Biased and Unbiased Samples

James J. Heckman
Econ 312, Spring 2019

May 13, 2019

Definitions and Some Examples of Biased Samples

- All sampling models can be described by the following set-up.
- Let $\mathbf{Y}$ be a vector of outcomes of interest and let $\mathbf{X}$ be a vector of "control" or "explanatory" variables.
- The population distribution of $(\mathbf{Y},\mathbf{X})$ is $F(y, x)$.
- Assume that the density is well defined and write it as $f(y, x)$.

- Any **sampling rule** can be interpreted as producing a non-negative weighting function of $\omega(\boldsymbol{y}, \boldsymbol{x})$ that alters the population density.
- Let $(\boldsymbol{Y}^*, \boldsymbol{X}^*)$ denote the sampled random variables.
- The density of the sampled data $g(\boldsymbol{y}^*, \boldsymbol{x}^*)$ may be written as

$$g(\boldsymbol{y}^*, \boldsymbol{x}^*) = \frac{\omega(\boldsymbol{y}^*, \boldsymbol{x}^*)\boldsymbol{f}(\boldsymbol{y}^*, \boldsymbol{x}^*)}{\int \omega(\boldsymbol{y}^*, \boldsymbol{x}^*)f(\boldsymbol{y}^*, \boldsymbol{x}^*)d\boldsymbol{y}^* d\boldsymbol{x}^*} \tag{1}$$

- The denominator of the expression introduced to make the density $g(\boldsymbol{y}^*, \boldsymbol{x}^*)$ integrate to one.

- Alternatively, the weight may be defined as

$$\omega^*(\boldsymbol{y}^*\boldsymbol{x}^*) = \frac{\omega(\boldsymbol{y}^*, \boldsymbol{x}^*)}{\int \omega(\boldsymbol{y}^*, \boldsymbol{x}^*)\boldsymbol{f}(\boldsymbol{y}^*, \boldsymbol{x}^*)d\boldsymbol{y}^* d\boldsymbol{x}^*}$$

so that

$$g(\boldsymbol{y}^*, \boldsymbol{x}^*) = \omega^*(\boldsymbol{y}^*, \boldsymbol{x}^*) f(\boldsymbol{y}^*, \boldsymbol{x}^*). \tag{2}$$

- Sampling schemes for which $\omega(y, x) = 0$ for some values of $(\boldsymbol{Y}, \boldsymbol{X})$ create special problems.
- For such schemes, not all values of $(\boldsymbol{Y}, \boldsymbol{X})$ are sampled.
- Let indicator variable $i(x, y) = 0$ if a potential observation at values **y**,**x** cannot be sampled and let $i(x, y) = 1$ otherwise.
- Let $\Delta = 1$ record the occurrence of the event "a potential observation is sampled, i.e., the value of **y**,**x** is observed" and let $\Delta = 0$ if it is not.
- In the population, the proportion that is sampled is

$$\Pr(\Delta = 1) = \int i(y, x) f(y, x) dy dx \tag{3}$$

- $$\Pr(\Delta = 0) = 1 - \Pr(\Delta = 1).$$

- Consider samples in which $\omega(\boldsymbol{y}, \boldsymbol{x}) = 0$ for a non-negligible proportion of the population ($\Pr(\Delta = 0) > 0$).

Two Cases

- A *truncated sample* is one for which $\Pr(\Delta = 1)$ is not known and cannot be identified.
- A *censored sample* is one for which $\Pr(\Delta = 1)$ is known or can be identified.
- Sampling rule in this case is such that frequency of **y**,**x** for which $\omega(\boldsymbol{y}, \boldsymbol{x}) = 0$ are not known.
- It is known whether or not $i(\boldsymbol{y}, \boldsymbol{x}) = 0$ for all values of **Y**,**X**.

- Notational convenience: define $(\boldsymbol{Y}^*, \boldsymbol{X}^*) = (\mathbf{0}, \mathbf{0})$ for values of $\mathbf{y}$,$\mathbf{x}$ such that $\omega(\boldsymbol{y}, \boldsymbol{x}) = i(\boldsymbol{y}, \boldsymbol{x}) = 0$.
- Such a definition is innocuous provided that in the population there is no point mass (concentration of probability mass) at $(\mathbf{0}, \mathbf{0})$.
- (Any value other than $(\mathbf{0}, \mathbf{0})$ can be selected provided that there is no point mass at that value).
- Given $\Delta = 0$, the distribution of $\boldsymbol{Y}^*, \boldsymbol{X}^*$ is Dirac Function:

$$G(\boldsymbol{y}^*, \boldsymbol{x}^*) = 1 \qquad \text{for} \qquad \Delta = \mathbf{0}$$

at

$$\boldsymbol{Y}^* = \mathbf{0} \qquad \text{and} \qquad \boldsymbol{X}^* = \mathbf{0}.$$

- The joint density of $\boldsymbol{Y}^*, \boldsymbol{X}^*, \Delta$ for the case of a censored sample is obtained by combining (1) and (3).
- Thus

$$
\begin{aligned}
g(\boldsymbol{y}^*, \boldsymbol{x}^*, \delta) &= \left[\frac{\omega(\boldsymbol{y}^*, \boldsymbol{x}^*) f(\boldsymbol{y}^*, \boldsymbol{x}^*)}{\int \omega(\boldsymbol{y}^*, \boldsymbol{x}^*) f(\boldsymbol{y}^*, \boldsymbol{x}^*) d\boldsymbol{y}^* d\boldsymbol{x}^*}\right]^{\delta} \qquad (4)\\
&\times \left[\int i(\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x}\right]^{\delta} \\
&\times [1]^{1-\delta} \left[\int (1 - i(\boldsymbol{y}, \boldsymbol{x})) f(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x}\right]^{1-\delta}.
\end{aligned}
$$

- First term on the right-hand side of (4): conditional density of $\boldsymbol{Y}^*, \boldsymbol{X}^*$ given $\Delta = 1$.
- Second term: probability that $\Delta = 1$.
- Third term: conditional density of $\boldsymbol{Y}^*, \boldsymbol{X}^*$ given $\Delta = 0$.
- Density assigns unit mass to $\boldsymbol{y}^* = 0, \boldsymbol{x}^* = 0$ when $\Delta = 0$.
- Fourth term: probability that $\Delta = 0$.
- Notice that when $\omega(\boldsymbol{y}, \boldsymbol{x}) > 0$ for all $\boldsymbol{y}, \boldsymbol{x}, \Delta = 1$.
- Then (4) is identical to (1).

- In a random sample $\omega(\boldsymbol{y}^*, \boldsymbol{x}^*) = 1$ (and so $\omega^*(\boldsymbol{y}^*, \boldsymbol{x}^*) = 1$).
- In a selected sample, the sampling rule weights the data differently.
- Values of $(\boldsymbol{Y}, \boldsymbol{X})$ are over-sampled or under-sampled relative to their occurrence in the population.
- In the case of **truncated samples**, the weight is zero for certain values of the outcome.

- In many problems in economics, attention focuses on $f(y|x)$, the conditional density of $\mathbf{Y}$ given $\boldsymbol{X} = x$.
- In such problems knowledge of the population distribution of $\mathbf{X}$ is of no direct interest.
- If samples are selected solely on the $\mathbf{x}$ variables ("selection on the exogenous variables"), $\omega(\boldsymbol{y}, \boldsymbol{x}) = \omega(\boldsymbol{x})$ and there is no problem about using selected samples to make valid inference about the population conditional density.
- These are stratified samples.

- Selection on the exogenous variables:

$$g(y^*, x^*) = f(y^*|x^*)\frac{\omega(x^*)f(x^*)}{\int \omega(x^*)f(x^*)dx}$$

and

$$g(x^*) = \frac{\omega(x^*)f(x^*)}{\int \omega(x^*)f(x^*)dx^*}.$$

- Thus
$$g(\mathbf{y}^*|\mathbf{x}^*) = \frac{g(\mathbf{y}^*, \mathbf{x}^*)}{g(\mathbf{x}^*)} = f(\mathbf{y}^*|\mathbf{x}^*).$$
- For such problems, sample selection distorts inference only if selection occurs on **y** (or **y** and **x**).

General Stratified Sampling

- Sampling on both **y** and **x**.

