Semiparametric regression models for indirectly observed outcomes

Jan De Neve

Department of Data Analysis
Ghent University
Jan.DeNeve@UGent.be

In several applications the outcome of interest is not measured directly, but instead a proxy (or multiple proxies) is (are) used. Examples include the body mass index as a proxy for body fat percentage, fluorescence intensity as a proxy for gene expression and the proportion of words correctly recalled as a proxy for the information stored in the memory. We illustrate by examples that the relationship between the outcome of interest and the proxy can be non-linear. The majority of the available methods (e.g. standard structural equation models), however, typically assume that this relationship is linear. We illustrate how slight deviations from linearity can have a substantial impact on the validity of these inferential procedures (Wagenmakers et al. 2012).

We therefore present a semiparametric regression strategy that quantifies the effect of observed covariates on a summary measure of the unobserved outcome without assuming linearity (but assuming monotonicity). We use the probabilistic index as a summary measure, i.e. the probability that the outcome of one subject exceeds the outcome of another subject, conditional on covariates (Thas et al., 2012, De Neve and Thas, 2015). Since this effect measure is invariant under monotone transformations, we do not need to model the relationship between the unobserved outcome and the proxy. By considering this relationship as nuisance, we can apply the inferential procedure to settings where the outcome of interest cannot be observed and hence only weak assumptions about the relationship can be imposed. The estimation strategy makes use of semiparametric linear transformation models (Cheng et al., 1995) due to their invariance under monotone transformations of the outcome. Since the relationship between the proxy and outcome of interest is typically subject to noise, we then extend these models to account for outcome measurement error. This extension uses an approach similar to the pseudo-value idea of Andersen et al. (2003).

References

Andersen, P. K., Klein, J. P., & Rosthoj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multistate models. *Biometrika*, 90(1), 15-27.

Cheng, S. C., Wei, L. J., & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4), 835-845.

De Neve, J., & Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511), 1276–1283.

Thas, O., De Neve, J., Clement, L., & Ottoy, J. P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4), 623–671.

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & cognition*, 40(2), 145-160.