Selection Bias in the
Social Sciences\textsuperscript{1}

James J. Heckman
The University of Chicago
January 16, 2004

\textsuperscript{1}This research was supported by NSF SES-0241858.
Outline

I. Weighted Distributions

II. A Regression Representation of the Selection Problem When There is Selection on Unobservables

III. Identification

IV. Bounding and Sensitivity Analysis
Glossary

Selection Bias  Bias arising in general from nonrandom sampling.

Roy Model  A model of sectoral choice in which agents choose their sector of employment on the basis of where they get the highest income.

Generalized Roy Model  A model of sectoral choice with a more general sectoral selection criterion than the Roy Model.

Weighted Distribution  A distribution derived from a population probability distribution by weighting the sampling at different coordinates of the population distribution differently than simple random sampling.

Population Distribution  A mathematical description of the probability the outcomes of random variables can assume.
A random sample of a population produces a description of the population distribution of characteristics. A sample selected by any rule not equivalent to random sampling produces a description of the population distribution of characteristics that does not accurately describe the true population distribution of characteristics no matter how big the sample size.

The problem of selection bias arises when a rule other than simple random sampling is used to sample the underlying population that is the object of interest. The distorted representation of a true population in a sample as a consequence of a sampling rule is the essential source of the selection problem. The identification problem is to recover features of a hypothetical population from an observed sample. (See Figure 1). The hypothetical population can refer to the potential wages of all persons whether or not they work (and wages are observed for them) or to the potential outcomes of any choice problem where only actual choices are observed. Distorting selection rules may arise from decisions of sample survey statisticians, or as a consequence of self selection, so we only observe a subset of possible outcomes.
Two characterizations of the selection problem are fruitful. The first, which originates in statistics, involves characterizing the sampling rule depicted in Figure 1 as applying a weighting to hypothetical population distributions to produce observed distributions. The second, which originates in econometrics, explicitly treats the selection problem as a missing data problem and uses observables to impute the relevant unobservables.
I. Weighted Distributions

Any selection bias model can be described in terms of weighted distributions. Let $Y$ be a vector of outcomes of interest and let $X$ be a vector of “control” or “explanatory” variables. The population distribution of $(Y,X)$ is $F(y,x)$. To simplify the exposition assume that the density is well defined and write it as $f(y,x)$. 
Any sampling rule is equivalent to a non-negative weighting function \( \omega(y, x) \) that alters the population density. People are selected into the sampled population by a rule that differs, in general, from random sampling. Let \((Y^*, X^*)\) denote the random variables produced from sampling. The density of the sampled data \(g(y^*, x^*)\) may be written as

\[
g(y^*, x^*) = \frac{\omega(y^*, x^*) f(y^*, x^*)}{\int \omega(y^*, x^*) f(y^*, x^*) dy^* dx^*}
\]

where the denominator of the expression is introduced to make the density \(g(y^*, x^*)\) integrate to one as is required for proper densities. Simple random sampling corresponds to the case where \(\omega(y, x) = 1\). Sampling schemes for which \(\omega(y, x) = 0\) for some values of \((Y, X)\) create special problems because not all values of \((Y, X)\) are sampled.\(^1\)

\(^1\) For samples in which \(\omega(y, x) = 0\) for a non-negligible proportion of the population, it is useful to consider two cases. A *truncated sample* is one for which the probability of observing the sample from the larger random sample is not known. For such a sample, (1) is the density of all the sampled \(Y\) and \(X\) values. A *censored sample* is one for which the probability is known or can be consistently estimated.
In many problems in economics, attention focuses on \( f(y \mid x) \), the conditional density of \( Y \) given \( X = x \). If samples are selected solely on the \( x \) variables (“selection on the exogenous variables”), \( \omega(y, x) = \omega(x) \) and there is no problem about using selected samples to make valid inference about the population conditional density. Sampling on both \( y \) and \( x \) is termed general stratified sampling, and a variety of different sampling schemes can be characterized by the structure they place on the weights (Heckman, 1987).
From a sample of data, it is not possible to recover the true density \( f(y, x) \) without knowledge of the weighting rule. On the other hand, if the weight \( \omega(y^*, x^*) \) is known and the support of \((y, x)\) is known and \( \omega(y, x) \) is nonzero, then \( f(y, x) \) can always be recovered because