- From this sample, it is not possible to recover the true density $f(\boldsymbol{y}, \boldsymbol{x})$ without knowledge of the weighting rule.
- **If the weighting rule is known** $(\omega(\boldsymbol{y}^*, \boldsymbol{x}^*))$, the density of the sampled data is known $(g(\boldsymbol{y}^*, \boldsymbol{x}^*))$, the support of $(\mathbf{y}, \mathbf{x})$ is known.
- If $\omega(\boldsymbol{y}, \boldsymbol{x})$ is nonzero and known, $f(\boldsymbol{x}, \boldsymbol{y})$ can be recovered:
- Why?
- $$\frac{g(\boldsymbol{y}^*, \boldsymbol{x}^*)}{\omega(\boldsymbol{y}^*, \boldsymbol{x}^*)} = \frac{f(\boldsymbol{y}^*, \boldsymbol{x}^*)}{\int \omega(\boldsymbol{y}^*, \boldsymbol{x}^*) f(\boldsymbol{y}^*, \boldsymbol{x}^*) d\boldsymbol{y}^* d\boldsymbol{x}^*} \tag{5}$$
- By hypothesis both the numerator and denominator of the left-hand side are known, and nonzero.

- The requirement that $(\boldsymbol{y}^*, \boldsymbol{x}^*)$ has a well defined density$\Rightarrow$

$$\int f(\boldsymbol{y}^*, \boldsymbol{x}^*) d\boldsymbol{y}^* d\boldsymbol{x}^* = 1.$$

- Integrating the left-hand side of (5) it is possible to determine $\int \omega(\boldsymbol{y}^*, \boldsymbol{x}^*) f(\boldsymbol{y}^*, \boldsymbol{x}^*) d\boldsymbol{y}^* d\boldsymbol{x}^*$.
- Hence can use (5) to recover the population density of the data.

- Requirements that
  1. the support of ($\mathbf{y}$,$\mathbf{x}$) is known
  2. $\omega(\boldsymbol{y}, \boldsymbol{x})$ is nonzero and known, are not innocuous
- In many important problems in economics requirement (b) is not satisfied.
- If it fails it is impossible without invoking further assumptions to determine the population distribution of ($\mathbf{Y}$,$\mathbf{X}$) 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 $\mathbf{y}$,$\mathbf{x}$ values is due to the sampling plan or that the population density has no support at those values.
- Some specific sampling plans of interest in economics.

- *Example 1.* **Truncated Sample/Truncated Random Variable.** Data are collected on incomes of individuals whose income **Y** exceeds a certain value c (for cutoff value).
- Observe Y if $Y > c$.
- Thus $\omega(y) = 1$ if $y > c$ and $\omega(y) = 0$ if $y \leq c$.
- Knowledge of the sampling rule does *not* suffice to recover the population distribution.
- From a random sample of the entire population, the social scientist can identify
  - the sample distribution of Y above c
  - but not the proportion of the original random sample with income below c ($F(c)$ where F is the distribution function of Y).
- Does not observe values of Y below c.

- Y: *truncated random variable*.
- The point of truncation is c.
- If the proportion of the original random sample with income below c is not known and cannot be identified, the *sample* is *truncated*.
- In a truncated sample, nothing is known about the proportion of the underlying population that can appear in the sample.

- A sample truncated only if $\omega(\mathbf{y}) = 0$ for some intervals of $\mathbf{y}$ (for $\mathbf{y}$ continuous) or if $\omega(\mathbf{y}) = 0$ at values of $\mathbf{y}$ at which there is finite probability mass.
- Censored sample: the proportion of the underlying population that can appear in the sample is known.
- Still don't know support of Y.

- Let $Y^* = Y$ if $Y > c$.
- Define $Y^* = 0$ otherwise (the choice of value for Y^* when Y is not observed is inessential and any value can be used in place of 0 provided that the true distribution places no mass at the selected value).
- Indicator variable $\Delta = 1$ if $Y > c$.
- $\Delta = 0$ otherwise.

- Distribution of Y^* is

$$G(y^*|Y > c) = F(y^*|Y > c) = F(y^*|\Delta = 1) \tag{6a}$$
$$= \frac{F(y^*)}{1 - F(c)}, y^* > c.$$

Point mass at $Y^* = 0$ (Convention) for $Y^* = 0$ $(\Delta = 0)$. (6b)

- Observe that (6a) is obtained from (1) by setting $\omega(y^*) = 1$ if $y > c$, and $\omega(y^*) = 0$ otherwise, and integrating up with respect to y^*.
- The distribution of Δ is

$$\Pr(\Delta = \delta) = [1 - F(c)]^{\delta}[F(c)]^{1-\delta}, \delta \in \{0, 1\}.$$

- The joint distribution of (Y^*, Δ) for a censored sample:

$$\begin{aligned} F(y^*, \delta) &= F(y^*|\delta)\Pr(\delta) \\ &= \left\{\frac{F(y^*)}{(1 - F(c))}\right\}^{\delta} [1 - F(c)]^{\delta} (1)^{1-\delta}[F(c)]^{1-\delta} \\ &= [F(y^*)]^{\delta}[F(c)]^{1-\delta}. \end{aligned} \tag{7}$$

- (7) is obtained from (4) by setting $\omega(y) = 0$
- $y < c, \omega(y) = 1$ otherwise, by setting $i(y) = \omega(y)$, and by integrating up with respect to y^*.
- For normally distributed Y: (7) is “Tobit” model.

- More information in a censored sample than in a truncated sample because one can obtain (6a) from (7) (by conditioning on $\Delta = 1$) but not vice versa.

- Inferences about the population distribution based on assuming that $F(y^*|Y > c)$ closely approximates $F(y)$ are potentially very misleading.
- A description of population income inequality based on a subsample of high income people may convey no information about the true population distribution.

- Without further information about F and its support, it is not possible to recover F from $G(y^*)$ from either a censored or a truncated sample.
- Access to a censored sample enables the analyst to recover $F(y)$ for $y > c$ but obviously does not provide any information on the shape of the true distribution for values of $y \leq c$.

- Problem is routinely "solved" by assuming that F is of a known functional form.
- This solution strategy does not always work.
- If F is normal, then it can be recovered from a censored or truncated sample (Pearson, 1900).
- If F is Pareto, F cannot be recovered from either a truncated or a censored sample (see Flinn and Heckman, 1982b).
- Show this.
- If F is real analytic (i.e., possesses derivatives of all order) and the support of Y is known, then F can be recovered (Heckman and Singer, 1986).

- *Example 2.* **Expand the previous discussion to a linear regression setting**.
- Let
$$Y = \boldsymbol{X}\boldsymbol{\beta} + U \tag{8}$$
be the population earnings function where Y is earnings.
- "$\boldsymbol{\beta}$": suitably dimensioned parameter vector.
- $\boldsymbol{X}$ is a regressor vector assumed to be distributed independently of mean zero disturbance U.
- $U \perp\!\!\!\perp X$; $E(XX')$ full rank, $E(U) = 0$.

- Data are collected on incomes of persons for whom Y exceeds c.
- Weight depends solely on y: $\omega\left(y, \boldsymbol{x}\right) = 0, y \leq c, \omega(y, \boldsymbol{x}) = 1, y > c$.
- Can identify
  - the sample distribution of Y above c
  - the sample distribution of $\boldsymbol{X}$ for Y above c and
  - the proportion of the original random sample with income below c.
- Do not know Y below c.

- As before, let $Y^* = Y$ if $Y > c$.
- Define $Y^* = 0$ otherwise.
- $\Delta = 1$ if $Y > c$, $\Delta = 0$ otherwise.
- The probability of the event $\Delta = 1$ given $\boldsymbol{X} = x$ is

$$
\begin{aligned}
\Pr(\Delta = 1 |\ \boldsymbol{X} = x) &= \Pr(Y > c |\ \boldsymbol{X} = x) \\
&= \Pr(U > c - x\beta |\ \boldsymbol{X} = x).
\end{aligned}
$$

- Invoke independence between U and $\boldsymbol{X}$ and letting F_u denote the distribution of U,

$$\Pr(\Delta = 1 | \ \boldsymbol{X} = x) = 1 - F_u(c - x\beta) \tag{9a}$$

and

$$\Pr(\Delta = 0 | \ \boldsymbol{X} = x) = F_u\left(c - x\beta\right). \tag{9b}$$

- The distribution of Y^* conditional on $\boldsymbol{X}$:

$$\begin{aligned} G(y^* \mid Y > \mathbf{0}, \boldsymbol{X} = x) &= F(y^* \mid X = x, Y > c) \qquad (10a) \\ &= F(y^* \mid \boldsymbol{X} = x, \Delta = 1) \\ &= \frac{F_u(y^* - x\beta)}{1 - F_u(c - x\beta)}, \qquad y^* > c. \end{aligned}$$

$$G(y^* \mid Y \leq 0) = 1 \text{ for } Y^* = 0 \ (\Delta = 0). \qquad (10b)$$

- The joint distribution of (Y^*, Δ) given $\boldsymbol{X} = x$ is

$$\begin{aligned} F(y*, \delta | \ \boldsymbol{X} &= x) = F\left(y^* | \ \delta, x\right) \Pr\left(\delta | \ x\right) \\ &= \left\{F_u\left(y^* - x\boldsymbol{\beta}\right)\right\}^{\delta} \left\{F_u\left(c - x\boldsymbol{\beta}\right)\right\}^{1-\delta}. \end{aligned} \tag{11}$$

- In particular,

$$\begin{aligned} E(Y^* \mid \boldsymbol{X} = x, \Delta = 1) &= x\boldsymbol{\beta} + E\left(U \mid \boldsymbol{X} = x, \delta = 1\right) \\ &= x\boldsymbol{\beta} + \int_{c - x\boldsymbol{\beta}}^{\infty} \frac{z \, \mathrm{d}F_u\left(z\right)}{\left(1 - F_u\left(c - x\boldsymbol{\beta}\right)\right)} \end{aligned} \tag{12}$$

- z: dummy variable of integration.

