Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang SAS Macros[1]

Dan McCaffrey, Lane Burgette, Beth Ann Griffin, Craig Martin, and Greg Ridgeway*

April 24, 2016

*The development of this software and tutorial was funded by National Institute of Drug Abuse grant number 1R01DA034065 (PIs: Griffin/McCaffrey).

[1] Much of the content in this tutorial is taken from "Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang," (Ridgeway, G., D. McCaffrey, A. Morral, L. Burgette and B.A. Griffin), the tutorial for the twang package in R environment for statistical computing and graphics. The material has been modified to reflect use of SAS macros to implement the twang tools.

1 Introduction

The Toolkit for Weighting and Analysis of Nonequivalent Groups, TWANG, contains a set of macros to support causal modeling of observational data through the estimation and evaluation of propensity scores and associated weights (Ridgeway et al., 2013). The macros call functions from the twang package in the R environment for statistical computing and graphics (R Core Team, 2013). The twang package was developed in 2004. After extensive use, it received a major update in 2012. The SAS twang macros were developed in 2014 to support the use of the twang tools without requiring analysts to learn R. This tutorial provides an introduction to twang and demonstrates its use through illustrative examples.

The foundation to the methods supported by twang is the propensity score. The propensity score is the probability that a particular case would be assigned or exposed to a treatment condition. Rosenbaum & Rubin (1983) showed that knowing the propensity score is sufficient to separate the effect of a treatment on an outcome from observed confounding factors that influence both treatment assignment and outcomes, provided the necessary conditions hold. The propensity score has the balancing property that given the propensity score the distribution of features for the treatment cases is the same as that for the control cases. While the treatment selection probabilities are generally not known, good estimates of them can be effective at diminishing or eliminating confounds between pretreatment group differences and treatment outcomes in the estimation of treatment effects.

There are now numerous propensity score methods in the literature. They differ in how they estimate the propensity score (e.g. logistic regression, CART), the target estimand (e.g. treatment effect on the treated, population treatment effect), and how they utilize the resulting estimated propensity scores (e.g. stratification, matching, weighting, doubly robust estimators). We originally developed the twang package with a particular process in mind, namely, generalized boosted regression to estimate the propensity scores and weighting of the comparison cases to estimate the average treatment effect on the treated (ATT). However, we have updated the package to also meaningfully handle the case where interest lies in using the population weights (e.g., weighting of comparison and treatment cases to estimate the population average treatment effect, ATE). The main workhorse of twang is the %ps macro which implements generalized boosted regression modeling to estimate the propensity scores and other tools that allow users to assess the success of the resulting weights at obtaining equivalence (or "balance") in the pretreatment covariate distributions of treatment and control groups. However, the framework and functions of the package are flexible enough to allow the user to use propensity score estimates from other methods and to assess the usefulness of those estimates for ensuring balance between the treatment and control groups using tools from the package. The same set of macros is also useful for other tasks such as non-response weighting, as discussed in Section 4.

The twang macros aim to (i) compute from the data estimates of the propensity scores which yield accurate causal effect estimates, (ii) check the quality of the resulting propensity score weights by assessing whether or not they have the balancing properties that we expect in theory, and (iii) use them in computing treatment effect estimates.

2 An ATT example to start

2.1 Set-up

If you have not already done so, you will need to download the `twang` macro file ("twang_mac_v1.0.sas"– where the version number v3.1.0 is replaced with the current version of the macros) and supporting documents to a folder on your computer. The files include:

- twang_mac_v3.1.0.sas – file containing the SAS macro code (where 3.1.0 is replaced by the current version number of the macro)
- twang_mac_help.txt – help file with details on implementing the macros
- Start-up.pdf – step by step details on installing R
- lalonde.sas7bdat – Example SAS dataset from the Lalonde Study
- lindner.sas7bdat – Example SAS dataset from the Lindner Study
- egsingle.sas7bdat – Example SAS dataset from the Raudenbush and Bryk Study
- tutorial_code.sas – SAS code from examples presented in this tutorial

The SAS datasets and example code will be useful for trying the code presented in this tutorial. Those files are not necessary for you to run your own applications.

To use the macros, you will need to include them in your SAS scripts. This is accomplished using a "%include" statement such as

%include "C:\Users\uname\SASFile\twang_mac.sas";

where the path, "C:\Users\uname\SASFile", is replaced by the path to where you have stored the macro file.[1]

The macros will run code in R and import the results into your SAS session. The macros will export the users' data to a .csv file that can be read into R. They will also create an R script file that is run

[1] Users might be tempted to copy and paste code from this PDF document into an editor to run this example code. We advise against this. Text from the PDF file may not appear the same in a text editor as it does in the PDF file; symbols or spaces may be added. The file "tutorial_code.sas" file, that is available with the `twang` macro contains all the code from this tutorial in text file. Analysts can use that file to copy the code and run the examples.

in R batch mode. The script file exports weights and diagnostic information in .csv files that are then ported back into SAS. By default all the files created by the macro are stored in "SAStemp" directory that is deleted at the end of the SAS session. Users can redirect the storage of the intermediate output files through parameters in the macros. Details are presented below. To manage file transfer between SAS and R and ensure all the necessary R functions are working, the macros use features of the Window operating system only available in Windows Vista or later.

The macros rely on the `twang` package in R. You will need to install R from from The Comprehensive R Archive Network (http://cran.us.r-project.org/). The software can be installed by clicking on the link for the users computer platform (e.g., Windows users would click on "Download R for Windows" and then click on the "base" link to download the standard R software). For assistance on installing R please see the Start-up file "Toolkit for Weighting and Analysis of Nonequivalent Groups SAS Macros Start-Up" (Start-up.pdf) or you can view tutorial videos on Youtube such as https://www.youtube.com/watch?v=PwfVCaMCO8U.

Users will need to note the directory where the R software is installed and the name of the executable file. For Windows users with a 64-bit processor the directory information for the standard installment is

C:\Program FilesR\R-3.0.2\bin\x64

where 3.0.2 is replaced by the current version of R at the time of installation.[2] For users with a 32-bit processor the directory information for the standard installment is

C:\Program FilesR\R-3.0.2\bin\i386

Again 3.0.2 is replaced by the current version of R at the time of installation. For both 64 or 32-bit processors, the executable is R.exe for batch implementation. The SAS macros can install the package if the users does not.

The macros also create a "TWANG" folder in the user's `AppData\Local` folder (C:\Users\username \AppData\Local\TWANG would be the default for a user with the "username" as his or her username). The macros may use this folder to store files used in running the `twang` package in R or other files needed by the programs. This folder will remain on the user's hard drive until it is removed. Users can remove the folder using any method they would use for removing a folder or they can run the `%remove_twang_folder` macro when they no longer plan to use TWANG.

2.2 Estimating propensity scores with the %ps() macro

To demonstrate the package we utilize data from Lalonde's National Supported Work Demonstration analysis (Lalonde, 1986, Dehejia & Wahba, 1999, http://users.nber.org/ rdehejia/nswdata2.html). This dataset is provided with the `twang` macros (lalonde.sas7bdat).

For the `lalonde` dataset, the variable treat is the 0/1 treatment indicator, 1 indicates "treatment" by

[2] Users might want to add the directory for R to their PATH environmental variable. Doing so would simplify the use of the `TWANG` macros. See Footnote 3 for more details.

being part of the National Supported Work Demonstration and 0 indicates “comparison”” cases drawn from the Current Population Survey. In order to estimate a treatment effect for this demonstration program that is unbiased by pretreatment group differences on other observed covariates, we include these covariates in a propensity score model of treatment assignment: age, education, Black, Hispanic, having no degree, married, earnings in 1974 (pretreatment), and earnings in 1975 (pretreatment). Note that we specify no outcome variables at this time. The `%ps` macro is the primary method in `twang` for estimating propensity scores. This step is computationally intensive and can take a few minutes.

```
%ps(treatvar=treat,
    vars=age educ black hispan nodegree married re74 re75,
    class=,
    dataset=sasin.lalonde,
    ntrees=5000,
    intdepth=2,
    shrinkage=0.01,
    permtestiters=0,
    stopmethod=es.mean ks.max,
    sampw =,
    estimand = ATT,
    output_dataset=lalonde_att_wgts,
    Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe,
    plotname=lalonde_twang.pdf,
    objpath=C:\Users\uname\twang_example
    )
```

The arguments to `%ps` require some discussion. The first argument is the treatment variable, which as noted earlier is “treat”. The next, `vars`, lists the names of the covariates to be used in model. Following the common syntax for SAS macros, the variable names are separated by spaces and not in quotation marks. The `class` argument list categorical variables. These variables should also be specified in the `vars` argument.

The next argument, `dataset`, indicates the dataset to be used in the analysis. It can be a permanent or temporary SAS dataset. In this example it is assumed to be stored in a permanent SAS dataset named “lalonde” in a folder identified by the `libname` “sasin” (code for specifying the libname is not presented). R, and in particular, the `twang` package in R, can be slow and require large amounts of memory to process large datasets. The user might get better computational performance by including only the variables to be used by `%ps` in the dataset specified by the `data` parameter.

ntrees, intdepth, and shrinkage are parameters for the GBM that %ps computes. The argument ntrees is the maximum number of iterations that the GBM will run. There will be a warning in the SAS log if the estimated optimal number of iterations is too close to the bound selected in this argument because it indicates that balance may improve if more complex models (i.e., those with more trees or a larger value for ntrees) are considered. The user should increase ntrees or decrease shrinkage and rerun the macro if this warning appears. The argument intdepth controls the level of interactions allowed in the GBM. The default is 3 and we typically use the default in our analyses. The GBM estimation algorithm uses shrinkage to enhance the smoothness of resulting model. The shrinkage argument controls the amount of shrinkage. Small values such as 0.005 or 0.001 yield smooth fits but require greater values of ntrees to achieve adequate fits. Computational time increases inversely with shrinkage argument. Additional details on ntrees, intdepth, and shrinkage can be found McCaffrey, Ridgeway, and Morral (2004).

permtestiters specifies whether p-values for KS statistics should be calculated using Monte Carlo methods, which is slow but accurate, or estimated using an analytic approximation that is fast, but produces poor estimates in the presence of many ties. If permtestiters=0 (default) is called, then analytic approximations are used. If permtestiters=500 is called, then 500 Monte Carlo trials are run to establish the reference distribution of KS statistics for each covariate. Higher numbers of trials will produce more precise p-values for the test of the KS statistic. Specifying permtestiters greater than zero can greatly slow down the twang computations. We tend to rely on the approximations (permtestiters=0) when using twang in practice.

The stopmethod argument specifies a set (or sets) of rules and measures for assessing the balance, or equivalence, established on the pretreatment covariates of the weighted treatment and control groups. The iterations used in the GBM minimize the differences between the treatment and control groups as measured by the balance statistics specified by values given to stopmethod argument. The package includes four built-in stop methods. They are "es.mean", "es.max", "ks.mean", and "ks.max". The four stopping rules are defined by two components: a balance metric for covariates and rule for summarizing across covariates. The balance metric summarizes the difference between two univariate distributions of a single pre-treatment variable (e.g., age). The default stopping rules in twang use two balance metrics: absolute standardized bias (also referred to as the absolute standardized mean difference or the *E*ffect *S*ize) and the Kolmogorov-Smirnov (KS) statistic. The stopping rules use two different rules for summarizing across covariates: the mean of the covariate balance metrics ("mean") or the maximum of the individual covariate balance metrics ("max"). The first piece of the stopping rule name identifies the balance metric ("es" for the effect size or standardized bias or "ks" for the KS statistic) and the second piece specifies the method for summarizing across balance metrics ("mean" or "max"). For instance, "es.mean" uses the effect size or the absolute standardized bias and summarizes across variables with the mean and the "ks.max" uses the KS statistics to assess balances and summarizes using the maximum across variables and the other two stopping rules use the remaining two combinations of balance metrics and summary statistics. The balance metrics depend on the estimand and correct specification of the metrics is set automatically by the specification of the `estimand` argument in the `%ps` macro.

The `sampw` argument is the name of the variable that contains sampling weights if they exist. If there are no sampling weights the parameter can be left unspecified as it is in this example.

The `estimand` argument is used to indicate whether the analyst is interested in estimating the average treatment effect (ATE) or the average treatment effect on the treated (ATT), as we do above. ATE addresses the question of how outcomes would differ if everyone in the sample were given the treatment versus everyone being given the control (Wooldridge, 2002). ATT, on the other hand, estimates the analogous quantity averaging only over the subjects who were actually treated.

The primary results of the `%ps` macro are the weights which can be used for estimating effects. It also produces checks of the balance of covariates returned to the SAS lst in the form of balance tables and an overall summary table. It will also produce diagnostic plots to help assess the balance and the GBM fit, if the user requests them by specifying the `plotname` argument.