\[
\frac{g(y^*, x^*)}{\omega(y^*, x^*)} = \frac{f(y^*, x^*)}{\int \omega(y^*, x^*) f(y^*, x^*) dy^* dx^*}
\]

and by hypothesis both the numerator and denominator of the left-hand side are known, and we know \( \int f(y^*, x^*) dy^* dx^* = 1 \), so it is possible to determine \( \int \omega(y^*, x^*) f(y^*, x^*) dy^* dx^* \).
It is fundamentally easier to correct for sampling plans with known non-negative weights or weights that can be estimated separately from the full model than it is to correct for selection where the weights are not known, and must be estimated jointly with the model.\footnote{Selection with known weights has been studied under the rubric of the Horvitz-Thompson estimates since the mid 50s. Rao (1965, 1985) summarizes this research in statistics. Contributions to the choice based sampling literature in economics were made by Manski and McFadden (1981). Length biased sampling is analytically equivalent to choice based sampling and has been studied since the late 19th Century by Danish actuaries. See Sheps and Menken (1973). Heckman and Singer (1985) extend the classical analysis of length biased sampling in duration analysis to consider models with unobservables dependent across spells and time varying variables. In their more general case, simple weighting methods with weights determined independently from the model are not available.} Choice based sampling, length biased sampling and size biased sampling are examples of the former. Sampling arising from more general selection models, described in Heckman (2001) and below, cannot be put in this form because the weights require that the model be known in advance of any analysis of the data.
The requirements that (a) the support of \((y, x)\) is known and (b) \(\omega(y, x)\) is nonzero are not innocuous. In many important problems in economics, requirement (b) is not satisfied: the sampling rule excludes observations for certain values of \((y, x)\) and hence it is impossible without invoking further assumptions to determine the population distribution of \((Y, X)\) at those values. If neither the support nor the weight is known, it is impossible, without invoking strong assumptions, to determine whether the fact that data are missing at certain \((y, x)\) values is due to the sampling plan or that the population density has no support at those values. Using this framework, Heckman (1987) analyzes a variety of sampling plans of interest in economics, showing what assumptions they make about the weights and the model to solve the inferential problem of going from the observed population to the hypothetical population.
II. A Regression Representation of the Selection Problem

When There is Selection on Unobservables

A regression version of the selection problem when the weights $\omega(y, x)$ cannot be estimated independently of the model originates in the work of Gronau (1974) and Heckman (1974, 1976, 1979). Let there be two outcomes $Y_1$ and $Y_0$, corresponding to outcomes in sector 1 and outcomes in sector 0. We write outcomes as

\begin{align*}
Y_1 &= \mu_1(X) + U_1; \\
Y_0 &= \mu_0(X) + U_0;
\end{align*}

The decision rule that characterizes the sector of choice is based on $I$, a net utility, and

\begin{align*}
I &= \mu_I(Z) + U_I; \\
D &= 1[I \geq 0].
\end{align*}

The special case where $\mu_I(Z) = \mu_1(X) - \mu_0(X)$ and $U_I = U_1 - U_0$ so $I = Y_1 - Y_0$ is the Roy Model.
In this model, selection only occurs on outcomes. \((Y_1, Y_0)\) are potential outcomes. For simplicity we assume that \((U_1, U_0, U_I)\) are statistically independent of \((X, Z)\) and that \((U_1, U_0, U_I)\) have mean zero. We observe \(Y = DY_1 + (1 - D)Y_0\) and \(Y_1\) or \(Y_0\) but not both. In some applications, \(Y_0\) (or \(Y_1\)) is never observed. In general

\[
E(Y_1 \mid X, D = 1) \neq \mu_1(X),
\]

\[
E(Y_0 \mid X, D = 0) \neq \mu_0(X).
\]

The observed outcomes are a nonrandom sample of potential outcomes,

\[
E(Y \mid X, Z, D = 1) = E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + E(U_1 \mid X, Z, D = 1) \quad (4a)
\]

and

\[
E(Y \mid X, Z, D = 0) = E(Y_0 \mid X, Z, D = 0) = \mu_0(X) + E(U_0 \mid X, Z, D = 1). \quad (4b)
\]

In some cases only (4a) or (4b) can be constructed since we only observe \(Y_1\) or \(Y_0\). The conditional
means of $U_0$ and $U_1$ are the “control functions” or bias functions as introduced and defined in Heckman and Robb (1985; 1986, reprinted 2000). The mean observed outcomes (the left hand side variables) are generated by the mean of the potential outcomes plus a bias term. The control function is the bias term.

