Overview of the Causal Inference Methods

Johnson Zhang, Diane Lu

1. Causal Effects

Firstly, we introduce the definition of causal effects. Suppose we have a random sample of size N from a large population. For each unit i in the random sample, we use $T_i \in \{0, 1\}$ to denote whether the unit i received the treatment of interest. Let $Y_i(0)$ indicates the outcome that the unit i was under control while $Y_i(1)$ indicates the outcome under treatment. For unit i the treatment effect is $Y_i(1) - Y_i(0)$. We are interested in the average effect of the treatment in the whole population (to simply the notation, we suppress subscript i for unit):

$$\Delta = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

We call the above effect as Average Treatment Effect(ATE). Sometimes we also discuss methods of estimating the Average Treatment Effect on the Treated(ATT):

$$\Delta_t = E(Y_1 - Y_0|T = 1)$$

In project 4, we mainly focus on ATE given that the dataset we used is for ATE instead of ATT.

2. Propensity Scores

We define the propensity score as

$$e(x) = \Pr(T = 1|X = x)$$

we assume that

$$0 < e(x) < 1$$

for all x, here we denote X as the covariates of p-dimensional vector of pre-treatment variables.

3. Calculation for Propensity Scores

(See Westreich, Lessler, and Funk (2010), Friedman, Hastie, and Tibshirani (2010) and Hastie, Tibshirani, and Friedman (2009))

3.1 Logistic Regression:

The logistic regression model represents the class-conditional probabilities through a linear function of the predictors:

$$logit[\Pr(T = 1|X)] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

$$\Pr(T = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}}$$

3.2 L1 Penalized Logistic Regression

To avoid overfitting of the logistic regression model, we introduce regularization term to decrease the model variance in the loss function Q. In order to achieve this, we modifying the loss function with a penalty term which effectively shrinks the estimates of the coefficients. In this case, the penalty term is 'L1 norm':

$$Q = -\frac{1}{n}\sum_{i=1}^{n}[y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) + \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))] + \lambda\sum_{j=1}^{p}|\beta_j|$$

where $Y \in \{0, 1\}$

3.3 L2 Penalized Logistic Regression

The only difference between L1 and L2 penalized logistic regression is for the regularization term, where L2 is the 'L2 norm':

$$Q = -\frac{1}{n}\sum_{i=1}^{n}[y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) + \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}))] + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

3.4 Regression Trees(CART)

Decision trees are classification algorithms, which specify a 'tree' of cut points that minimize some measures of diversity in the final nodes once the the tree is complete. We can apply most of the decision tree models into constructing propensity categories. What's more, if we need a probability of class membership, classification and regression trees(CART) can provide such probabilities. In this case, we choose CART for propensity scores calculation.

For CART method, we first split the space into two regions, and model the response by the mean of Y in each region. We choose the variable and split-point to achieve the best fit. Then one or both of these regions are split into two more regions, and this process is continued, until some stopping rule is applied. The corresponding regression model predicts Y with a constant c_m in region R_m, that is,

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

3.5 Boosting Stumps

Boosting algorithms can have several advantages including high performance and possibility outcome for calculating propensity scores. We can use a boosting stumps to calculate the propensity score. We can represent the boosting Stumps model as an additive model:

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m)$$

where $T(x; \theta)$ is the stump, Θ_m is the parameter of the tree stump, M is the number of tree stumps.

4. Propensity Score Matching(Full matching)

Full matching creates a series of matched sets, where each matched set contains at least one treated individual and at least one control individual (and each matched set may have many from either group). Full matching forms these matched sets in an optimal way, such that treated individuals who have many comparison individuals who are similar (on the basis of the propensity score) will be grouped with many comparison individuals, whereas treated individuals with few similar comparison individuals will be grouped with relatively fewer comparison individuals.(See Stuart (2010))

4.1 Mahalanobis Metric (Do not need Propensity Score)

The Mahalanobis distance is defined as:

$$D_{ij} = (X_i - X_j)^T \sum^{-1} (X_i - X_j)$$

$\sum$ is the variance covariance matrix of X in the pooled treatment and full control groups. If X contains categorical variables they should be converted to a series of binary indicators, although the distance works best with continuous variables.

4.2 Propensity Score

The distance of Propensity Score is defined as:

$$D_{ij} = |e_i - e_j|$$

where e_k is the propensity score for individual k

4.3 Linear Propensity Score

$$D_{ij} = |logit(e_i) - logit(e_j)|$$

After the matched sets are obtained, calculate a "subclass effects" for each matched set/subclass, and then estimate overall ATE by an weighted average of the subclass effects where weights would be the number of individuals in each subclass.

