Toolkit for Weighting and Analysis of Nonequivalent Groups

A Tutorial on the TWANG Shiny App for Two Treatments

BETH ANN GRIFFIN, RICARDO SANCHEZ, MATTHEW CEFALU, LYNSAY AYER, DONNA L. COFFMAN, BRIAN GARRETT VEGETABLE, JOSEPH D. PANE, LANE F. BURGETTE, DANIEL F. MCCAFFREY

Prepared for the National Institute on Drug Abuse
This tutorial describes how to use a menu-driven Shiny application (app) based on the Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG) R package. The TWANG package was first developed in 2004 by RAND Corporation researchers for the R statistical computing language and environment (RAND Corporation, undated a). The Shiny software development package allowed the TWANG project team to develop a menu-driven app that can be used to perform analyses using the TWANG package’s suite of commands without requiring a user to learn R. This tutorial provides an introduction on how to use the TWANG Shiny app to estimate propensity score weights and related treatment effects for comparing two treatment groups when using observational data. The tutorial demonstrates use of the Shiny app through an illustrative example and explains key inputs and outputs of the TWANG Shiny app.

Development of the TWANG Shiny app and this associated tutorial was supported by funding from grant R01DA045049 from the National Institute on Drug Abuse. The overarching goals of the grant are to develop statistical methods and tools that will provide addiction health services researchers (and beyond) with the tools and training they need for studying the effectiveness of treatments using observational data. The proposed project will conduct an extensive series of outreach and educational training efforts to ensure the tools and methods are being used appropriately. In doing so, the proposed research aims to improve the statistical practices of addiction researchers and consequently greatly strengthen the scientific information upon which decisions are made to improve care in our country. The intended audience for this tool includes both quantitative researchers (like statisticians and analysts) as well as substantive researchers with an interest in learning more about statistical methods. The work forms an integral part of a rich body of research done at RAND to improve causal inference and addiction science (RAND Corporation, undated b).

RAND Social and Economic Well-Being is a division of the RAND Corporation that seeks to actively improve the health and social and economic well-being of populations and communities throughout the world. This research was conducted in the Justice Policy Program within RAND Social and Economic Well-Being. The program focuses on such topics as access to justice, policing, corrections, drug policy, and court system reform, as well as other policy concerns pertaining to public safety and criminal and civil justice. For more information, email justicepolicy@rand.org.
Contents

Preface ........................................................................................................................................... iii
Figures ............................................................................................................................................. v
Acknowledgments .......................................................................................................................... vi
Abbreviations ............................................................................................................................... vii
1. Introduction ................................................................................................................................. 1
2. Setting Up .................................................................................................................................... 2
   Launching the Stand-Alone Version of the TWANG Shiny App ..................................................... 2
   Launching the TWANG Shiny App from R ....................................................................................... 3
   Reading in the Data Set .................................................................................................................... 5
3. Estimating Propensity Scores ...................................................................................................... 8
4. Evaluating the Quality of the Propensity Score Weights .......................................................... 18
   Assessing “Balance” Using Graphics ........................................................................................... 20
   Assessing “Balance” Using Balance Tables ................................................................................... 24
5. Analysis of Outcomes ................................................................................................................ 31
6. Data Outputs .............................................................................................................................. 35
References ..................................................................................................................................... 36
Figures

Figure 2.1. Stand-Alone TWANG Shiny App Launch Screen ..........................................................3
Figure 2.2. R TWANG Shiny App Launch Screen ......................................................................5
Figure 2.3. Uploading the Data to the App ...............................................................................6
Figure 2.4. Successfully Uploaded Data ....................................................................................7
Figure 3.1. Propensity Score Model ..........................................................................................8
Figure 3.2. Computing the KS Statistic ....................................................................................13
Figure 3.3. Propensity Score Model—Example Settings ............................................................15
Figure 3.4. Propensity Score Model—Calculating Window ........................................................16
Figure 3.5. Propensity Score Model—Summary Table ...............................................................17
Figure 4.1. Model Evaluation—Convergence Plot ....................................................................18
Figure 4.2. Model Evaluation—Propensity Score .......................................................................20
Figure 4.3. Model Evaluation—Balance Plot .............................................................................21
Figure 4.4. Model Evaluation—ES P-Values .............................................................................23
Figure 4.5. Model Evaluation—KS P-Values .............................................................................24
Figure 4.6. Model Evaluation—Balance Table (es.max) ............................................................25
Figure 4.7. Model Evaluation—Balance Table (ks.max) ............................................................26
Figure 4.8. Propensity Score Model—Summary Table ...............................................................28
Figure 4.9. Model Evaluation—Relative Influence ....................................................................30
Figure 5.1. Treatment Effect Estimation ....................................................................................31
Figure 5.2. Treatment Effect Estimation—Results ....................................................................32
Figure 5.3. Treatment Effect Estimation—Adjusted Results .......................................................33
Figure 5.4. Treatment Effect Estimation—Doubly Robust ........................................................34
Figure 6.1. Saving Propensity Score Weights ...........................................................................35
Acknowledgments