By default, the propensity score based weights and the propensity scores created by the `%ps` macro are saved in a temporary SAS dataset (`_inputds`) in the work folder. Users can save this file in permanent dataset using a datastep or they can specify an alternative temporary or a permanent dataset using the `output_dataset` argument. In the example, we save the output dataset in "lalonde_att_wgts". The output dataset contains the data from the dataset specified in `dataset` with the weight and propensity score variables attached. There is one weight variable for each stopping rule specified in `stopmethod`. The weight variables are named according to the stopping rule and estimand so that in this example there is a weight variable `es_mean_ATT` with the weights from a GBM with the iterations chosen to minimize the mean standardized bias (effect size) and a second weight variable `ks_max_ATT` with the weights from a GBM with the iterations chosen to minimize the maximum KS statistic. Because the estimand is ATT, the weight equals 1 for every individual in the treatment group. There is also a propensity score variable for each stopping rule. The names for the propensity scores follow the same conventions as those of the weights, except that the string "ps_" is appended to the start of each name. This means that, in this example, the propensity scores are named `ps_es_mean_ATT` and `ps_ks_max_ATT` for the `es_mean` and `ks_max` stopping rules, respectively. The returned output dataset can be saved and used to estimate the treatment effects.

The argument `Rcmd` specifies the R program executable file for running R. The location of the file is determined through the installation of R. The default the default setup of R Version 3.0.2 on Windows 7 resulted in the executable being "C:\Program Files\R\R-3.0.2\bin\x64\R.exe" For other versions of R "3.0.2" will be replaced by the version number. If the analyst has added R to the path environmental variable then the path does not need to be included in the `Rcmd` argument, “Rcmd=R” will work.[3] (Similarly specifying the “.exe” extension is not necessary.)

The argument `plotname` gives the name for a pdf file of default diagnostic plots that `twang` creates. Creation of the plots is optional. If `plotname` is not given or if left blank, then no plots are created. If the argument contains a path, then the pdf file with plots will be stored in the folder specified by it. Otherwise the file will be stored in the folder specified by `objpath`, if it is specified. If a path is not include in `plotname` value and `objpath` is not set, then the file with plots is stored in the directory where SAS was launched (BATCH mode SAS) or the user’s home directory (interactive SAS). In our call to `%ps` we do not include the path in the `plotname` (lalonde_twang.pdf) but we do specify the `objpath` as “C:\Users\uname\twang_example” so the plots will be stored in that folder. User will need to specify an appropriate folder where they can write file to use for the `objpath` if the specify it.

The final argument `objpath` specifies a folder where files created by the macro to run the `twang` functions in R and return the results to SAS are stored. Namely, an “R object” (“ps.RData”) with the GBM and a log of the R session (“ps.Rout”). The R object is required for running the `%plot` macro. `objpath` is optional. If it is not specified, then the macro writes intermediate files and R object and log the SAStmp directory and all the files are deleted at the end of the SAS session. If the analysts wants to review the log or use the R object, then `objpath` must be specified. If `objpath` does not include a full path then the specified folder is assumed to exist in the folder where SAS was launched(BATCH SAS) or the user’s home directory (interactive SAS).

Having fit the GBM, the analyst should perform several diagnostic checks before estimating the causal effect in question. The first of these diagnostic checks makes sure that the specified value of `ntrees` allowed the GBM to explore sufficiently complicated models. We can do this quickly using the “convergence” or

[3] To add the R directory to the PATH the user can open a command (cmd) window and type:

```
setx PATH ‘‘%PATH%;C:\Program Files\R\R-3.0.2\bin\x64
```

where “C:\Program FilesR\R-3.0.2\bin\x64” should be replaced by actual directory where R is stored. Alternatively, the R directory can be added to the PATH by

1. Click *Start*
2. Right click *Computer*
3. Select *Properties*
4. Select *Advanced system settings*
5. Select *Environmental Variables...*
6. In the upper window double click *PATH*
7. Add “;C:\Program Files\R\R-3.0.2\bin\x64” to the end
8. Click OK until no longer prompted and close out windows that were opened.

Again ‘C:\Program FilesR\R-3.0.2\bin\x64” should be replaced by actual directory where R is stored.

“optimization” plot created by the R functions. There are two ways to generate this plot. The first way to obtain the plot is to specify `plotname` argument in `%ps`. This will create all the default diagnostics plots available in `twang`. They will be stored in the single pdf file specified by the argument. In this example the file is: “C:\Users\uname\twang_example\lalonde_twang.pdf”. The default file created by specifying the `plotname` contains one page for each type of diagnostic plot and each page contains a multi-panel plot with one panel for each stopping rule specified in the `stopmethod` argument. If only one stopping rule is specified each page contains a single panel. Figure 1 presents the plots for the first page of the `%ps` call.

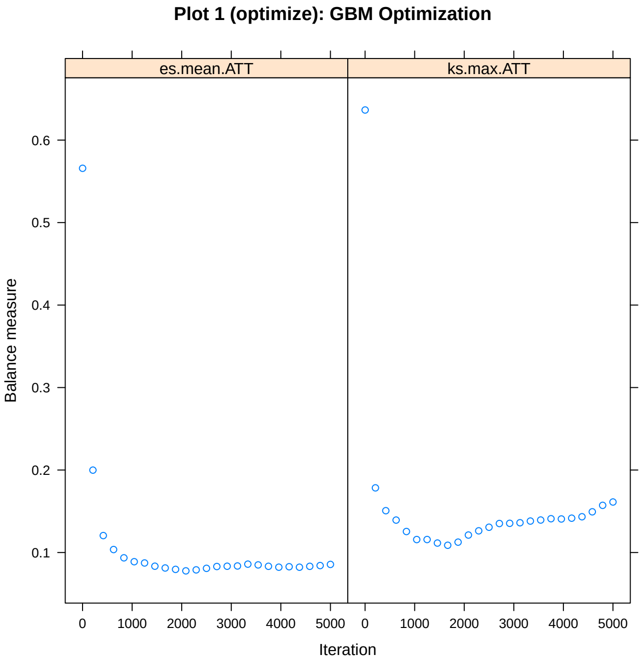


Figure 1: Example of an optimization plot for two stopping rules (“es.mean,” and “ks.max”) for estimating ATT weights for the Lalonde dataset. It was generated by setting the `plotname` argument in the %ps() macro and appears as the first page of a multiple page document of diagnostics plots.

The second way to obtain the plot is to run the `%plot` macro which can create specific plots and store them using pdf or other file formats. As noted above, the `objpath` argument must be specified in `%ps` to use the `%plot` macro. The `inputobj` argument specifies the R object created by the `%ps` macro. If the argument does not include a path the file is assumed be stored in the folder where SAS was launched (BATCH SAS) or the user’s home directory (interactive SAS). `plotname` gives the file name for the plot and `plotformat` specifies the file format. Allowable file formats are: jpg - JPEG, pdf - PDF, png - PNG, wmf - Windows enhanced metafile, and ps - postscript. The argument `plots` specifies the diagnostic plot to be created. Only one type of diagnostic plot can be created by a `%plot` call.[4] The plot can be specified by number or name and the names should not be in quotation marks. The convergence plot is specified by "1" or "optimize". The results of the following code produce the same plot as Figure 1, except there are no titles on the plots produced by the plot macro.

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde_opt.pdf,
      plotformat=pdf,
      plots=1,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

The `%plot` macro also allows for restricting the plot to results for a single stopping rule by specifying it by number in the `subset` argument. Stopping rules are numbered by alphabetical order; not the order in which they are specified. See Figure 2.

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde\opt_ks.pdf,
      plotformat=pdf,
      plots=1,
      subset=2,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

The convergence plot plots the balance measures as a function of the number of iterations in the GBM algorithm, with higher iterations corresponding to more complicated fitted models. In this example, 2127 iterations minimized the average effect size difference and 1756 iterations minimized the largest of the eight Kolmogorov-Smirnov (KS) statistics computed for the covariates. This can be observed in Figure 1. The maximum of KS statistics starts large, decreases and then increases somewhere between 1000 and 2000 iterations. The plot suggest `ntrees=5000` was sufficient. However, if it had appeared that additional iterations would be likely to result in lower values of the balance statistic – for instance, if the maximum was still declining without appearing to have attained a minimum by the maximum number of iterations, `ntrees` should be increased. As shown in the plot, after a point, additional complexity typically makes the balance worse. This figure also gives information on how compatible two or more stopping rules are: if the minima for multiple stopping rules under consideration are near one another, the results should not be sensitive to which stopping rule one uses for the final analysis. See Section 5.3 for a discussion of these and other balance measures.

[4] The value of Rcmd in the following example will need to be modified to specify the user's version of R.

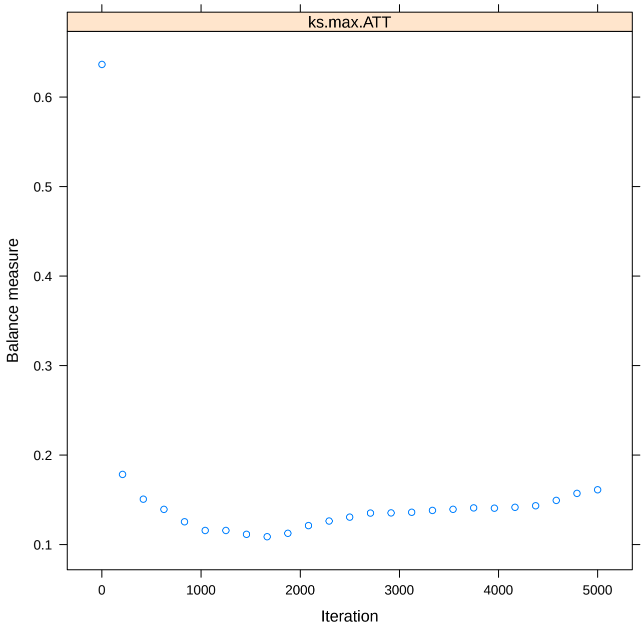


Figure 2: Example of an optimization plot for a single stopping rule (ks.max) for estimating ATT weights for the Lalonde dataset.

2.3 Assessing "balance" using balance tables

The `%ps` macro returns a balance table in the temporary SAS data set `_baltab`, in the work folder, and writes the table to the list file or window.[5] This table shows how well the resulting weights succeed in manipulating the control group so that its weighted pretreatment characteristics match, or balance, those of the unweighted treatment group if `estimand` = "ATT" or in adjusting both the control and treatment groups so that their weighted pretreatment characteristics match, or balance, with one another if `estimand` = "ATE". The balance table includes information on the pretreatment covariates before and after weighting. The results for the unweighted analysis (are identified by "unw"). Those for each of the specified values to `stopmethod` are identified by the stop method label with the specified `estimand` appended, here "es.mean.ATT" and "ks.max.ATT." There is no research on which stopping rule is best and the choice is likely to depend on the application. McCaffrey et al. (2004) essentially used "es.mean" for the analyses, but our more recent work has sometimes used "ks.max". See McCaffrey et al. (2013) for a greater details on stopping rules.

If there are missing values (represented as NA) in the covariates, `twang` will attempt to construct weights that also balance rates of missingness in the treatment and control arms. In this case, the balance

[5] If `objpath` is specified then all the files produced by the R scripts, including "baltab.cvs," are stored in the specified folder.

table will have an extra row for each variable that has missing entries. The columns of the table consist of the following items:

row_name The covariate variable name with the stopping rule or unw added as a prefix

tx_mn, **ct_mn** The treatment means and the control means for each of the variables. The unweighted table (unw) shows the unweighted means. For each stopping rule the means are weighted using weights corresponding to the gbm model selected by `%ps` using the stopping rule. When `estimand` = "ATT" the weights for the treatment group always equal 1 for all cases and there is no difference between unweighted and propensity score weighted tx.mn

tx_sd, **ct_sd** The propensity score weighted treatment and control groups' standard deviations for each of the variables. The unweighted table (unw) shows the unweighted standard deviations

std_eff_sz The standardized effect size, defined as the treatment group mean minus the control group mean divided by the treatment group standard deviation if `estimand` = "ATT" or divided by the pooled sample (treatment and control) standard deviation if `estimand` = "ATE". (In discussions of propensity scores this value is sometimes referred to as "standardized bias".) Occasionally, lack of treatment group or pooled sample variance on a covariate results in very large (or infinite) standardized effect sizes. For purposes of analyzing mean effect sizes across multiple covariates, we set all standardized effect sizes larger than 500 to NA (missing values)

stat, **p** Depending on whether the variable is continuous or categorical, stat is a t-statistic or a χ^2 statistic. **p** is the associated p-value

ks, **ks.pval** The Kolmogorov-Smirnov test statistic and its associated p-value. P-values for the KS statistics are either derived from Monte Carlo simulations or analytic approximations, depending on the specifications made in the `permtestiters` argument of the `%ps` macro. For categorical variables this is just the χ^2 test p-value.

