

Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 20 : A Brief Introduction To Three More Topics

Propensity Scores
Log Links And Identity Links
Slope Ratio Assays

Propensity Scores

Propensity score [PS] methods have many proponents and many detractors.

For illustration here, we will suppose that we have a cohort study and that the primary objective is the relationship between a dichotomous outcome (D) and a dichotomous exposure (E).

A PS analysis typically has two steps. In the first step, one begins with modelling the exposure in terms of explanatory variables [but not including the outcome (D) as an explanatory variable]. For example, using logistic regression :

$$\epsilon = Pr(E) \quad \log\left(\frac{\epsilon}{1-\epsilon}\right) = \sum_{j=0}^I \alpha_j x_j \quad \log\left(\frac{e}{1-e}\right) = \sum_{j=0}^I a_j x_j$$

In the PS literature, the probabilities of exposure ϵ are called the propensity scores. The estimates e of ϵ are the fitted values on the probability scale.

There are a number of strategies for the second step suggested using this model and its fit. With these strategies, the analyst is using the estimated propensity score as an explanatory variable. These estimates all have standard errors. These standard errors are functions of the explanatory variables in the first step. In principle, all results from the second step need to accommodate variation implicit in the first step.

a) stratified analysis

One might construct k strata based on the estimated propensity scores. One then has k stratified 2x2 tables with D [yes, no] as columns and E [yes, no] as rows. One could expand the stratification with other dichotomous explanatory variables.

b) unmatched model based analysis

One might construct logistic regressions. There are many variations. For example, assuming linearity of the log odds of disease with the log odds of estimated propensity, one might have :

$$p = Pr(D) \\ \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 \log\left(\frac{e}{1-e}\right) + \beta_3 E \log\left(\frac{e}{1-e}\right)$$

Notice that this model is :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 E + \beta_2 \sum_{j=0}^l a_j x_j + \beta_3 E \sum_{j=0}^l a_j x_j = \beta_0 + \beta_1 E + \sum_{j=0}^l \beta_2 a_j x_j + \sum_{j=0}^l \beta_3 a_j E x_j$$

Notice that this model shares some of the features of a model that addresses the explanatory variables x_1, x_2, \dots, x_l as potential modifiers. The terms involving $E x_j$ are all implicitly included but with a substantial set of constraints and assumptions.

c) matched analysis

One might establish matched pairs. Each pair contains a person exposed [$E=1$] with a person unexposed [$E=0$]. Each member in the pair has a 'similar' estimated propensity score. Then one constructs the 2x2 table, the concordant and discordant pairs. One might establish matched sets and try conditional logistic regression.

King & Nielsen [2016] argue against PS Matching.

d) other more elaborate methods