by Thad Dunning, Political Science Department
and David Freedman, Statistics Department
UC Berkeley, CA 94720

1. Introduction

Selection bias is a pervasive issue in social science. Three research topics illustrate the point.

(i) What are the returns to education? College graduates earn more than high school graduates, but the difference could be due to factors—like intelligence and family background—that lead some persons to get a college degree while others stop after high school.

(ii) Are job training programs effective? If people who take the training are relatively ambitious and well organized, any direct comparison is likely to over-estimate program effectiveness, because participants are more likely to find employment anyway. (See references below.)

(iii) Do boot camps for prisoners prevent recidivism? Possibly, but prisoners who want to go straight are more likely to participate, and less likely to find themselves in jail again—even if boot camp has no effect.

These questions could be settled by experiment, but experimentation in such contexts is expensive at best, impractical or unethical at worst. Investigators rely, therefore, on observational (non-experimental) data, with attendant difficulties of confounding.

In brief, comparisons can be made between a treatment group and a control group that does not get the treatment. But there are likely to be differences between the groups, other than the treatment. Such differences are called "confounding factors." Differences on the response variable of interest (income, employment, recidivism) may be due to treatment, or confounding factors, or both. Confounding is especially troublesome when subjects select themselves into one group or another, rather than being assigned to different regimes by the investigator. Self-selection is the hallmark of an observational study; assignment by the investigator is the hallmark of an experiment.

This article will review one of the most popular models for selection bias. The model, due to Heckman, will be illustrated on the relationship between admissions tests and college grades. Causal inference will be mentioned. There will be some pointers to the literature on selection bias, including critiques and alternative models. The intention-to-treat principle for clinical trials will be discussed, by way of counterpoint.

Model-based corrections for selection bias turn out to depend strongly on the assumptions built into the model. Thus, caution is in order. Sensitivity analysis is highly recommended: try different models with different assumptions. Alternative research designs should also be considered: stronger designs may permit data analysis with weaker assumptions.

2. Admissions data

In the US, many colleges and universities require applicants to take the SAT (Scholastic Achievement Test). Admission is based in part on SAT scores and in part on other evidence—high school GPA (grade point average), essays, recommendations, interviews by admissions officers. Figure 1 shows a somewhat hypothetical scatter diagram. Each student is represented by a dot. The

response variable is first-year college GPA, plotted on the vertical axis. The explanatory variable is the SAT score, plotted on the horizontal axis. The correlation between the two variables is about 0.5, which is fairly realistic. The "regression line," which slopes across the diagram from lower left to upper right, estimates the average GPA at each level of SAT. GPAs are between 0 and 4. If the college requires two SATs, the combined score will be between 400 and 1600, as in the diagram.

The data in Figure 1 would be available only for a college that takes all comers. If the college rejects applicants with an SAT below 800, we get a truncated scatter diagram, as in Figure 2. Truncation reduces the correlation coefficient. The reduction is called "attenuation due to restriction of range." The slope of the regression line, however, is largely unaffected. Selecting on values of the explanatory variable need not bias the slope of the regression line. Truncation has one impact on correlation and quite another on slope.