SAS Output: Balance Table Results

```
                                                    Balance table: unw                            16:28 Monday, January 6, 2014   2

Obs row_name                         tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

  1 unw.age                         25.816        7.155        28.03       10.787       -0.309       -2.994        0.003        0.158        0.003 unw
  2 unw.educ                        10.346        2.011       10.235        2.855        0.055        0.547        0.584        0.111        0.074 unw
  3 unw.black                        0.843        0.365        0.203        0.403        1.757       19.371            0         0.64            0 unw
  4 unw.hispan                       0.059        0.237        0.142         0.35       -0.349       -3.413        0.001        0.083        0.317 unw
  5 unw.nodegree                     0.708        0.456        0.597        0.491        0.244        2.716        0.007        0.111        0.074 unw
  6 unw.married                      0.189        0.393        0.513          0.5       -0.824       -8.607            0        0.324            0 unw
  7 unw.re74                      2095.574      4886.62     5619.237     6788.751       -0.721       -7.254            0        0.447            0 unw
  8 unw.re75                      1532.055     3219.251     2466.484     3291.996        -0.29       -3.282        0.001        0.288            0 unw

                                                 Balance table: ks.max.ATT                        16:44 Monday, January 6, 2014   3

Obs row_name                         tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

 17 ks.max.ATT.age                  25.816        7.155       25.764        7.408        0.007        0.055        0.956        0.107        0.919 ks.max.ATT
 18 ks.max.ATT.educ                 10.346        2.011       10.572         2.14       -0.113       -0.712        0.477        0.107        0.919 ks.max.ATT
 19 ks.max.ATT.black                 0.843        0.365        0.835        0.371        0.022        0.187        0.852        0.008            1 ks.max.ATT
 20 ks.max.ATT.hispan                0.059        0.237        0.043        0.203        0.069        0.779        0.436        0.016            1 ks.max.ATT
 21 ks.max.ATT.nodegree              0.708        0.456        0.601         0.49        0.235          1.1        0.272        0.107        0.919 ks.max.ATT
 22 ks.max.ATT.married               0.189        0.393        0.199          0.4       -0.024       -0.169        0.866         0.01            1 ks.max.ATT
 23 ks.max.ATT.re74               2095.574      4886.62     1673.666       3944.6        0.086          0.8        0.424        0.054            1 ks.max.ATT
 24 ks.max.ATT.re75               1532.055     3219.251     1257.242     2674.922        0.085        0.722        0.471        0.094        0.971 ks.max.ATT

                                                 Balance table: es.mean.ATT                       16:44 Monday, January 6, 2014   4

Obs row_name                         tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

  9 es.mean.ATT.age                 25.816        7.155       25.802        7.279        0.002        0.015        0.988        0.122        0.892 es.mean.ATT
 10 es.mean.ATT.educ                10.346        2.011       10.573        2.089       -0.113       -0.706         0.48        0.099        0.977 es.mean.ATT
 11 es.mean.ATT.black                0.843        0.365        0.842        0.365        0.003        0.027        0.978        0.001            1 es.mean.ATT
 12 es.mean.ATT.hispan               0.059        0.237        0.042        0.202        0.072        0.804        0.421        0.017            1 es.mean.ATT
 13 es.mean.ATT.nodegree             0.708        0.456        0.609        0.489        0.218        0.967        0.334        0.099        0.977 es.mean.ATT
 14 es.mean.ATT.married              0.189        0.393        0.189        0.392        0.002        0.012         0.99        0.001            1 es.mean.ATT
 15 es.mean.ATT.re74              2095.574      4886.62      1556.93     3801.566         0.11        1.027        0.305        0.066            1 es.mean.ATT
 16 es.mean.ATT.re75              1532.055     3219.251     1211.575     2647.615          0.1        0.833        0.405        0.103        0.969 es.mean.ATT
```

The %ps macro also returns a compact summary of the sample sizes of the groups and the balance measures in a temporary SAS data file _summ and writes it to the list file or window.[6] The summary table includes one row for the results using the weights produced by each stopping rule specified by the stopmethod argument and one row for the unweighted data. Each row contains: the row names which specifies the stopping rule used for the weights or that the data are unweighted ("unw"), the estimand is also appended to the label; n_treat and n_ctrl, the treatment and control group sample sizes, respectively; n_ess and n_ctrl, the effective sample sizes of the treatment and control groups, n_ess equals the treatment group sample size for ATT because the weights are one (additional details on the effective sample size follow); max_es and mean_es, and max_ks and mean_ks, the maximum and mean or average of the standardized effect sizes or KS statistics for the covariates, respectively; max_ks_p, the p-value for testing the the maximum of the KS statistics is greater than zero; and iter, the number of iterations or trees in the GBM that minimizes the stopping rule, missing for unw. The max_ks_p is only produced when permtestiters> 0; otherwise it is missing for all rows of the summary table, as it is for our example. If

[6] If objpath is specified the summary table is saved to the specified directory in the file “‘summary.csv”

permtestiters> 0 was used in the call to %ps, then Monte Carlo simulation is used to estimate p-values for the maximum KS statistic that would be expected across the covariates, had individuals with the same covariate values been assigned to groups randomly. Thus, a p-value of 0.04 for max.ks.p indicates that the largest KS statistic found across the covariates is larger than would be expected in 96% of trials in which the same cases were randomly assigned to groups.

Output: Summary Table Results

```
                                                        Summary table                                  16:44 Monday, January 6, 2014   1

Obs  row_name          n_treat     n_ctrl   ess_treat     ess_ctrl       max_es      mean_es       max_ks    max_ks_p      mean_ks         iter

  1  unw                   185        429         185          429 1.7567745147 0.5687258912 0.6404460404           . 0.2702450702            .
  2  es.mean.ATT           185        429         185 22.964299329 0.2177817453 0.0774617496 0.1223383986           . 0.0636102136         2127
  3  ks.max.ATT            185        429         185 27.054721987 0.2348846317 0.0802599384  0.107076053           . 0.0628243238         1756
```

In general, weighted means can have greater sampling variance than unweighted means from a sample of equal size. The effective sample size (ESS) of the weighted comparison group captures this increase in variance as

$$\mathrm{ESS} = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2}, \tag{1}$$

where summation is over cases in the control group. The ESS is approximately the number of observations from a simple random sample that yields an estimate with sampling variation equal to the sampling variation obtained with the weighted comparison observations. Therefore, the ESS will give an estimate of the number of comparison participants that are comparable to the treatment group when estimand = “ATT”. When the estimand of interest is “ATE”, there is an analogous ESS for the treatment group because the weights are no longer equal to one for that group. The ESS is an accurate measure of the relative size of the variance of means when the weights are fixed or they are uncorrelated with outcomes. Otherwise the ESS underestimates the effective sample size (Little & Vartivarian, 2004). With propensity score weights, it is rare that weights are uncorrelated with outcomes. Hence the ESS typically gives a lower bound on the effective sample size, but it still serves as a useful measure for choosing among alternative models and assessing the overall quality of a model, even if it provides a possibly conservative picture of the loss in precision due to weighting.

2.4 Graphical assessments of balance

The %plot macro can generate useful diagnostic plots to evaluate the propensity scores. The full set of plots available in twang and the argument value of plot to produce each one are given in Table 1. The convergence or optimization plot was discussed above. Other diagnostic plots are specified by the value of the plots argument. For example, specifying plots = 2 or plots = boxplot produces boxplots illustrating the spread of the estimated propensity scores in the treatment and comparison groups (Figure 3). Whereas propensity score stratification requires considerable overlap in these spreads, excellent covariate balance can often be achieved with weights, even when the propensity scores estimated for the treatment and control groups show little overlap.

Table 1: Available options for plots argument to `%plot` macro.

Descriptive	Numeric	Description
argument	argument	
“optimize”	1	Balance measure as a function of GBM iterations
“boxplot”	2	Boxplot of treatment/control propensity scores
“es”	3	Standardized effect size of pretreatment variables
“t”	4	t-test p-values for weighted pretreatment variables
“ks”	5	Kolmogorov-Smirnov p-values for weighted pretreatment variables

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde_box.pdf,
      plotformat=pdf,
      plots=2,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

(Recall that wherever the example code has “C:\Program Files\R\R-3.0.2\bin\x64” it must be replaced by that actual path where the R.exe is stored.)

The effect size plot (see Figure 4) illustrates the effect of weights on the magnitude of differences between groups on each pretreatment covariate. These magnitudes are standardized using the standardized effect size described earlier. In these plots, substantial reductions in effect sizes are observed for most variables (blue lines), with only one variable showing an increase in effect size (red lines), but only a seemingly trivial increase. Closed red circles indicate a statistically significant difference, many of which occur before weighting, none after. In some analyses variables can have very little variance in the treatment group sample or the entire sample and group differences can be very large relative to the standard deviations. In these situations, the user is warned that some effect sizes are too large to plot.

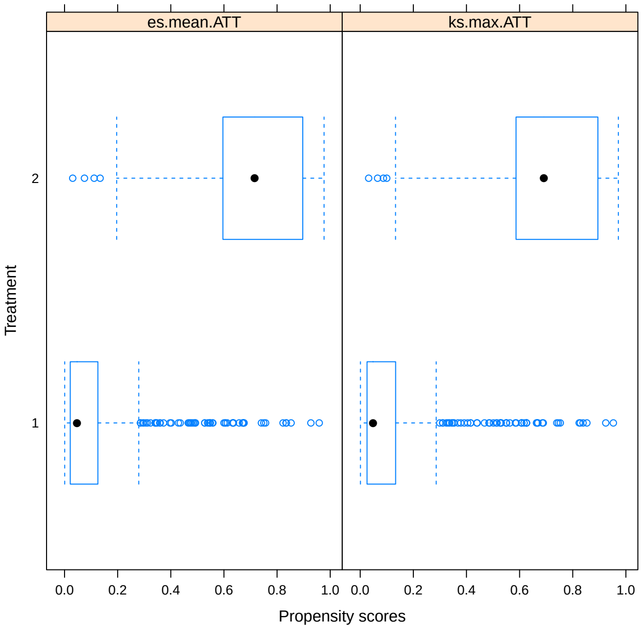


Figure 3: Example of the default diagnostic boxplot of propensity scores from the `%plot` macro for estimating ATT weights for the Lalonde dataset.

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde_es.pdf,
      plotformat=pdf,
      plots=3,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

When many of the p-values testing individual covariates for balance are very small, the groups are clearly imbalanced and inconsistent with what we would expect had the groups been formed by random assignment. After weighting we would expect the p-values to be larger if balance had been achieved. We use a QQ plot comparing the quantiles of the observed p-values to the quantiles of the uniform distribution (45 degree line) to conduct this check of balance. Ideally, the p-values from independent tests in which the null hypothesis is true will have a uniform distribution. Although the ideal is unlikely to hold even if we had random assignment (Bland, 2013), severe deviation of the p-values below the diagonal suggests lack of balance and p-values running at or above the diagonal suggests balance might have been achieved. The p-value plot (`plots`=4) allows users to visually to inspect the p-values of the t-tests for group differences in the covariate means.

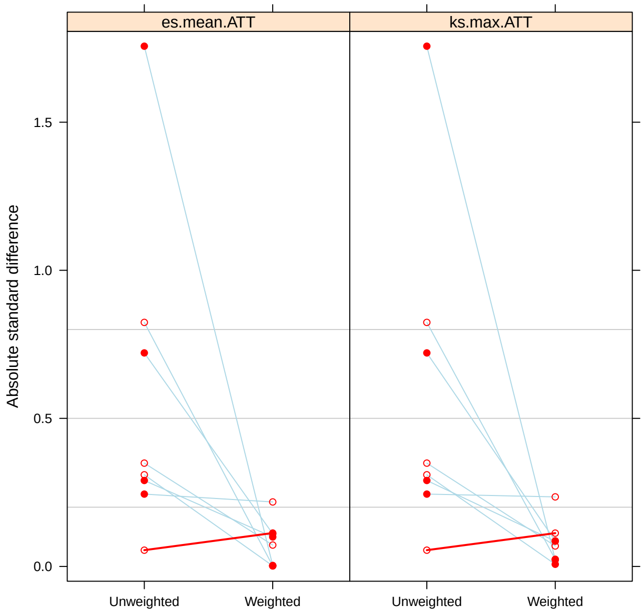


Figure 4: Example of the default diagnostic standardized effect size plot from the `%plot` macro for estimating ATT weights for the Lalonde dataset.

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde_p.pdf,
      plotformat=pdf,
      plots=4,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

Figure 5, presents the t-test p-value plot for the Lalonde example. Before weighting (closed circles), the groups have statistically significant differences on many variables (i.e., p-values are near zero). After weighting (open circles) the p-values are generally above the 45-degree line, which represents the cumulative distribution of a uniform variable on [0,1]. This indicates that the p-values are even larger than would be expected in an ideal randomized study, so that balance is generally good. One can inspect similar plots for the KS statistic with the argument `plots` = 5 or "ks" (see Figure 6).