- Population mean regression function is

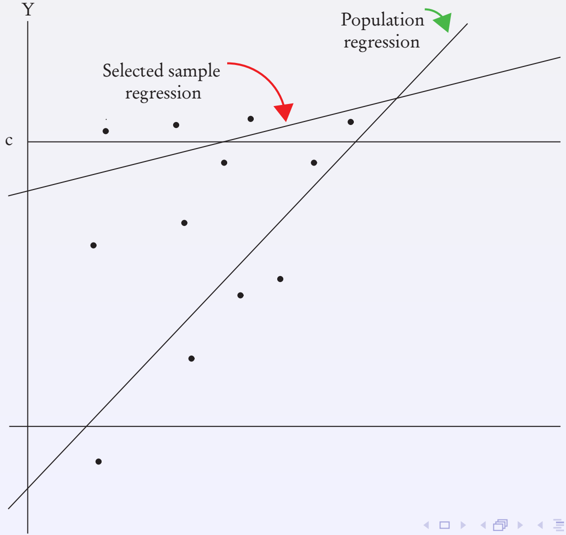
$$E(Y \mid \boldsymbol{X} = x) = x\boldsymbol{\beta}. \tag{13}$$

- Contrast between (12) and (13) illuminating.
- When theoretical model is estimated on a selected sample ($\Delta = 1$), the true conditional expectation is (12) not (13).

- The conditional mean of U depends on $\boldsymbol{x}$.
- Omitted variable analysis, $E(U \mid \boldsymbol{X} = \boldsymbol{x}, \Delta = 1)$: omitted from the regression.
- Likely to be correlated with $\boldsymbol{x}$.
- Least squares estimates of β obtained on selected samples which do not account for selection are biased and inconsistent.

- Illustrate the nature of the bias, it is useful to draw on the work of Cain and Watts (1973).
- Suppose that X is a scalar random variable (e.g., education) and that its associated coefficient is positive ($\beta > 0$).
- Under conventional assumptions about U (e.g., mean zero, independently and identically distributed and distributed independently of X), the population regression of Y on X is a straight line.
- The scatter about the regression line and the regression line are given in Figure 1.

Figure 1:



- When $Y > c$ is imposed as a sample inclusion requirement, lower population values of U are excluded from the sample in a way that systematically depends on x.
- ($Y > c$ or $U > c - x\beta$).
- As x increases and $\beta > 0$, the conditional mean of U: $[E(U \mid X = x, \Delta = 1)]$ decreases.
- Regression estimates of β that do not correct for sample selection (i.e., include $E(U \mid X = x, \Delta = 1)$
- Downward biased because of the negative correlation between x and $E(U \mid X = x, \Delta = 1)$.
- Flattened regression line for the selected sample in Figure 1.

- In models with more than one regressor, no sharp result on the sign of the bias in the regression estimate that results from ignoring the selected nature of the sample is available.
- Conventional least squares estimates of β obtained from selected samples are biased and inconsistent remains true.

- Fruitful to distinguish between the case of a truncated sample and the case of a censored sample.
- In the truncated sample case, no information is available about the fraction of the population that would be allocated to the truncated sample $[\Pr(\Delta = 1)]$.
- In the censored sample case, this fraction is known or can be consistently estimated.
- Fruitful to distinguish two further cases:
- Case (a), the case in which $\boldsymbol{X}$ is not observed when $\Delta = 0$.
- Case (b) is the one most fully developed in the literature: X observed when $D = 0$.

- Conditional mean $E(U \mid \boldsymbol{X} = x, \Delta = 1)$ is a function of $c - \boldsymbol{x\beta}$ solely through $\Pr(\Delta = 1 \mid \boldsymbol{x})$.
- Since $\Pr(\Delta = 1 \mid \boldsymbol{x})$ is monotonic in $c - \boldsymbol{x\beta}$.
- The conditional mean depends solely on $\Pr(\Delta = 1 \mid \boldsymbol{x})$ and the parameters F_u i.e., since

$$F_u^{-1}(1 - \Pr(\Delta = 1 \mid x)) = c - \boldsymbol{x\beta}$$

$$E(U \mid X = x, \Delta = 1) = \int_{F_u^{-1}[1-\Pr(\Delta=1|\boldsymbol{x})]}^{\infty} \frac{z dF_u(z)}{\Pr(\Delta = 1 \mid \boldsymbol{x})}$$

$$= K(P(\Delta = 1|x))$$

$$\ln P(\Delta = 1|x) \to 1, K(P(\Delta = 1|x)) = 0.$$

- This relationship demonstrates that the conditional mean is a function of the probability of selection.
- As the probability of selection goes to 1, the conditional mean goes to zero.
- For samples chosen so that the values of x are such that the observations are certain to be included the sample, there is no problem in using ordinary least squares on selected samples to estimate β.
- Thus in Figure 1, ordinary least squares regressions fit on samples selected to have large x values closely approximate the true regression function and become arbitrarily close as x becomes large.

- The conditional mean in (12) is a surrogate for $\Pr(\Delta = 1 \mid x)$.
- As this probability goes to one, the problem of sample selection in regression analysis becomes negligibly small.
- Much more general idea
- Heckman (1976) demonstrates that β and F_u are identified if U is normally distributed and standard conditions invoked in regression analysis are satisfied.
- In Newey; Gallant and Nycha, Powell, etc., F_u is consistently nonparametrically estimated.

- *Example 3*: *censored random variables*.
- This concept extends the notion of a truncated random variable by letting a more general rule than truncation on the outcome of interest generate the selected sample.
- Because the sample generating rule may be different from a simple truncation of the outcome being studied, the concept of a censored random variable in general requires at least two distinct random variables.

- Let Y_1 be the outcome of interest.
- Let Y_2 be another random variable.
- Denote observed Y_1 by Y_1^*.
- If $Y_2 < c$, Y_1 is observed.
- Otherwise Y_1 is not observed and we can set $Y_1^* = 0$ or any other convenient value (assuming that Y_1 has no point mass at $Y_1 = 0$ or at the alternative convenient value).
- In weighting function ω;
- $\omega(y_1, y_2) = 0$ if $y_2 > c$.
- $\omega(y_1, y_2) = 1$ if $y_2 \leq c$.

- Selection rule $Y_2 < c$ does not necessarily restrict the range of Y_1.
- Thus Y_1^* is not in general a truncated random variable.
- Define $\Delta = 1$ if $Y_2 < c$; $\Delta = 0$ otherwise.

- If $F(y_1, y_2)$ is the population distribution of (Y_1, Y_2), the distribution of Δ is

$$\Pr(\Delta = \delta) = [1 - F_2(c)]^{1-\delta}[F_2(c)]^{\delta}, \qquad \delta = 0, 1,$$

- F_2 is the marginal distribution of Y_2.

- The distribution of Y_1^* is

$$G(y_1^*) = F(y_1^*; \delta = 1) = \frac{F(y_1^*; c)}{F_2(c)}, \qquad \Delta = 1, \tag{14a}$$

$$G(y_1^* = 0) = 1, \qquad \Delta = 0. \tag{14b}$$

- (14a): the distribution function corresponding to the density in (1) when $\omega(y_1, y_2) = 1$ if $y_2 \leq c$ and $\omega(y_1, y_2) = 0$ otherwise.

- The joint distribution of (Y_1^*, Δ) is

$$G(y_1^*, \delta) = [F(y_1^*; c)]^{\delta}[1 - F_2(c)]^{1-\delta}. \tag{15}$$

- This is the distribution function corresponding to density (4) for the special weighting rule of this example.
- In a censored sample, under general conditions it is possible to consistently estimate $\Pr(\Delta = \delta)$ and $G(y_1^*)$.

- In a truncated sample, only conditional distribution (14a) can be estimated.
- A degenerate version of this model has $Y_1 \equiv Y_2$.
- In that case, censored random variable Y_1 is also a truncated random variable.
- Note that a *censored random variable* may be defined for a truncated or censored sample.

