Tool

Toolkit for Weighting and Analysis of Nonequivalent Groups

A Tutorial on the Twang Commands for Stata Users

Matthew Cefalu, Shuangshuang Liu, Craig Marti

RAND Justice, Information, and Environment

TL-170-NIDA
September 2015
Prepared for the National Institute on Drug Abuse

For more information on this publication, visit www.rand.org/t/tl170

Published by the RAND Corporation, Santa Monica, Calif.

Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial on the twang commands for Stata users

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 commands 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, and after extensive use, it received a major update in 2012. The Stata twang macros were developed in 2015 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 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 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** command, which implements generalized boosted regression modeling (GBM) to estimate the propensity scores. Other tools are included 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** Stata package aims 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 **twang** ado files and supporting documents to a folder on your computer. The files are available at http://www.rand.org/statistics/twang/downloads.html. They include:

- **twang** Stata package – files containing the Stata program and help files with details on implementing the commands
- Stata_Start-up.pdf - step by step details on installing R
- lalonde.dta - Example dataset from the Lalonde Study
- lindner.dta - Example dataset from the Lindner Study
- egsingle.dta - Example dataset from the Raudenbush and Bryk Study
- tutorial_code.do - Stata code from examples presented in this tutorial
- tutorial_code_using_macros.do - Stata code from examples presented in this tutorial using global macros to keep track of the location of adofiles, data, and output

The 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 ado files, you could place the files in the PERSONAL ado-path directory, which can be identified using **adopath**. Alternately, you can place the files in any directory, and add the directory to ado-path using the command

adopath + "C:\Users\uname\adofile"

The help files should also be placed in the same folder. [1] After adding the directory to ado-path, the help files can be opened using **help** command together with the command that you need help with (e.g., **help ps**). Note that **adopath** temporarily adds the directory to the ado-path — you must rerun the command each time Stata is opened.

The ado files will run code in R and import the results into your Stata session. The ado files 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 in R batch mode. The script file exports weights and diagnostic information in .csv files that are then ported back into Stata. All files created by the macro are stored in the directory specified by the user in the **objpath** option as seen in subsequent sections. Any files in this directory created from previous calls of **ps** will be overwritten.

[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.do" file, that is available with the twang ado files contains all the code from this tutorial in text file. Analysts can use that file to copy the code and run the examples.

To manage file transfer between Stata and R and ensure all the necessary R functions are working, we require the operating system to be Windows Vista or later, Mac OS-X or later, or UNIX.

The ado files rely on the **twang** package in R. You will need to install R 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 Stata Commands Start-Up" (Stata_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 Files\R\R-3.0.2\bin\x64

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

C:\Program Files\R\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 Rscript.exe for batch implementation. Although not necessary, users with some familiarity with R might also want to install the **twang** package in R. The ado files will install the package if the users do not.

The ado files also create a “TWANG” folder in a standard location based on the operating system. In Windows, the folder is 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). In Mac OS-X, the folder is in the user’s Library folder (/Users/username/Library/Twang). This folder will remain on the user's hard drive until it is removed. Users can remove the folder using any method they would normally use for removing a folder when they no longer plan to use **twang.**

2.2 Estimating propensity scores with the ps command

To demonstrate the package we utilize data from Lalonde's National Supported Work Demonstration (NSWD) analysis (Lalonde, 1986, Dehejia & Wahba, 1999, http://users.nber.org/~rdehejia/nswdata2.html). The NSWD was a temporary employment program that gave work experience and counseling service to disadvantaged workers to help them move into the labor market. A comparison group that did not participate in the NSWD was constructed using the Current Population Survey (CPS) for the same years as the NSWD.

In this example, we will estimate the causal effect of the NSWD on earnings among those who participated in the program, or, in other words, the average treatment effect on the treated (ATT). Pretreatment covariates include age, education, race, ethnicity, education level, marital status, earnings in 1974 (pretreatment), and earnings in 1975 (pretreatment). As we will show, the challenge in this analysis is that the distribution of these pretreatment covariates differs between the individuals

who participated in the NSWD and those who did not. The dataset is provided with the **twang** Stata package (lalonde.dta).

In the **lalonde** dataset, the variable **treat** is the 0/1 treatment indicator, 1 indicates “treatment” by being part of the NSWD and 0 indicates “comparison" cases drawn from the CPS. In order to estimate a treatment effect for this demonstration program that is unbiased by pretreatment group differences on other observed covariates, we include the pretreatment covariates listed above in a propensity score model of treatment assignment. The **ps** command is the primary method in twang for estimating propensity scores. This step is computationally intensive and can take a few minutes.

```
· use lalonde.dta, clear
· ps treat age educ black hispan nodegree married re74 re75, ///
      ntrees(5000) intdepth(2) shrinkage(0.01) ///
      permtestiters(0) stopmethod(es.mean ks.max) estimand(ATT) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output) ///
      plotname(C:\Users\uname\twang\output\lalonde_twang.pdf)
· cd "C:\Users\uname\twang\output"
· save lalonde_att_wgts,replace
```

The syntax of the **ps** command requires some discussion. All of the variables that are to be used in the model are listed after the command name and before the comma. The first variable is the treatment variable, which as noted earlier is “treat". It is followed by the names of the covariates.

All the options of the command go after the comma. **ntrees**, **intdepth**, and **shrinkage** are parameters for the GBM that **ps** computes. The option **ntrees** is the maximum number of iterations that the GBM will run. There will be a warning if the estimated optimal number of iterations is too close to the bound selected in this option 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 it if this warning appears. The option **intdepth** controls the level of interactions allowed in the GBM, with larger values specifying more complex models. We specified a value of 2, indicating that the algorithm will consider all two-way interactions between covariates. The GBM estimation algorithm uses shrinkage to enhance the smoothness of resulting model. The **shrinkage** option 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. Additional details on **ntrees**, **intdepth**, and **shrinkage** can be found in McCaffrey, Ridgeway, and Morral (2004).

permtestiters specifies whether p-values for KS statistics are 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** option 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** option. 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 stopping rules in **twang** use two balance metrics: absolute standardized bias (also referred to as the absolute standardized mean difference or the Effect Size) 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** option.

The **sampw** option 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** option 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** command are the weights which can be used for estimating effects. It also produces checks of the balance of covariates 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** option.

The propensity score based weights created by the **ps** command could be found in the two new variables at the end of the original dataset. In this example, we save this new dataset as “lalonde_att_wgts” in the output folder specified by the **cd** command (**save lalonde_att_wgts, replace**). 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 ‘esmeanatt’ with the weights from a GBM with the iterations chosen to minimize the mean standardized bias (effect size) and a second weight variable ‘ksmaxatt’ 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. This dataset can be saved and used to estimate the treatment effects.

The options **rcmd** specifies the R program executable file for running R. The location of the file is determined through the installation of R. The specification of **rcmd** is not necessary for Mac OS-X, but the default location of the executable is “/usr/bin/Rscript”. 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\Rscript.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** option, "**rcmd(Rscript)**" will work.[2] (Similarly specifying the ".exe" extension is not necessary.)

The option **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, then no plots are created. If the option 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 in the option **objpath.** In our example, we specify **plotname** as "C:\Users\uname\twang\output\lalonde_twang.pdf" so the plots will be stored in that folder. Users will need to specify an appropriate folder where they can write file to use if they specify it.