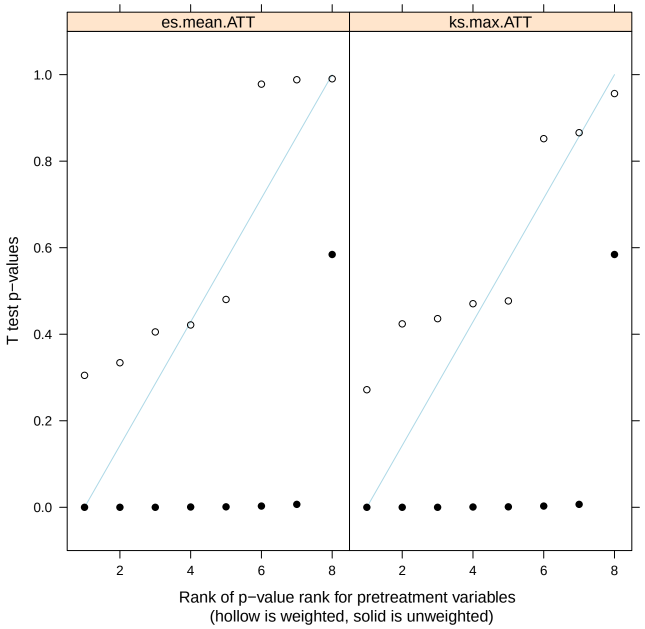


Figure 5: Example of the default diagnostic t-test p-value plot from the `%plot` macro for estimating ATT weights for the Lalonde dataset.

```
%plot(inputobj=C:\Users\uname\twang_example\ps.RData,
      plotname=lalonde_ks.pdf,
      plotformat=pdf,
      plots=5,
      Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe);
```

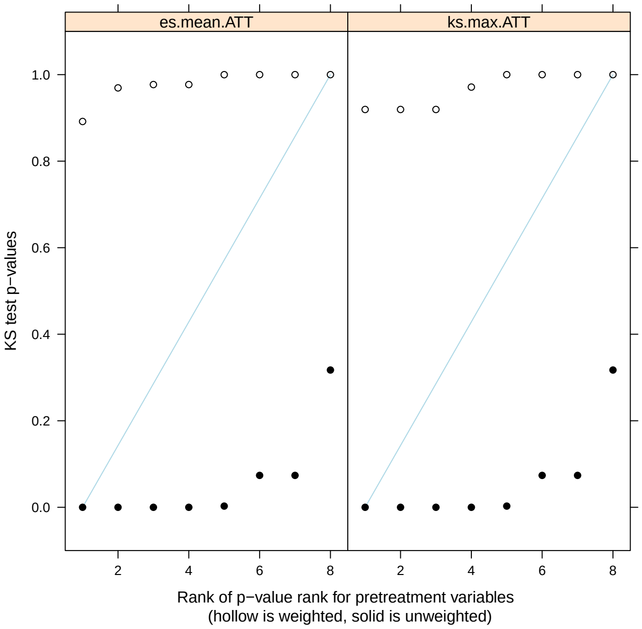


Figure 6: Example of the default diagnostic Kolmogorov-Smirnov test p-value plot from the `%plot` macro for estimating ATT weights for the Lalonde dataset.

2.5 Analysis of outcomes

The aim of the National Supported Work Demonstration analysis is to determine whether the program was effective at increasing earnings in 1978. We will estimate this effect as the difference in the treatment and weighted control group means and test that it is not zero using a Wald test. The propensity score adjusted test can be computed with PROC SURVEYREG. We start with an analysis using the weights derived from the GBM selected to minimize the mean standardized bias ("es.mean" stopping rule).

```
proc surveyreg data=lalonde_att_wgts;
   model re78 = treat;
   weight es_mean_att;
run;
```

SAS Output: PROC SURVEYREG Results

```
                                              13:21 Friday, January 3, 2014   8


                          The SURVEYREG Procedure
```

```
        Regression Analysis for Dependent Variable re78

                          Data Summary

              Number of Observations           614
              Sum of Weights             329.58393
              Weighted Mean of re78         6027.8
              Weighted Sum of re78       1986665.6

                         Fit Statistics

                   R-square          0.002650
                   Root MSE           7062.97
                   Denominator DF         613

                     Tests of Model Effects

            Effect       Num DF    F Value    Pr > F


            Model             1       0.48    0.4888
            Intercept         1      40.22    <.0001
            treat             1       0.48    0.4888

NOTE: The denominator degrees of freedom for the F tests is 613.

               Estimated Regression Coefficients
                               Standard
  Parameter      Estimate         Error    t Value    Pr > |t|


  Intercept    5616.62695     885.64351       6.34      <.0001
  treat         732.51658    1057.45754       0.69      0.4888

NOTE: The denominator degrees of freedom for the t tests is 613.
```

The analysis estimates an increase in earnings of $733 for those that participated in the NSW compared with similarly situated people observed in the CPS. The effect, however, does not appear to be statistically significant.

Some authors have recommended utilizing both propensity score adjustment and additional covariate adjustment to minimize mean square error or to obtain “doubly robust”” estimates of the treatment effect (Huppler-Hullsiek & Louis 2002, Bang & Robins 2005). These estimators are consistent if either the propensity scores are estimated correctly or the regression model is specified correctly. For example, note that the balance table for ks.max.ATT made the two groups more similar on nodegree, but still some differences remained, 70.8% of the treatment group had no degree while 60.9% of the comparison group had no degree. While linear regression is sensitive to model misspecification when the treatment and comparison groups are dissimilar, the propensity score weighting has made them more similar, perhaps enough so that additional modeling with covariates can adjust for any remaining differences. In addition to potential bias reduction, the inclusion of additional covariates can reduce the standard error of the treatment effect if some of the covariates are strongly related to the outcome.

```
proc surveyreg data=lalonde_att_wgts;
   model re78 = treat nodegree;
   weight es_mean_att;
run;
```

SAS Output: PROC SURVEYREG Results, One Covariate

```
                                               13:21 Friday, January 3, 2014  10

                            The SURVEYREG Procedure

                Regression Analysis for Dependent Variable re78

                                  Data Summary

                      Number of Observations           614
                      Sum of Weights             329.58393
                      Weighted Mean of re78         6027.8
                      Weighted Sum of re78       1986665.6
```

```
                          Fit Statistics


                   R-square            0.01848
                   Root MSE            7012.43
                   Denominator DF          613



                        Tests of Model Effects


               Effect       Num DF    F Value    Pr > F


               Model             2       1.89    0.1520
               Intercept         1      21.10    <.0001
               treat             1       0.72    0.3965
               nodegree          1       2.24    0.1350

        NOTE: The denominator degrees of freedom for the F tests is 613.



                  Estimated Regression Coefficients

                                         Standard
          Parameter      Estimate           Error    t Value    Pr > |t|


          Intercept     6768.4113    1473.38264       4.59      <.0001
          treat          920.3338    1084.58659       0.85      0.3965
          nodegree     -1891.8037    1264.00123      -1.50      0.1350

        NOTE: The denominator degrees of freedom for the t tests is 613.
```

Adjusting for the remaining group difference in the nodegree variable slightly increased the estimate of the program's effect to $920, but the difference is still not statistically significant. We can further adjust for the other covariates, but that too in this case has little effect on the estimated program effect.

```
proc surveyreg data=lalonde_att_wgts;
   model re78 = treat age educ black hispan nodegree married re74 re75;
   weight es_mean_att;
run;
```

SAS Output: PROC SURVEYREG Results, All Covariates

```
                                               13:21 Friday, January 3, 2014  12

                            The SURVEYREG Procedure

                Regression Analysis for Dependent Variable re78

                                  Data Summary

                      Number of Observations           614
                      Sum of Weights             329.58393
                      Weighted Mean of re78         6027.8
                      Weighted Sum of re78       1986665.6


                                 Fit Statistics

                           R-square            0.05578
                           Root MSE            6917.62
                           Denominator DF          613


                             Tests of Model Effects

                    Effect       Num DF    F Value    Pr > F

                    Model             9       2.49    0.0086
                    Intercept         1       0.32    0.5696
                    treat             1       0.55    0.4601
                    age               1       0.00    0.9572
```

```
            educ              1       8.20    0.0043
            black             1       0.56    0.4549
            hispan            1       0.13    0.7232
            nodegree          1       0.11    0.7441
            married           1       0.21    0.6490
            re74              1       0.10    0.7535
            re75              1       0.64    0.4242

NOTE: The denominator degrees of freedom for the F tests is 613.

              Estimated Regression Coefficients

                              Standard
 Parameter      Estimate         Error    t Value    Pr > |t|

 Intercept    -2458.7982    4321.26255      -0.57      0.5696
 treat          758.4891    1026.06245       0.74      0.4601
 age              3.0047      55.99413       0.05      0.9572
 educ           748.8193     261.55961       2.86      0.0043
 black         -762.6813    1019.88558      -0.75      0.4549
 hispan         610.6271    1723.24917       0.35      0.7232
 nodegree       535.0055    1638.20712       0.33      0.7441
 married        491.7587    1079.99447       0.46      0.6490
 re74             0.0570       0.18142       0.31      0.7535
```

2.6 Estimating the program effect using linear regression

Users may be wondering whether using twang and weighting to adjust for differences between groups yields different results than the more familiar regression approaches to adjusting for group differences on observed covariates. We now compare our weighted estimates of the program effect to results from a more traditional analysis in which the program effect is estimated by a linear model with a treatment indicator and linear terms for each of the covariates. PROC REG is the standard procedure for fitting such models in SAS. Recall that "sasin" is the libname for the folder where the Lalonde dataset set is stored.

```
proc reg data=sasin.lalonde;
   model re78 = treat age educ black hispan nodegree married re74 re75;
run;
```

SAS Output: PROC REG Results, All Covariates

```
                                              13:37 Friday, January 3, 2014  16

                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: re78

                   Number of Observations Read         614
                   Number of Observations Used         614


                             Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     9     5055423967      561713774      11.64    <.0001
Error                   604    29157220815       48273544
Corrected Total         613    34212644782


             Root MSE           6947.91655    R-Square     0.1478
             Dependent Mean     6792.83448    Adj R-Sq     0.1351
             Coeff Var           102.28302


                             Parameter Estimates

                          Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1       66.51452     2436.74580       0.03      0.9782
     treat         1     1548.24380      781.27930       1.98      0.0480
     age           1       12.97763       32.48891       0.40      0.6897
```

```
educ           1      403.94123      158.90624       2.54      0.0113
black          1    -1240.64408      768.76441      -1.61      0.1071
hispan         1      498.89686      941.94255       0.53      0.5966
nodegree       1      259.81737      847.44205       0.31      0.7593
married        1      406.62085      695.47232       0.58      0.5590
re74           1        0.29638        0.05827       5.09      <.0001
re75           1        0.23153        0.10462       2.21      0.0273
```

This model estimates a rather strong treatment effect, estimating a program efect of \$1548 with a p-value=0.048. Several variations of this regression approach also estimate strong program effects. For example using square root transforms on the earnings variables yields a p- value=0.016. These estimates, however, are very sensitive to the model structure since the treatment and control subjects differ greatly as seen in the unweighted balance comparison (unw) from the balance table.