Development of the TWANG Shiny app and associated tutorial was supported by funding from grant R01DA045049 from the National Institute on Drug Abuse. We wish to acknowledge our TWANG Shiny app beta testers (Sarah Hunter, Bryan Garner, Marika Suttorp Booth, Heather M. Buzick, and Hyunsuk Lee) and our RAND quality assurance reviewers, Geoff Grimm and Sarah Meadows, for their constructive feedback on the Shiny app and its tutorial. We also thank Bergen Goesch for her extensive formatting of this tutorial.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>app</td>
<td>application</td>
</tr>
<tr>
<td>ATE</td>
<td>average treatment effect on the population</td>
</tr>
<tr>
<td>ATT</td>
<td>average treatment effect on the treated</td>
</tr>
<tr>
<td>CPS</td>
<td>Current Population Survey</td>
</tr>
<tr>
<td>CSV</td>
<td>comma-separated value</td>
</tr>
<tr>
<td>ECDF</td>
<td>empirical cumulative distribution function</td>
</tr>
<tr>
<td>ES</td>
<td>effect size</td>
</tr>
<tr>
<td>ESS</td>
<td>effective sample size</td>
</tr>
<tr>
<td>GBM</td>
<td>generalized boosted model</td>
</tr>
<tr>
<td>KS</td>
<td>Kolmogorov-Smirnov</td>
</tr>
<tr>
<td>NSWD</td>
<td>National Supported Work Demonstration</td>
</tr>
<tr>
<td>QQ</td>
<td>quantile-quantile</td>
</tr>
<tr>
<td>SB</td>
<td>standardized bias</td>
</tr>
<tr>
<td>TWANG</td>
<td>Toolkit for Weighting and Analysis of Nonequivalent Groups</td>
</tr>
</tbody>
</table>
1. Introduction

The Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG) was first developed in 2004 by RAND Corporation researchers for the R statistical computing language and environment. The R version of the package contains functions for creating high-quality propensity score weights that can be used to estimate treatment effects with two or more treatment groups and time-varying treatments (RAND Corporation, undated a; Ridgeway et al., 2017). The Shiny software development package allowed the TWANG project team to develop the TWANG Shiny app, a menu-driven application (app) that can be used to perform analyses using the TWANG package’s suite of commands. This tutorial provides an introduction on how to use the TWANG Shiny app to estimate propensity score weights for two treatment groups when using observational data and to estimate treatment effect estimates using those propensity score weights. The tutorial demonstrates use of the app through illustrative examples.

The foundational method of this app is the propensity score. The propensity score is the probability that an individual or unit would be assigned or exposed to a treatment condition conditional on a set of observed covariates. The ability of the propensity score to serve as a balancing score (Rosenbaum and Rubin, 1983) implies that, conditional on the propensity score, the distribution of the observed covariates should be the same for the treatment and control groups. Thus, one can use propensity score weights to balance, or make the treatment groups under consideration comparable to the distributions of the observed covariates used in the model of the propensity score.

The TWANG Shiny app can be used to

- compute the needed propensity score weights for an analysis
- check the quality of the resulting propensity score weights by assessing whether they have good balancing properties
- use the propensity score weights to estimate causal treatment effect estimates, assuming that key assumptions of methods hold (e.g., no unobserved confounding or overlap concerns exist in the data).
2. Setting Up

There are two ways to launch the TWANG Shiny app. The first is to download and run the stand-alone version of the app. The second is to download the source code and run it from R.

The stand-alone version of the app is a single, binary, executable file that contains a copy of R and all the packages required to launch the TWANG Shiny app. On the website, we provide executables for both Mac and PC, in addition to several sample data sets available for download (including the one used in this tutorial), as follows:

- stand-alone version of the TWANG Shiny app
  - Windows (PC): TWANG-Shiny-1.0-win.exe
  - OS X (Mac): TWANG-Shiny-1.0

- example data sets, including lalonde_racecats.csv from the LaLonde (1986) study, with race derived as a categorical variable. (This data set is useful for trying the menu-driven commands presented in this tutorial, but it is not necessary for you to run your own applications.)