The final option **objpath** specifies a folder where files created by the command to run the **twang** functions in R and return the results to Stata are stored. Namely, an "R object" ("ps.RData") with the GBM fit information and a log of the R session ("ps.Rout"). The **objpath** option is required for running the **ps** command and must reference an existing directory.

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 "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** option in **ps** command. This will create all the default diagnostics plots available in **twang**. They will be stored in the single pdf file specified by the option. In this example the file is: "C:\Users\uname\twang\output\lalonde_twang.pdf". The default file created by specifying **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** option. If only one stopping rule is specified each page contains a single panel. Figure 1 presents the plots for the first page of the **lalonde_twang.pdf** file.

[2] 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 Files\R\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 [Or appropriate path for executable]
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.

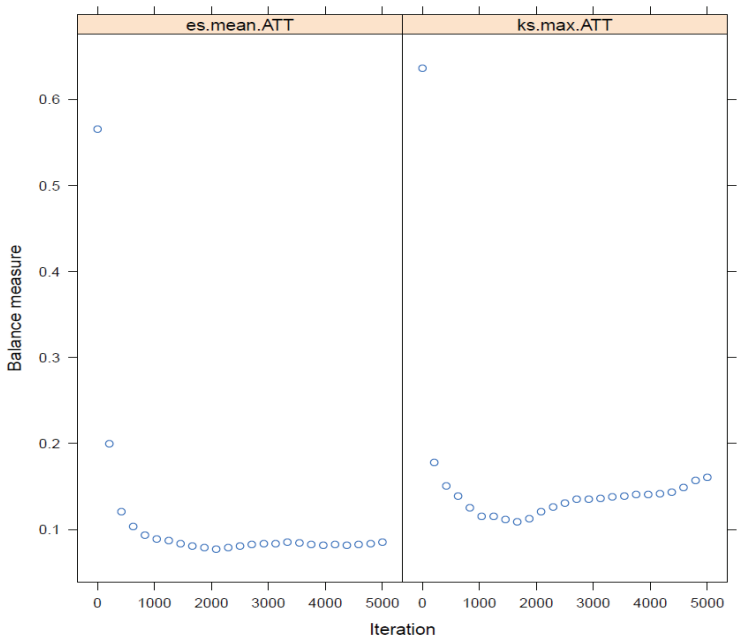

Figure 1: Example of an optimization plot for two stopping rules ("es.mean," and "ks.max") for estimating ATT weights for the **Lalonde** dataset. It was generated by setting the **plotname** option in the **ps** command and appears as the first page of a multiple page document of diagnostics plots.

The second way to obtain the plot is to run the **psplot** command which can create specific plots and store them using pdf or other file formats. The **inputobj** option specifies the R object created by the **ps** command. **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 **plots** option specifies the diagnostic plot to be created. Only one type of diagnostic plot can be created by a **psplot** command.[3] 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 **psplot** command.

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_opt.pdf) plotformat(pdf)  ///
      plots(1) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output)
```

The **psplot** command also allows for restricting the plot to results for a single stopping rule by specifying it by number in the subset option. Stopping rules are numbered by alphabetical order; not the order in which they are specified. See Figure 1.

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_opt_ks.pdf) plotformat(pdf) ///
```

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

```
plots(1) subset(2) ///
rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
objpath(C:\Users\uname\twang\output)
```

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 and **ps** rerun. 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.

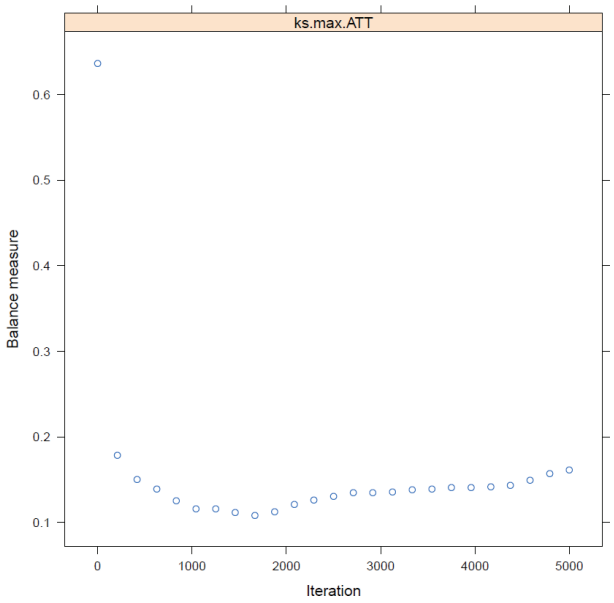

Figure 2: Example of an optimization plot for a single stopping rule (ks.max) for estimating **ATT** weights for the **Lalonde** dataset.

2.3 Assessing "balance" using balance tables

The **ps** command generates a "balance table" which provides a tabular summary of the balance between the covariate distributions for the treatment and control groups. The table created by the **ps** command could be found in a csv file "**baltab**", in the folder specified by **objpath**, or can be printed by the use of the post-estimation **balance** command. The **balance** command can create three types of balance tables:

a summary table generated by

- **balance, summary**

an unweighted table generated by

· **balance, unweighted**

and a weighted table generated by

· **balance, weighted**

Multiple tables can be created with a single syntax. For instance,

· **balance, summary unweighted weighted**

would create all three types of tables.

The weighted tables show 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 unweighted table provides the same information without weighting to give a sense of the general imbalance in the groups. For both the weighted and unweighted tables, balance is assessed separately for each covariate by each stop method. Statistics 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 in the covariates, **twang** will attempt to construct weights that also balance rates of missingness in the treatment and control groups. In this case, the balance table will have an extra row for each variable that has missing entries. The columns of the table consist of the following items:

txmn, ctmn The treatment means and the control means for each of the variables. The unweighted table shows the unweighted means. For each stopping rule the means are weighted using weights corresponding to the gbm model selected by **ps** command 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 txmn

txsd, ctsd The propensity score weighted treatment and control groups' standard deviations for each of the variables. The unweighted table shows the unweighted standard deviations

stdeffsz 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, kspval 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** option of the **ps** command. For categorical variables this is just the χ^2 test p-value.