2.7 Propensity scores estimated from logistic regression

Propensity score analysis is intended to avoid problems associated with the misspecification of covariate adjusted models of outcomes, but the quality of the balance and the treatment effect estimates can be sensitive to the method used to estimate the propensity scores. For instance, we consider estimating the propensity scores using logistic regression instead of the `%ps` macro and compare the results to the weights from `twang`.

```
proc logistic data=sasin.lalonde descending;
   model treat = age educ black hispan nodegree married re74 re75;
   output out=llogit pred=phat;
run;
```

PROC LOGISTIC fits the logistic regression model. The `descending` option specifies that the model fit the probability that `treat`=1. The default is to model the probability that it equals zero. The `output` statement specifies that predicted probabilities (keyword `pred`) be saved in a variable named `phat` to a temporary SAS dataset `llogit`. The dataset includes all the variables in the input dataset and appends the predicted probabilities. We can create the ATT weights in a datastep as shown below with weights for the treatment group equal to one and weights for the control group equal to the odds of treatment.

```
data llogit;
   set llogit;
   w_logit_att = treat + (1-treat)*phat/(1-phat);
run;
```

The %dxwts macro provides the balance assessment tools of twang for weights generated using any method, not just by %ps. The arguments are similar to those in %ps except weight variables are now specified and must by part of the dataset specified by the dataset argument. Multiple weights can be assessed but they must all be set for a common estimand. The macro produces summary and balance tables that are just like those produced by the %ps macro except there is no iter variable in the summary because the weights might not come from GBM model.

```
%dxwts(treatvar=treat,
       vars=age educ black hispan nodegree married re74 re75,
       dataset=llogit,
       weightvars=w_logit_att,
       estimand=ATT,
       permtestiters=0,
       Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe)
```

SAS Output: %dxwts Summary and Balance Tables

```
                                                     dxwts Summary table                           16:48 Monday, January 6, 2014  19

    Obs    type                 n_treat        n_ctrl     ess_treat      ess_ctrl        max_es       mean_es        max_ks       mean_ks

     1     unw                      185           429           185           429  1.7567745147  0.5687258912  0.6404460404  0.2702450702
     2     w_logit_att              185           429           185    99.8153887  0.1188495948  0.0318840912  0.3078038985  0.0930231893

                                                 dxwts Balance table: w_logit_att                  16:48 Monday, January 6, 2014  20

Obs row_name                     tx_mn       tx_sd       ct_mn       ct_sd  std_eff_sz        stat           p          ks     ks_pval table_name

  9 w_logit_att.age             25.816       7.155      24.966      10.535       0.119       0.739        0.46       0.308           0 w_logit_att
 10 w_logit_att.educ            10.346       2.011      10.403       2.459      -0.028      -0.219       0.827       0.036           1 w_logit_att
 11 w_logit_att.black            0.843       0.365       0.845       0.362      -0.006      -0.069       0.945       0.002           1 w_logit_att
 12 w_logit_att.hispan           0.059       0.237       0.059       0.236       0.001       0.008       0.993           0           1 w_logit_att
 13 w_logit_att.nodegree         0.708       0.456        0.69       0.463        0.04       0.332        0.74       0.018           1 w_logit_att
 14 w_logit_att.married          0.189       0.393       0.171       0.377       0.047       0.456       0.649       0.019           1 w_logit_att
 15 w_logit_att.re74          2095.574     4886.62    2106.045    4235.833      -0.002      -0.022       0.983       0.228       0.002 w_logit_att
 16 w_logit_att.re75          1532.055    3219.251    1496.541    2716.258       0.011       0.107       0.915       0.133       0.185 w_logit_att

                                                     dxwts Balance table: unw                      16:48 Monday, January 6, 2014  21

Obs row_name                     tx_mn       tx_sd       ct_mn       ct_sd  std_eff_sz        stat           p          ks     ks_pval table_name

  1 unw.age                     25.816       7.155       28.03      10.787      -0.309      -2.994       0.003       0.158       0.003 unw
  2 unw.educ                    10.346       2.011      10.235       2.855       0.055       0.547       0.584       0.111       0.074 unw
  3 unw.black                    0.843       0.365       0.203       0.403       1.757      19.371           0        0.64           0 unw
  4 unw.hispan                   0.059       0.237       0.142        0.35      -0.349      -3.413       0.001       0.083       0.317 unw
  5 unw.nodegree                 0.708       0.456       0.597       0.491       0.244       2.716       0.007       0.111       0.074 unw
  6 unw.married                  0.189       0.393       0.513         0.5      -0.824      -8.607           0       0.324           0 unw
  7 unw.re74                  2095.574     4886.62    5619.237    6788.751      -0.721      -7.254           0       0.447           0 unw
  8 unw.re75                  1532.055    3219.251    2466.484    3291.996       -0.29      -3.282       0.001       0.288           0 unw
```

For weights estimated with logistic regression, the largest KS statistic was reduced from the unweighted sample's largest KS of 0.64 to 0.31, which is still quite a large KS statistic. The means of the two groups appear to be quite similar while the KS statistic shows substantial differences in their distributions.

Table 2 compares the balancing quality of the weights directly with one another.

Table 2: Summary of the balancing properties of logistic regression and gbm

	n_treat	ess_ctrl	max_es	mean_es	max_ks	mean_ks
unw	185	429.00	1.76	0.57	0.64	0.27
logit	185	99.82	0.12	0.03	0.31	0.09
es.mean.ATT	185	22.96	0.22	0.08	0.12	0.06
ks.max.ATT	185	27.05	0.23	0.08	0.11	0.06

```
proc surveyreg data=llogit;
   model re78 = treat;
   weight w_logit_att;
run;
```

The analysis estimates an increase in earnings of $1214 for those that participated in the NSW compared with similarly situated people observed in the CPS. We can also run a model with linear adjustments for the covariates combined with the logistic regression weights. Table 3 compares all of the treatment effect estimates.

Table 3: Treatment effect estimates by various methods

Treatment effect	PS estimate	Linear adjustment
$733	GBM, minimize es	none
$920	GBM, minimize es	nodegree
$758	GBM, minimize es	all
$1548	None	all
$1214	Logistic regression	none
$1237	Logistic regression	all

3 An ATE example

In the analysis of Section 2, we focused on estimating ATT for the `lalonde` dataset. In that example, the ATE is not of great substantive interest because not all people who are offered entrance into the program could be expected to take advantage of the opportunity. Further, there is some evidence that the treated subjects were drawn from a subset of the covariate space. In particular, in an ATE analysis, we see that we are unable to achieve balance, especially for the `black` indicator.

We now turn to an ATE analysis that is feasible and meaningful. We focus on the `lindner` dataset, which was included in the `USPS` package in R (Obenchain 2011), and is included with the `twang` macros in the SAS data set `lindner.sas7bdat`. A tutorial by Helmreich and Pruzek (2009; HP) for the `PSAgraphics` package also uses propensity scores to analyze a portion of these data. HP describe the data as follows on p. 3 with our minor recodings in square braces:

> The `lindner` data contain data on 996 patients treated at the Lindner Center, Christ Hospital, Cincinnati in 1997. Patients received a Percutaneous Coronary Intervention (PCI). The data consists of 10 variables. Two are outcomes: [`sixMonthSurvive`] ranges over two values... depending on whether patients surved to six months post treatment [denoted by `TRUE`] or did not survive to six months [`FALSE`]... Secondly, `cardbill` contains the costs in 1998 dollars for the first six months (or less if the patient did not survive) after treatment... The treatment variable is `abcix`, where 0 indicates PCI treatment and 1 indicates standard PCI treatment and additional treatment in some form with abciximab. Covariates include `acutemi`, 1 indicating a recent acute myocardial infarction and 0 not; `ejecfrac` for the left ventricle ejection fraction, a percentage from 0 to 90; `ves1proc` giving the number of vessels (0 to 5) involved in the initial PCI; `stent` with 1 indicating coronary stent inserted, 0 not; `diabetic` where 1 indicates that the patient has been diagnosed with diabetes, 0 not; `height` in centimeters and `female` coding the sex of the patent, 1 for female, 0 for male.

HP focus on `cardbill` — the cost for the first months after treatment — as their outcome of interest. However, since not all patients survived to six months, it is not clear whether a lower value of `cardbill` is good or not. For this reason, we choose six-month survival (`sixMonthSurvive`) as our outcome of interest.

Ignoring pre-treatment variables, we see that `abcix` is associated with lower rates of 6-month mortality:

```
proc freq data=sasin.lindner;
   table sixMonthSurvive * abcix / chisq;
title "ATE Tutorial";
run;
```

SAS Output: Frequency Table

```
                              ATE Tutorial                              1
```

```
                  The FREQ Procedure

          Table of sixMonthSurvive by abcix

         sixMonthSurvive     abcix

         Frequency|
         Percent  |
         Row Pct  |
         Col Pct  |       0|       1|  Total
         ---------+--------+--------+
         FALSE    |     15 |     11 |     26
                  |   1.51 |   1.10 |   2.61
                  |  57.69 |  42.31 |
                  |   5.03 |   1.58 |
         ---------+--------+--------+
         TRUE     |    283 |    687 |    970
                  |  28.41 |  68.98 |  97.39
                  |  29.18 |  70.82 |
                  |  94.97 |  98.42 |
         ---------+--------+--------+
         Total         298      698      996
                     29.92    70.08   100.00


   Statistics for Table of sixMonthSurvive by abcix

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      9.8207    0.0017
Likelihood Ratio Chi-Square    1      8.8530    0.0029
Continuity Adj. Chi-Square     1      8.5077    0.0035
Mantel-Haenszel Chi-Square     1      9.8108    0.0017
```

```
Phi Coefficient                        0.0993
Contingency Coefficient                0.0988
Cramer's V                             0.0993
```

The question is whether this association is causal. If health care policies were to be made on the basis of these data, we would wish to elicit expert opinion as to whether there are likely to be other confounding pretreatment variables. For this tutorial, we simply follow HP in choosing the pre-treatment covariates. The `twang` model is fit as follows

```
%ps(treatvar=abcix,
     vars=stent height female diabetic acutemi ejecfrac ves1proc,
     dataset=sasin.lindner,
     estimand = ATE,
     Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe,
     plotname=abcix_twang.pdf);
```

We set `estimand = ''ATE''` because we are interested in the effects of abciximab on everyone in the population. We do not specify the stopping rules. Consequently `%ps` uses the defaults: "es.mean" and "ks.mean". We then inspect pre- and post-weighting balance using the balance table.

SAS Output: Balance Tables for `abcix`

```
                                                     Balance table: unw                               22:59 Monday, January 6, 2014   3

Obs row_name                      tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

  1 unw.stent                     0.705        0.456        0.584        0.494        0.257        3.624            0        0.121        0.004 unw
  2 unw.height                  171.443       10.695      171.446       10.589            0       -0.005        0.996        0.025        0.999 unw
  3 unw.female                    0.331        0.471        0.386        0.488       -0.115       -1.647          0.1        0.055        0.531 unw
  4 unw.diabetic                  0.205        0.404        0.268        0.444       -0.152       -2.127        0.034        0.064        0.349 unw
  5 unw.acutemi                   0.179        0.384         0.06        0.239        0.338        5.923            0        0.119        0.005 unw
  6 unw.ejecfrac                 50.403       10.419       52.289       10.297       -0.181        -2.64        0.008        0.114        0.008 unw
  7 unw.ves1proc                  1.463        0.706        1.205         0.48        0.393        6.693            0        0.188            0 unw

                                                 Balance table: ks.mean.ATE                           22:59 Monday, January 6, 2014   4

Obs row_name                      tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

  8 ks.mean.ATE.stent             0.683        0.466        0.657        0.475        0.054        0.718        0.473        0.025            1 ks.mean.ATE
  9 ks.mean.ATE.height           171.47        10.55      171.591       10.589       -0.011       -0.155        0.877        0.015            1 ks.mean.ATE
 10 ks.mean.ATE.female            0.338        0.473        0.345        0.476       -0.015         -0.2        0.841        0.007            1 ks.mean.ATE
 11 ks.mean.ATE.diabetic          0.215        0.411        0.229        0.421       -0.033       -0.432        0.666        0.014            1 ks.mean.ATE
 12 ks.mean.ATE.acutemi           0.148        0.355        0.107         0.31         0.12        1.331        0.183         0.04        0.935 ks.mean.ATE
 13 ks.mean.ATE.ejecfrac         51.052       10.333       51.604         9.11       -0.056       -0.798        0.425        0.027        0.999 ks.mean.ATE
 14 ks.mean.ATE.ves1proc          1.395        0.666        1.337        0.573        0.094        1.203        0.229        0.028        0.999 ks.mean.ATE

                                                 Balance table: es.mean.ATE                           22:59 Monday, January 6, 2014   5

Obs row_name                      tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

 15 es.mean.ATE.stent             0.683        0.466        0.653        0.477        0.063        0.849        0.396         0.03        0.996 es.mean.ATE
 16 es.mean.ATE.height          171.468       10.534      171.501        11.04       -0.003       -0.039        0.969        0.018            1 es.mean.ATE
```

```
17 es.mean.ATE.female        0.337      0.473       0.35      0.478     -0.028      -0.38      0.704      0.013          1 es.mean.ATE
18 es.mean.ATE.diabetic      0.214      0.411      0.237      0.426     -0.055     -0.719      0.473      0.023          1 es.mean.ATE
19 es.mean.ATE.acutemi       0.148      0.355       0.11      0.313      0.114      1.303      0.193      0.038       0.95 es.mean.ATE
20 es.mean.ATE.ejecfrac     51.008     10.389     51.386      9.396     -0.038     -0.532      0.595      0.027      0.999 es.mean.ATE
21 es.mean.ATE.ves1proc      1.398       0.67       1.35      0.589      0.076      0.951      0.342      0.023          1 es.mean.ATE
```

This balance table shows that `stent`, `acutemi`, and `ves1proc` were all significantly imbalanced before weighting. After weighting (using either `stop.method` considered) we do not see problems in this regard. Examining the diagnostic plots created by the specification of the `plotname` does not reveal problems, either. In regard to the optimize plot, we note that the scales of the KS and ES statistics presented in the optimize plots are not necessarily comparable. The fact that the KS values are lower than the ES values in the optimize plot does not suggest that the KS stopping rule is finding superior models. Each panel of the optimize plot indicates the gbm that minimizes each stopping rule. The panels should not be compared other than to compare the number of iterations selected by each rule.

