Centaur

Sara E. Dempster

May 10, 2017
OHDSI Population level estimation WG

Package 'Centaur'

April 5, 2017

Type Package

Title Centaur Propensity Score Balancing Workflow and Toolkit

Version 1.0.0

Date 2017-04-06

Author Sara Dempster
Stephen Kottmann
Alex Bayeh

Maintainer Alex Bayeh <alex.bayeh.centaur@gmail.com>

Description Performs propensity score based population balancing. This package is a toolkit to calculate propensity scores, balancea population dataset via either weighting or matching, and perform a variety of diagnostics to assess the scientific validity of the approach. The authors acknowledge the following team from AstraZeneca Pharmaceuticals, Robert LoCasale, Michael Goodman, Ramin Arani, Yiduo Zhang, and Sudeep Karve for contributing to the requirements with their expertise in epidemiology, safety informatics, health economics and biostatics and for reviewing the final product. The authors also acknowledge Jonathan Herz and Pramod Kumar for help with testing early versions of the package.

License Apache License 2.0

LazyData no

RoxygenNote 6.0.1

Imports AUC, broom, data.table, dplyr, ff, gtools, Hmisc, MASS

Background

- started as a component of an internal project at AZ for overall platform development to standardize and scale up observational data analysis

- team wanted a propensity score/cohort method workflow in R to validate against existing SAS code sets

- team wanted recommended workflows and parameter settings, but also flexible options for advanced users

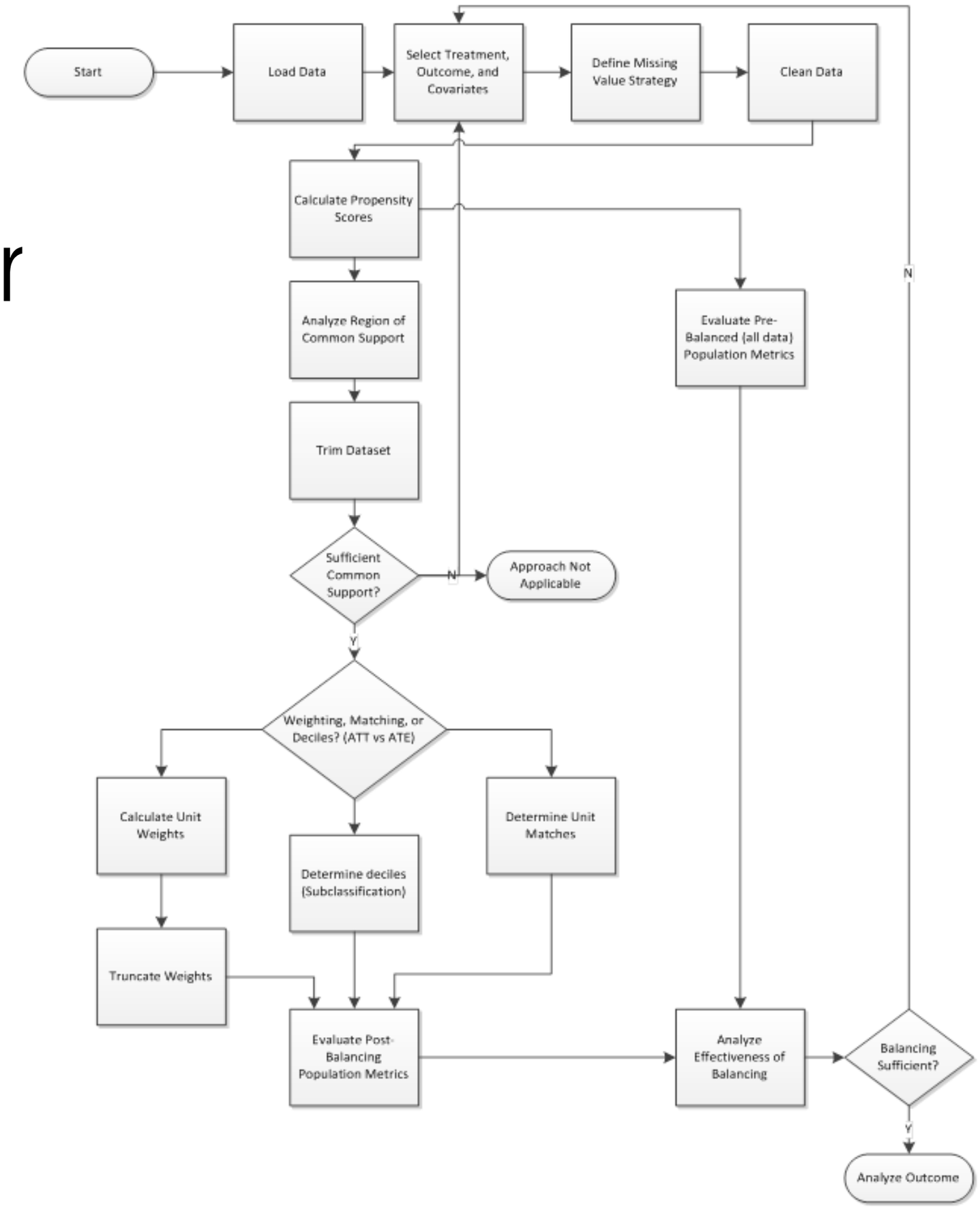
- team wanted to draw on commonly used R packages i.e Twang, MatchIT, but have all integrated into one framework and workflow.