- Example 3:
- Let Y_1 be the wage of a woman.
- Wages of women are observed only if women work.
- Let Y_2 be an index of a woman's propensity to work.

- Y_2 is postulated as the difference between reservation wages (the value of time at home determined from household preference functions) and potential market wages Y_1.
- Then if $Y_2 < 0$, the woman works.
- Otherwise, she does not.
- $Y_1^* = Y_1$ if $Y_2 < 0$ is the observed wage.

- If Y_1 is the offered wage of an unemployed worker, and Y_2 is the difference between reservation wages (the return to searching) and offered market wages, $Y_1^* = Y_1$ if $Y_2 < 0$ is the accepted wage for an unemployed worker (see Flinn and Heckman, 1982a).
- If Y_1 is the potential output of a firm and Y_2 is its profitability, $Y_1^* = Y_1$ if $Y_2 > 0$.
- If Y_1 is the potential income in occupation one and Y_2 is the potential income in occupation two.

- $Y_1^* = Y_1$ if $Y_1 - Y_2 < 0$ while $Y_2^* = Y_2$ if $Y_1 - Y_2 \geq 0$.

- *Example 4*. Builds on example 3 by introducing regressors.
- This produces the *censored regression model* Heckman (1976, 1979).
- In example 3 set

$$Y_1 = \boldsymbol{X}_1 \beta_1 + U_1 \tag{16a}$$

$$Y_2 = \boldsymbol{X}_2 \beta_2 + U_2 \tag{16b}$$

where $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ are distributed independently of (U_1, U_2), a mean zero, finite variance random vector.

- Conventional assumptions are invoked to ensure that if Y_1 and Y_2 can be observed, least squares applied to a random sample of data on $(Y_1, Y_2, \boldsymbol{X}_1, \boldsymbol{X}_2)$ would consistently estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.
- $Y_1^* = Y_1$ if $Y_2 < 0$.
- If $Y_2 < 0, \Delta = 1$.
- Regression function for the selected sample is

$$E(Y_1^* \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, Y_2 < 0) = E(Y_1^* \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \Delta = 1) = \boldsymbol{X}_1\beta_1 + E(U_1 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1, \Delta = 1) \tag{17}$$

- Regression function for the population is

$$E(Y_1 \mid \boldsymbol{X}_1 = x_1) = \boldsymbol{X}_1\boldsymbol{\beta}_1. \tag{18}$$

- The conditional mean is a surrogate for the probability of selection $[\Pr(\Delta = 1 \mid x_2)]$.
- As $\Pr(\Delta = 1 \mid x_2)$ goes to one, the problem of sample selection bias becomes negligible.
- In the censored regression case, a new phenomenon appears.
- If there are variables in $\boldsymbol{X}_2$ not in $\boldsymbol{X}_1$, such variables may appear to be statistically important determinants of Y_1 when ordinary least squares is applied to data generated from censored samples.

- Example: suppose that survey statisticians use some extraneous (to X_1) variables to determine sample enrollment.
- Such variables may appear to be important determinants of Y_1 when in fact they are not.
- They are important determinants of Y_1 when in fact they are not.
- They are important determinants of Y_1^*.

- In an analysis of self-selection, let Y_1 be the wage that a potential worker could earn were they to accept a market offer.
- Let Y_2 be the difference between the best non-market opportunity available to the potential worker and Y_1.
- If $Y_2 < 0$, the agent works.
- The conditional expectation of observed wages ($Y_1^* = Y$, if $Y_2 < 0$) given x_1 and x_2 will be a non-trivial function of x_2.

- Thus variables determining non-market opportunities will determine Y_1^*, even though they do not determine Y_1.
- For example, the number of children less than six may appear to be significant determinants of Y_1 when inadequate account is taken of sample selection, even though the market does not place any value or penalty on small children in generating wage offers for potential workers.

- *Example 5. Length biased sampling.*
- Let T be the duration of an event such as a completed unemployment spell or a completed duration of a job with an employer.
- The population distribution of T is $F(t)$ with density $f(t)$.
- The sampling rule is such that *individuals* are sampled at random.
- Data are recorded on a completed spell *provided that at the time of the interview the individual is experiencing the event.*
- Such sampling rules are in wide use in many national surveys of employment and unemployment.

- In order to have a sampled completed spell, a person must be in the state at the time of the interview.
- Let "0" be the date of the survey.
- Decompose any completed spell T into a component that occurs before the survey T_b and a component that occurs after the survey T_a.
- Then $T = T_a + T_b$.
- For a person to be sampled, $T_b > 0$.
- The density of T given $T_b = t_b$ is

$$f(t|t_b) = \frac{f(t)}{1 - F(t_b)}, t \geq t_b. \tag{19}$$

- This is the **hazard rate**.

- Suppose that the environment is stationary.
- The population entry rate into the state at each instant of time is k.
- From each vintage of entrants into the state distinguished by their distance from the survey date t_b, only $1 - F(t_b) = \Pr(T > t_b)$ survive.
- Aggregating over all cohorts of entrants, the population proportion in the state at the date of the interview is P where

$$P = \int_0^{\infty} k(1 - F(t_b))dt_b \tag{20}$$

which is assumed to exist.

- In a duration of unemployment example, P is the unemployment rate.

- The density of T_b^*, sampled **presurvey** duration, is

$$g(t_b^*|t_b^* > 0) = \frac{k(1 - F(t_b^*))}{P}. \tag{21}$$

- The density of sampled **completed** durations is thus

$$\begin{aligned} g(t^*) &= \int_0^{t^*} f(t^*|t_b^*) g(t_b^*|t_b^* > 0) dt_b^* \\ &= k \frac{f(t^*)}{1 - F(t_b^*)} \frac{1 - F(t_b^*)}{P} \int_0^{t^*} dt_b^* \\ &= k \frac{t^* f(t^*)}{P}. \end{aligned}$$

- **Length biased sampling**.

- Integration by parts:

$$P = k\int_0^\infty (1 - F(z))dz = k\int_0^\infty z dF(z) = kE(T).$$

- Note that

$$g(t^*) = \frac{t^* f(t^*)}{E(T)}. \tag{22}$$

- We know that $g(t^*)$.
- Can form $\frac{g(t^*)}{t^*}$, $t^* > 0$.
- $\therefore$ we know $\frac{f(t^*)}{E(T)}$.
- Apply analysis of (5): $\int_0^\infty \frac{g(t^*)}{t^*} dt^* = \frac{\overbrace{\int_0^\infty f(t^*)dt^*}^{=1}}{E(T)}$.
- $\therefore$ know $f(t^*)$.

- In this form (22) is equivalent to (1) with $\omega(t) = t$.
- $E(T)$.
- **Length biased sampling.**
- Intuitively, longer spells are oversampled when the requirement is imposed that a spell be in progress at the time the survey is conducted ($T_b > 0$).
- Suppose, instead, that individuals are randomly sampled and data are recorded on the **next** spell of the event (after the survey date).

- As long as successive spells are independent, such a sampling frame does not distort the sampled distribution because no requirement is imposed that the sampled spell be in progress at the date of the interview.
- It is important to notice that the source of the bias is the requirement that $T_b > 0$ (i.e., sampled spells are in progress), not that only a fraction of the population experiences the event $(P < 1)$.

- The simple length weight ($\omega(t) = t$) that produces (22) is an artefact of the stationarity assumption.
- Heckman and Singer (1986): non-stationarity and unobservables when there is selection on the event that a person be in the state at the time of the interview.
- They also demonstrate the bias that results from estimating parametric models on samples generated by length biased sampling rules when inadequate account is taken of the sampling plan.

- The probability that a spell lasts until t_c given that it has lasted t_b

$$f(t_c|t_b) = \frac{f(t_c)}{1 - F(t_b)}$$

- So the density of a spell that lasts for t_c is

$$\begin{aligned} g(t_c) &= \int_0^{t_c} f(t_c|t_b) g(t_b) dt_b \\ &= \int_0^{t_c} \frac{f(t_c)}{m} dt_b = \frac{f(t_c) t_c}{m} \end{aligned}$$

- Likewise, the density of a spell that lasts until t_a is

$$
\begin{aligned}
g(t_a) &= \int_0^{\infty} f(t_a + t_b | t_b) g(t_b) dt_b \\
&= \int_0^{\infty} \frac{f(t_a + t_b)}{m} dt_b \\
&= \frac{1}{m} \int_{t_a}^{\infty} f(t_b) dt_b \\
&= \frac{1 - F(t_a)}{m}
\end{aligned}
$$

- So the functional form of $g(t_b) = g(t_a)$.
- Stationarity$\Rightarrow$backward and forward densities same.
- Mirror images.
- Back to the future.