The files are available on the downloads page: www.rand.org/statistics/twang.

To download the TWANG Shiny app source code, please visit RAND’s Corporate GitHub Account (GitHub, undated).

Launching the Stand-Alone Version of the TWANG Shiny App

If you have not already done so, please download the TWANG Shiny app executable from the TWANG website. Once the download is complete, double-click the executable to launch the Shiny TWANG app. If successful, you will see the “Introduction” tab to the TWANG Shiny app (Figure 2.1).
Launching the TWANG Shiny App from R

These instructions assume you have downloaded and installed R on your computer. Next, we need to make sure the following packages are installed:

- shiny
- shinyjs
- bsplus
- tidyverse
- data.table
- DT
- haven
- knitr
- kableExtra
- margins
- readxl
- twang.
To install these packages, run the following command in your R terminal.

```r
install.packages(
  c(
    "shiny",
    "shinyjs",
    "bsplus",
    "tidyverse",
    "data.table",
    "DT",
    "haven",
    "knitr",
    "kableExtra",
    "margins",
    "readxl",
    "twang"
  )
)
```

The app can now be launched with the command:

```r
shiny::runUrl('https://github.com/RANDCorporation/TWANG-PS/archive/master.zip')
```

This command will download the source code from the RAND Corporate GitHub account (GitHub, undated) and launch the TWANG Shiny app. If successful, you will see another introductory screen (Figure 2.2).
Here, we walk through an example. From the “Introduction” tab, click next and go to the “File Upload” tab (Figure 2.3).
This is where you will read in your data set. For reading in the LaLonde data set, which is saved as a comma-separated-value (CSV) file delimited by commas, select “CSV/Text,” comma, and double-quote options, then browse to where you saved the data and click on lalonde_racecats.csv to upload your data to the app.

After successfully uploading the data, a subset of the data is shown on the right-hand side of the app (Figure 2.4).
Figure 2.4. Successfully Uploaded Data
3. Estimating Propensity Scores

Once the data are read into the app, click on the “Propensity Score Model” tab to estimate the propensity score and the associated propensity score weights (Figure 3.1).

Figure 3.1. Propensity Score Model

To demonstrate the usage of the TWANG Shiny app, we use data from LaLonde's National Supported Work Demonstration (NSWD) analysis (Dehejia and Wahba, 1999; LaLonde, 1986). 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 average causal effect of the NSWD on earnings on the entire sample (or the average treatment on the population [ATE]). Pretreatment covariates include age, education, race, ethnicity, education level, marital status, earnings in 1974, and earnings in 1975. 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.

In the LaLonde data set, the variable “treat" is the 0/1 treatment indicator, where 0 indicates comparison cases drawn from the CPS and 1 indicates treatment as part of the NSWD. 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 tab in Figure 3.1 implements the primary method in TWANG for estimating propensity scores. The TWANG package estimates propensity score weights for binary treatments through the use of a generalized boosted model (GBM), a flexible, nonparametric estimation technique that can regress the treatment indicator (here, whether someone was a part of the work demonstration program) onto a large number of confounding (baseline) covariates (Ridgeway, 2001). GBM was first proposed for the estimation of propensity score weights by McCaffrey, Ridgeway, and Morral (2004) as a way to circumvent the long and tedious process of fitting and refitting logistic regression models to obtain high-quality propensity score weights (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999). Burgette, McCaffrey, and Griffin (2015) provide a brief overview of GBM, and we refer interested readers to that document for more extensive details. In brief, GBM adaptively captures the functional form of the relationship between the covariates and the treatment indicator using what is known as a “forward stage-wise additive algorithm” or a regression tree approach whereby, at each step of the algorithm, a new simple regression tree is added to the model from the previous steps as a way to improve the model’s predictive ability. Thus, when the algorithm has run only a few iterations, the model will be quite simple; as the algorithm runs out further (e.g., thousands of iterations), it will become increasingly complex.