Unweighted

	txmn	txsd	ctmn	ctsd	stdeffsz	stat	p	ks	kspval
age	25.82	7.155	28.03	10.79	-.309	-2.994	.003	.158	.003
black	.843	.365	.203	.403	1.757	19.37	0	.64	0
educ	10.35	2.011	10.23	2.855	.055	.547	.584	.111	.074
hispan	.059	.237	.142	.35	-.349	-3.413	.001	.083	.317
married	.189	.393	.513	.5	-.824	-8.607	0	.324	0
nodegree	.708	.456	.597	.491	.244	2.716	.007	.111	.074
re74	2096	4887	5619	6789	-.721	-7.254	0	.447	0
re75	1532	3219	2466	3292	-.29	-3.282	.001	.288	0

Weighted: esmean

	txmn	txsd	ctmn	ctsd	stdeffsz	stat	p	ks	kspval
age	25.82	7.155	25.8	7.279	.002	.015	.988	.122	.892
black	.843	.365	.842	.365	.003	.027	.978	.001	1
educ	10.35	2.011	10.57	2.089	-.113	-.706	.48	.099	.977
hispan	.059	.237	.042	.202	.072	.804	.421	.017	1
married	.189	.393	.189	.392	.002	.012	.99	.001	1
nodegree	.708	.456	.609	.489	.218	.967	.334	.099	.977
re74	2096	4887	1557	3802	.11	1.027	.305	.066	1
re75	1532	3219	1212	2648	.1	.833	.405	.103	.969

Weighted: ksmax

	txmn	txsd	ctmn	ctsd	stdeffsz	stat	p	ks	kspval
age	25.82	7.155	25.76	7.408	.007	.055	.956	.107	.919
black	.843	.365	.835	.371	.022	.187	.852	.008	1
educ	10.35	2.011	10.57	2.14	-.113	-.712	.477	.107	.919
hispan	.059	.237	.043	.203	.069	.779	.436	.016	1
married	.189	.393	.199	.4	-.024	-.169	.866	.01	1
nodegree	.708	.456	.601	.49	.235	1.1	.272	.107	.919
re74	2096	4887	1674	3945	.086	.8	.424	.054	1
re75	1532	3219	1257	2675	.085	.722	.471	.094	.971

The **balance** command with the **summary** option returns a compact summary of the sample sizes of the groups and the balance measures in the results window as well as in a csv file "**summary**" in the same folder specified by **objpath**. The summary table includes one row for the results using the weights produced by each stopping rule specified by the **stopmethod** option and one row for the unweighted data. Each row contains: the row name which specifies the stopping rule used for the weights or that the data are unweighted ("unw"), the estimand is also appended to the label; **ntreat** and **nctrl**, the treatment and control group sample sizes, respectively; **esstreat** and **essctrl**, the effective sample sizes of the treatment and control groups, **esstreat** equals the treatment group sample size for ATT because the weights are one (additional details on the effective sample size follow); **maxes** and **meanes**, and **maxks** and **meanks**, the maximum and mean or average of the

standardized effect sizes or KS statistics for the covariates, respectively; **maxksp**, the p-value for testing 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 **maxksp** is only produced when **permtestiters**> 0; otherwise it is missing for all rows of the summary table, as it is for our example. If **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 **maxksp** 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.

```
Summary
-------------------------------------------------------------------------------
        ntreat  nctrl esstreat  essctrl  maxes  meanes   maxks  maxksp  meanks    iter
-------------------------------------------------------------------------------
unw        185     429     185      429  1.757   .5687   .6404       .   .2702       .
es_mean    185     429     185    22.96  .2178   .0775   .1223       .   .0636    2127
ks_max     185     429     185    27.05  .2349   .0803   .1071       .   .0628    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

$$ESS = \frac{(\sum_{i \in C} w_i)^2}{\sum_{i \in C} w_i^2} \qquad (1)$$

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 **psplot** command can generate useful diagnostic plots to evaluate the propensity scores. The full set of plots available in twang and the option 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** option[4]. 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

[4] Recall that, for Windows, 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 Rscript.exe is stored. **rcmd** is optional on Mac OS-X.

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 in **psplot** command.

Descriptive option	Numeric option	Description
“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

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_box.pdf) plotformat(pdf) ///
      plots(2) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output)
```

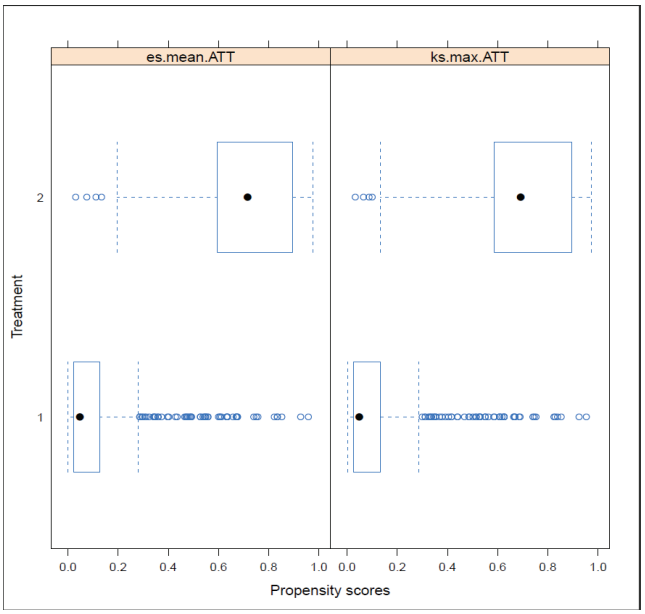

Figure 3: Example of the default diagnostic boxplot of propensity scores from the **psplot** command for estimating ATT weights for the **Lalonde** dataset.

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.

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_es.pdf) plotformat(pdf) ///
      plots(3) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output)
```

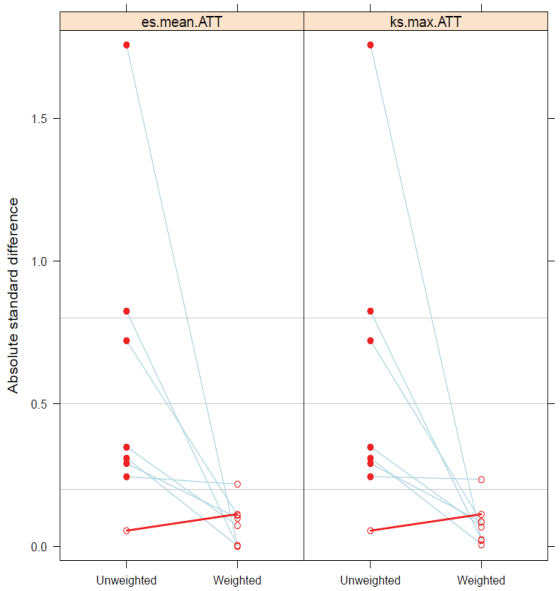

Figure 4: Example of the default diagnostic standardized effect size plot from the **psplot** command for estimating ATT weights for the **Lalonde** dataset.

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 inspect the p-values of the t-tests for group differences in the covariate means.

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_p.pdf) plotformat(pdf) ///
      plots(4) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output)
```

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 option **plots** = 5 or "ks" (see Figure 6).

