Remarks on Inferences Under Selection

The examples on pages 66 and 71 discuss how inferences can be affected when data are non-randomly sampled. This note provides a little backround.

Consider two random vectors $x \in \Omega$ and $y \in \Phi$ where Ω and Φ are the respective sample spaces. The joint density of x and y is

$$p(x, y) = p(y|x)\, p(x), \quad x \in \Omega,\, y \in \Phi, \tag{1}$$

where the parameters of these distributions are omitted from the notation. The likelihood is proportional to (1).

Imagine that x is non-randomly sampled and that $S \subset \Phi$, a subset of Φ, is the set that includes the non-randomly sampled vector x. What are the consequences of selection operating on x, when inferences are based on the conditional likelihood $p(y|x)$? The conditional likelihood of y given that $x \in S$ is

$$\begin{aligned} p(y|x, x \in S) &= \frac{p(y, x|x \in S)}{p(x|x \in S)} \\ &= \frac{p(y, x)}{\Pr(x \in S)} \frac{\Pr(x \in S)}{p(x)}, \\ &= p(y|x), \quad x \in S,\, y \in \Phi, \end{aligned} \tag{2}$$

that takes the same mathematical form as if selection on x had not taken place. Selection on x can be ignored.

Let $z \in \Psi$ represent a third random vector with sample space Ψ. Assume that z is selected and $S \subset \Psi$, a subset of Ψ, includes the non-randomly sampled vector z. What are the consequences of non-random sampling of z if inferences are to be based on the conditional likelihood $p(y|x)$? The joint likelihood is $p(x, y, z)$ and the conditional likelihood is

$$\begin{aligned} p(y|x, z \in S) &= \frac{\int_S p(x, y, z)\, dz}{\int_S \int_\Phi p(x, y, z)\, dy dz} \\ &= \frac{\int_S p(x, y, z)\, dz}{\int_S p(x, z)\, dz}, \quad x \in \Omega,\, y \in \Phi, \end{aligned} \tag{3}$$

which is different from (2). For correct inferences, selection on z cannot be ignored and must be incorporated as part of the likelihood.

If z is independent of (x, y) so that $p(x, y|z) = p(x, y)$ and $p(x|z) = p(x)$, factorising numerator and denominator in (3):

$$\begin{aligned} p(y|x, z \in S) &= \frac{\int_S p(x, y|z)\, p(z)\, dz}{\int_S p(x|z)\, p(z)\, dz} \\ &= \frac{p(x, y)}{(x)} \frac{\int_S p(z)\, dz}{\int_S p(z)\, dz} \\ &= p(y|x), \quad x \in \Omega,\, y \in \Phi \end{aligned} \tag{4}$$

and the likelihood based on (4) is equal to the conditional likelihood of y given x as if selection on z had not occurred. These are special cases of a more general theory of inferences under missing data (Rubin, 1976; Little and Rubin, 1987).

References

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.