- Some useful results that follow from this model:
  1. If $f(t) = \theta e^{-t\theta}$, then $g(t_b) = \theta e^{-t_b\theta}$ and $g(t_a) = \theta e^{-t_a\theta}$.
  2. **Proof**:

$$
\begin{aligned}
f(t) &= \theta e^{-t\theta} \rightarrow m = \frac{1}{\theta}, \\
F(t) &= 1 - e^{-t\theta} \rightarrow g(t_a) = \frac{1 - F(t)}{m} = \theta e^{-t\theta}
\end{aligned}
$$

1. $E(T_a) = \frac{m}{2}(1 + \frac{\sigma^2}{m^2})$.

Proof:

$$\begin{aligned} E(T_a) &= \int t_a f(t_a) dt_a = \int t_a \frac{1 - F(t_a)}{m} dt_a \\ &= \frac{1}{m}\left[\frac{1}{2} t_a^2 (1 - F(t_a))\Big|_0^\infty - \int \frac{1}{2} t_a^2 d(1 - F(t_a))\right] \\ &= \frac{1}{m}\int \frac{1}{2} t_a^2 f(t_a) dt_a = \frac{1}{2m}[var(t_a) + E^2(t_a)] \\ &= \frac{1}{2m}[\sigma^2 + m^2] \end{aligned}$$

1. $E(T_b) = \frac{m}{2}(1 + \frac{\sigma^2}{m^2})$.
2. **Proof**: See proof of Proposition 2.
3. $E(T_c) = m(1 + \frac{\sigma^2}{m^2})$.
4. **Proof**:

$$E(T_c) = \int \frac{t_c^2 f(t_c)}{m} dt_c = \frac{1}{m}(var(t_c) + E^2(t_c))$$

$$\rightarrow E(T_c) = 2E(T_a) = 2E(T_b), E(T_c) > m \text{ unless } \sigma^2 = 0$$

Examples

Specification of the Distribution

Weibull Distribution

- Parameters: $\lambda > 0, k > 0$
- Probability Density Function (PDF):

$$\frac{\lambda}{k}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{k}\right)^{k}\right)$$

- Cumulative Density Function:

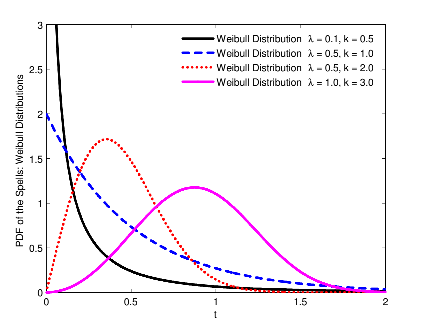
$$1 - \exp\left(-\left(\frac{t}{k}\right)^{k}\right)$$

- Set of Parameters:

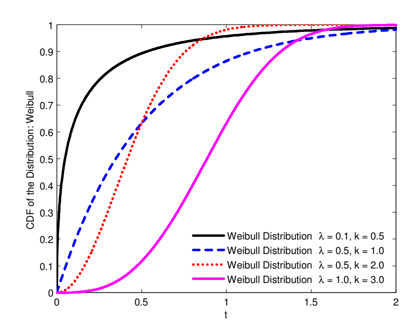
$$\begin{pmatrix} \lambda_1, k_1 = 0.5 \\ \lambda_2, k_1 = 1.0 \\ \lambda_3, k_1 = 2.0 \\ \lambda_3, k_1 = 3.0 \end{pmatrix}, \quad \text{respectively}$$

Basic Distribution Graphs

PDF for Weibull Distribution

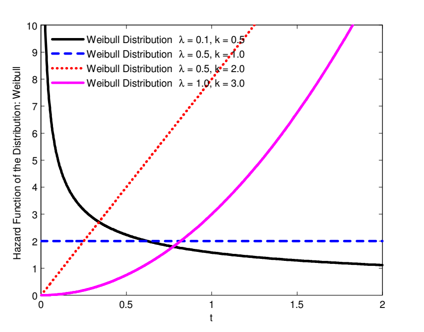


CDF of Weibull Distribution

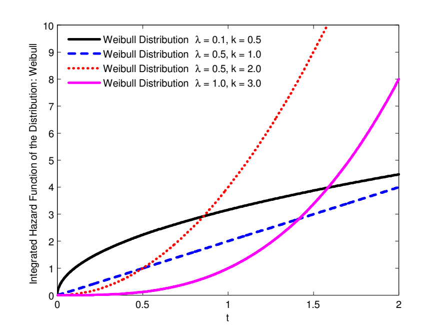


Basic Duration Graphs

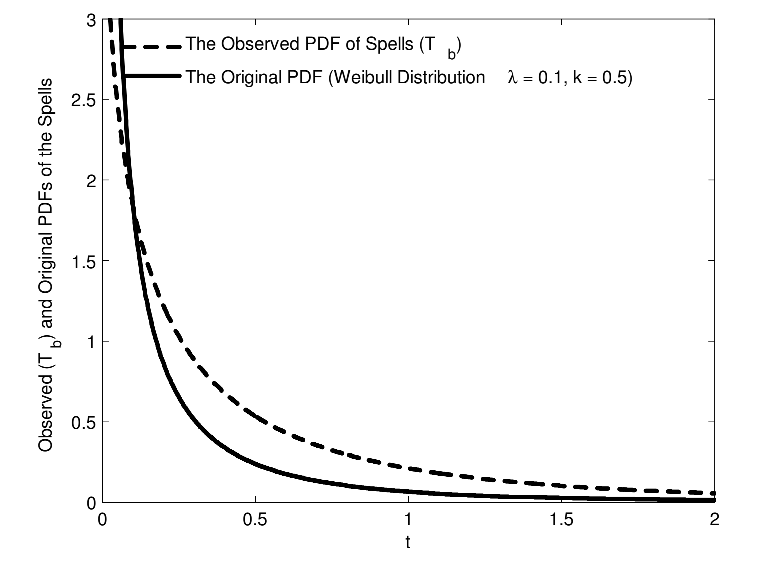
Hazard Function for Weibull Distribution



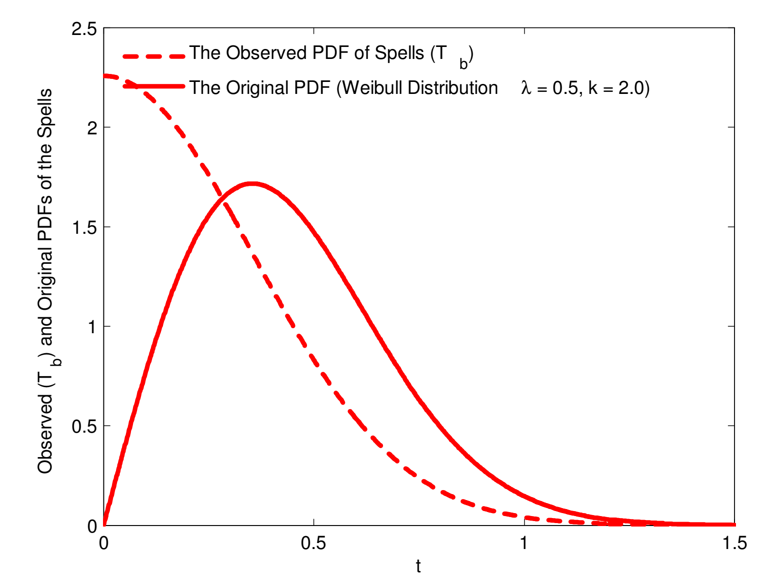
Integrated Hazard Function for Weibull



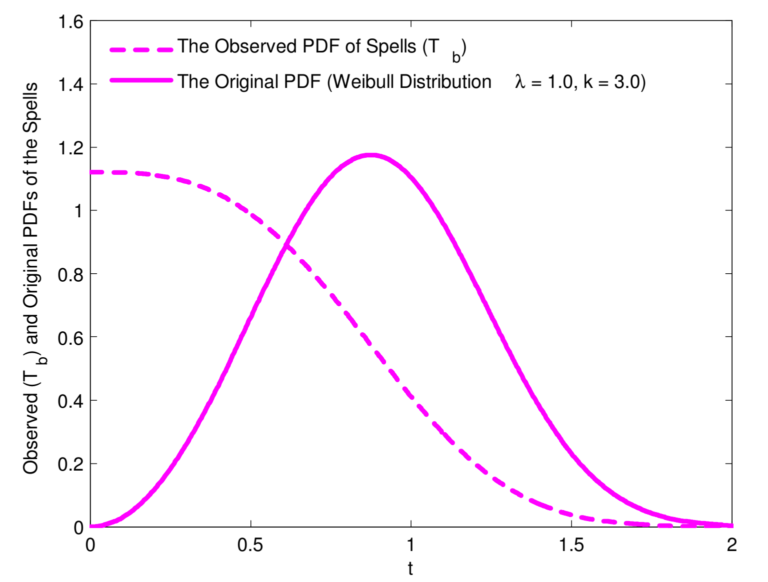
Observed and Original Distribution for T_b (Example 1)