Figure 1. No selection

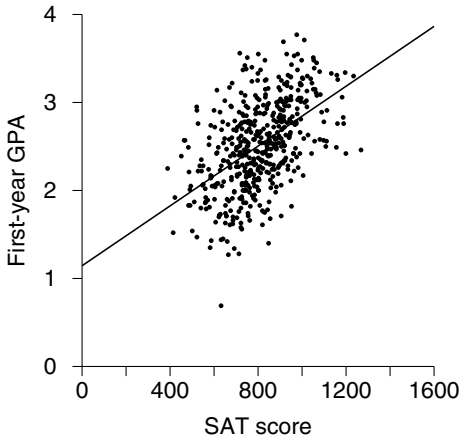


Figure 2. Selection on X

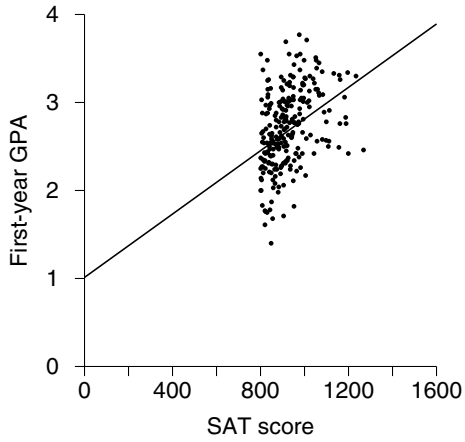
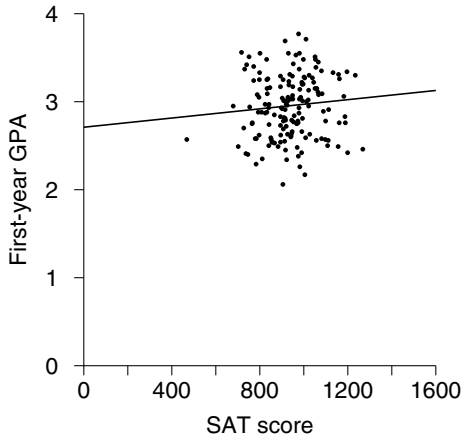


Figure 3. Selection on Y



Suppose now that the admissions office selects students who will get good grades despite low SAT scores. (This is hypothetical; there is little empirical evidence to suggest that admissions offices have that ability, beyond using high school GPA—which, like the SAT, is a good predictor

of college GPA—to help guide the decisions.) We might get a scatter diagram like the one shown in Figure 3. The correlation is much reduced. Correspondingly, the regression line is much shallower than the line at the top left. This kind of selection impacts correlation and slope in similar ways.

Selecting on the response variable—or more generally on variables correlated with the error term in the regression—is likely to bias the slope of the line. *If* we have a valid model for the selection process—and that is a big if—the bias can be corrected; details are given, below. For educational policy analysis, the scatter diagrams have a clear message. Highly selective institutions cannot expect to see any substantial correlation between (i) variables that drive admissions decisions and (ii) measures of student performance—a point that often gets lost in debates over "high-stakes testing."

3. Association versus causation

In the admissions example, there is no implication that SAT scores cause GPA. In many other examples, selection models are used to draw causal inferences from observational data. This raises additional questions; see, for instance, Heckman (1989, 2000), Briggs (2004), or Freedman (2005). Briggs discusses the effect of coaching programs on SAT scores. As the admissions example shows, however, selection bias is a problem even when causation is not in the picture.

4. Some pointers to the literature

Heckman (1976, 1978, 1979) proposed formal statistical models for dealing with selection bias. However, the model—like other such models—is rather sensitive to specification error (Briggs 2004, Breen 1996, Copas and Li 1997, Hartman 1991, Lalonde 1986, Nawata 1993, 1994, Stolzenberg and Relles 1990, Vella 1998, Zuehlke and Zeman 1991). Estimates may be more stable if the selection equation includes some explanatory variables that can be excluded a priori from the response equation.

Lalonde (1986) and Fraker and Maynard (1987) contrast the effects of job training programs, as estimated from observational data, with results from experiments. Heckman and Hotz (1989) try to reconcile the estimates. Other methods for handling selection bias include weighting (Scharfstein, Rotnitzky, and Robins 1999), and modeling based on conditional independence assumptions (Little and Rubin 2002). In the health sciences, selection effects are often handled using proportional-hazard models (Lawless 2003).

Scharfstein, Rotnitzky, and Robins (1999) quantify the (substantial) extent to which inferences depend on unidentifiable parameters; also see Robins (1999) and Manski (1995). There is a lively discussion from various perspectives in Wainer (1989).

5. Intention to treat

Randomized controlled experiments generally give the best evidence on causation, because they minimize problems created by confounding and self-selection. However, experiments on people cannot be immune from difficulty. By way of example, consider the first randomized controlled experiment on mammography, that is, screening for breast cancer by x-rays (Shapiro et al 1988). This trial started in the 1960s, when mammography was very unusual. Some women were randomized to screening, and others (the controls) were randomized to usual medical care without screening. There was, however, "crossover:" many women assigned to screening declined

to be screened. Subjects who cross over are very different from compliers, which raises the issue of selection bias—even in an experimental setting.

