StatNews #80

Propensity Score Analysis

March 2011

Randomized trials are considered to be the gold standard in study design. Random assignment of participants to treatment groups ensures that the groups are similar in their characteristics, thus removing any confounding. However, true experimental designs are not always practical, or even ethical to carry out. Study designs that lack randomization are called quasiexperimental designs. In such designs it is often the case that the experimental groups are not homogeneous on measured or unmeasured background variables that can be associated with the probability of being in a given experimental group (e.g. treatment vs. control) as well as with the outcome of interest. Confounding is also a potential problem in the observational studies, where assignment into treatment and control groups is outside the control of the investigator. When experimental groups differ on observed covariates in ways that affect study outcomes, there is a possibility of *overt* selection bias. Overt bias can be accounted for in analysis, allowing researchers to draw valid conclusions. However, because it is often not possible to measure all potential confounders, the experimental groups may be different in terms of unobserved characteristics, which may lead to *hidden* selection bias.

Controlling for confounding variables in statistical models by including them as predictors does not always eliminate bias, especially if there are a large number of such variables. Matching treatment and control group subjects on their background variables has been one of the earliest attempts to eliminate overt bias. However, matching on confounders falls short when confounding variables are continuous, and it is harder to match treatment and control subjects in the presence of multiple confounders. In contrast to overt bias, hidden bias cannot be accounted for, and sensitivity analysis is recommended for assessing the sensitivity of the model to hidden bias.

Better techniques to reduce overt selection bias in studies that lack randomization have been developed by statisticians and econometricians. The most widely used methods are: (a) Propensity score matching model (Rosenbaum & Rubin, 1983), which summarizes all confounding variables into one propensity score that is later used for matching, (b) Heckman's sample selection model (1978, 1979) and the related treatment effect model, which explicitly models the structure of selection into study groups, and (c) Matching estimators (Abadie & Imbens, 2002), which directly impute counterfactuals for participants in all groups. A counterfactual for a participant in a treatment group is the unobserved outcome for that participant under the control condition. Likewise, a counterfactual for a participant in the control group is the unobserved outcome for that participant under the treatment condition. Average treatment effect (ATE) can be defined as the difference between the observed outcome and the counterfactual, averaged

over all participants. The true value of the counterfactual is never observed, and, therefore, is approximated.

A propensity score is the predicted probability that a participant is assigned to a treatment group. This probability for each participant is usually obtained from a logistic regression model where a set of background variables thought to affect the probability of selection into the treatment group are the predictors, and the treatment is the outcome variable. Matching the treatment and control participants on their propensity scores approximates the counterfactual by choosing for each treated participant, one or more control participants with the similar values of the confounding variables. Within each matched pair, the treatment effect is the difference in the outcome of the treated and the control participants. In addition to matching on the propensity score, other methods such as weighting the participants by the reciprocal of their propensity score when modeling the outcome of interest, blocking on the propensity score, and regression on the propensity score have also been developed. Following propensity score matching, standard statistical methods such as General Linear Model (GLM), Generalized Linear Models, Survival Analysis, and Hierarchical Linear Model (HLM) can be used to estimate the treatment effect.

The most comprehensive statistical software packages that allow implementation of a variety of propensity score methods are STATA (StataCorp LP, 2010, College Station, TX, USA), and R (R Foundation for Statistical Computing, 2011, Vienna, Austria).

If you need assistance with a propensity score analysis problem, do not hesitate to contact a statistical consultant at the Cornell Statistical Consulting Unit.

References:

Heckman, J.J. (1978). Dummy endogenous variables in a simultaneous equations system. Econometrica, 46, 931-960.
Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica, 47, 153-161.
Rosenbaum, P.R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.
Rosenbaum, P.R. Design of Observational Studies. Springer, New York, 2010
Abadie, A., Imbens, G.W. (2002). Simple and bias-corrected matching estimators for average treatment effects [Technical Report]. http://www.nber.org/papers/t0283.pdf

Author: Shamil Sadigov.