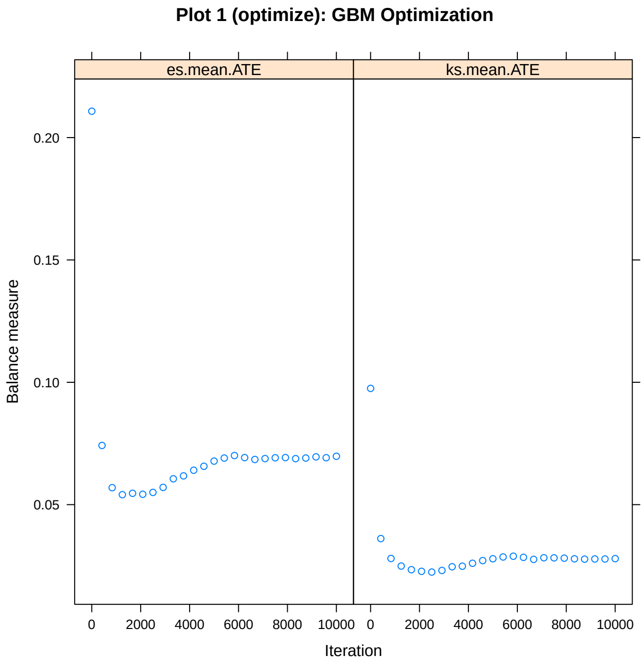


Figure 7: Example of the default diagnostic optimization plot from the specifying the `plotname` argument of the `%ps` macro for estimating ATE weights for the Lindner dataset.

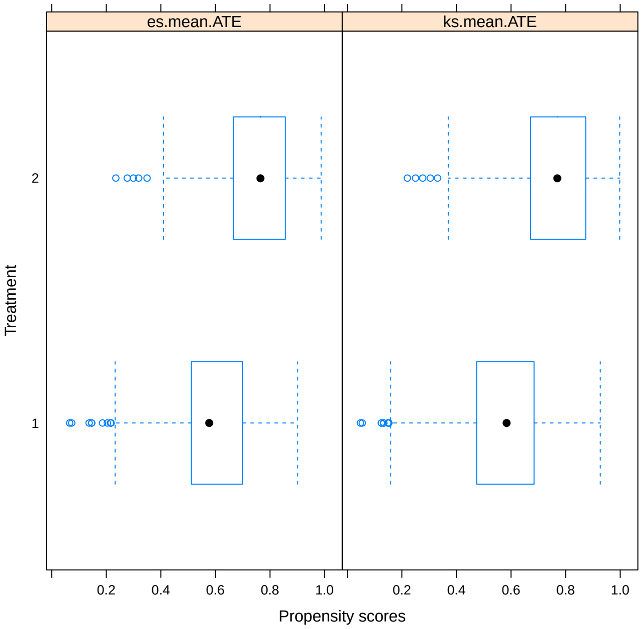


Figure 8: Example of the default diagnostic boxplot of propensity scores from the specifying the `plotname` argument of the `%ps` macro for estimating ATE weights for the Lindner dataset.

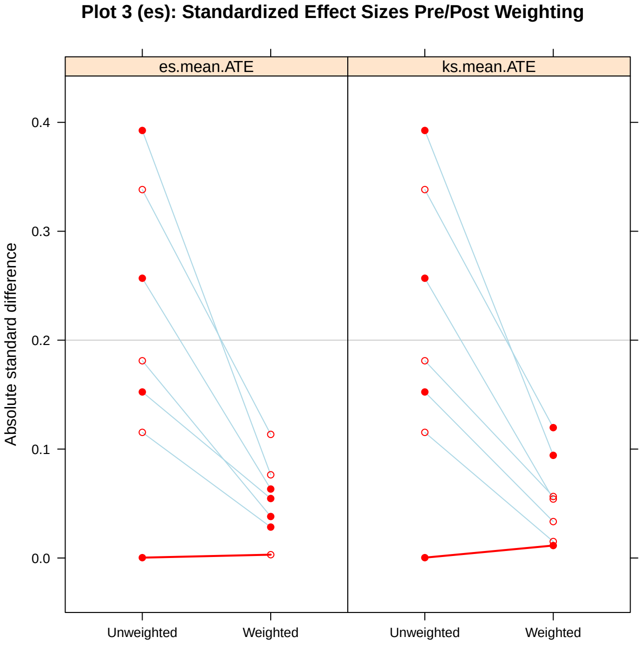


Figure 9: Example of the default diagnostic standardized effect size from the specifying the `plotname` argument of the `%ps` macro for estimating ATE weights for the Lindner dataset.

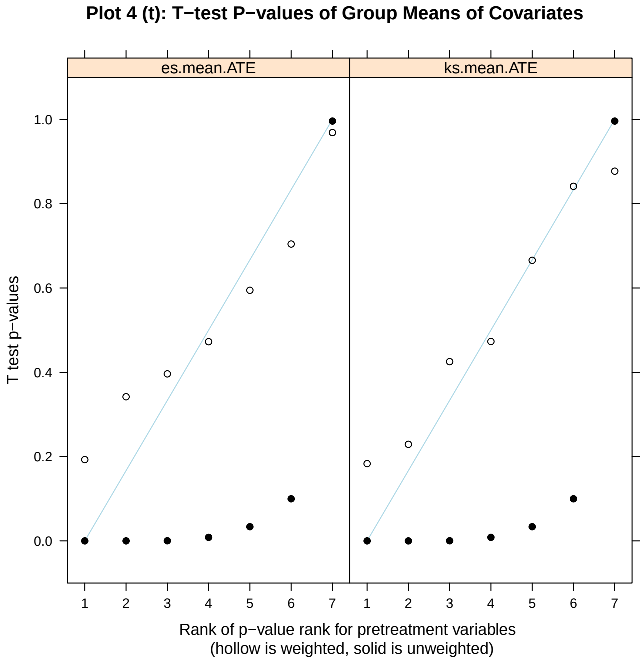


Figure 10: Example of the default diagnostic t-test p-value plot from the specifying the `plotname` argument of the `%ps` macro for estimating ATE weights for the Lindner dataset.

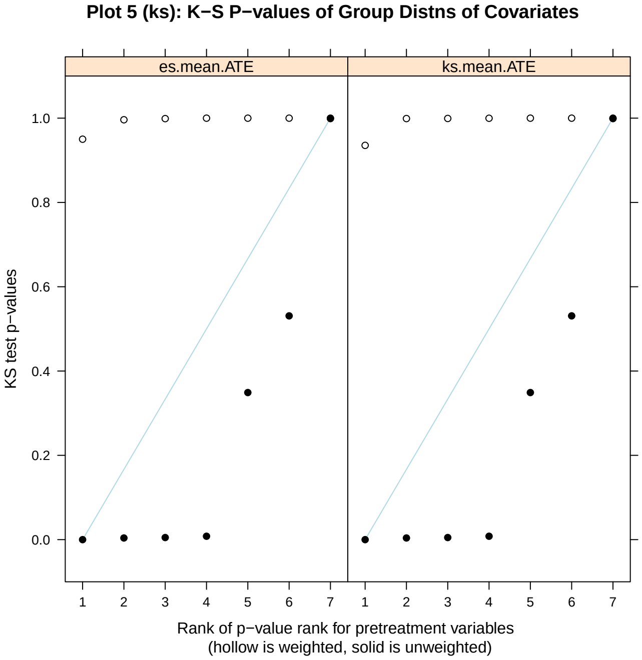


Figure 11: Example of the default diagnostic Kolmogorov-Smirnov test p-value plot from the specifying the `plotname` argument of the `%ps` macro for estimating ATE weights for the Lindner dataset.

From the summary table generated by `%ps`, we see that the "es.mean.ATE" stopping rule results in a slightly higher ESS with comparable balance measures, so we proceed with those weights. Also, we note that `ess.treat` is no longer equal to `n.treat` since we are focusing on ATE rather than ATT.

SAS Output: Summary Table for `abcix`

```
                                                            Summary table                                06:50 Tuesday, January 7, 2014   2

Obs  row_name            n_treat        n_ctrl     ess_treat      ess_ctrl        max_es       mean_es        max_ks      max_ks_p       mean_ks          iter

  1  unw                     698           298           698           298  0.3925636731  0.2052894288  0.1884194535             .  0.0979184466             .
  2  ks.mean.ATE             698           298  655.67541985  228.85005087  0.1197286705  0.0549551078  0.0401275716             .   0.022359505          2603
  3  es.mean.ATE             698           298  663.84002267  237.14437956  0.1135056019  0.0539249993  0.0381939155             .  0.0246343475          1303
```

Because the outcome is dichotomous, we use the `PROC SURVEYFREQ` procedure to obtain a weighted proportions and test for differences. The `chisq` option requests the Rao-Scott chi-square test and the `row` option requests the row percentages.

```
proc surveyfreq data=_inputds;
   tables sixMonthSurvive * abcix / chisq col;
   weights es_mean_ate;
run;
```

SAS Output: Weighted Frequencies and Chi-Square Test

```
                                             07:07 Tuesday, January 7, 2014   6

                          The SURVEYFREQ Procedure

                                Data Summary

                      Number of Observations           996
                      Sum of Weights            1838.76433



                       Table of sixMonthSurvive by abcix

 six
 Month                            Weighted   Std Dev of             Std Err of
 Survive   abcix    Frequency    Frequency     Wgt Freq   Percent      Percent
 -----------------------------------------------------------------------------
```

```
FALSE                     15     47.93424     14.02204    2.6069       0.7530
                          11     14.24984      4.34049    0.7750       0.2372

          Total           26     62.18409     14.63162    3.3818       0.7853
-----------------------------------------------------------------------------
TRUE                     283    823.91801     48.24287   44.8082       1.9149
                         687    952.66223     21.87682   51.8099       1.8755

          Total          970         1777     35.04639   96.6182       0.7853
-----------------------------------------------------------------------------
Total                    298    871.85225     49.44297   47.4151       1.8872
                         698    966.91207     21.68290   52.5849       1.8872

          Total          996         1839     34.93242   100.000
-----------------------------------------------------------------------------

                      Table of sixMonthSurvive by abcix

                   six
                   Month             Column     Std Err of
                   Survive   abcix   Percent    Col Percent
                   ----------------------------------------
                   FALSE             5.4980         1.5674
                                     1.4737         0.4483

                             Total
                   ----------------------------------------
                   TRUE             94.5020         1.5674
                                    98.5263         0.4483

                             Total
                   ----------------------------------------
                   Total            100.000
                                    100.000
                         Total
                 ---------------------------------------
```

The reweighting does not diminish the association between the treatment and the outcome. Indeed, it is marginally more significant after the reweighting.

4 Non-response weights

The `twang` macros were designed to estimate propensity score weights for the evaluation of treatment effects in observational or quasi-experimental studies. However, we find that the package includes functions and diagnostic tools that are highly valuable for other applications, such as for generating and diagnosing nonresponse weights for survey nonresponse or study attrition. We now present an example that uses the tools in `twang`. This example uses the subset of the US Sustaining Effects Study data distributed with the HLM software (Bryk, Raudenbush, Congdon, 1996), also available in the R package `mlmRev`, and included with the `twang` macros in the SAS data set `egsingle.sas7bdat`. The data include mathematics test scores for 1721 students in kindergarten to fourth grade. They also include student race (black, Hispanic, or other), gender, an indicator for whether or not the student had been retained in grade, the percent low income students at the school, the school size, the percent of mobile students, the students' grade-levels, student and school IDs, and grades converted to year by centering. The study analysis plans to analyze growth in math achievement from grade 1 to grade 4 using only students with complete data. However, the students with complete data differ from other students. To reduce bias that could potentially result from excluding incomplete cases, our analysis plan is to weight complete cases with nonresponse weights.

The goal of nonresponse weighting is to develop weights for the respondents that make them look like the entire sample — both the respondents and nonrespondents. Since the respondents already look like themselves, the hard part is to figure out how well each respondent represents the nonrespondents. Nonresponse weights equal the reciprocal of the probability of response and are applied only to respondents.

Note that the probability of response is equivalent to the propensity score if we consider subjects with an observed outcome to be the "treated" group, and those with an unobserved outcome to be the "controls". We wish to reweight the sample to make it equivalent to the population from which the sample was drawn, so ATE weights are appropriate in this case. Further, recall that the weights for the treated subjects are $1/p$ in an ATE analysis. Therefore we can reweight the sample of respondents using the weights returned from the `%ps` macro.