Define $P(z) = \Pr(D = 1 \mid Z = z)$. As a consequence of decision rule (3d), Heckman (1980) demonstrates that under general conditions we may always write these expressions as

\[
E(Y \mid X, Z, D = 1) = \mu_1(X) + K_1(P(Z)) \tag{5a}
\]

and

\[
E(Y \mid X, Z, D = 0) = \mu_0(X) + K_0(P(Z)) \tag{5b}
\]

where $K_1(P(Z))$ and $K_0(P(Z))$ are control functions and depend on $Z$ only through $P$. The functional forms of the $K$ depend on specific distributional assumptions. See Heckman and MaCurdy (1985) for a catalogue of examples.
The value of $P$ is related to the magnitude of the selection bias. As samples become more representative, $P(Z) \rightarrow 1$, $K_1(P) \rightarrow 0$. See Figure 2 which plots control function $K_1(P)$ versus $P$. As $P \rightarrow 1$, the sample becomes increasingly representative since the probability of any type of person being included in the sample is the same (and $P = 1$). The bias function declines with $P$. We can compute the population mean of $Y_1$ in samples with little selection. (High $P$). In general, regressions on selected samples are biased for $\mu_1(X)$. We conflate the selection bias term with the function of interest. If there are variables in $Z$ not in $X$, regressions on selected samples would indicate that they “belong” in the regression. Representation (5a)–(5b) is the basis for an entire econometric literature on selection bias in regression functions.$^3$ The key idea in all this literature is to control for the effect of $P$ on fitted relationships.$^4$

$^3$Heckman, Ichimura, Smith and Todd (1998) present methods for testing the suitability of this expression in a semiparametric setting.

$^4$Heckman (1980) suggests a series expansion of the $K_1$ and $K_0$ functions in terms of polynomials in $P$ and suggests that a test for the absence of selection can be based on a test of whether the joint set of polynomials is statistically significant in an outcome equation. Andrews (1991) provides a more general analysis.
The control functions relate the missing data (the $U_0$ and $U_1$) to observables. Under a variety of assumptions, it is possible to form these functions up to unknown parameters and identify the $\mu_0(X), \mu_1(X)$ and the unknown parameters from regression analysis, and control for selection bias (See Heckman, 1976; Heckman and Robb, 1985; 1986, reprinted 2000; and Heckman and Vytlacil, 2005).

In the early literature, specific functional forms for (4) and (5) were derived assuming that the $U$ were joint normally distributed:

**Assumption 1** $(U_0, U_1, U_2) \sim N(0, \Sigma)$.

**Assumption 2** $(U_0, U_1, U_2) \perp \perp (X, Z)$. 
Assumption 1 coupled with assumption 2 produces precise functional forms for $K_1$ and $K_0$. For censored samples, a two step estimation procedure was developed. (1) Estimate $P(Z)$ from data on the decision to work and (2) using an estimated $P(Z)$, form $K_1(P(Z))$ and $K_0(P(Z))$ up to unknown parameters. Then (5a) and (5b) can be estimated using regression. This produces a convenient expression linear in the parameters when $\mu_1(X) = X\beta_1$ and $\mu_0(X) = X\beta_0$. A direct one step regression procedure was developed for truncated samples. (See Heckman and Robb, 1985; 1986, reprinted 2000.) Equations (5a) and (5b) are the basis for an entire literature which generalizes and extends the early models, and remains active to this day (See Heckman and Vytlacil, 2005).