The mammography experiment was therefore analyzed according to the "intention-to-treat" principle: deaths from breast cancer among those assigned to treatment—whether or not they accepted—were counted in the treatment arm. Similarly, deaths among women assigned to the control condition would have been charged to the control arm, even if these women sought out screening. Intention-to-treat gives an unbiased estimate for the effect of assignment, and (in many situations) a conservative estimate for the effect of treatment. Intention-to-treat is the standard analysis for clinical trials.

Despite occasional bursts of controversy, the experiments gave solid evidence for the efficacy of mammography: screening cuts the death rate from breast cancer by a factor of about two. See International Agency for Research on Cancer (2002), Health Council of the Netherlands (2002), Freedman, Petitti, and Robins (2004).

When there is crossover from the treatment arm to the control arm, and little if any crossover in the other direction, there are robust estimates for the effect of treatment on the treated (Freedman, Petitti, and Robins, 2004, p. 73). When there is crossover in both directions, estimating the effect of treatment on the treated requires additional modeling assumptions. Under some circumstances, econometric techniques like instrumental-variables regression may be helpful.

Intention-to-treat, and related analyses, may be useful alternatives for handling selection effects, because they are relatively simple, and depend on minimal assumptions about selection mechanisms. Such techniques are best applied to natural experiments, and data collection is likely to be expensive. On the other hand, with a strong research design, causal inference can be persuasive. There is an informative survey in Angrist and Krueger (2001).

6. A formal model

In the admissions study discussed earlier, GPA is observed only for subjects in the sample—the ones who go to the college where the study is done. We present Heckman's model in that context. Subjects are indexed by i. Let $C_i = 1$ if subject i is in the sample, else $C_i = 0$. Let X_i be the SAT score for subject i, and let Y_i be the GPA. Assume that X_i is observed for all subjects (e.g., all applicants) but Y_i is observed only if $C_i = 1$. The model has two equations:

$$Y_i = a + bX_i + \sigma U_i, \tag{1}$$

$$C_i = 1 \text{ if } c + dX_i + V_i > 0, \text{ else } C_i = 0. \tag{2}$$

The pairs (U_i, V_i) are assumed to be independent and identically distributed across subjects i, and independent of the X's. The common distribution of (U_i, V_i) is assumed to be bivariate normal, with expected values equal to 0 and variances equal to 1; the correlation is ρ. The parameters in the model are a, b, c, d, σ, ρ. The U_i and V_i are "latent" (unobserved) variables, which represent unmeasured characteristics of the subjects.

Equation (1) is the "response equation:" it explains how Y_i is related to X_i. The error term is σU_i, with expectation 0 and variance σ^2. Equation (2) is the "selection equation:" it explains how subjects come to be in the sample. This equation involves the latent variable V_i. The two equations are connected by the correlation ρ between U_i and V_i.

The response equation may look like an ordinary regression equation, but there is a crucial difference. The variable Y_i is observed only for i in the sample. If i is in the sample, then U_i has a non-zero conditional expectation: $E(U_i|C_i = 1) \neq 0$, and $E(U_i|C_i = 1)$ depends on i. Ordinary least squares therefore gives biased estimates for a and b.

Using the two equations together leads to estimates that, with large samples, are nearly unbiased. This works because equation (2) assumes a very particular mechanism for selection into the sample: i is selected if $c + dX_i + V_i > 0$. Correspondingly, the expected value of U_i changes in a very special way, controlled by the correlation ρ between U_i and V_i. If $\rho = 0$, then selection bias is not an issue after all, and the second equation is unnecessary. Further details on the model and estimation procedures will be found in the next section.