Before we can generate nonresponse weights, we need to prepare the data using the following commands. The data contain zero, one or two observations for students from grade "0" (kindergarten) and zero or one observation for each of grades 1 to 4. Only students with data from each of grades 1 to 4 will be included so we need to identify those students. We also need to collapse the data back to a single record per student and create a race variables with values 1 if “Black”, 2 if “Hispanic” and 3 otherwise.

```
proc sort data=sasin.egsingle out=egsingle;
   by childid grade;
run;

data egsingle;
   set egsingle;
   by childid;
   retain gpatt;
   length gpatt $6.;
   if black = 1 then race = 1;
   else if hispanic = 1 then race = 2;
   else race = 3;
   if first.childid then gpatt = put(grade, best1.);
   else gpatt = compress(gpatt || put(grade, best1.));
   if last.childid then do;
      resp = index(gpatt, "1234") > 0;
      output;
      end;
run;

proc freq data=egsingle;
   table resp race;
run;
```

SAS Output: Frequency Counts of Respondents and Racial Codes

```
                                 The SAS System                                3
                                                08:55 Tuesday, January 7, 2014

                              The FREQ Procedure

                                             Cumulative    Cumulative
          resp    Frequency     Percent     Frequency      Percent
```

```
-----------------------------------------------------------
   0         878       51.02            878        51.02
   1         843       48.98           1721       100.00



                                   Cumulative    Cumulative
race     Frequency      Percent     Frequency      Percent
-----------------------------------------------------------
   1        1195       69.44           1195        69.44
   2         250       14.53           1445        83.96
   3         276       16.04           1721       100.00
```

There are 1721 children in the study and 843 (49%) have the necessary four years of outcome data.

As discussed above, to use %ps to estimate nonresponse, we let respondents be the treatment group by modeling an indicator of response.

```
%ps(treatvar=resp,
    vars=race female size lowinc mobility,
    class=race,
    dataset=egsingle,
    stopmethod=es.mean ks.max,
    ntrees=2500,
    estimand = "ATE",
    Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe,
    plotname=egsingle_twang.pdf);
```

SAS Output: Balance Table for Nonresponse Analysis

```
                                                   Balance table: unw                                10:13 Tuesday, January 7, 2014   5

Obs row_name                      tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name

  1 unw.race:1                    0.656        0.475        0.731        0.443       -0.158       18.402            0        0.075            0 unw
  2 unw.race:2                    0.129        0.336        0.161        0.367       -0.093            .            .        0.031            0 unw
  3 unw.race:3                    0.215        0.411        0.108        0.311        0.259            .            .        0.107            0 unw
  4 unw.female:Female             0.466        0.499        0.517          0.5       -0.102        4.454        0.035        0.051        0.035 unw
  5 unw.female:Male               0.534        0.499        0.483          0.5        0.102            .            .        0.051        0.035 unw
  6 unw.size                    750.222      316.351      761.328      312.387       -0.035       -0.733        0.464        0.066        0.043 unw
  7 unw.lowinc                   75.492       28.581       80.745       24.082       -0.198       -4.116            0          0.1            0 unw
  8 unw.mobility                 32.662       14.044       36.436       13.701        -0.27       -5.642            0        0.122            0 unw

                                                Balance table: ks.max.ATE                            10:13 Tuesday, January 7, 2014   6

Obs row_name                      tx_mn        tx_sd        ct_mn        ct_sd   std_eff_sz         stat            p           ks      ks_pval table_name
```

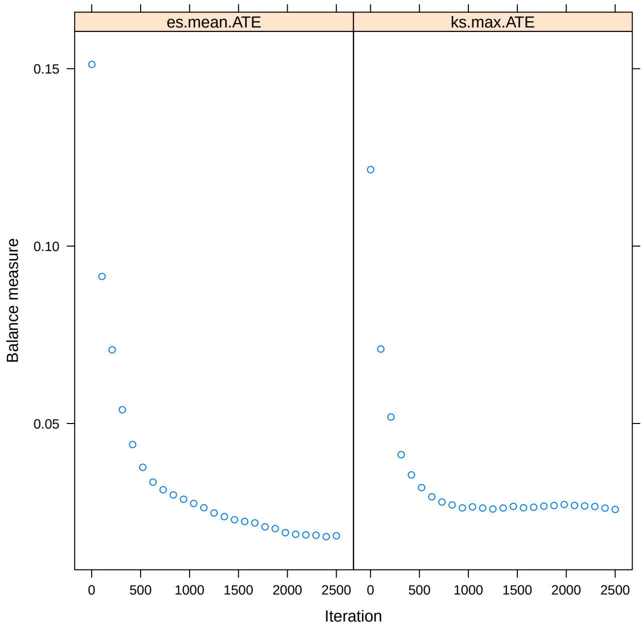


Figure 12: Optimization of `es.mean.ATE` and `ks.max.ATE` for nonresponse weighting of egsingle data. The horizontal axes indicate the number of iterations and the vertical axes indicate the measure of imbalance between the two groups. For `es.mean.ATE` the measure is the average effect size difference between the two groups and for `ks.max.ATE` the measure is the largest of the KS statistics

```
 17 ks.max.ATE.race:1               0.689      0.463      0.704      0.456     -0.032      0.386      0.679      0.015      0.679 ks.max.ATE
 18 ks.max.ATE.race:2               0.142      0.349      0.144      0.351     -0.005          .          .      0.002      0.679 ks.max.ATE
 19 ks.max.ATE.race:3               0.169      0.374      0.152      0.359      0.044          .          .      0.017      0.679 ks.max.ATE
 20 ks.max.ATE.female:Female        0.487        0.5      0.491        0.5     -0.008      0.024      0.877      0.004      0.877 ks.max.ATE
 21 ks.max.ATE.female:Male          0.513        0.5      0.509        0.5      0.008          .          .      0.004      0.877 ks.max.ATE
 22 ks.max.ATE.size               756.722    313.079    758.894    314.197     -0.007     -0.138       0.89      0.024      0.974 ks.max.ATE
 23 ks.max.ATE.lowinc              78.517     27.164     78.479     26.479      0.001      0.027      0.979      0.026      0.947 ks.max.ATE
 24 ks.max.ATE.mobility            34.227     13.696     34.787     13.972      -0.04     -0.791      0.429      0.022      0.991 ks.max.ATE

                                                  Balance table: es.mean.ATE                              10:13 Tuesday, January 7, 2014   7

Obs row_name                        tx_mn      tx_sd      ct_mn      ct_sd  std_eff_sz       stat          p         ks      ks_pval table_name

  9 es.mean.ATE.race:1              0.689      0.463      0.704      0.457     -0.031       0.39      0.676      0.014      0.676 es.mean.ATE
 10 es.mean.ATE.race:2              0.142      0.349      0.144      0.351     -0.006          .          .      0.002      0.676 es.mean.ATE
 11 es.mean.ATE.race:3              0.169      0.374      0.152      0.359      0.045          .          .      0.017      0.676 es.mean.ATE
 12 es.mean.ATE.female:Female       0.487        0.5      0.491        0.5     -0.007      0.017      0.896      0.003      0.896 es.mean.ATE
 13 es.mean.ATE.female:Male         0.513        0.5      0.509        0.5      0.007          .          .      0.003      0.896 es.mean.ATE
 14 es.mean.ATE.size              756.727    313.071    758.829    314.231     -0.007     -0.133      0.894      0.024      0.971 es.mean.ATE
 15 es.mean.ATE.lowinc             78.518     27.158     78.467     26.482      0.002      0.037      0.971      0.026       0.94 es.mean.ATE
 16 es.mean.ATE.mobility           34.226     13.697     34.781     13.972      -0.04     -0.784      0.433      0.022       0.99 es.mean.ATE
```

By default the balance table generated by `%ps` compares the weighted treatment group (respondents) to the weighted comparison group (nonresponders) – both groups weighted to equal the overall population.

However, the goal is to weight the respondents to match the population not to compare the weighted respondents and nonrespondents. The default balance table may be useful for evaluating the propensity scores, but it does not directly assess the quality of the weights for balancing the weighted respondents with the overall population.

We can “trick” the `%dxwts` macro in `twang` into making the desired comparison. We want to compare the weighted respondents to the unweighted full sample. When evaluating ATT weights, we compare the weighted comparison group with the unweighted treatment group. If we apply `%dxwts` to a data set where the “treatment” group is the entire `esingle` sample and the “control” group is the `esingle` respondents and the weights equal one for every student in the pseudo-treatment group and equal the weights from `%ps` for every student in the pseudo-control group, we can obtain the balance statistics we want.

We begin by setting up the data with the pseudo-treatment and control groups. We add ATE weights from the “ks.max” stopping rule as our nonresponse weights.

```
proc sort data=_inputds out=_inputds;
   by childid;
run;

proc sort data=sasin.egsingle out=egsingle;
   by childid;
run;

data egsingle_nrespwt;
   merge _inputds(keep=childid ks_max_ate resp) egsingle;
   by childid;
   if resp = 1;
   wgt = ks_max_ate;
run;

proc sort data=egsingle_nrespwt;
   by childid grade;
run;
```

We now stack the full sample and the respondents. The variable “nr2” is the pseudo-treatment indicator. We set it equal to one for the full sample and 0 for the respondents. Similarly, “wgt2” is the pseudo-ATT weight which is set equal to one for the full sample and equal to the nonresponse weights for the respondents.

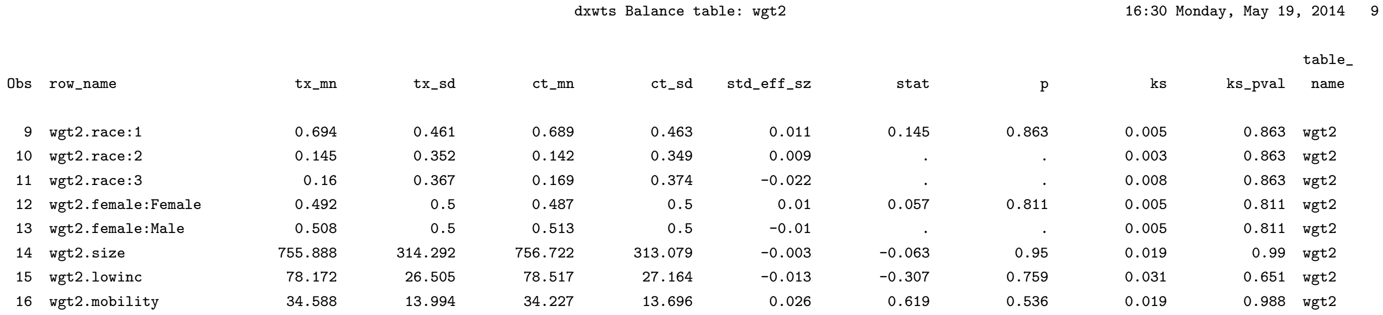
```
data egtmp;
```

```
   set egsingle(in=_full) egsingle_nrespwt;
   if _full then do;
      nr2 = 1;
      wgt2 = 1;
      end;
   else do;
      nr2 = 0;
      wgt2 = wgt;
      end;
run;
```

We now run `%dxwts` to obtain the balance statistics.

```
%dxwts(treatvar=nr2,
           vars=race female size lowinc mobility,
          class=race female,
        dataset=egtmp,
     weightvars=wgt2
       estimand=ATT,
  permtestiters=0,
           Rcmd=C:\Program Files\R\R-3.0.2\bin\x64\R.exe)
```

SAS Output: Balance Table for Nonresponse Analysis
Comparing Weighted Respondents to Overall Sample



```
                                              dxwts Balance table: wgt2                              16:30 Monday, May 19, 2014   9

                                                                                                                             table_
 Obs  row_name                   tx_mn       tx_sd       ct_mn       ct_sd    std_eff_sz        stat           p          ks     ks_pval   name

   9  wgt2.race:1                0.694       0.461       0.689       0.463         0.011       0.145       0.863       0.005       0.863  wgt2
  10  wgt2.race:2                0.145       0.352       0.142       0.349         0.009           .           .       0.003       0.863  wgt2
  11  wgt2.race:3                 0.16       0.367       0.169       0.374        -0.022           .           .       0.008       0.863  wgt2
  12  wgt2.female:Female         0.492         0.5       0.487         0.5          0.01       0.057       0.811       0.005       0.811  wgt2
  13  wgt2.female:Male           0.508         0.5       0.513         0.5         -0.01           .           .       0.005       0.811  wgt2
  14  wgt2.size                755.888     314.292     756.722     313.079        -0.003      -0.063        0.95       0.019        0.99  wgt2
  15  wgt2.lowinc               78.172      26.505      78.517      27.164        -0.013      -0.307       0.759       0.031       0.651  wgt2
  16  wgt2.mobility             34.588      13.994      34.227      13.696         0.026       0.619       0.536       0.019       0.988  wgt2
```