---

*Corrections for using estimated $P(Z)$ in first stage estimation are given in Heckman (1979) and Newey and McFadden (1994). Assumptions 1 and 2 were also used to estimate the model by maximum likelihood as in Heckman (1974).*
III. Identification

Much of the econometric literature on the selection problem combines discussions of identification (going from populations generated by selection rules back to the source population) with discussions of estimation in solving the inferential problem of going from observed samples to hypothetical populations. It is analytically useful to distinguish the conditions required to identify the selection model from ideal data from the numerous practical and important problems of estimating the model. Understanding the sources of identification of a model are essential to understanding how much of what we are getting out of an empirical model is a consequence of what we put into it.

---

6See Heckman (2000) for one precise definition of identification.
Holland (1986) used the law of iterated expectations to write the conditional distribution of an outcome, say $Y_1$ on $X$ in the following form:

$$F(Y_1 \mid X) = F(Y_1 \mid X, D = 1) \Pr(D = 1 \mid X) + F(Y_1 \mid X, D = 0) \Pr(D = 0 \mid X).$$

(6)

We observe $Y_1$ only if $D = 1$. In a censored sample, we can identify $F(Y_1 \mid X, D = 1)$, $\Pr(D = 1 \mid X)$ and hence $\Pr(D = 0 \mid X)$. We do not observe $Y_1$ when $D = 0$. Hence, we do not identify $F(Y_1 \mid X)$. In independent work, James Smith and Finis Welch (1986) made a similar decomposition of conditional means (replacing $F$ with $E$).

Holland questioned how one could identify $F(Y_1 \mid X)$ and compared selection models with other approaches. Smith and Welch (1986) discussed how to bound $F(Y_1 \mid X)$ (or $E(Y_1 \mid X)$) by placing bounds on the missing components $F(Y_1 \mid X, D = 0)$ ($E(Y_1 \mid X, D = 0)$, respectively). A clear precedent for this idea was the work of Peterson (1976) who developed nonparametric bounds for the competing risk model of duration analysis which is mathematically identical to the Roy Model which is the model of equations (3a)–(3d) when $I = Y_1 - Y_0$.8

---

7 Smith and Welch use their analysis to bound the effects of dropping out on the black-white wage gap.

8 The competing risks model replaces $\max(Y_0, Y_1)$ with $\min(Y_0, Y_1)$ for selecting outcomes. Thus, in this model, $D = 1$ if $I = Y_0 - Y_1 > 0$. 
The normality assumption widely made in the early literature has been called into question. Arabmazar and Schmidt (1981) presented Monte Carlo analyses of models showing substantial bias for models with continuous outcomes when normality was assumed but the true model was non-normal. The empirical evidence is more mixed. Normality is not a bad assumption for analyzing models of self selection for log wage outcomes once allowance is made for truncation and self selection.\footnote{Normality of latent variables turns out to be an acceptable assumption for discrete choice models except under extreme conditions (Todd, 1996). See Heckman (2001) for a discussion on the evidence of the performance of the normality assumption in data.}
Heckman and Honoré (1990) consider identification of the Roy Model \( I = Y_1 - Y_0 \) in the notation of equations (3a)–(3d) under a variety of conditions. They establish that under normality, the model is identified even if there are no regressors so there are no exclusion restrictions. They further establish that the model is identified (up to subscripts) even if one observes only \( Y \), but do not know if it is \( Y_1 \) or \( Y_0 \). The original normality assumption used in selection models was based on powerful functional form assumptions.\(^\text{10}\)

\(^{10}\) Powerful, but testable. The model is over-identified. See for example the tests by Bera, Jarque and Lee (1984) for the tests of distributional assumptions within a class of limited dependent variable models.
They develop a nonparametric Roy Model and establish conditions under which variation in regressors over time or across people can identify the model nonparametrically. One can replace distributional assumptions with different types of variation in the data to identify the Roy version of the selection model. Heckman and Smith (1998) extend this line of analysis to the Generalized Roy Model where the decision is based on a more general $I$. It turns out that the decision rule with $I = Y_1 - Y_0$ plays a crucial role in securing identification of the selection model. In a more general case, where $I$ may depend on $Y_1 - Y_0$ but on other unobservables as well, even with substantial variation in regressors across persons or over time, only partial identification of the full selection model is possible. When the models are not identified, it is still possible to bound crucial parameters and an entire literature has grown up elaborating this idea. Heckman, Ichimura, Smith and Todd (1998), among others, discuss semiparametric estimation of selection models. See also Ahn and Powell (1993).
IV. Bounding and Sensitivity Analysis

Starting from equation (6) or its version for conditional means, the papers by Smith and Welch, (1986), Holland (1986) and Glynn, Laird and Rubin (1986) characterize the selection problem more generally without the structure of equations (3a)–(3d), and offer Bayesian and classical methods for performing sensitivity analyses for the effects of different identifying assumptions on inferences about population mean.

