

PS C236A/ Stat C239A
Problem Set 6
Due: October 16, 2009

Problem 1 An alternative distance metric to the propensity score is Mahalanobis distance. This metric reduces the multidimensional problem of multivariate matching to a unidimensional problem. Although Mahalanobis distance was originally developed for use with multivariate Normal data, we often encounter covariates that are not normally distributed. This problem will explore the implications of these non-normal variables on this distance metric.

- a. When including a binary variable in a Mahalanobis distance metric, will a binary variable with $p = 1/2$ or a binary variable with p near zero be given greater weight by this distance metric? Prove why this is true mathematically.
- b. How will this distance metric treat covariates with outliers? How about covariates that have long-tailed distributions?
- c. Should we or shouldn't we be concerned by the behavior of the Mahalanobis distance metric for the covariate distributions described in parts a and b? Why?

Problem 2 When analyzing data from a regression discontinuity design, our desired estimand is the $\tau_{RD} = E[Y(1) - Y(0)|X = c]$. Remember that X is the “forcing variable” and “ c ” is the point at which units switch from control to treatment. If the discontinuity design is valid, we would like to simply estimate $E[Y|T = 1, X = c] - E[Y|T = 0, X = c]$, but due to the absence of data immediately at the cutpoint, we are forced to extrapolate by using data in a window around c . The size of that window is determined by h , which is known as the “bandwidth”.

Because choice of the bandwidth is somewhat arbitrary, Imbens (2009) recommends combining local linear regression (discussed in section) with a “cross-validation” procedure for choosing h . The basic idea behind this approach is the following. Consider an observation i . To see how well a linear regression with a bandwidth h fits the data, we run a regression with observation i left out and use the estimates to predict the value of Y at $X = x_i$. To emulate the fact that RD estimates are based on regression estimates at the boundary, the regression is estimated using only observations with values of X on the left of X_i ($X_i - h \leq X < X_i$) for observations on the left of the cutpoint ($X_i < c$). For observations on the right of the cutoff point ($X_i \geq c$), the regression is estimated using only the observations with values of X on the right of X_i ($X_i < X \leq X_i + h$).

After repeating this procedure for each and every observation, we will have a collection of predicted values of Y that can be compared to the actual values of Y . The optimal bandwidth can be picked by choosing the value of h that minimizes the mean square of the difference between the predicted and actual value of Y .

Formally, let $\hat{Y}(X_i)$ be the predicted value of Y obtained using the regressions described above. The cross validation criterion is defined as

$$CV_Y(h) = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}(X_i))^2$$

with the corresponding cross-validation choice for the bandwidth

$$h_{CV}^{opt} = \arg \min_h CV_Y(h)$$

For a more detailed discussion of this method, see: <http://www.econ.ubc.ca/lemieux/papers/designs.pdf>

- a. Use the cross-validation procedure described above to calculate h_{CV}^{opt} for a trimmed subset of the Brazilian mayoral election data discussed in section (posted on the course website). The forcing variable is `vote.margin`, the treatment indicator is `treat`, and the outcome variable is `PMDB.win.04`. Select a range of possible h to check in your procedure (say $h = .01, .02, .03, \dots, .3$.)
- b. Using local linear regression, estimate the local average treatment effect (τ_{rd}) and its associated standard error with your h_{CV}^{opt} calculated in part a.