Other explanatory variables could be entered into (1) and (2), e.g., high school GPA, denoted by Z:

$$Y_i = a + bX_i + cZ_i + \sigma U_i, \tag{3}$$

$$C_i = 1 \text{ if } d + eX_i + fZ_i + V_i > 0, \text{ else } C_i = 0. \tag{4}$$

In typical applications, the choice of explanatory variables may seem a little arbitrary. So is the functional form. Why linearity? Why are the coefficients the same for all subjects? The statistical assumptions might raise other questions. Why do the latent variables have the same distribution for all subjects? Why normality? Even the independence assumption may seem questionable in competitive situations like college admissions: if one applicant gets in, another must be excluded.

7. Mathematical details

Our object here is to sketch Heckman's two-stage estimation procedure, illustrated on equations (1) and (2). Recall that (U_i, V_i) were assumed to be bivariate normal with $E(U_i) = E(V_i) = 0$, $\text{var}(U_i) = \text{var}(V_i) = 1$, and the correlation is ρ; the U's and V's were assumed to be independent of the X's, and independent across subjects.

As a preliminary mathematical fact, there is a random variable W_i with the following properties:

(i) W_i is normal with expectation 0 and variance 1,
(ii) W_i is independent of V_i and the X's,
(iii) $U_i = \rho V_i + \sqrt{1 - \rho^2}\, W_i$.

Indeed, we can set $W_i = (U_i - \rho V_i)/\sqrt{1 - \rho^2}$ and verify (i)-(ii)-(iii). In (ii), for instance, W_i is independent of the X's because $W_i = (U_i - \rho V_i)/\sqrt{1 - \rho^2}$, and (U_i, V_i) is independent of the X's by assumption. Moreover, W_i is independent of V_i because the correlation between these two variables is 0, and they are jointly normal. That in turn is because (U_i, V_i) were assumed to be jointly normal.

We turn now to estimation. Equation (2) is a probit model, which can be estimated by maximum likelihood. Actually, equations (1) and (2) could be estimated together using maximum likelihood. However, Heckman suggested estimating (1) on its own, after putting in a new variable M_i to mop up $\sigma E(U_i|C_i = 1)$:

$$Y_i = a + bX_i + qM_i + \text{error}, \quad \text{where error} = \sigma U_i - qM_i. \tag{5}$$

Besides the intercept, this equation has two explanatory variables, X_i and M_i. The equation can be estimated by ordinary least squares, although generalized least squares might be preferable.

The new explanatory variable needs to be put into a more explicit form. Condition on the X's, which can then be treated as constant:

$$\begin{aligned}\sigma E(U_i|C_i = 1) &= \sigma E(U_i|V_i > -c - dX_i) \\ &= \sigma\rho E(V_i|V_i > -c - dX_i) \\ &= \sigma\rho M(c + dX_i),\end{aligned} \tag{6}$$

where

$$M(v) = \phi(v)/\Phi(v), \tag{7}$$

Φ being the standard normal distribution function, and $\phi = \Phi'$ its density. "Mills' ratio" is $\Phi(x)/\phi(x)$, which is the inverse of M.

The normal distribution is relevant because, by assumption, U_i and V_i are standard normal variables: $P(U_i < x) = P(V_i < x) = \Phi(x)$. The first equality in (6) comes from the selection equation (2). To get the second equality, substitute $U_i = \rho V_i + \sqrt{1-\rho^2}W_i$, then use properties (i)-(ii) of W_i: $E(W_i|V_i > -c - dX_i) = E(W_i)$ by independence, and $E(W_i) = 0$. To get the last equality, we must compute $E(V_i|V_i > -v)$. This is an exercise in calculus, although the signs are confusing. To begin with,

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{8}$$

so $x\phi(x)$ is the derivative of $-\phi(x)$. Now

$$\begin{aligned}E(V_i|V_i > -v) &= \frac{1}{P(V_i > -v)}\int_{-v}^{\infty} x\phi(x)dx \\ &= \frac{\phi(-v)}{P(V_i > -v)} \\ &= \frac{\phi(v)}{P(-V_i < v)} \\ &= \frac{\phi(v)}{\Phi(v)} = M(v).\end{aligned} \tag{9}$$