- main original use case = quick feasibility analysis on patient balance (exclude outcome analysis)

- package including outcome analysis was used for internal validation of CVD-REAL results

Centaur

Workflow diagram



Whitepaper coming soon

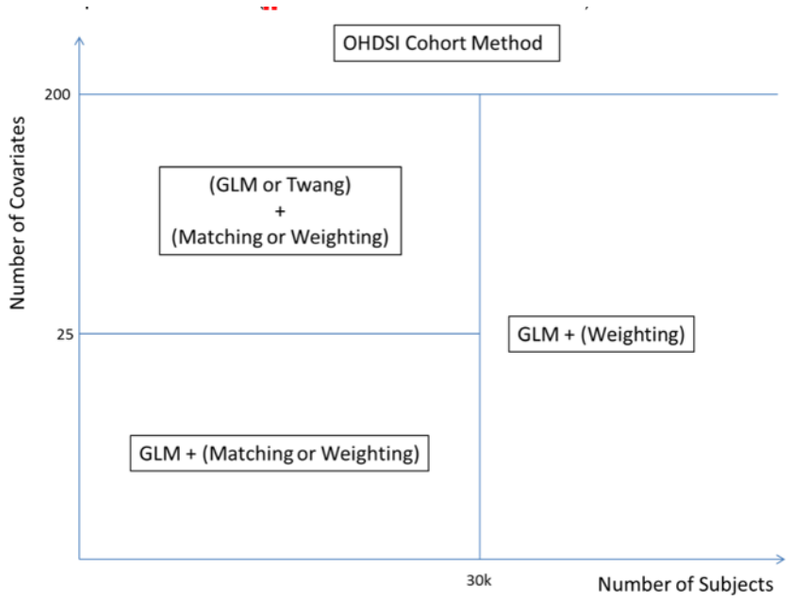


Figure 2 - "Phase Diagram" of available methods

Table 1 – Twang Parameter Analysis

Trial	Number of Trees	Depth	Shrinkage	Bag Fraction	Estimand	Time to Run	Avg. % Reduction in Std. Diff. of Means
1	10000	3	0.01	1	ATT	508 s	82.11439
2	5000	3	0.01	1	ATT	469 s	81.98465
3	2000	3	0.01	1	ATT	378 s	81.36216
4	1000	3	0.01	1	ATT	347 s	77.14936
5	10000	3	0.005	1	ATT	499 s	81.22527
6	10000	3	0.05	1	ATT	514 s	81.22527
7	10000	3	0.1	1	ATT	508 s	82.62994
8	10000	2	0.01	1	ATT	456 s	88.67967
11	10000	3	0.01	1	ATE	538 s	82.05354
12	5000	2	0.01	1	ATT	405 s	92.02401

