tag{1}$$

where we made use of the conditional independence

$$Y \perp\!\!\!\perp S|(X, W_1, W_2)$$

confirmed by the corresponding *d*-separation condition in the graph. Since the first factor of Eq. (1) is estimable in the study, and the second term from external sources, the expression above constitutes a solution to question (a).

2.2.2 Solution to Question (b)

This exercise also provides a simple solution to the more general question (b). For any selection set V_S, measurement of the set $M = \{Y, X, V_S\}$ should allow us to correct for selection bias, provided the marginal probability $P(x, v_S)$ is estimable. The correction is given by:

$$P(y|x) = \sum_{v_s} P(y|x, v_s, S = 1)P(v_s|x) \tag{2}$$

Again, the first factor is estimable in the study, while the second is estimable from external sources.

A degenerate, yet frequently analyzed case of (2) occurs when $V_S = X$, namely, selection is determined by "treatment" alone. In this case we have

$$P(y|x) = P(y|x, S = 1) \tag{3}$$

and, notably, selection bias is removed without resorting to external information.

These simple solutions are predicated on the assumption that V_S is in the measurement set M, and begs a generalization to cases where some elements of V_S are not measured.

Assume, for example, that $V_S = \{W_1, W_2\}$, but W_1 can not be measured; can we still recover our target quantity $Q = P(y|x)$ and, if so, what set of variables M, need to be measured? Clearly, M should contain X and Y, but what other elements should be measured? This question can be expressed as a requirement that a subset M' of M will be found that satisfies the following condition:

$$\begin{aligned} P(y|x) &= \sum_{m'} P(y|x, m')P(m'|x) \\ &= \sum_{m'} P(y|x, m', S = 1)P(m'|x) \end{aligned} \tag{4}$$

Clearly, the condition

$$Y \perp\!\!\!\perp S | (X, M') \tag{5}$$

would satisfy the equality above and leads to a general solution to question (b).

Theorem 1 *Given a DAG G_S in which a node S indicates selection, a sufficient condition for bias free estimation of $P(y|x)$ is the existence of a subset M' of variables such that:*

(i) M', X, Y is measured is the study.

(ii) The marginal probability of $\{M', X\}$ is estimable.

(iii) $\{M', X\}$ separates S from Y in the augmented graph G_S, i.e. $(Y \perp\!\!\!\perp S | M', X)_{G_S}$.

Moreover, conditions (i)–(iii) lead to

$$P(y|x) = \sum_{m'} P(y|m', x, S = 1) P(m'|x). \tag{6}$$

2.2.3 Illustrating Theorem 1

In our example of Fig. 2, it is trivial to confirm that any (pre-treatment) set M containing W_2 and Z_3 would satisfy the conditions of Theorem 1. In particular, $M' = \{W_2, Z_3\}$ is such a set, and it allows us to recover Q without measuring W_2, via

$$Q = P(y|x) = \sum_{w_2, z_3} P(y|x, w_2, z_3, S = 1) P(w_2, z_3|x).$$

Note that the set $M' = \{W_2, Z_1, Z_2\}$ will not be sufficient for bias correction. It fails condition (iii) of Theorem 1 because conditioning on $\{X, W_2, Z_1, Z_2\}$ leaves an unblocked path between S and Y, i.e., $(S \leftarrow W_1 \rightarrow X \leftarrow Z_3 \rightarrow Y)$.

3 Generalizations

So far, we have assumed that external data is available on all variables measured in the study, with the exception of Y. The fact is, however, that the type of variables we can measure in a carefully conducted study is usually different from that which we can measure in the population at large. In our example, assuming again $V_S = \{W_1, W_2\}$, we may have external data on the set $T = \{X, W_2, Z_1, Z_2\}$ which does not include Z_3, a variable found to be essential for satisfying the conditions of Theorem 1. The question arises whether measurements taken in the study, which are all conditioned on $S = 1$, can help extend T into a larger set that includes Z_3 and thus enables the removal of selection bias as authorized by Theorem 1.

Whether this can be accomplished depends on whether a subset T' of T can be measured in the study such that the following d-separation condition holds in the graph:

$$Z_3 \perp\!\!\!\perp S | T'$$

because that would allow us to write $P(z_3, t')$ in terms of estimable probabilities:

$$\begin{aligned} P(z_3, t') &= P(z_3|t')P(t') \\ &= P(z_3|t', S = 1)P(t') \end{aligned}$$

Obviously, this d-separation does not hold in our graph, but the requirement can be turned into a formal condition for bias removal.

