Efficiency of two-phase designs to correct for measurement error in regression

Roseanne McNamee[1] and Evridiki Batistatou[1]

[1]Biostatistics, Health Methodology Research Group, University of Manchester, UK

Measurement error can lead to substantial bias in regression problems but can be overcome if replicate measures are available; however replication may be expensive. Although there is a substantial literature on use of 2-phase case control studies to address measurement error problems (eg McNamee 2005), less attention has been paid to 2-phase designs for estimation of β in the Normal regression model $E[Y]=\alpha+\beta X$ where Y and X are continuous and X can only be measured with error by W. Here we consider the efficiency of four 2-phase designs for this problem. In design A, both Y and $W_1=X+\varepsilon_1$ –where ε_1 denotes random error - are measured on n 1st phase subjects and $W_2=X+\varepsilon_2$ on a fraction p of these in the 2nd phase. In design B, W_1 alone is measured for n 1st phase subjects and Y and W_2 in the second phase fraction. Variants of A and B allow 2nd phase subjects to be chosen randomly or from the extremes of the 1st phase W_1 distribution ('extreme sampling'). Berglund et al (2005) showed that, for design A, extreme sampling is more efficient than random sampling, but the overall efficiency of these 2-phase designs compared to usual, 'single-phase' designs has not been evaluated previously. Therefore it is not clear whether 2-phase designs should be recommended over usual practice.

We evaluated the efficiency of the four 2-phase (2P) designs compared to a single phase (1P) study - of the same overall cost in which Y, W_1 and W_2 are measured for all subjects - by the variance ratio $V_{1P}/V_{2P.}$ We also derived expression for the optimal values of p and n in the random sampling designs. In general, 2P efficiency depends on (i) the reliability λ of W, (ii) the cost ratio c_W/c_Y where c_W and c_Y are costs per subject of measuring Y and W respectively, and (iii) a function S=residual variance/$[V(X)\beta^2]$.

Design B with 2nd phase random sampling was always inefficient. Design A with random sampling was preferable to 1P designs when S, c_W/c_Y , and λ were all high but could be worse for low c_W/c_Y; optimal p varied strongly. Both designs A and B, when used with extreme sampling, were able to demonstrate marked increases in efficiency (egV_{1P}/V_{2P} =2), with design A being favoured when c_W/c_Y >>1 and design B when c_W/c_Y <<1; both could also be markedly inefficient in same cases. Results from further work involving multiple replicates and less extreme designs will also be presented and the implications of the results discussed.

1. McNamee R. Optimal design and efficiency of 2-phase case control studies with error-prone and error-free exposure measures. *Biostatistics* 2005; 6: 590-603.
2.Berglund, L., Garmo, H., Lindback, J., & Zethelius, B. "Correction for regression dilution bias using replicates from subjects with extreme first measurements", *Stat. Med 2006; 26: 2246-2257..*