Selection on observables solves the problem of selection by assuming that $Y_1$ is independent of $D$ given $X$ so $F(Y_1 \mid X, D = 1) = F(Y_1 \mid X)$. This is the assumption that drives matching models. It is inconsistent with the Roy Model of self selection (Heckman and Vytlacil, 2005).
Various approaches to bounding this distribution, or moments of the distribution, have been proposed in the literature all building on insights by Holland (1986) and Peterson (1976). To illustrate these ideas in the simplest possible setting, let $g(Y_1 \mid X, D = 1)$ be the density of outcomes (e.g., wages) for persons who work ($D = 1$ corresponds to work). Assume censored samples. Missing is $g(Y \mid X, D = 0)$ (e.g., the density of the wages of non-workers).
In order to estimate \( E(Y_1 \mid X) \), Smith and Welch (1986) use the law of iterated expectations to obtain

\[
E(Y_1 \mid X) = E(Y_1 \mid X, D = 1) \Pr(D = 1 \mid X) + E(Y_1 \mid X, D = 0) \Pr(D = 0 \mid X).
\]

To estimate the left hand side of this expression, it is necessary to obtain information on the missing component \( E(Y_1 \mid X, D = 0) \). Smith and Welch propose and implement bounds on \( E(Y_1 \mid X, D = 0) \), for example,

\[
Y_L \leq E(Y_1 \mid X, D = 0, Z) \leq Y_U
\]

where \( Y_U \) is an upper bound and \( Y_L \) is a lower bound.\(^{11}\) Using this information, they construct the bounds

\[
E(Y_1 \mid X, D = 1) \Pr(D = 1 \mid X) + Y_L \Pr(D = 0 \mid X) \leq E(Y_1 \mid X)
\]

\[
 \leq E(Y_1 \mid X, D = 1) \Pr(D = 1 \mid X) + Y_U \Pr(D = 0 \mid X).
\]

By doing a sensitivity analysis, they produce a range of values for \( E(Y \mid X) \) that are explicitly

\(^{11}\)In their problem there are plausible ranges of wages which dropouts can earn.
dependent on the range of values assumed for $E(Y \mid X, D = 0)$. Later work by Manski (1995) and Robins (1989) develops this type of analysis more systematically for a variety of models.

Glynn, Laird and Rubin (1986) present a sensitivity analysis for distributions using Bayesian methods under a variety of different assumptions about $F(Y_1 \mid X, D = 0)$ to determine a range of values of $F(Y \mid X)$. Holland (1986) proposes a more classical sensitivity analysis that varies the ranges of parameters of models. Rosenbaum (1995) discusses a variety of sensitivity analyses.
The objective of these analyses of bounds and the Bayesian and classical sensitivity analyses is to clearly separate what is known from what is conjectured about the data, and to explore the sensitivity of reported estimates to the assumptions used to secure them. Manski (1994) and Heckman and Vytlacil (2000, 2005) demonstrate the extra restrictions that come from using the structure of model (3a)–(3d) to produce bounds on outcomes.

Much of the theoretical analysis presented in the recent literature is nonparametric although in practice, much practical experience in statistics and econometrics demonstrates that high-dimensional nonparametric estimation is not feasible for most sample sizes available in cross sectional econometrics. Some form of structure must be imposed to get any reliable nonparametric estimates. However, feasible parametric versions of these methods run the risk of imposing false parametric structure.12

---

12 The methods used in the bounding literature depend critically on the choice of conditioning variables $X$. In principle, all possible choices of the conditioning variables should be explored especially in computing bounds for all possible models, although in practice this is never done. If it were, the range of estimates produced by the bounds would be substantially larger than the wide bounds already reported in this literature.