Week 8 Propensity Scores

Let $z = 1, 0$ T/C $\underset{\sim}{X}$ vector of covariates

propensity score $e(\underset{\sim}{X}) = Pr(z=1 | \underset{\sim}{X})$ scalar $\hat{e}(\underset{\sim}{X})$

cond'l prob unit w vector $\underset{\sim}{X}$ observed cov. assigned to $T (z=1)$

Thm Balancing score $b(\underset{\sim}{X})$ s.t. conditional distrib of $\underset{\sim}{X}$ given $b(\underset{\sim}{X})$ same of treated and control units

$\underset{\sim}{X} \perp\!\!\!\perp z | b(\underset{\sim}{X})$. Coarsest (low dimen) balancing score is propensity score. $Pr(X, z|e) = Pr\{\underset{\sim}{X}|e\} Pr(z|e)$

Thm (result) Approx 90% reduction in bias for subclassifying at quintiles of population propensity score. $B_I = E(f(\underset{\sim}{x})|z=1) - E(f(\underset{\sim}{x})|z=0)$, B_S after strat, treatm percent reduction in bias $100(1 - B_S/B_I) \approx 90\%$

Ros Rubin
1983 Biometrika
1984 JASA

(i) The propensity score is a balancing score.
(ii) Any score that is 'finer' than the propensity score is a balancing score; moreover, x is the finest balancing score and the propensity score is the coarsest.
(iii) If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score
(iv) At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects.
(v) Using sample estimates of balancing scores can produce sample balance on x.

Applications: Rubin Breast Cancer, Love (RR '84) CAD, Love Aspirin, Hansen SAT coaching, Substance Abuse (UNC)
Rosenbaum, Danish downers

Rubin Ann Int Medicine

Table 3: Estimated 5-year Survival Rates for Node-Negative Patients in SEER from Tables 5 and 7 in U.S. GAO Report (1994). AIM pub

Propensity Score

Subclass	Treatment	n	Estimate	n*	Estimate*
1	Breast Conservation	56	85.6%	54	88.8%
	Mastectomy	1,008	86.7%	966	90.5%
2	Breast Conservation	106	82.8%	102	86.0%
	Mastectomy	964	82.8%	917	87.7%
3	Breast Conservation	193	85.2%	184	89.4%
	Mastectomy	866	88.8%	841	91.4%
4	Breast Conservation	289	88.7%	279	92.0%
	Mastectomy	978	87.3%	742	91.5%
5	Breast Conservation	462	89.0%	453	90.7%
	Mastectomy	604	88.5%	589	90.7%

* omitting patients whose deaths were unrelated to cancer.

Lalonde data

Lab 4 stratification

```
> table(propbin, treat)
                 treat
propbin            0   1
  (0,0.0401]     122   1
  (0.0401,0.0872] 116   7
  (0.0872,0.27]  101  21
  (0.27,0.671]    53  71
  (0.671,1]       37  85
> tapply(re78, list(propbin, treat),mean)
                    0    1
(0,0.0401]      10467    0
(0.0401,0.0872]  5797 7919
(0.0872,0.27]    6043 9211
(0.27,0.671]     4977 5819
(0.671,1]        4666 6030
```

counts

means re78

Stat 209 [redacted]

p. 2

matchit package:MatchIt R Documentation

redacted by drr

MatchIt: Matching Software for Causal Inference

Description: 'matchit' is the main command of the package _MatchIt_, which enables parametric models for causal inference to work better by selecting well-matched subsets of the original treated and control groups. MatchIt implements a wide range of sophisticated matching methods, Matched data sets created by MatchIt can be entered easily in Zelig (<URL: http://gking.harvard.edu/zelig>) for subsequent parametric analyses. Full documentation is available online at <URL: http://gking.harvard.edu/matchit>, and help for specific commands is available through 'help.matchit'.

Usage: matchit(formula, data, method = "nearest", distance = "logit", distance.options = list(), discard = "none", reestimate = FALSE, ...)

Arguments: formula: This argument takes the usual syntax of R formula, 'treat ~ x1 + x2', where 'treat' is a binary treatment indicator and 'x1' and 'x2' are the pre-treatment covariates. Both the treatment indicator and pre-treatment covariates must be contained in the same data frame, which is specified as 'data' (see below). All of the usual R syntax for formula works. For example, 'x1:x2' represents the first order interaction term between 'x1' and 'x2', and 'I(x1^2)' represents the square term of 'x1'.

data: This argument specifies the data frame containing the variables called in 'formula'

method: This argument specifies a matching method. Currently, '"exact"' (exact matching), categorical vars '"full"' (full matching), Ben Hansen optmatch (gender equity) '"genetic"' (genetic matching), Sekhon fancy '"nearest"' (nearest neighbor matching), historical method '"optimal"' (optimal matching), and Ben Hansen optmatch, makes (2:1) '"subclass"' (subclassification) are available. The default is '"nearest"'. Note that within each of these matching methods, _MatchIt_ offers a variety of options. See <URL: http://gking.harvard.edu/matchit/docs/Inputs.html> for the complete list

References: Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart (2004) `Matching as Nonparametric Preprocessing for Improving Parametric Causal Inference,'' preprint available at <URL: http://gking.harvard.edu/files/abs/matchp-abs.shtml>

See Also: Please use 'help.matchit' to access the matchit reference manual. The complete document is available online at <URL: http://gking.harvard.edu/matchit>.

match.data package:MatchIt R Documentation

Output Matched Data Sets

get the list of matches

Description: 'match.data' outputs matched data sets from 'matchit()'.

Usage: match.data <- match.data(object, group="all", distance = "distance", weights = "weights", subclass = "subclass")

Arguments:

object: The output object from {\tt matchit()}. This is an required input.

group: This argument specifies for which matched group the user wants to extract the data. Available options are '"all"' (all matched units), '"treat"' (matched units in the treatment group), and '"control"' (matched units in the control group). The default is '"all"'.

Value: Returns a subset of the original data set sent to 'matchit()', with just the matched units. The data set also contains the additional variables 'distance', 'weights', and 'subclass'. The variable 'distance' gives the estimated distance measure, and 'weights' gives the weights for each unit, generated in the matching procedure. The variable 'subclass' gives the subclass index for each unit (if applicable). See the <URL: http://gking.harvard.edu/matchit/> for the complete documentation and type 'demo(match.data)' at the R prompt to see a demonstration of the code. :

pdf or html manual