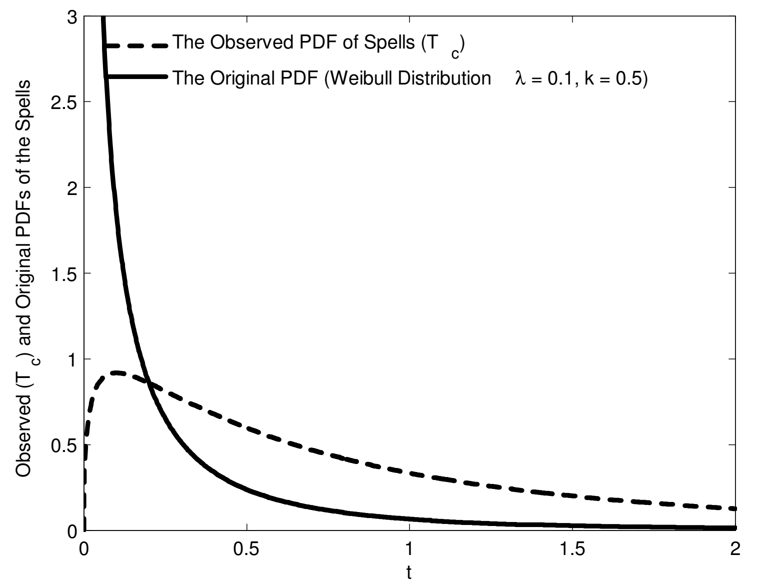
Observed and Original Distribution for T_b (Example 2)



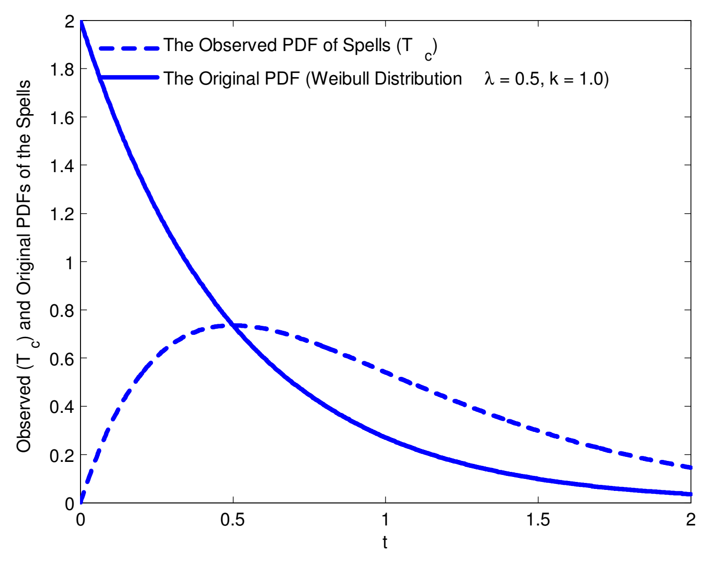
Observed and Original Distribution for T_b (Example 3)



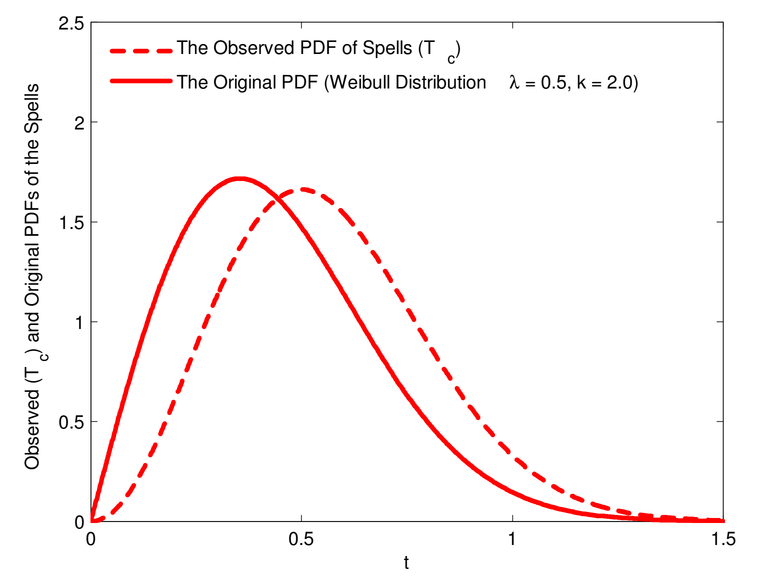
Observed and Original Distribution for T_c (Example 1)



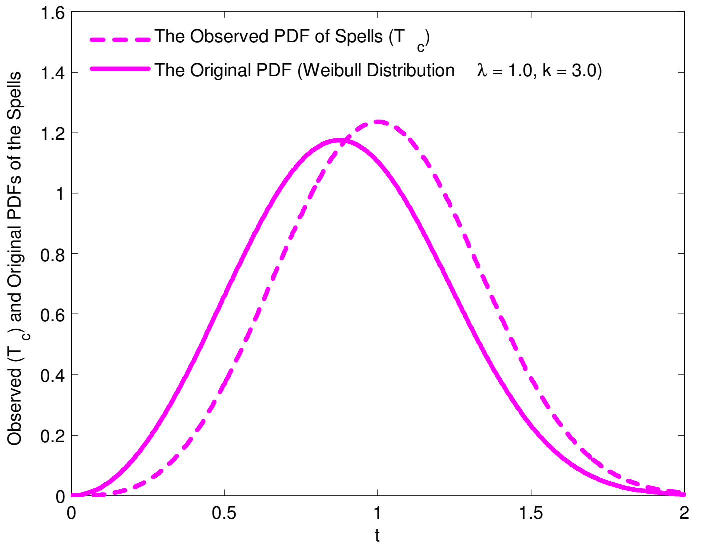
Observed and Original Distribution for T_c (Example 2)



Observed and Original Distribution for T_c (Example 3)



Observed and Original Distribution for T_c (Example 4)



- *Example 6. Choice based sampling.* Let D be a discrete valued random variable which assumes a finite number of values I.
- Discrete choice model.
- $D = i, i = 1, \dots, I$ corresponds to the occurrence of state i.
- States are mutually exclusive.
- In the existing literature the states may be modes of transportation choice for commuters (Domencich and McFadden, 1975), occupations, migration destinations, financial solvency status of firms, schooling choices of students, etc.

- Interest centers on estimating a population choice model

$$\Pr(D = i | \boldsymbol{X} = x), \quad i = 1, \ldots, I. \tag{23}$$

- The population density of $(D, \boldsymbol{X})$ is

$$f(d, x) = \Pr(D = d | \boldsymbol{X} = x) h(x) \tag{24}$$

where, in this example, $h(x)$ is the density of the data.

- For example, interviews about transportation preferences conducted at train stations tend to over-sample train riders and under-sample bus riders.
- Interviews about occupational choice preferences conducted at leading universities over-sample those who select professional occupations.

- In choice based sampling, selection occurs solely on the D coordinate of $(D, \boldsymbol{X})$.
- In terms of (1) (extended to allow for discrete random variables), $\omega(d, \boldsymbol{X}) = \omega(d)$.
- Then sampled $(D^*, \boldsymbol{X}^*)$ has density

$$g\left(d^*, \boldsymbol{x}^*\right) = \frac{\omega(d^*) f(d^*, \boldsymbol{x}^*)}{\sum_{i=1}^{I} \int \omega(i) f(i, x^*) \mathrm{d}x^*}. \tag{25}$$

- Notice that the dominator can be simplified to

$$\sum_{i=1}^{I} \omega(i) f(i)$$

- $f(d^*)$ is the marginal distribution of D^* so that

$$g(d^*, \boldsymbol{x}^*) = \frac{\omega(d^*) f(d^*, \boldsymbol{x}^*)}{\sum_{i=1}^{I} \omega(i) f(i)}. \tag{26}$$

- Integrating (25) with respect to x using (26) we obtain

$$g(d^*) = \frac{\omega(d^*)f(d^*)}{\sum_{i=1}^{I} \omega(i)f(i)} \tag{27}$$

- Sampling rule causes the sampled proportions to deviate from the population proportions.

- Note further that as a consequence of sampling only on D, the population conditional density

$$h(\boldsymbol{x}^*|d^*) = \frac{f\left(d^*, x^*\right)}{f(d^*)} \tag{28}$$

can be recovered from the choice based sample.