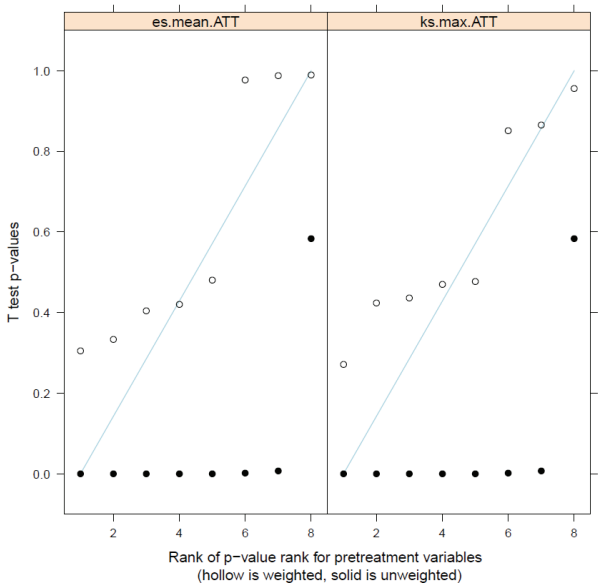

Figure 5: Example of the default diagnostic t-test p-value plot from the **psplot** command for estimating ATT weights for the **Lalonde** dataset.

```
· psplot, ///
      inputobj(C:\Users\uname\twang\output\ps.RData) ///
      plotname(lalonde_ks.pdf) plotformat(pdf) ///
      plots(5) ///
      rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
      objpath(C:\Users\uname\twang\output)
```

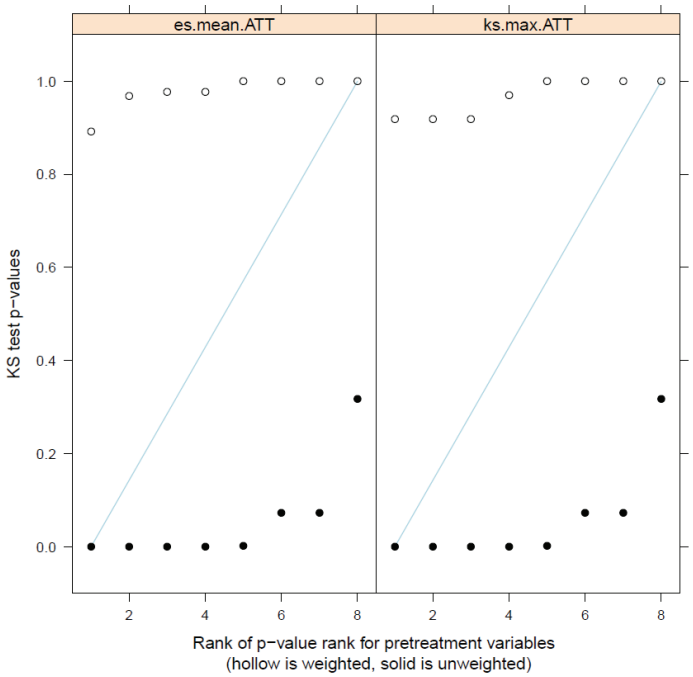

Figure 6: Example of the default diagnostic Kolmogorov-Smirnov test p-value plot from the **psplot** command for estimating ATT weights for the Lalonde dataset.

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 using **regress** along with Stata's built in weighting features (**svyset** followed by the **svy:** prefix). We start with an analysis using the weights derived from the GBM selected to minimize the mean standardized bias ("es.mean" stopping rule).We use **svyset** to declare that our weighting variable is **esmeanatt.** The propensity score adjusted results are then estimated using the **svy** prefix with **regress.**

```
. use lalonde_att_wgts, clear
. svyset [pweight=esmeanatt]
. svy: regress re78 treat
```

```
Survey: Linear regression

Number of strata   =         1                  Number of obs      =       614
Number of PSUs     =       614                  Population size    = 329.58393
                                                Design df          =       613
                                                F(   1,    613)    =      0.48
                                                Prob > F           =    0.4884
                                                R-squared          =    0.0027

------------------------------------------------------------------------------
             |             Linearized
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   732.5166   1056.595     0.69   0.488    -1342.468    2807.501
```

```
       _cons |   5616.627   884.9208     6.35   0.000     3878.783    7354.471
------------------------------------------------------------------------------
```

The analysis estimates an increase in earnings of $733 for those that participated in the NSWD 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 60.1% 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.

```
. svy: regress re78 treat nodegree

Survey: Linear regression

Number of strata   =         1                  Number of obs      =       614
Number of PSUs     =       614                  Population size    = 329.58393
                                                Design df          =       613
                                                F(   2,    612)    =      1.89
                                                Prob > F           =    0.1515
                                                R-squared          =    0.0185

------------------------------------------------------------------------------
             |             Linearized
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   920.3338   1082.816     0.85   0.396    -1206.145    3046.812
    nodegree |  -1891.804   1261.938    -1.50   0.134    -4370.049    586.4416
       _cons |   6768.411   1470.977     4.60   0.000     3879.645    9657.177
------------------------------------------------------------------------------
```

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.

```
. svy: regress re78 treat age educ black hispan nodegree married re74 re75

Survey: Linear regression

Number of strata   =         1                  Number of obs      =       614
Number of PSUs     =       614                  Population size    = 329.58393
                                                Design df          =       613
                                                F(   9,    605)    =      2.49
                                                Prob > F           =    0.0085
                                                R-squared          =    0.0558

------------------------------------------------------------------------------
             |             Linearized
```

```
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   758.4891   1018.502     0.74   0.457    -1241.688    2758.666
         age |   3.004667   55.58156     0.05   0.957    -106.1487     112.158
        educ |   748.8193   259.6324     2.88   0.004     238.9424    1258.696
       black |  -762.6813   1012.371    -0.75   0.452    -2750.817    1225.455
      hispan |   610.6271   1710.552     0.36   0.721    -2748.626     3969.88
    nodegree |   535.0055   1626.137     0.33   0.742    -2658.469     3728.48
     married |   491.7587   1072.037     0.46   0.647    -1613.552    2597.069
        re74 |   .0569882    .180081     0.32   0.752    -.2966623    .4106387
        re75 |    .156755   .1945705     0.81   0.421    -.2253507    .5388607
       _cons |  -2458.798   4289.423    -0.57   0.567    -10882.55    5964.949
------------------------------------------------------------------------------
```

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. **regress** is the standard procedure for fitting such models in Stata.

```
. regress re78 treat age educ black hispan nodegree married re74 re75

      Source |       SS       df       MS              Number of obs =     614
-------------+------------------------------           F(9, 604)     =   11.64
       Model |  5.0554e+09     9   561713775            Prob > F      =  0.0000
    Residual |  2.9157e+10   604  48273544.4            R-squared     =  0.1478
-------------+------------------------------           Adj R-squared =  0.1351
       Total |  3.4213e+10   613  55811818.5            Root MSE      =  6947.9

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1548.244   781.2793     1.98   0.048     13.88986    3082.598
         age |   12.97763   32.48891     0.40   0.690    -50.82731    76.78258
        educ |   403.9412   158.9062     2.54   0.011     91.86538    716.0171
       black |  -1240.644   768.7644    -1.61   0.107     -2750.42    269.1318
      hispan |   498.8968   941.9425     0.53   0.597    -1350.984    2348.777
    nodegree |   259.8174   847.4421     0.31   0.759    -1404.474    1924.108
     married |   406.6208   695.4723     0.58   0.559    -959.2168    1772.458
        re74 |   .2963774   .0582726     5.09   0.000     .1819358     .410819
        re75 |   .2315259   .1046199     2.21   0.027      .026063    .4369888
       _cons |   66.51451   2436.746     0.03   0.978    -4719.009    4852.038
------------------------------------------------------------------------------
```

This model estimates a rather strong treatment effect, estimating a program effect 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** command and compare the results to the weights from **twang**.

```
· logit treat age educ black hispan nodegree married re74 re75
· predict phat
```

logit fits the logistic regression model. The default is to model the probability that treat=1. The **predict** command used after running **logit** generates the predicted probabilities, which are then saved in a variable named **phat**. The dataset includes all the variables and appends the predicted probabilities. We can create the ATT weights as shown below with weights for the treatment group equal to one and weights for the control group equal to the odds of treatment.

```
· gen w_logit_att=treat+(1-treat)*phat/(1-phat)
```