GBM has been shown in the literature to estimate causal effects with less bias than traditional regression approaches, such as a logistic regression model (Harder, Stuart, and Anthony, 2010; Lee, Lessler, and Stuart, 2010; McCaffrey, Ridgeway, and Morral, 2004). GBM, via its use of regression trees, provides several notable advantages to analysts interested in estimating propensity score weights. First, it can handle all different types of covariates (continuous, discrete, ordinal, etc.). Second, it will automatically accommodate nonlinear and interactive effects, meaning the analyst or researcher does not need to know a priori whether there should be higher-order terms in the model. Finally, it can include variables with missing values. When there are missing values for a covariate, the algorithm will create a binary indicator for missingness and use the binary indicator as a control covariate in its estimation.
Now, let’s turn to fitting the GBM for the LaLonde data set. Please note this step is computationally intensive and can take a few minutes. Prior to running our GBM, here are instructions on each box (similar details can be found by clicking on the blue information icon to the right of each option as shown in Figure 3.1):

- **Treatment**: This is where you select the treatment variable. It is important to note that the format of the binary treatment variable should be 0/1 so that 0 denotes the control group and 1 denotes the treatment group. In the LaLonde example, the treatment variable is called “treat.”

- **Outcomes**: This is where you select the primary outcome variables; you can select as many outcomes as you’d like to analyze; for illustration, we use “re78” which represents the wages earned in 1978 after the work demonstration program.

- **Numerical Covariates**: This is where you select the observed covariates that will go into the propensity score model. These are the factors on which you are aiming to balance the treatment and control groups. This box is specifically for binary, continuous, and ordinal covariates; categorical covariates, if included, should be inputted into the next box. For the LaLonde example, we select the following variables: age, educ, married, nnodegree, re74, and re75.

- **Categorical Covariates**: This is where you select the observed categorical covariates that will go into the propensity score model (e.g., any grouped variable in the data that contains more than two groups, such as race or education in many applications); this option is needed to allow the Shiny app to tell TWANG to treat these covariates more carefully (specifically, as factor variables with more than two categories); for the LaLonde example, we have only race as a categorical covariate (black = 1, Hispanic = 2, and other = 3).

- **GBM Iterations**: This option sets the maximum number of iterations that the GBM will run. As noted above, with each iteration of the algorithm, the GBM becomes more complex, fitting additional features of the data. With too few iterations, it will miss salient features of the data. With too many, it will overfit the data. This is where you must select the number of iterations GBM should be allowed to run through. We suggest that people start with 5,000 iterations and increase if the package indicates the need to do so. There will be a warning if the estimated optimal number of iterations is not large enough (e.g., because the GBM algorithm has not yet converged to the optimal iteration). This indicates that balance could improve if more-complex models are considered. Increase GBM iterations and rerun it if this warning appears. For our example, we set this equal to 5,000.

- **Interaction Depth**: This option controls the level of interactions allowed between covariates used in the propensity score model. When the value is set equal to 1, no interactions are allowed in the GBM and thus regression trees will only split on the
values of the covariates themselves. This is similar to including only main effects for the control covariates into a logistic regression model. If we set interaction depth equal to 2, then all possible two-way interactions are allowed into the model for the treatment indicator (i.e., the propensity score model). This means the GBM will be able to include regression trees that are defined based on interaction terms (e.g., age X gender). The maximum value for interaction depth is 3. We specified a value of 2 here and generally think this is a reasonable place for you to start. If balance is not easily obtained, it might be worth changing this to 3 to allow for all possible three-way interactions among the control covariates being used in the propensity score model.

- **Shrinkage**: This option controls the amount of shrinkage. The GBM estimation algorithm uses shrinkage to enhance the smoothness of the resulting model. Small values, such as 0.005 or 0.001, yield smooth fits but require greater values of GBM iterations to achieve adequate fits. Computational time increases inversely with shrinkage. Additional details on the number of GBM iterations, interaction depth, and shrinkage can be found in McCaffrey, Ridgeway, and Morral (2004). For this example, we set this equal to 0.01.

- **Estimand**: The estimand option is used to indicate whether the analyst is interested in estimating the ATE or the average treatment effect on the treated (ATT). 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, 2010). With ATT weights, one is weighting the group of control individuals (here, those who did not participate in the work demonstration program) to look like the treatment group (here, those that did participate). When there is a highly select set of groups like these, ATT weights tend to be more successful than ATE weights because it can be easier to weight the control group individuals to look like treated individuals than to make the treated individuals look like the overall population. It is important for a study to carefully consider what is of most interest in a given application. For this illustrative example, we will focus only on ATE.