- The density of x in the sample is thus

$$g\left(x^*\right) = \sum_{i=1}^{I} h(x^*|\ i)g(i). \tag{29}$$

- Then using (26)-(29) we reach

$$g(d^*|x^*) = f(d^*|x^*) \times \left\{ \left[\frac{\omega(d^*)}{\sum_{i=1}^{I} \omega(i) f(i)} \right] \left[\frac{1}{\sum_{i=1}^{I} f(i|x^*) \frac{g(i)}{f(i)}} \right] \right\}. \tag{30}$$

- The bias that results from using choice based samples to make inference about $f(d^*|x^*)$ is a consequence of neglecting the terms in braces on the right-hand side of (30).

- Notice that if the data are generated by a random sampling rule, $\omega(d^*) = 1, g(d^*) = f(d^*)$ and the term in braces is one.

Further Discussion of Choice Based Samples

- Pick D first (*e.g.* travel mode).
- Probability of selecting D is $C(D)$.
- $f(D,X)$ is the joint density of D and X in the population.

$$\begin{aligned} f(D,X\,|\theta) &= g(D\,|X,\ \theta)h(X) = \varphi(X \mid D)f(D\,|\theta) \\ f(D\,|\theta) &= \int g(D\,|X,\ \theta)h(X)dX \end{aligned}$$

- Given D we observe X (the implicit assumption is that we are sampling only on D, not on D and X).
- Probability of *sampled* X, D is $\varphi(X \mid D)C(D)$.

- A fact we use later is

$$\begin{aligned}\varphi(X \mid D)C(D) &= \left\{\frac{g(D \mid X)h(X)}{f(D)}\right\}C(D)\\ &= \frac{g(D \mid X)h(X)C(D)}{\left[\int g(D \mid X)h(X)dX\right]}.\end{aligned}$$

- When $C(D) = f(D) = \int g(D \mid X)h(X)dX$, choice based sampling is random sampling.

- Note, the likelihood function in an exogenous sampling scheme is

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{I} f(D_i, X_i) = \prod_{i=1}^{I} f(D_i \mid X_i, \theta) h(X_i) \\
\ln \mathcal{L} &= \sum_{i=1}^{I} \ln f(D_i \mid X_i) + \sum \ln h(X_i).
\end{aligned}
$$

- By exogeneity, we get the lack of dependence of distribution of X on θ.

- Likelihood function for a choice-based sampling scheme is

$$\ln \mathcal{L} = \sum_{i=1}^{I} \left[\ln g(D_i \mid X_i) + \ln h(X_i) - \ln f(D_i) + \ln C(D_i) \right].$$

- Suppose $f(D)$ depends on parameters θ. $\therefore$ Max with θ.

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \sum_{i=1}^{I} \frac{\partial \ln g(D_i \mid X_i)}{\partial \theta} - \underbrace{\sum_{i=1}^{I} \frac{\partial \ln f(D_i)}{\partial \theta}}_{\text{source of bias}}.$$

- We neglect the second term in forming the usual estimators using only the first term.
- That is the source of the inconsistency.

Further Analysis of Choice Based Samples:

- An example in discrete choice.
- (c) Draw d by $\varphi(d)$.
- (d) Draw X by $f(X \mid d = 1)$.
- Joint density of data:

$$\begin{aligned} & \varphi(d=1) f(X \mid d=1, \theta) \\ = \; & \varphi(d=1)\left[\frac{\Pr(d=1 \mid X, \theta) f(X)}{\Pr(d=1 \mid \theta)}\right] \end{aligned}$$

- Now in a choice-based sample

$$\Pr^*(d = 1 \mid X) = \frac{f(X \mid d = 1, \theta)\varphi(d = 1)}{h^*(X)}$$

where $g^*(X)$ is the sampled X data.

- Joint density of *data* X is given by:

$$h^*(X) = f(X \mid d = 1, \theta)\varphi(d = 1) + f(X \mid d = 0, \theta)\varphi(d = 1)$$

and

$$\Pr(D = 1 \mid X) = \frac{f(X \mid d = 1) \Pr(d = 1)}{f(X)}$$

- Assume $f(X) > 0$.
- Using Bayes' theorem for Y write:
- $\Pr^*(D = 1 \mid X) =$

$$
= \frac{\dfrac{\Pr(D=1 \mid X,\theta) f(X)}{\Pr(D=1 \mid \theta)}\varphi(D=1)}{\dfrac{\Pr(D=1 \mid X,\theta) f(X)}{\Pr(D=1 \mid \theta)}\varphi(D=1) + \dfrac{\Pr(D=0 \mid X,\theta) f(X)}{\Pr(D=0 \mid \theta)}\varphi(D=0)}
$$

$$
= \frac{\Pr(D=1 \mid X,\theta)\varphi(D=1)/\Pr(D=1 \mid \theta)}{\Pr(D=1 \mid X,\theta)\dfrac{\varphi(D=1)}{\Pr(D=1 \mid \theta)} + \Pr(D=0 \mid X,\theta)\dfrac{\varphi(D=0)}{\Pr(D=0 \mid \theta)}}.
$$

- Now we missample the population with density $f(X \mid D=1)$ in a choice based sample:

$$
\begin{aligned}
\Pr^*(D &= 1 \,|\, X) = \frac{f(X \mid D=1,\theta)\varphi(D=1)}{f(X \mid D=1,\theta)\varphi(D=1) + f(X \mid D=0,\theta)\varphi(D=0)} \\
&= \frac{\dfrac{f(X)\Pr(D=1 \mid X)}{\Pr(D=1)}\varphi(D=1)}{\dfrac{f(X)\Pr(D=1 \mid X)}{\Pr(D=1)}\varphi(D=1) + \dfrac{f(X)\Pr(D=0 \mid X)}{\Pr(D=0)}\varphi(D=0)} \\
&= \frac{\Pr(D=1 \mid X)}{\Pr(D=1 \mid X) + \Pr(D=0 \mid X)\dfrac{\varphi(D=0)}{\varphi(D=1)} \cdot \dfrac{\Pr(D=1)}{\Pr(D=0)}} \\
&= \frac{1}{1 + \left[\dfrac{\Pr(D=0 \mid X)}{\Pr(D=1 \mid X)}\right] \cdot \dfrac{\varphi(D=0)}{\varphi(D=1)} \cdot \dfrac{\Pr(D=1)}{\Pr(D=0)}}
\end{aligned}
$$

- With logit we get

$$\Pr^*(D = 1 \mid X) = \frac{1}{1 + e^{-(\alpha_0 + X\beta) + \ln\left[\frac{\varphi(D=0)}{\varphi(D=1)} \cdot \frac{\Pr(D=1)}{\Pr(D=0)}\right]}}.$$

This goes into an intercept term:

$$= \frac{e^{\alpha^* + X\beta}}{1 + e^{\alpha^* + X\beta}}$$

$$\alpha^* = \alpha_0 - \ln\left[\frac{\varphi(D=0)}{\varphi(D=1)} \cdot \frac{\Pr(D=1)}{\Pr(D=0)}\right].$$

- How to solve problem: Reweight data by relative frequency in population.
- (Idea due to C.R. Rao, 1965, 1986.)
- Joint density of the data is

$$f(X \mid D = 1)\varphi(D = 1).$$

Use Bayes' rule to obtain

$$\frac{P(D = 1 \mid X)f(X)}{P(D = 1)}\varphi(D = 1).$$

- Now weight by

$$\frac{P(D = 1)}{\varphi(D = 1)}.$$

- Solution: Reweight the data to form the following weighted likelihood:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\Pr(D_i=1)}{\varphi(D_i=1)}(D_i^*)\ln\Pr(D_i=1\mid X,\theta)+\frac{\Pr(D_i=0)}{\varphi(D_i=0)}(1-D_i^*)\ln\Pr(D_i=0\mid X,\theta)\right]$$

$$P\int\{[\Pr(D=1\mid X,\theta_0)f(X\mid\theta_0)]\ln\Pr(D=1\mid X,\theta)+$$

$$\int[\Pr(D=0\mid X,\theta_0)f(X\mid\theta_0)]\ln\Pr(D=0\mid X,\theta)\}\,f(X\mid D)DX$$

- This step uses the result that reweighting the data gives us the true density.
- Better way to see what is giving on:

$$\frac{f(X \mid D=1)\varphi(D=1)}{g^*(X)} = \frac{\Pr(D=1 \mid X)f(X)}{g^*(X)} \frac{\varphi(D=1)}{\Pr(D=1)}.$$

- Reweight the data: when we reweight the data, g^* is restored to f.

$$f(X) = f(X \mid D=1)\varphi(D=1)\left[\frac{P(D=1)}{\varphi(D=1)}\right] + f(X \mid D=0)\varphi(D=0)\frac{\Pr(D=0)}{\varphi(D=0)}.$$

- *Example 7. Size biased sampling.* Let N be the number of children in a family.
- $f(N)$ is the density of discrete random variable N.
- Suppose that family size is recorded only when at least one child is interviewed.
- Suppose further that each child has an independent and identical chance β of being interviewed.