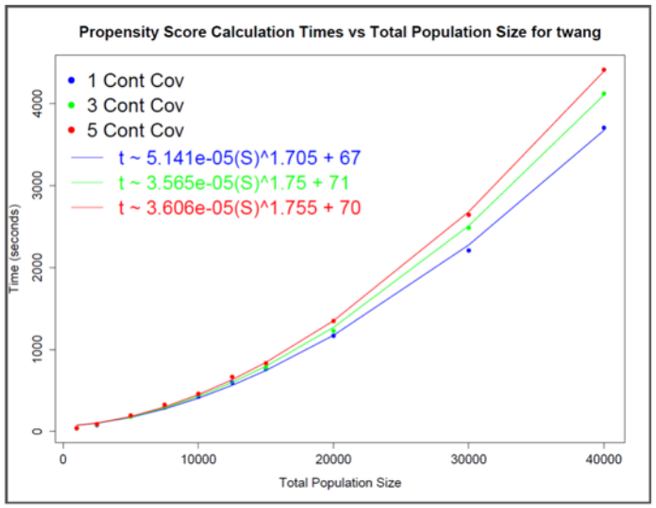


Figure 3 – Computational time for PS calculation using Twang

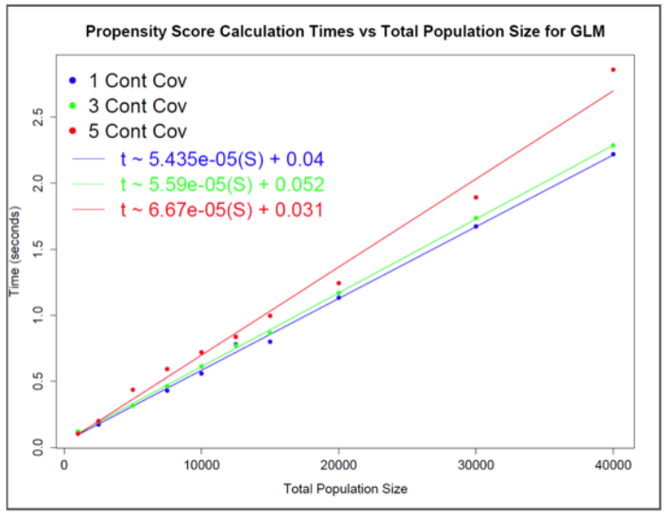


Figure 4 – Computational time for PS calculation using GLM

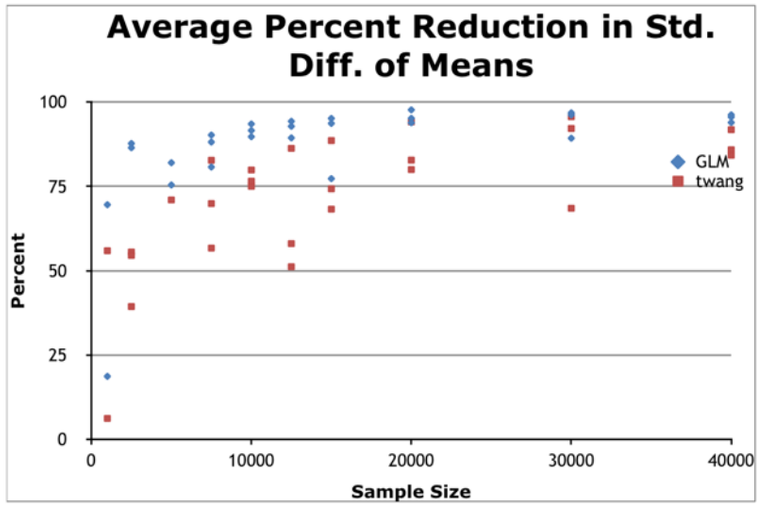


Figure 5 – Performance comparison of GLM and twang

Vignette coming soon

Load Data

A simple dataset has been included in the `drive.ps` package to support this vignette. The full T2DM cohort has been downsampled to include 40k samples for each of the drug classes (with the exception of AGI). The data is included as an internal resource, and the details of the data can be viewed with:

```
ps.getDataAvailability()
```

```
##            DRUGCLASS  freq
## 1                AGI   810
## 2          Biguanide 40000
## 3       Combinations 40000
## 4               DPP4 40000
## 5            Insulin 40000
## 6       No T2DM Drug 40000
## 7              Other 40000
## 8      Sulfonylureas 40000
## 9 Thiazolidinediones 40000
```