We cannot set $M_i = M(c + dX_i)$ in (5), because c and d are unknown. Heckman's estimation procedure begins by fitting the selection equation (2) to the data, using maximum likelihood. This gives estimated values $\tilde{c}$ for c and $\tilde{d}$ for d. Next, set $M_i = M(\tilde{c} + \tilde{d}X_i)$, and fit

$$Y_i = a + bX_i + qM_i + \text{error} \tag{10}$$

to the data using least squares. That gives $\hat{a}, \hat{b}, \hat{q}$. The estimates of main interest are usually $\hat{a}$ and $\hat{b}$, but $\hat{q}$ would estimate $\sigma\rho$. When Heckman published his papers, estimating two equations by maximum likelihood would have been a major-league enterprise: fitting one equation by maximum likelihood and the other by least squares was a real simplification. Today, computers are much faster....

Heckman developed models to cover a variety of situations. Variables can be binary (yes/no), or continuous; the response variables might be observed for all subjects, or just for subjects in the

sample. In a study that compares incomes for college and high school graduates, the key explanatory variable is binary, indicating whether the subject did or did not graduate from college. The response variable (income) is continuous. Both variables are observed for all subjects in the study. Other control variables could be added to the equations. In the admissions study, the explanatory variable (SAT) and the response variable (GPA) are continuous; GPA is observed only for subjects in the sample, as noted above. Other cases will not be discussed here.

References

Angrist JD, Krueger AB (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 19: 2–16.

Breen R (1996). *Regression Models: Censored, Sample Selected, or Truncated Data*. Sage.

Briggs DC (2004). Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics* 29: 397–420.

Copas JB, Li HG (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B* 59: 55–77.

Fraker T, Maynard R (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources* 22: 194–217.

Freedman DA (2005). Statistical models for causation. This Handbook.

Freedman DA, Petitti DM, Robins JM (2004). On the efficacy of screening for breast cancer. *International Journal of Epidemiology* 33: 43–73 (with discussion). Correspondence, pp. 1404–6.

Hartman RS (1991). A Monte Carlo analysis of alternative estimators in models involving selectivity. *Journal of Business and Economic Statistics* 9 41–9.

Health Council of the Netherlands (2002). *The Benefit of Population Screening for Breast Cancer with Mammography*. The Hague,

Heckman JJ (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–92.

Heckman JJ (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46: 931–959

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

Heckman JJ (1989). Causal inference and nonrandom samples. *Journal of Educational Statistics* 14: 159–68. Reprinted in J. Shaffer, ed., *The Role of Models in Nonexperimental Social Science*, AERA/ASA.

Heckman JJ (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics* CVX: 45–97.

Heckman JJ, Hotz VJ (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84: 862–80 (with discussion).

International Agency for Research on Cancer (2002). *Breast Cancer Screening*. Lyon: IARC. IARC Handbooks of Cancer Prevention, Vol. 7.

Lalonde RJ (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76: 604–20.

Lawless JF (2003). *Statistical Models and Methods for Lifetime Data*. 2nd ed. Wiley-Interscience.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data*. Wiley.

Manski CF (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Nawata K (1993). A note on the estimation of models with sample selection biases. *Economics Letters* 42: 15–24.

Nawata K (1994). Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator. *Economics Letters* 45: 33–40.

Robins JM (1999). Association, causation, and marginal structural models. *Synthese* 121: 151–79.

Scharfstein DO, Rotnitzky A, Robins JM (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association* 94: 1096-1146.

Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986*. Baltimore: Johns Hopkins, 1988.

Stolzenberg RM, Relles DA (1990). Theory testing in a world of constrained research design. *Sociological Methods & Research* 18: 395–415.

Vella F (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources* 33 127–169.

Wainer H (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics* 14: 121–99 (with discussion). Reprinted in J. Shaffer, ed., *The Role of Models in Nonexperimental Social Science*, AERA/ASA.

Zuehlke TW, Zeman AR (1991). A comparison of two-stage estimators of censored regression models. *Review of Economics and Statistics* 73: 185–88.