

Problem 1: An alternative distance metric to the propensity score is the Mahalanobis distance. This metric also reduces a multi-dimensional problem into a uni-dimensional problem. The Mahalanobis distance was originally developed for use with multivariate Normal data, however, we often use covariates that are not normally distributed in our matching methods. This problem will explore what the implications of these non-normal variables are for this distance metric.

- a. When including a binary variable in a Mahalanobis distance metric, will a binary variable with $p = \frac{1}{2}$ or a binary variable with p near zero be given greater weight by this distance metric? Prove why this is true mathematically.

The Mahalanobis distance is defined as:

$$D_m(X_i, X_j) = \{(X_i - X_j)^T S^{-1} (X_i - X_j)\}^{\frac{1}{2}}$$

Where S^{-1} is the sample covariance matrix of X .

A binary variable with probability of success p has variance $p(1 - p)$. A variable with $p = \frac{1}{2}$ therefore has variance of $p(1 - p) = \frac{1}{4}$, whereas a variable with p near 0 would have variance near 0. Since we take the inverse of the sample covariance matrix, therefore dividing by the variance, a variable with $p = \frac{1}{2}$ will be given less weight than a variable with p near 0 (or, similarly a variable with p near 1). By FOC, we can show that the variance of a binary variable is greatest when $p = \frac{1}{2}$, so a binary variable with $p = \frac{1}{2}$ will be given less weight than any binary variable with $p \neq \frac{1}{2}$.

- b. How will this distance metric treat covariate distributions that have outliers? How about covariates that have long-tailed distributions?

Variables with long tails or extreme outliers tend to have inflated variances, and by the same logic as above, any variable with larger variance will be given relatively less weight.

- 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?

We should be concerned. Outliers and long tails do not make a covariate unimportant, so we may not wish to downweight it relative to other covariates. Binary variables that are very rare may not be of overriding importance, so it may not be wise to give them significantly higher weight than binary variables with p closer to $\frac{1}{2}$. However, if it is a rare binary event, then we might want to treat a difference in outcome as worse than a difference in outcome for a covariate where p is closer to $\frac{1}{2}$. Overall, we should be concerned that Mahalanobis distance exhibits these behaviors for variables for which the theory was not designed.

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.

Solution: See code for Section 7.