

## Lab3: Using Targeted MLE to estimate the marginal effect of a risk factor - due Nov. 20

October 30, 2006

This lab involves analysis of a data set collected on old guys called the Western Collaborative Group Study (WCGS). The goal was to examine what risk factors caused heart attacks in old guys and not so old guys. Specifically, WCGS was a prospective cohort study, which recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured to provide the best possible assessment of the CHD risk associated with behavior type. Additional variables collected include age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and corneal arcus (a fatty deposit in the eye). The goal is to use target MLE (as discussed in class) to estimate the marginal difference in the probabilities of CHD in type A vs. type B subjects:

$$\Psi(p) = E_p(E_p(Y | A = 1, W) - E_p(Y | A = 0, W))$$

which, if the randomization assumption (as well as others) is satisfied, this equals:  $EY_1 - EY_0$ . In words, one is modeling the difference of the probability that a man (in this target population) has a heart attack during the study period given every in the population has  $A = 1$  versus everyone has  $A = 0$ . In our case,  $A = 1$  is a type A personality (driven, stressed out);  $A = 0$  is a type B (mellow dude). Note, the data file on the website (`wcgs.csv`) has the variables labeled as to their roles. There is one labeled  $V$  (age) - ignore that for now and treat it as another  $W$ .

Note, to implement this you will need to estimate two nuisance parameters:  $g(A | W) \equiv P(A | W)$  and  $Q(A, W) \equiv E(Y | A, W)$  - for now, I leave how to estimate these up to you (e.g., simple logistic regression models with only main effects). In a future assignment, we will explore using likelihood based cross-validation to choose among competing models of nuisance parameters.

Include in your report:

- Your annotated code,
- the naive estimate (unadjusted) with standard inference (standard errors),
- the targeted MLE estimate with inference from the nonparametric bootstrap,
- a short description of the results which contrasts the naive and targeted MLE estimate.