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

```
· dxwts treat age educ black hispan nodegree married re74 re75, ///
     weightvars(w_logit_att) estimand(ATT) permtestiters(0) ///
     rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
     objpath(C:\Users\uname\twang\output)
· balance, summary unweighted weighted
```

```
Summary
------------------------------------------------------------------------------------------
             ntreat    nctrl esstreat  essctrl    maxes   meanes    maxks   meanks   iter
------------------------------------------------------------------------------------------
unw             185      429      185      429    1.757    .5687    .6404    .2702      .
w_logit_~t      185      429      185    99.82    .1188    .0319    .3078     .093      .
------------------------------------------------------------------------------------------


Unweighted
------------------------------------------------------------------------------------------
               txmn     txsd     ctmn     ctsd stdeffsz     stat        p       ks kspval
------------------------------------------------------------------------------------------
age           25.82    7.155    28.03    10.79    -.309   -2.994     .003     .158   .003
black          .843     .365     .203     .403    1.757    19.37        0    .64        0
educ          10.35    2.011    10.23    2.855     .055     .547     .584     .111   .074
hispan         .059     .237     .142      .35    -.349   -3.413     .001     .083   .317
married        .189     .393     .513       .5    -.824   -8.607        0     .324      0
nodegree       .708     .456     .597     .491     .244    2.716     .007     .111   .074
re74           2096     4887     5619     6789    -.721   -7.254        0     .447      0
re75           1532     3219     2466     3292     -.29   -3.282     .001     .288      0
------------------------------------------------------------------------------------------

Weighted: w_logit_att
------------------------------------------------------------------------------------------
               txmn     txsd     ctmn     ctsd stdeffsz     stat        p       ks kspval
```

```
-------------------------------------------------------------------------------
age          25.82    7.155    24.97    10.53     .119     .739      .46     .308       0
black         .843     .365     .845     .362    -.006    -.069     .945     .002       1
educ         10.35    2.011     10.4    2.459    -.028    -.219     .827     .036       1
hispan        .059     .237     .059     .236     .001     .008     .993        0       1
married       .189     .393     .171     .377     .047     .456     .649     .019       1
nodegree      .708     .456      .69     .463      .04     .332      .74     .018       1
re74          2096     4887     2106     4236    -.002    -.022     .983     .228    .002
re75          1532     3219     1497     2716     .011     .107     .915     .133    .185
-------------------------------------------------------------------------------
```

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

```
. svyset [pweight=w_logit_att]
. svy: regress re78 treat
```

The analysis estimates an increase in earnings of $1214 for those that participated in the NSWD 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 Stata package as **lindner.dta**. 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 survived 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 patient, 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:

```
· use " C:\Users\uname\twang\lindner.dta",clear
· tab sixmonthsurvive abcix,chi2 lrchi2 V cell row column
```

```
+--------------------+
| Key                |
|--------------------|
| frequency          |
| row percentage     |
| column percentage  |
| cell percentage    |
+--------------------+

sixMonthSu |          abcix
     rvive |      0            1 |     Total
-----------+----------------------+----------
     FALSE |      15         11   |        26
           |   57.69         42.31 |    100.00
           |    5.03         1.58  |      2.61
           |    1.51         1.10  |      2.61
-----------+----------------------+----------
       TRUE|     283         687   |       970
           |   29.18         70.82 |    100.00
           |   94.97         98.42 |     97.39
           |   28.41         68.98 |     97.39
-----------+----------------------+----------
     Total |     298         698   |       996
           |   29.92         70.08 |    100.00
           |  100.00        100.00 |    100.00
           |   29.92         70.08 |    100.00
```

```
Pearson chi2(1) = 9.8207              Pr = 0.002
likelihood-ratio chi2(1) = 8.8530     Pr = 0.003
Cramér'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 abcix stent height female diabetic acutemi ejecfrac ves1proc, ///
     stopmethod(es.mean ks.mean) estimand(ATE) ///
     rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
     objpath(C:\Users\uname\twang\output) ///
     plotname(C:\Users\uname\twang\output\abcix_twang.pdf)
```

We set **estimand** = "**ATE**" because we are interested in the effects of **abciximab** on everyone in the population. We specify the stopping rules to be **es.mean** and **ks.mean**. We then inspect pre- and post-weighting balance using the balance table.

```
.balance, unweighted weighted
Unweighted
------------------------------------------------------------------------------
             txmn     txsd     ctmn     ctsd stdeffsz     stat      p       ks   kspval
------------------------------------------------------------------------------
--
acutemi       .179     .384      .06     .239     .338    5.923      0     .119     .005
diabetic      .205     .404     .268     .444    -.152   -2.127   .034     .064     .349
ejecfrac      50.4    10.42    52.29     10.3    -.181    -
2.64     .008     .114     .008
female        .331     .471   .386     .488    -.115   -1.647       .1     .055     .531
height       171.4    10.69    171.4    10.59        0    -.005   .996     .025     .999
stent         .705     .456     .584     .494     .257    3.624      0     .121     .004
ves1proc     1.463     .706    1.205      .48     .393    6.693      0     .188        0
------------------------------------------------------------------------------

Weighted: esmean
------------------------------------------------------------------------------
             txmn     txsd     ctmn     ctsd stdeffsz     stat      p       ks   kspval
------------------------------------------------------------------------------
acutemi       .148     .355      .11     .313     .114    1.303   .193     .038      .95
diabetic      .214     .411     .237     .426    -.055    -.719   .473     .023        1
ejecfrac     51.01    10.39    51.39    9.396    -.038    -.532   .595     .027     .999
female        .337     .473      .35     .478    -.028     -.38    .704     .013
1
height       171.5    10.53    171.5    11.04    -.003    -.039   .969     .018        1
stent         .683     .466     .653     .477     .063     .849   .396      .03     .996
ves1proc     1.398      .67     1.35     .589     .076     .951   .342     .023        1
------------------------------------------------------------------------------

Weighted: ksmean
------------------------------------------------------------------------------
             txmn     txsd     ctmn     ctsd stdeffsz     stat      p       ks   kspval
------------------------------------------------------------------------------
acutemi       .148     .355     .107      .31      .12    1.331   .183      .04     .935
diabetic      .215     .411     .229     .421    -.033    -.432   .666     .014        1
ejecfrac     51.05    10.33     51.6     9.11    -.056    -.798   .425     .027     .999
female        .338     .473     .345     .476    -.015      -.2   .841     .007        1
height       171.5    10.55    171.6    10.59    -.011    -.155   .877     .015        1
stent         .683     .466     .657     .475     .054     .718   .473     .025        1
ves1proc     1.395     .666    1.337     .573     .094    1.203   .229     .028     .999
------------------------------------------------------------------------------
```

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.