5. Inverse Propensity Weighting

Propensity scores can also be used directly as inverse weights in estimates of the ATE, known as inverse probability of treatment weighting (IPTW). The weight w_i is:

$$w_i = \frac{T_i}{\hat{e}_i} + \frac{1 - T_i}{1 - \hat{e}_i}$$

where $\hat{e}_i$ is the estimated propensity score for individual i. This weighting serves to weight both the treated and control groups up to the full sample, in the same way that survey sampling weights weight a sample up to a population. (See Stuart (2010) and Austin (2011))

The estimate ATE using IPW is:

$$\hat{\Delta}_{IPW} = N^{-1}(\sum_{i \in treated} w_i Y_i - \sum_{i \in controlled} w_i Y_i)$$

6. Doubly Robust Estimation

Doubly robust estimator has the smallest asymptotic variance.

$$\hat{\Delta}_{DR} = N^{-1} \sum_{i=1}^{N} \frac{T_i Y_i - (T_i - \hat{e_i})\hat{m}_1(X_i)}{\hat{e_i}} - N^{-1} \sum_{i=1}^{N} \frac{(1 - T_i)Y_i + (T_i - \hat{e_i})\hat{m}_0(X_i)}{1 - \hat{e_i}}$$

where $\hat{m_t}(X)$ is a consistent estimate for $E(Y|T = t, X)$ and is usually obtained by regressing the observed response Y on X in group t (where t = 0,1).

"Doubly robust" in the sense that the estimator remains consistent if either (i) if the propensity score model is correctly specified but the two regression models m_0 and m_1 are not or (ii) the two regression models are correctly specified but the propensity score model is not, although under these conditions it might not be the most efficient.(See Lunceford and Davidian (2004))

Motivation of how this estimator is derived to reduce large-sample variance:

$$\begin{aligned} E[Y_1 - Y_0] &= E[E(Y|T = 1, X) - E(Y|T = 0, X)] + E[(Y_1 - E(Y|T = 1, X)) - (Y_0 - E(Y|T = 0, X))] \\ &= E[E(Y|T = 1, X) - E(Y|T = 0, X)] + E[(\frac{I[T = 1]}{\Pr(T = 1|X)} - \frac{I[T = 0]}{\Pr(T = 0|X)})(Y_T - E(Y|T, X))] \end{aligned}$$

The key idea for the motivation is that $Y_T - E(Y|T, X)$ usually have smaller variability than Y_T. The second equality comes from using inverse propensity weighting to obtain an unbiased estimate for the second term. Then, estimating $E(Y|T = t, X)$ by $\hat{m_t}(X)$, $\Pr(T = 1|X = x)$ by $\hat{e}(x)$, and $\Pr(T = 0|X = x)$ by $1 - \hat{e}(x)$ could give you $\hat{\Delta}_{DR}$.

7. Regression Estimate

A simple regression estimate that doesn't make use of the propensity score:

$$\hat{\Delta}_{reg} = N^{-1} \sum_{i=1}^{N} (\hat{m}_1(X_i) - \hat{m}_0(X_i))$$

(Similar to the regression estimate in Chan et al. (2010), except here we're interested in estimating ATE instead of ATT)

8. Stratification

A common approach to estimate ATE using stratification based on propensity scores.

$$\hat{\Delta}_S = \sum_{j=1}^{K} \frac{N_j}{N} \{N_{1j}^{-1} \sum_{i=1}^{N} T_i Y_i I(\hat{e}_i \in \hat{Q}_j) - N_{0j}^{-1} \sum_{i=1}^{N} (1 - T_i) Y_i I(\hat{e}_i \in \hat{Q}_j)\}$$

where K is the number of strata, some literature have advocate to use quintiles (K=5). N_j is the number of individuals in stratum j. N_{1j} is the number of "treated" individuals in stratum j, while N_{0j} is the number of "controlled" individuals in stratum j. $\hat{Q}_j = (\hat{q}_{j-1}, \hat{q}_j]$ where $\hat{q}_j$ is the jth sample quantile of the estimated propensity scores. (See Lunceford and Davidian (2004))

9. Regression Adjustment

Regress the outcome variable Y on treatment indicator variable T and the estimated propensity score. Then, the estimated coefficient on the treatment indicator variable would be an estimate of ATE.(See Austin (2011))

10. Weighted Regression

Weighted least square estimation of the regression function:

$$Y_i = \alpha_0 + \tau \cdot T_i + \alpha_1' \cdot Z_i + \alpha_2' \cdot (Z_i - \bar{Z}) \cdot T_i + \varepsilon_i$$

The weights are the ones mentioned in "Inverse Propensity Weighting". The Z_i are a subset of the covariates X_i; with sample average $\bar{Z}$. τ is an estimate for ATE.

How to select Z?

Estimate linear regressions:

$$Y_i = \beta_{k0} + \beta_{k1} \cdot T_i + \beta_{k2} \cdot X_{ik} + \varepsilon_i$$

We calculate the t-statistic for the test of the null hypothesis that the slope coefficient β_{k2} is equal to zero in each of these regressions, and now select for Z all the covariates with a t-statistic larger in absolute value than t_{reg}. Thus, we include in the final regression all covariates which have substantial correlation with the outcome conditional on the treatment. (See Hirano and Imbens (2001))

Reference

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3): 399–424.

Chan, David, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. "Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale." In *Proceedings of the 16th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 7–16.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hirano, Keisuke, and Guido W Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2 (3-4): 259–78.

Lunceford, Jared K, and Marie Davidian. 2004. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects a Comparative Study." *Statistics in Medicine* 23 (19): 2937–60.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25 (1): 1.

Westreich, Daniel, Justin Lessler, and Michele Jonsson Funk. 2010. "Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (Cart), and Meta-Classifiers as Alternatives to Logistic Regression." *Journal of Clinical Epidemiology* 63 (8): 826–33.