In general, we can characterize the selection bias problem in terms of three subsets of variables, (T, M, S), where:

- T is a set of variables for which we have population data in the form of $P(T = t)$.
- M is a set of variables for which we have measurements in the study, and permit therefore, the estimation of $P(M = m|S = 1)$.
- V_S is the set which determines the selection of samples into the dataset.

Our problem now is one of identification: Identify $P(y|x)$ from two sources of information, $P(t)$ and $P(m|S = 1)$, given a graph G in which V_S satisfies $S \perp\!\!\!\perp V | V_S$.

Theorem 1 offers a sufficient condition for this problem. It states that selection bias is removable if we can find a measured subset M of T that contains X, and separates Y from S. Now we are asking whether the condition can be relaxed to allow for a measurement set M that is NOT a subset of T. Namely some elements in M will not have population probabilities.

Thus formulated, the question can be given a simple answer: Let M be a minimal set satisfying (ii) and (iii) in Theorem 1. Is it possible to construct $P(m)$ from $P(t)$ and $P(m|S = 1)$ This would be feasible if there was a subset T' of T such that

$$\begin{aligned} P(m) &= \sum_{t'} p(m|t')p(t') \\ &= \sum_{t'} p(m|t', S = 1)p(t') \end{aligned}$$

which requires the independence $M \perp\!\!\!\perp S | T'$.

To summarize, we now we have two conditions for bias removal:

Theorem 2 *$P(y|x)$ is recoverable from sample-selection bias if there exists two measured sets, T and M, such that:*

(i) $P(T = t)$ *and* $P(M = m|S = 1)$ *are estimable.*

(ii) $(Y \perp\!\!\!\perp S | (M', X))$ *for some subset* M' *of* M*.*

(iii) $(M' \perp\!\!\!\perp S | (T', X))$ *for some subset* T' *of* T*.*

Moreover, when (i)-(iii) are satisfied, $P(y|x)$ is given by:

$$P(y|x) = \sum_{m'} P(y|x, m', S = 1)P(m'|x)$$

where $M' = M \backslash \{X, Y\}$ and $P(m'|x)$ is estimable through:

$$P(m'|x) = \sum_{t'} P(m'|t', x, S = 1)P(t'|x)$$

In the special case of M' being a subset of T, item (iii) is satisfied trivially by choosing $T' = M'$.

3.1 Illustrating Theorem 2

A graph satisfying Theorem 2 and not Theorem 1 is contrived in Fig. 3, where we let the path $W_1 \to X$ be mediated with the variable T_1. In this model, selection bias is removable using the sets

$$\begin{aligned} T &= \{Z_1, T_1, X, W_1, W_2\}, \\ T' &= \{Z_1, T_1, X, W_2\} \\ M &= \{Y, X, Z_1, Z_3, T_1, W_2\}. \end{aligned}$$

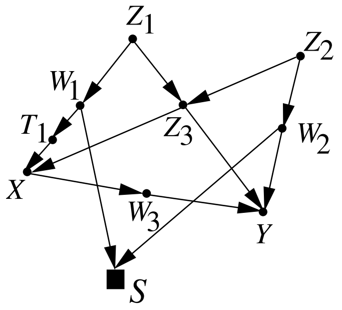

Figure 3: Selection bias can be removed if measurements on the set $M = \{Y, X, Z_1, Z_3, T_1, W_2\}$ are available in the study and the joint probability of $T = \{X, T_1, Z_1, W_2\}$ is estimable from external sources.

Note that, in violation of Theorem 1, Z_3 is not in T, and still we can remove the bias through the estimand

$$P(y|x) = \sum_{m'} P(y|x, m', S = 1)P(m'|x)$$

where $M' = M \backslash X$ and $P(m'|x)$ is estimable through:

$$\begin{aligned} P(z_3, z_1, t_1, w_2|x) &= P(z_3|z_1, t_1, w_2, x)P(z_1, t_1, w_2|x) \\ &= P(z_3|z_1, t_1, w_2, x, S = 1)P(z_1, t_1, w_2|x) \end{aligned}$$

An interesting case arises when $T = 0$, namely, we do not have any external information, save for that coming from the study. It is not hard to see that the choice $M = \{X, Y\}$ would satisfy the conditions of Theorem 2, provided $Y \perp\!\!\!\perp S | X$, namely, X separates Y from S. This coincide with the degenerate case of "treatment dependent selection" analyzed in Section 2.2.2 (Eq. (3)).