Plot 1 (optimize): GBM Optimization

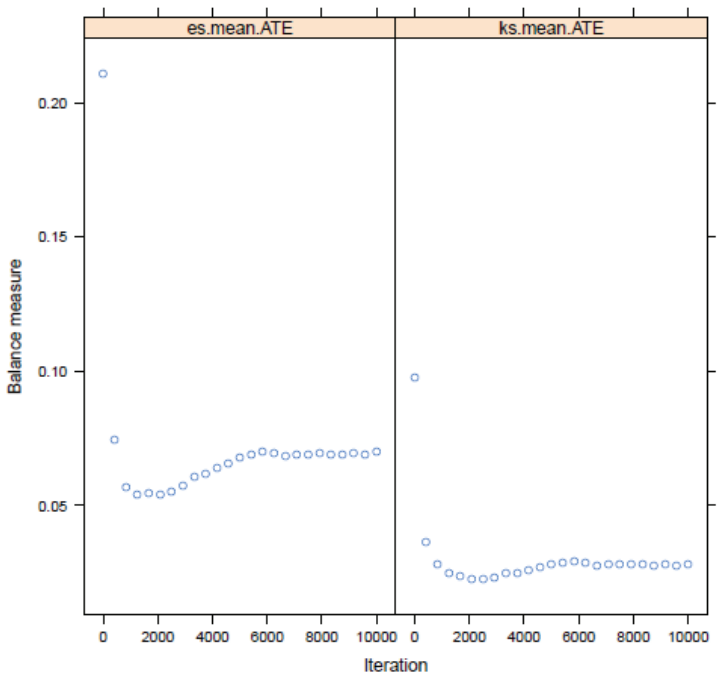

Figure 7: Example of the default diagnostic optimization plot from the specifying the **`plotname`** option of the **`ps`** command for estimating ATE weights for the **`Lindner`** dataset.

Plot 2 (boxplot): Boxplot of Propensity Scores

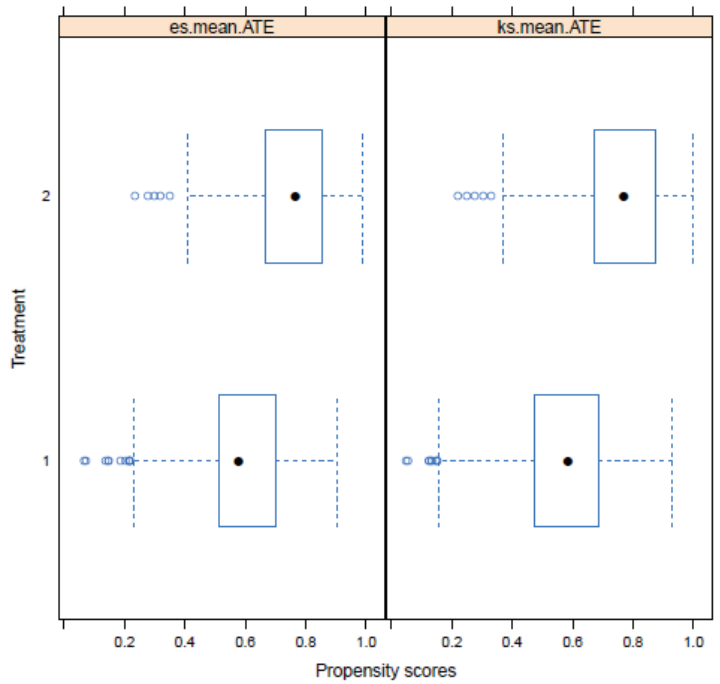

Figure 8: Example of the default diagnostic boxplot of propensity scores from the specifying the **`plotname`** option of the **`ps`** command for estimating ATE weights for the **`Lindner`** dataset.

Plot 3 (es): Standardized Effect Sizes Pre/Post Weighting

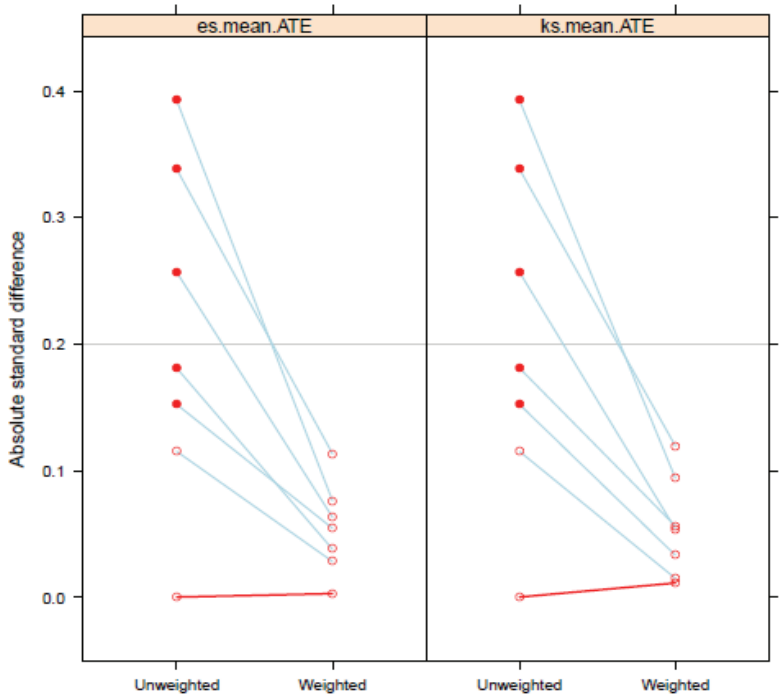

Figure 9: Example of the default diagnostic standardized effect size from the specifying the **`plotname`** option of the **`ps`** command for estimating ATE weights for the **`Lindner`** dataset.

Plot 4 (t): T−test P−values of Group Means of Covariates

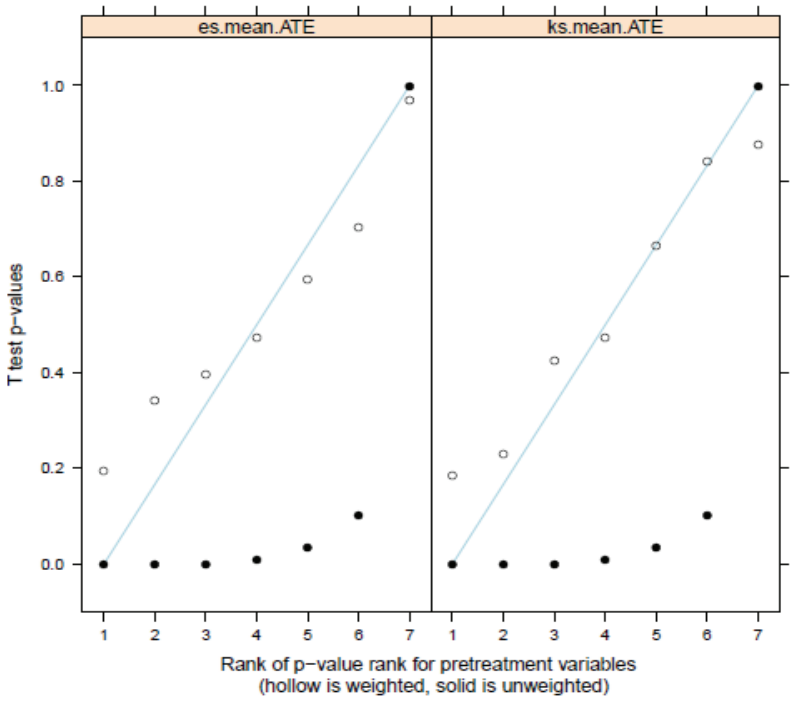

Figure 10: Example of the default diagnostic t-test p-value plot from the specifying the **`plotname`** option of the **`ps`** command for estimating ATE weights for the **`Lindner`** dataset.