- The probability of sampled family size of $N^* = n^*$ is

$$g(n^*) = \frac{\omega(n^*)f(n^*)}{E[\omega(N^*)]} \tag{31}$$

where $\omega(n^*) = 1 - (1-\beta)^{n^*}$ (the probability that at least one child from a family of size n^* will be sampled).

- Note $(1-\beta) =$ probability of sampling a child (assumed the same across all n^*).
- $$E[\omega(N^*)] = \sum_{n^*}(1-(1-\beta)^{n^*})f(n^*)$$

  is the probability of observing a family.
- In a large population $\beta \to 0$ with increasing population size.

- Using l'Hospital's rule, and assuming that passage to the limit under the summation sign is valid

$$\lim_{\beta \to 0} g(n^*) = \frac{n^* f(n^*)}{E(N^*)}. \tag{32}$$

- Thus the limit form of (31) is identical to (22).
- Larger families tend to be oversampled and hence a misleading estimate of family size will be produced from such samples

- Since the model is formally equivalent to the length biased sampling model, all references and statements about identification given in Example 6 apply with full force to this example.
- See the discussion in Rao (1965).

Appendix

- *Example 5.* This example demonstrates how self-selection bias affects the interpretation placed on estimated consumer demand functions when there is self-selection.
- We postulate a population of consumers with a quasi-concave utility function $U(Z, E)$ which depends on the consumption of goods and preference shock E which represents heterogeneity in preferences among consumers.
- The support of E is $\boldsymbol{E}$.
- For price vector P and endowment income M, the consumer's problem is to

- *Max* $U(\boldsymbol{Z}, E)$ subject to $\boldsymbol{P}'\boldsymbol{Z} \leq M$.
- In the population $\boldsymbol{P}$ and M are distributed independently of E.
- First order conditions for this problem are

$$\frac{\partial U(\boldsymbol{Z}, E)}{\partial \boldsymbol{Z}} \leq \lambda \boldsymbol{P} \tag{33}$$

where λ is the Lagrange multiplier associated with the budget constraint.

- Focusing on the demand for the first good, Z_1, none of it is purchased if at zero consumption of Z_1

$$\frac{\partial U(\boldsymbol{Z}, E)}{\partial Z_1} |_{Z_1=0} \leq \lambda P_1 \tag{34}$$

i.e., marginal valuation is less than marginal cost in utility terms.

- Conventional interior solution demand functions for Z_1 are defined for a given $\boldsymbol{P}, M$ only for values of E such that

$$\frac{\partial U(Z, E)}{\partial Z_1} |_{Z_1=0} \geq \lambda P_1. \tag{35}$$

- Let the set of E for which conventional interior solution consumer demand functions for Z_1 are defined be denoted by $\underline{\underline{E}}$.
- Then

$$\underline{\underline{E}} = \left\{ E \left\| \frac{\partial U(\boldsymbol{Z}, E)}{\partial Z_1} \right|_{Z_1=0} \geq \lambda P_1 \text{ for given } \boldsymbol{P}, M \right\}.$$

- Let $\Delta_1 = 0$ if the consumer does not purchase Z_1.
- Let $\Delta_1 = 1$ otherwise.
- If $F(\varepsilon)$ is the population distribution of E, the proportion purchasing none of good Z_1 given $\boldsymbol{P}, M$ is

$$\Pr(\Delta_1 = 0 \mid \boldsymbol{P}, M) = 1 - \int_{\underline{\underline{E}}} dF(\varepsilon).$$

- Provided inequality (35) is satisfied, $\Delta_1 = 1$ and interior solution demand function

$$Z_1 = Z_1(\boldsymbol{P}, M, E) \tag{36}$$

  is well defined and $Z_1 = Z_1^*$.
- When $\Delta_1 = 0$, observed $Z_1 = Z_1^* = 0$.

- Equation (36) is the conventional object of interest in consumer theory.
- Partial derivatives of that function *holding E and the other arguments constant* have well defined economic interpretations.
- Suppose that some non-negligible proportion of the population buys none of the good Z_1.
- Regression estimates of the parameters of (36) using Z_1^* approximate the conditional expectation

$$E(Z_1 \mid \Delta_1 = 1, \boldsymbol{P}, M) = \int_{\underline{\underline{E}}} Z_1(\boldsymbol{P}, M, \varepsilon) dF(\varepsilon) \qquad (37)$$

- The derivatives of (37) are different from the derivatives of (36).
- In order to define these derivatives, it is helpful to define $I_{\underline{\underline{E}}}(E)$ as an indicator function for set $\underline{\underline{E}}$ which equals one if $E \in \underline{\underline{E}}$ and equals zero otherwise.
- When prices or income change, the set of values of E that satisfy inequality (I-21) changes.
- Let $\underline{\underline{E}} + \Delta \underline{\underline{E}}_{\boldsymbol{P}}$ be the set of E values that satisfy (1.21) when there is a finite price change $\Delta \boldsymbol{P}$.

- $I_{E+\Delta E_P}(E)$ is an indicator function which equals one when $E \in \underset{=}{E} + \Delta \underset{=\boldsymbol{P}}{E}$.
- Then the derivatives of (37) are, for the jth price

$$\begin{aligned} \frac{\partial E(Z_1 \mid \Delta = 1, \boldsymbol{P}, M)}{\partial P_j} &= \int_{\underset{=}{E}} \frac{\partial Z_1(\boldsymbol{P}, M, \varepsilon)}{\partial P_j} dF(\varepsilon) \qquad (38) \\ &+ \lim_{\Delta P_j \to 0} \int_{\underset{=}{E}} \frac{[(I_{\underset{=}{E}+\Delta \underset{=P_j}{E}}(\varepsilon) - I_{\underset{=}{E}}(\varepsilon)] Z(\boldsymbol{P}, M, \varepsilon)}{\Delta P_j} dF(\varepsilon). \end{aligned}$$

- When the limit in the second term does not exist, the derivative does not exist.
- We assume for expositional convenience that the limit is well defined.

- The first expression on the right-hand side of (38) is the *average effect* of price change on commodity demand.
- The second term on the right-hand side of (38) arises from the change in sample composition of E as the proportion of non-purchasers changes in response to price change.
- This term generates the selection bias.

- *Neither term* is the same as the price derivative of (36) for an arbitrary value of $E = \varepsilon$ although the first term on the right-hand side of (38) approximates the price derivative of (36) for some value of $E = \varepsilon$.

- Just as in the statistical sample selection bias problem, there is a population of interest.
- In this case, the population parameters of interest are the distribution of E and the parameters of $U(\boldsymbol{Z}, E)$.
- Those who buy Z_1 are a self-selected sample of the population.
- Estimates of population parameters estimated on self-selected samples are biased and inconsistent.

- There is a population distribution of $Z_1(\boldsymbol{P}, M, E)$ generated by the distribution of E.
- Observations of Z_1 are obtained only if $E \in \underline{\underline{E}}(\omega(E) = 1 \text{ if } E \in \underline{\underline{E}}, \omega(E) = 0 \text{ otherwise})$.
- Alternatively one can express the inclusion criteria in terms of the latent population distribution of Z_1 induced by E (given $\boldsymbol{P}$ and M) and write $\omega(z_1) = 1$ if $z_1 > 0, \omega(z_1) = 0$ if $z_1 \leq 0$.

Cain, G. G. and H. W. Watts (1973). Summary and overview. In *Income Maintenance and Labor Supply: Econometric Studies*. Chicago: Rand McNally College Publishing Company.

Domencich, T. and D. L. McFadden (1975). *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland. Reprinted 1996.

Flinn, C. and J. J. Heckman (1982a, January). New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics 18*(1), 115–168.

Flinn, C. J. and J. J. Heckman (1982b). New methods for analyzing individual event histories. *Sociological Methodology 13*, 99–140.

Heckman, J. J. (1976, December). 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*(4), 475–492.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J. and B. S. Singer (1986). Econometric analysis of longitudinal data. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume 3, Chapter 29, pp. 1690–1763. Amsterdam: North-Holland.

Pearson, K. (1900). Mathematical contributions to the theory of evolution, VII: On the correlation of characters not quantitatively measureable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 195*(262–273), 1–47.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In G. Patil (Ed.), *Classical and Contagious Discrete Distributions; Proceedings*. New York: Pergamon Press.