A utility method has been included in the package to create new datasets for comparing any two of these treatment groups. For this vignette, create a dataset comparing Biguanides to a No Drug control group:

```
myData <- ps.createDataset("Biguanide", 10000, control.name = "No T2DM Drug", control.number
= 20000)
```

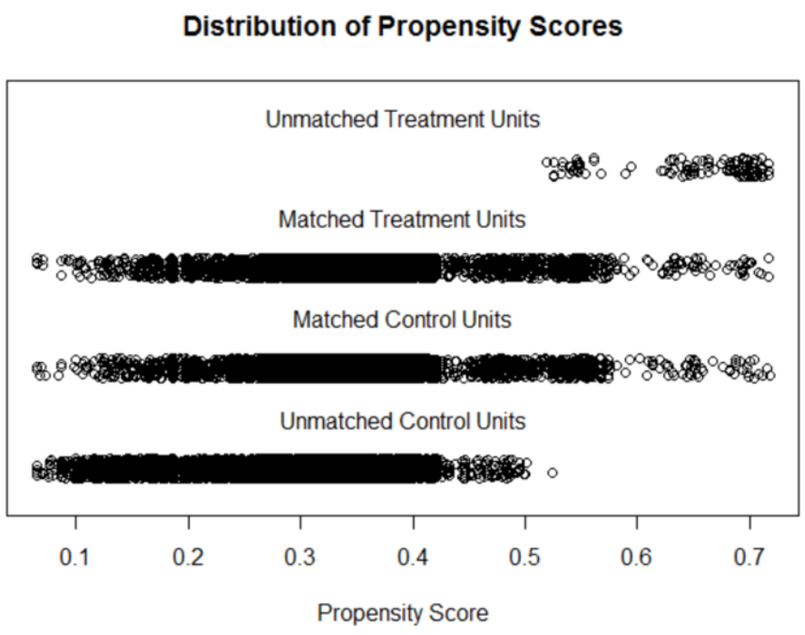
Calculate Propensity Scores

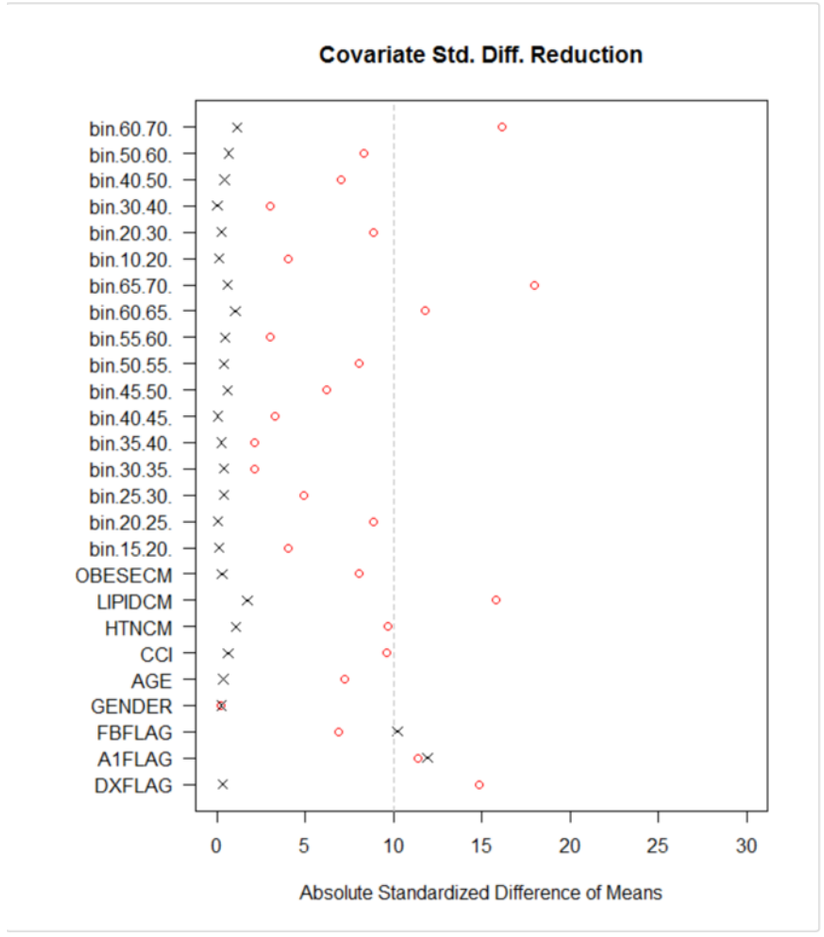
Propensity scores are calculated via the `ps.score` method. The primary inputs to this method are the dataset, the list of covariates to include in the calculation, and the propensity score method. The two methods available are 'glm' and 'twang' (twang is, by default, only available for datasets with less than 30k samples).

```
myData <- ps.score(myData, T2DM.covariates, ps.method = "glm")
```

```
##
## Call:  glm(formula = formula, family = binomial(), data = data, control = list(maxit =
100))
##
## Coefficients:
## (Intercept)       DXFLAG       A1FLAG       FBFLAG       GENDER
##   -1.247274     1.811913     0.618708     0.660038     0.007213
##         AGE          CCI        HTNCM      LIPIDCM      OBESECM
##   -0.026565    -0.097182    -0.079160    -0.239265    -0.142157
##  bin.15.20.   bin.20.25.   bin.25.30.   bin.30.35.   bin.35.40.
##   -1.198418    -1.322439    -0.713089    -0.152973    -0.048070
##  bin.40.45.   bin.45.50.   bin.50.55.   bin.55.60.   bin.60.65.
##    0.108488     0.279788     0.435822     0.480809     0.426177
##  bin.65.70.   bin.10.20.   bin.20.30.   bin.30.40.   bin.40.50.
##          NA           NA           NA           NA           NA
##  bin.50.60.   bin.60.70.
##          NA           NA
##
## Degrees of Freedom: 29999 Total (i.e. Null);  29980 Residual
## Null Deviance:       38190
## Residual Deviance: 37430     AIC: 37470
```

The data frame returned by ps.score includes the original T2DM cohort data frame, but has added a new variable `ps_values`.





Centaur