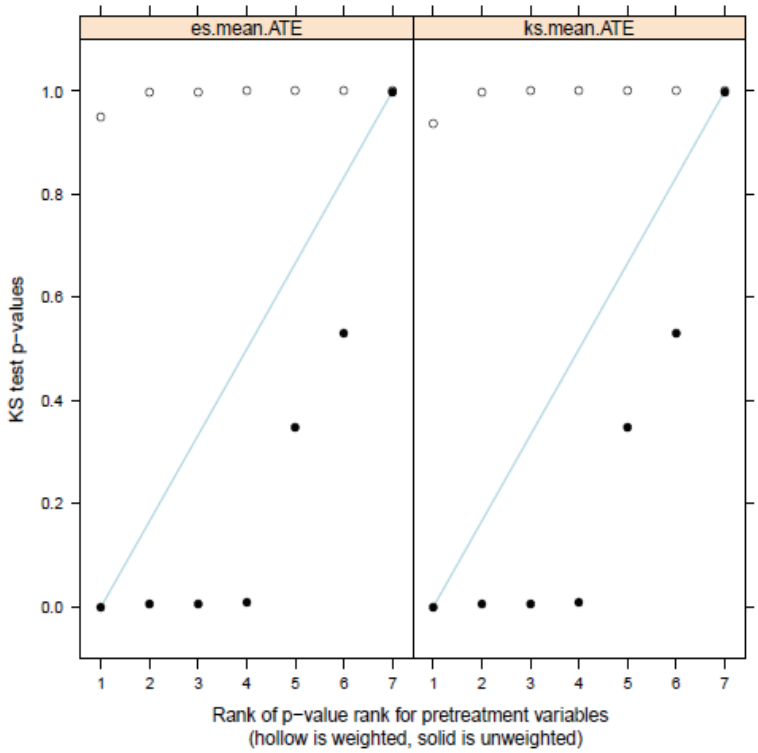

Figure 11: Example of the default diagnostic Kolmogorov-Smirnov test p-value plot from the specifying the **plotname** option of the **ps** command for estimating ATE weights for the **Lindner** dataset.

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 **esstreat** is no longer equal to **ntreat** since we are focusing on ATE rather than ATT.

```
.balance, summary
```

```
Summary
------------------------------------------------------------------------------
        ntreat  nctrl esstreat  essctrl  maxes  meanes  maxks  maxksp  meanks  iter
------------------------------------------------------------------------------
unw        698    298       698      298  .3926   .2053  .1884       .   .0979     .
es_mean    698    298     663.8    237.1  .1135   .0539  .0382       .   .024   1303
ks_mean    698    298     655.7    228.9  .1197    .055  .0401       .   .0224  2603
------------------------------------------------------------------------------
```

Because the outcome is dichotomous, we use **svyset** with **svy:tab** to obtain a weighted proportions and test for differences. The **obs**, **count**, **cell** and **col** options request number of observations, weighted counts, cell proportions and column proportions, respectively. The **se** option requests the corresponding standard errors for weighted counts, cell proportions and column proportions.

- **svyset[pweight=esmeanate]**
- **svy: tab sixmonthsurvive abcix, obs count se**
- **svy: tab sixmonthsurvive abcix, cell se**
- **svy: tab sixmonthsurvive abcix, col se**

```
Number of strata =   1                    Number of obs   = 996
Number of PSUs   = 996                    Population size = 1838.7643
                                          Design df       = 995

----------------------------------------
 sixMonthS |       abcix
    urvive |     0        1        Total
-----------+----------------------------
     FALSE |   47.93    14.25      62.18
           | (14.02)   (4.34)    (14.63)
           |      15       11         26
           |
      TRUE |   823.9     952.7      1777
           | (48.24)   (21.88)   (35.05)
           |     283       687        970
           |
     Total |  871.9       966.9     1839
           | (49.44)    (21.68)
           |     298        698      996
----------------------------------------
  Key:  weighted counts
        (linearized standard errors of weighted counts)
        number of observations

  Pearson:
  Uncorrected chi2(1)    =   12.3081
  Design-based F(1, 995) =   11.3780   P = 0.0008

 ----------------------------------------
  sixMonthS |        abcix
    urvive  |     0          1       Total
 -----------+----------------------------
      FALSE |   .0261      .0077     .0338
            | (.0075)    (.0024)   (.0079)
            |
       TRUE |   .4481     .5181      .9662
            | (.0191)   (.0188)    (.0079)
            |
      Total |   .4742     .5258        1
            | (.0189)    (.0189)
  ----------------------------------------
    Key:  cell proportions
          (linearized standard errors of cell proportions)

 ----------------------------------------
 sixMonthS |        abcix
    urvive |    0          1        Total
 ----------+----------------------------
     FALSE |    .055     .0147      .0338
           | (.0157)   (.0045)    (.0079)
           |
      TRUE |    .945     .9853      .9662
           | (.0157)   (.0045)    (.0079)
           |
     Total |       1         1          1
           |
  ----------------------------------------
    Key:  column proportions
          (linearized standard errors of column proportions)
```

The reweighting does not diminish the association between the treatment and the outcome. Indeed, it is marginally more significant after the reweighting. Alternatively, one could use logistic regression to assess the relationship between the dichotomous outcome and dichotomous treatment.

4 Non-response weights

The **twang** commands 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** Stata package as **egsingle.dta**. The data include mathematics test scores for 1,721 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** command.

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.

```
· use " C:\Users\uname\twang\egsingle.dta",clear

· bysort childid: egen sum=sum(grade)
· gen resp=1 if sum==10|sum==15
· replace resp=0 if resp !=1

· duplicates drop childid,force
· tab1 resp race

-> tabulation of resp

  resp |  Freq.   Percent     Cum.
-------+-----------------------
     0 |   878      51.02    51.02
```

```
     1 |   843      48.98   100.00
-------+-----------------------
 Total | 1,721     100.00

-> tabulation of race

  race |  Freq.   Percent      Cum.
-------+-----------------------
 black| 1,195      69.44     69.44
  hisp|   250      14.53     83.96
 other|   276      16.04    100.00
-------+-----------------------
 Total | 1,721     100.00
```

There are 1,721 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 resp i.race female size lowinc mobility, ///
     stopmethod(es.mean ks.max) estimand(ATE) ntrees(5000) ///
     rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
     objpath(C:\Users\uname\twang\output) ///
     plotname(C:\Users\uname\twang\output\egsingle_twang.pdf)

.balance, unweighted weighted
.save inputds, replace

Unweighted
------------------------------------------------------------------------------
--
                txmn     txsd     ctmn     ctsd stdeffsz     stat      p       ks   kspval
------------------------------------------------------------------------------