The resulting balance table includes a table for an unweighted comparison the the respondents with the overall sample and a weighted comparison. We reproduce only the weighted comparison here. In the table, the columns for the treatment group mean and standard deviation (“tx_mn” and “tx_sd”) contain the sample statistics for the full sample (`egsingle`) and the columns for the comparison group (“ct_mn” and “ct_sd”) contain the weighted respondent The following codes prepares an analysis file with all of the data from the respondents with the nonresponse weights included.

```
proc sort data=sasin.egsingle out=egsingle;
   by childid;
run;

data egsingle_analysis;
   merge egsingle_nrespwt(keep=childid wgt in=_nr) egsingle;
   by childid;
   if _nr;
run;

proc sort data=egsingle_analysis;
   by childid grade;
run;
```

5 The details of twang

5.1 Propensity scores and weighting

Propensity scores can be used to reweight comparison cases so that the distribution of their features match the distribution of features of the treatment cases, for ATT, or cases from both treatment and control groups to match each other, for ATE (Rosenbaum 1987, Wooldridge 2002, Hirano and Imbens 2001, McCaffrey *et al.* 2004) Let $f(\mathbf{x}|t=1)$ be the distribution of features for the treatment cases and $f(\mathbf{x}|t=0)$ be the distribution of features for the comparison cases. If treatments were randomized then we would expect these two distributions to be similar. When they differ for ATT we will construct a weight, $w(\mathbf{x})$, so that

$$f(\mathbf{x}|t=1) = w(\mathbf{x})f(\mathbf{x}|t=0). \tag{2}$$

For example, if $f(\text{age=65}, \text{sex=F}|t=1) = 0.10$ and $f(\text{age=65}, \text{sex=F}|t=0) = 0.05$ (i.e. 10% of the treatment cases and 5% of the comparison cases are 65 year old females) then we need to give a weight of 2.0 to every 65 year old female in the comparison group so that they have the same representation as in the treatment group. More generally, we can solve (2) for $w(\mathbf{x})$ and apply Bayes Theorem to the numerator and the denominator to give an expression for the propensity score weight for comparison cases,

$$w(\mathbf{x}) = K\frac{f(t=1|\mathbf{x})}{f(t=0|\mathbf{x})} = K\frac{P(t=1|\mathbf{x})}{1-P(t=1|\mathbf{x})}, \tag{3}$$

where K is a normalization constant that will cancel out in the outcomes analysis. Equation (3) indicates that if we assign a weight to comparison case i equal to the odds that a case with features $\mathbf{x}_i$ would be exposed to the treatment, then the distribution of their features would balance. Note that for comparison cases with features that are atypical of treatment cases, the propensity score $P(t = 1|\mathbf{x})$ would be near 0 and would produce a weight near 0. On the other hand, comparison cases with features typical of the treatment cases would receive larger weights.

For ATE, each group is weighted to match the population. The weight must satisfy:

$$f(\mathbf{x}|t = 1) = w(\mathbf{x})f(\mathbf{x}), \text{ and} \tag{4}$$

$$f(\mathbf{x}|t = 0) = w(\mathbf{x})f(\mathbf{x}), \text{ and} \tag{5}$$

Again using Bayes Theorem we obtain $w(\mathbf{x}) \propto 1/f(t = 1|\mathbf{x})$ for the treatment group and $w(\mathbf{x}) \propto 1/f(t = 0|\mathbf{x})$ for the control group.

5.2 Estimating the propensity score

In randomized studies $P(t = 1|\mathbf{x})$ is known and fixed in the study design. In observational studies the propensity score is unknown and must be estimated, but poor estimation of the propensity scores can cause just as much of a problem for estimating treatment effects as poor regression modeling of the outcome. Linear logistic regression is the common method for estimating propensity scores, and can suffice for many problems. Linear logistic regression for propensity scores estimates the log-odds of a case being in the treatment given $\mathbf{x}$ as

$$\log \frac{P(t = 1|\mathbf{x})}{1 - P(t = 1|\mathbf{x})} = \beta'\mathbf{x} \tag{6}$$

Usually, β is selected to maximize the logistic log-likelihood

$$\ell\beta = \frac{1}{n}\sum_{i=1}^{n} t_i\beta'\mathbf{x}_i - \log\big(1 + \exp(\beta'\mathbf{x}_i)\big) \tag{7}$$

Maximizing (7) provides the maximum likelihood estimates of β. However, in an attempt to remove as much confounding as possible, observational studies often record data on a large number of potential confounders, many of which can be correlated with one another. Standard methods for fitting logistic regression models to such data with the iteratively reweighted least squares algorithm can be statistically and numerically unstable. To improve the propensity score estimates we might also wish to include non-linear effects and interactions in $\mathbf{x}$. The inclusion of such terms only increases the instability of the models.

One increasingly popular method for fitting models with numerous correlated variables is the lasso (least absolute subset selection and shrinkage operator) introduced in statistics in Tibshirani (1996). For logistic regression, lasso estimation replaces (7) with a version that penalizes the absolute magnitude of the coefficients

$$\ell\beta = \frac{1}{n}\sum_{i=1}^{n} t_i\beta'\mathbf{x}_i - \log\left(1+\exp(\beta'\mathbf{x}_i)\right) - \lambda\sum_{j=1}^{J}|\beta_j| \qquad (8)$$

The second term on the right-hand side of the equation is the penalty term since it decreases the overall of $\ell\beta$ when there are coefficients that are large in absolute value. Setting $\lambda = 0$ returns the standard (and potentially unstable) logistic regression estimates of β. Setting λ to be very large essentially forces all of the β_j to be equal to 0 (the penalty excludes β_0). For a fixed value of λ the estimated $\hat{\beta}$ can have many coefficients exactly equal to 0, not just extremely small but precisely 0, and only the most powerful predictors of t will be non-zero. As a result the absolute penalty operates as a variable selection penalty. In practice, if we have several predictors of t that are highly correlated with each other, the lasso tends to include all of them in the model, shrink their coefficients toward 0, and produce a predictive model that utilizes all of the information in the covariates, producing a model with greater out-of-sample predictive performance than models fit using variable subset selection methods.

Our aim is to include as covariates all piecewise constant functions of the potential confounders and their interactions. That is, in $\mathbf{x}$ we will include indicator functions for continuous variables like $I(\text{age} < 15), I(\text{age} < 16), \ldots, I(\text{age} < 90)$, etc., for categorical variables like $I(\text{sex} = \text{male}), I(\text{prior MI} = \text{TRUE})$, and interactions among them like $I(\text{age} < 16)I(\text{sex} = \text{male})I(\text{prior MI} = \text{TRUE})$. This collection of basis functions spans a plausible set of propensity score functions, are computationally efficient, and are flat at the extremes of $\mathbf{x}$ reducing the likelihood of propensity score estimates near 0 and 1 that can occur with linear basis functions of $\mathbf{x}$. Theoretically with the lasso we can estimate the model in (8), selecting a λ small enough so that it will eliminate most of the irrelevant terms and yield a sparse model with only the most important main effects and interactions. Boosting (Friedman 2001, 2003, Ridgeway 1999) effectively implements this strategy using a computationally efficient method that Efron *et al.* (2004) showed is equivalent to optimizing (8). With boosting it is possible to maximize (8) for a range of values of λ with no additional computational effort than for a specific value of λ. We use boosted logistic regression as implemented in the generalized boosted modeling (`gbm`) package in R (Ridgeway 2005).

5.3 Evaluating the weights

As with regression analyses, propensity score methods cannot adjust for unmeasured covariates that are uncorrelated with the observed covariates. Nonetheless, the quality of the adjustment for the observed covariates achieved by propensity score weighting is easy to evaluate. The estimated propensity score weights should equalize the distributions of the cases' features as in (2). This implies that weighted statistics of the covariates of the comparison group should equal the same statistics for the treatment group. For example, the weighted average of the age of comparison cases should equal the average age of the treatment cases. To assess the quality of the propensity score weights one could compare a variety of statistics such as means, medians, variances, and Kolmogorov-Smirnov statistics for each covariate as well as interactions. The `twang` package provides both the standardized effect sizes and KS statistics and p-values testing for differences in the means and distributions of the covariates for analysts to use in assessing balance.

5.4 Analysis of outcomes

With propensity score analyses the final outcomes analysis is generally straightforward, while the propensity score estimation may require complex modeling. Once we have weights that equalize the distribution of features of treatment and control cases by reweighting. For ATT, we give each treatment case a weight of 1 and each comparison case a weight $w_i = p(\mathbf{x}_i)/(1 - p(\mathbf{x}_i))$. To estimate the ATE, we give control cases weight $w_i = 1/p(\mathbf{x}_i)$ and we give the treatment cases $w_i = 1/(1 - p(\mathbf{x}_i)$. We then estimate the treatment effect estimate with a weighted regression model that contains only a treatment indicator. No additional covariates are needed if the weights account for differences in $\mathbf{x}$.

A combination of propensity score weighting and covariate adjustment can be useful for several reasons. First, the propensity scores may not have been able to completely balance all of the covariates. The inclusion of these covariates in addition to the treatment indicator in a weighted regression model may correct this if the imbalance is relatively small. Second, in addition to exposure, the relationship between some of the covariates and the outcome may also be of interest. Their inclusion can provide coefficients that can estimate the direction and magnitude of the relationship. Third, as with randomized trials, stratifying on covariates that are highly correlated with the outcome can improve the precision of estimates. Lastly, the some treatment effect estimators that utilize an outcomes regression model and propensity scores are "doubly robust" in the sense that if either the propensity score model is correct or the regression model is correct then the treatment effect estimator will be unbiased (Bang & Robins 2005).

References

[1] Bang H. and J. Robins (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics* 61:692–972.

[2] Bland M. (2013). "Do baseline p-values follow a uniform distribution in randomised trials?" *PLoS ONE* 8(10): e76010: 1–5.

[3] Dehejia, R.H. and S. Wahba (1999). "Causal effects in nonexperimental studies: re-evaluating the evaluation of training programs," *Journal of the American Statistical Association* 94:1053–1062.

[4] Efron, B., T. Hastie, I. Johnstone, R. Tibshirani (2004). "Least angle regression," *Annals of Statistics* 32(2):407–499.

[5] Friedman, J.H. (2001). "Greedy function approximation: a gradient boosting machine," *Annals of Statistics* 29(5):1189–1232.

[6] Friedman, J.H. (2002). "Stochastic gradient boosting," *Computational Statistics and Data Analysis* 38(4):367–378.

[7] Friedman, J.H., T. Hastie, R. Tibshirani (2000). "Additive logistic regression: a statistical view of boosting," *Annals of Statistics* 28(2):337–374.

[8] Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning.* Springer-Verlag, New York.

[9] Helmreich, J.E., and R.M. Pruzek (2009). "PSAgraphics: An R package to support propensity score analysis," *Journal of Statistical Software* 29(6):1–23.

[10] Hirano, K. and G. Imbens (2001). "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services and Outcomes Research Methodology* 2:259–278.

[11] Huppler-Hullsiek, K. and T. Louis (2002) "Propensity score modeling strategies for the causal analysis of observational data," *Biostatistics* 3:179–193.

[12] Lalonde, R. (1986). "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review* 76:604–620.

[13] Little, R. J. and S. Vartivarian (2004). "Does weighting for nonresponse increase the variance of survey means?" *ASA Proceedings of the Joint Statistical Meetings*, 3897-3904 American Statistical Association (Alexandria, VA) `http://www.bepress.com/cgi/viewcontent.cgi?article=1034&context=umichbiostat`.

[14] McCaffrey, D. F., B. A. Griffin, D. Almirall, M.E. Slaughter, R., Ramchand, L. Burgette (2013). "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, 32:3388-3414.

[15] McCaffrey, D., G. Ridgeway, A. Morral (2004). "Propensity score estimation with boosted regression for evaluating adolescent substance abuse treatment," *Psychological Methods* 9(4):403–425.

[16] Obenchain, B. (2011). *USPS 1.2 package manual.* `http://cran.r-project.org/web/packages/USPS/USPS.pdf`

[17] R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[18] Ridgeway, G. (1999). "The state of boosting," *Computing Science and Statistics* 31:172–181.

[19] Ridgeway, G. (2005). *GBM 1.5 package manual.* `http://cran.r-project.org/doc/packages/gbm.pdf`.

[20] Ridgeway, G. (2006). "Assessing the effect of race bias in post-traffic stop outcomes using propensity scores." *Journal of Quantitative Criminology* 22(1):1–29.

[21] Ridgeway, G., McCaffrey, D., Morral, A. Griffin, B.A., Burgette, L. (2013) *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups.* R package version 1.3-20. `http://CRAN.R-project.org/package=twang`.

[22] Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70(1):41–55.

[23] Rosenbaum, P. (1987). "Model-based direct adjustment," *Journal of the American Statistical Association* 82:387–394.

[24] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B* 58(1):267–288.

[25] Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*, MIT Press, Cambridge.