OHDSI Cohort Method

glm LR
twang (GBM)
MatchIt
violin plots
not dependent on CDM

PS plots
Balance diagnostics

regularized LR
dependent on CDM

variable selection

Variable Selection for Propensity Score Models

M. Alan Brookhart ✉, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, Til Stürmer

Am J Epidemiol (2006) 163 (12): 1149-1156. **DOI:** https://doi.org/10.1093/aje/kwj149
Published: 19 April 2006 **Article history** ▾

Abstract

Despite the growing popularity of propensity score (PS) methods in epidemiology, relatively little has been written in the epidemiologic literature about the problem of variable selection for PS models. The authors present the results of two simulation studies designed to help epidemiologists gain insight into the variable selection problem in a PS analysis. The simulation studies illustrate how the choice of variables that are included in a PS model can affect the bias, variance, and mean squared error of an estimated exposure effect. The results suggest that variables that are unrelated to the exposure but related to the outcome should always be included in a PS model. The inclusion of these variables will decrease the variance of an estimated exposure effect without increasing bias. In contrast, including variables that are related to the exposure but not to the outcome will increase the variance of the estimated exposure effect without decreasing bias. In very small studies, the inclusion of variables that are strongly related to the exposure but only weakly related to the outcome can be detrimental to an estimate in a mean squared error sense. The addition of these variables removes only a small amount of bias but can increase the variance of the estimated exposure effect. These simulation studies and other analytical results suggest that standard model-building tools designed to create good predictive models of the exposure will not always lead to optimal PS models, particularly in small studies.