_
female          .466     .499     .517       .5    -.102   -2.113   .035     .051     .206
lowinc         75.49    28.58    80.75    24.08    -.198   -4.116      0       .1        0
mobility       32.66    14.04    36.44     13.7     -.27   -5.642      0     .122        0
size           750.2    316.4    761.3    312.4    -.035    -.733   .464     .066     .043
------------------------------------------------------------------------------
race
black           .656     .475     .731     .443    -.158     18.4      0     .075        0
hispanic        .129     .336     .161     .367    -.093        .      .     .031        0
other           .215     .411     .108     .311     .259        .      .     .107        0
------------------------------------------------------------------------------


Weighted: esmean
------------------------------------------------------------------------------
                txmn     txsd     ctmn     ctsd stdeffsz     stat      p       ks   kspval
------------------------------------------------------------------------------

_
female          .487       .5      .49       .5    -.007    -.138    .89     .004        1
lowinc         78.51    27.19    78.48    26.45     .001     .024   .981     .026     .954
mobility       34.23    13.69    34.78    13.97     -.04    -.779   .436     .021     .994
size           756.8    312.8    758.7    314.2    -.006    -.119   .905     .022     .987
------------------------------------------------------------------------------
race
black           .689     .463     .705     .456    -.033     .325   .721     .015     .721
hispanic        .142      .35     .142     .349     .001        .      .        0     .721
other           .168     .374     .153      .36      .04        .      .     .015     .721
------------------------------------------------------------------------------

Weighted: ksmax
------------------------------------------------------------------------------
```

	txmn	txsd	ctmn	ctsd	stdeffsz	stat	p	ks	kspval
female	.486	.5	.49	.5	-.009	-.172	.864	.004	1
lowinc	78.51	27.21	78.51	26.42	0	-.005	.996	.025	.966
mobility	34.22	13.7	34.8	13.96	-.042	-.817	.414	.021	.994
size	756.5	312.9	758.9	314.2	-.008	-.151	.88	.023	.984
race									
black	.689	.463	.705	.456	-.034	.336	.713	.016	.713
hispanic	.142	.349	.141	.348	.002	.	.	.001	.713
other	.169	.374	.153	.36	.041	.	.	.015	.713

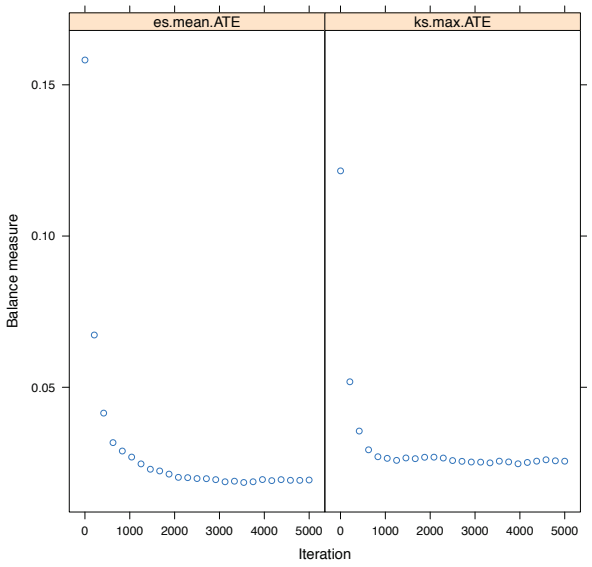

Figure 12: Optimization of es.mean.ATE and ks.max.ATE for nonresponse weighting of **egsingle** data. The horizontal axes indicate the number of iterations and the vertical axes indicate the measure of imbalance between the two groups. For es.mean.ATE the measure is the average effect size difference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

By default the balance table generated by **ps** and **balance** 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** command 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.

```
. use _inputds, clear
```

```
· keep childid ksmaxate resp
· sort childid
· save _inputds, replace

· use " C:\Users\uname\twang\egsingle.dta",clear
· duplicates drop childid, force
· sort childid
· save egsingle, replace

· merge childid using _inputds
· rename ksmaxate wgt
· keep if resp==1
· duplicates drop childid,force
· save egsingle_nrespwt, replace
```

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.

```
· use egsingle_nrespwt, clear
· append using egsingle, gen(nr2)
· gen wgt2=1 if nr2==1
· replace wgt2=wgt if nr2==0
```

We now run **dxwts** to obtain the balance statistics.

```
· dxwts nr2 i.race female size lowinc mobility, ///
weightvars(wgt2) estimand(ATT) permtestiters(0) ///
rcmd(C:\Program Files\R\R-3.0.3\bin\Rscript.exe) ///
objpath(C:\Users\uname\twang\output)

.balance, weighted
Weighted: wgt2
```

	txmn	txsd	ctmn	ctsd	stdeffsz	stat	p	ks	kspval
_									
female	.492	.5	.486	.5	.013	.306	.76	.007	1
lowinc	78.17	26.5	78.51	27.21	-.013	-.297	.766	.032	.631
mobility	34.59	13.99	34.22	13.7	.026	.625	.532	.019	.991
size	755.9	314.3	756.5	312.9	-.002	-.049	.961	.018	.991
_race									
black	.694	.461	.689	.463	.011	.148	.861	.005	.861
hispanic	.145	.352	.142	.349	.01	.	.	.003	.861
other	.16	.367	.169	.374	-.023	.	.	.008	.861

The resulting balance table includes a table for an unweighted comparison of 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.

```
· use " C:\Users\uname\twang\egsingle.dta",clear
· sort childid
· save egsingle, replace

· use egsingle_nrespwt, clear
· sort childid
· keep childid wgt
· merge childid using egsingle
· keep if _merge==3
· sort childid grade
· save egsingle_analysis, replace
```

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(x|t = 1) be the distribution of features for the treatment cases and f(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(x), so that

$$f(x|t = 1) = w(x)f(x|t = 0) \qquad (2)$$

For example, if f(age=65, sex=F|t = 1) = 0.10 and f(age=65; 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(x) and apply Bayes Theorem to the numerator and the denominator to give an expression for the propensity score weight for comparison cases,

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

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 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|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(x|t = 1) = w(x)f(x); and (4)

f(x|t = 0) = w(x)f(x); and (5)

Again using Bayes Theorem we obtain w(x) ∝1/f(t = 1|x) for the treatment group and w(x) ∝ 1/f(t =0|x) for the control group.

5.2 Estimating the propensity score

In randomized studies P(t = 1|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 x as

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

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

$$\ell\beta = \frac{1}{n}\sum_{i=1}^{n} t_i\beta' x_i - \log\left(1 + \exp\left(\beta' x_i\right)\right) \qquad (7)$$

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 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

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

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 λ = 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 x we will include indicator functions for continuous variables like ***I***(age <15); ***I***(age < 16),…, ***I***(age < 90), etc., for categorical variables like ***I***(sex = male), ***I***(prior MI = TRUE), and interactions among them like ***I***(age < 16)***I***(sex = male)***I***(prior MI = TRUE). This collection of basis functions spans a plausible set of propensity score functions, are computationally efficient,

and are at the extremes of x reducing the likelihood of propensity score estimates near 0 and 1 that can occur with linear basis functions of 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** commands provide 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 $w_i = p(x_i)/(1 - p(x_i))$. To estimate the ATE, we give control cases weight $w_i = 1/p(x_i)$ and we give the treatment cases $w_i = 1/(1-p(x_i))$. 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 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://biostats.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.rproject.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. https://cran.r-project.org/web/packages/gbm/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. https://cran.r-project.org/web/packages/twang/index.html

[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.