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, balance 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 Lo-Casale, Michael Goodman, Ramin Arani, Yiduo Zhang, and Sudeep Karve for contributing to the requirements with their expertise in epidemiology, safety informatics, health economics and biostatistics 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, ggplot2, 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

<table>
<thead>
<tr>
<th>Trial</th>
<th>Number of Trees</th>
<th>Depth</th>
<th>Shrinkage</th>
<th>Bag Fraction</th>
<th>Estimand</th>
<th>Time to Run</th>
<th>Avg. % Reduction in Std. Diff. of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10000</td>
<td>3</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>508 s</td>
<td>82.11439</td>
</tr>
<tr>
<td>2</td>
<td>5000</td>
<td>3</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>469 s</td>
<td>81.98465</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>3</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>378 s</td>
<td>81.36216</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>3</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>347 s</td>
<td>77.14936</td>
</tr>
<tr>
<td>5</td>
<td>10000</td>
<td>3</td>
<td>0.005</td>
<td>1</td>
<td>ATT</td>
<td>499 s</td>
<td>81.22527</td>
</tr>
<tr>
<td>6</td>
<td>10000</td>
<td>3</td>
<td>0.05</td>
<td>1</td>
<td>ATT</td>
<td>514 s</td>
<td>81.22527</td>
</tr>
<tr>
<td>7</td>
<td>10000</td>
<td>3</td>
<td>0.1</td>
<td>1</td>
<td>ATT</td>
<td>508 s</td>
<td>82.62994</td>
</tr>
<tr>
<td>8</td>
<td>10000</td>
<td>2</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>456 s</td>
<td>88.67967</td>
</tr>
<tr>
<td>9</td>
<td>10000</td>
<td>3</td>
<td>0.01</td>
<td>1</td>
<td>ATE</td>
<td>538 s</td>
<td>82.05354</td>
</tr>
<tr>
<td>10</td>
<td>5000</td>
<td>2</td>
<td>0.01</td>
<td>1</td>
<td>ATT</td>
<td>405 s</td>
<td>92.02401</td>
</tr>
</tbody>
</table>

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
**Load Data**

A simple dataset has been included in the `drive.psa` 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:

```r
ps.getDatasAvailibility()
```

<table>
<thead>
<tr>
<th>DRUGCLASS</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGI</td>
<td>810</td>
</tr>
<tr>
<td>Biguanide</td>
<td>40000</td>
</tr>
<tr>
<td>Combinations</td>
<td>40000</td>
</tr>
<tr>
<td>DPP4</td>
<td>40000</td>
</tr>
<tr>
<td>Insulin</td>
<td>40000</td>
</tr>
<tr>
<td>No T2DM Drug</td>
<td>40000</td>
</tr>
<tr>
<td>Other</td>
<td>40000</td>
</tr>
<tr>
<td>Sulfonylureas</td>
<td>40000</td>
</tr>
</tbody>
</table>

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:

```r
myData <- ps.createDataset("Biguanide", 30000, 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).

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

**Distribution of Propensity Scores**

A distribution of propensity scores is shown below.

**Covariate Std. Diff. Reduction**

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

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

OHDSI Cohort Method

- PS plots
- Balance diagnostics
- regularized LR
- dependent on CDM
Variable Selection for Propensity Score Models

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

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.