3.2 Recovering causal effects (Question (c))

A sufficient answer to question (c) (Section 2.1) follows in a fairly straightforward way from the discussion of Section 3. If our aim is to estimate the causal effect $P(Y = y|do(X = x))$ in the model of Fig. 2. We should seek a set T of variables such that $\{T, S\}$ blocks all back-door paths from X to Y. This can be accomplished using $T = \{Z_3, W_2\}$ or $T\{W_1, Z_3\}$, and would require a separate estimate of the marginal distribution $P(t)$ be estimable separately. This last requirement is not necessary if we can settle for the t-specific causal effect $P(Y = y|do(X = x), t)$ rather then the average effect. If S is a descendant of X then we need to block not only the back-door paths, but also paths containing arrows from X which carry spurious dependencies. For example, if $V_S = \{X, W_2\}$, we need to measure $T = \{Z_3, W_2\}$, which blocks the path $X \rightarrow S \leftarrow W_1 \rightarrow Y$ in addition to all back-door paths. The set $T = \{W_1, Z_3\}$ will not be sufficient. Lastly, if S is a descendant of any intermediate variable (e.g., W_3) on the path from X to Y, conditioning on S would unblock a "virtual collider" through $X \rightarrow W_3 \leftarrow U_{W_3}$. See (Pearl, 2009, pp. 339–340; Shpitser and VanderWeele, 2011). In summary, the set T should block all paths carrying non-causal dependencies between X and Y, and should be prevented from creating such paths. Daniel et al. (2011) operationalized these considerations in algorithms.

4 Conclusions

Theorem 1 and 2 give sufficient conditions for recovering the relationship $P(y|x)$ from selection biased data, provided that we know what variables V_S determine whether samples are selected or excluded. Extensions to more elaborate relationships, like $P(y|x, z)$ are straightforward, letting all expressions be conditioned on $Z = z$. Likewise, the causal effect $P(y|do(x))$ can be recovered when the conditions of Theorem 2 are satisfied with an added requirement that T' be an admissible set.

Economists associate the topic of "selection bias" with celebrated work of James Heckman (1979) which deals with outcome-dependent sampling, and relies on distributional assumptions. The results reported in this paper are orthogonal to these of Heckman's because they are applicable to the entire class of non parametric models. Angrist (1997) focused on the trivial case of X-dependent selection and on estimating Eq. (3) using propensity score.

Related works on selection bias can be found in Greenland and Pearl (2011) and Hernán et al. (2004), which give a classification of selection bias in various epidemiological scenarios and demonstrates cases where adjustment for the set V_S would remove such bias. Several of the graphical conditions formulated in Section 2 were also noted in Bareinboim and Pearl (2012) and Greenland and Pearl (2011). Bias caused by outcome-dependence sampling is treated in Bareinboim and Pearl (2012); Didelez et al. (2010); Geneletti et al. (2009). These papers focus primarily on recovering the odds-ratio which, due to its (X, Y) symmetry, can be recovered even under outcome-dependent sampling.

References

Angrist, J. D. (1997). Conditional independence in sample selection models. *Economics Letters* **54** 103–112.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*. La Palma, Canary Islands.

Daniel, R. M., Kenward, M. G., Cousens, S. N. and Stavola, B. L. D. (2011). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research* **21** 243–256.

Didelez, V., Kreiner, S. and Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science* **25(3)** 368–387.

Geneletti, S., Richardson, S. and Best, N. (2009). Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* **10(1)**.

Greenland, S. and Pearl, J. (2011). Adjustments and their consequences – collapsibility analysis using graphical models. *International Statistical Review* **79** 401–426.

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

Hernán, M., Hernández-Díaz, S. and Robins, J. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2012). Trygve Haavelmo and the emergence of causal calculus. Tech. Rep. R-391, <http://ftp.cs.ucla.edu/pub/stat_ser/r391.pdf>, University of California Los Angeles, Computer Science Department, CA.

Shpitser, I. and VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *The International Journal of Biostatistics* **7**. Article 16.