- **Stop Method**: One important component of the GBM is that you must determine which iteration should be used to produce the propensity score weights. Historically, the optimal iteration for GBM was selected as that which yields best model fit. In estimating propensity score weights, we use covariate balance to choose the iteration. This optimization process is directly incorporated by way of the *stop method*, which 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 optimal number of iterations used in the GBM for estimating the final propensity score weights will minimize the differences between the treatment and control groups as measured by the balance statistics specified by values given to the *stop method* option.
The ability of the propensity score to serve as a balancing score (Rosenbaum and Rubin, 1983) implies that, conditional on it, the distribution of the covariates should be the same for the treatment and control groups. Thus, because we are using propensity score weights to balance our groups, weighting by the propensity score should then balance the distributions of covariates between our groups of interest. Although this is true in theory, obtaining high-quality weights can be difficult in practice. Thus, an important step in any propensity score-weighted analysis is to assess the overall ability of the propensity score weights to create distributions of the covariates that are the same between the treatment and control groups. The TWANG package uses two balance measures to assess the performance of the propensity score weights for each candidate moderator: absolute effect size (ES) differences (also called the absolute standardized mean difference or standardized bias [SB]) and the Kolmogorov-Smirnov (KS) statistic (McCaffrey, Ridgeway, and Morral, 2004).

**ES difference:** For each covariate, the ES difference or SB equals the absolute value of the difference between the weighted mean for the treated group and the weighted mean for the control group divided by the unweighted standard deviation of the pooled sample for ATE or divided by the unweighted standard deviation of the treatment group for ATT. More specifically, for ATE, for covariate $k = 1, \ldots, K$,

$$ES_k = \frac{|\bar{x}_{k1} - \bar{x}_{k0}|}{s_k},$$

where $\bar{x}_{kz}$ is the weighted mean of the covariate for treatment ($z = 1$) or control ($z = 0$) and $s_k$ is the standard deviation of the covariate for the pooled sample. The weights are $1/\hat{p}(x)$ for members of the treatment group and $1/(1 - \hat{p}(x))$ for members of the control group, where $\hat{p}(x)$ is the GBM estimate of the propensity score. For ATT,

$$ES_k = \frac{|\bar{x}_{k1} - \bar{x}_{k0}|}{s_{k1}},$$

where the weights equal 1 for members of the treatment group and $\hat{p}(x)/(1 - \hat{p}(x))$ for members of the control group and $s_{k1}$ is the standard deviation of the covariate for the exposed group. Thus, an ES tells us how large the differences are between the two groups on the mean of a given covariate. An ES of 0 implies no difference, while an ES equal to 1 implies a difference of 1 standard deviation. The ES is also commonly called *Cohen’s d* (Cohen, 1992), and there are expected cut points for what constitutes a small (<0.20), moderate (0.40), or large (>0.60) difference between two groups. The field now recommends that all absolute ES differences be less than 0.10 to ensure that researchers are minimizing bias on the observed covariates used in the propensity score model.
**KS statistic:** The KS statistic depends on the weighted empirical distribution functions for the treatment and control groups. Figure 3.2 shows an illustration using a hypothetical example for how the KS statistic is computed.

**Figure 3.2. Computing the KS Statistic**

First, the empirical cumulative distribution function (ECDF) is plotted for the treated and comparison groups (e.g., those in the work demonstration program versus those who are not), and then the KS statistic computes the maximum vertical distance between the two lines. For covariate $k$, the $ECDF_{zk}(x)$ is defined as

$$ECDF_{zk}(x) = \frac{\sum_{i=1}^{N} w_i I(z_i = z) I(x_i \leq x)}{\sum_{i=1}^{N} w_i I(z_i = z)}$$

for $z = 0$ or 1, where $I(z_i = z)$ equals 1 if this is true and 0 otherwise and similarly for $I(x_i \leq x)$ and $w_i = 1$ for the unweighted case. For ATE, the weights are $1/\hat{p}(x)$ and $1/(1 - \hat{p}(x))$ for the treatment and the control groups, respectively, and for ATT they are 1 for the treatment group and $p(x)/(1 - \hat{p}(x))$ for the control group. Conditional on the propensity score, the ECDF curves for the different groups should be identical, because the distribution function for a given covariate should be the same after conditioning. Thus, the KS statistic for each covariate is
The KS is a useful metric to have in addition to the ES differences because it compares entire distributions rather than just means. It is possible for weights to balance on the means but not balance well on the entire distribution function, so the KS statistic provides a way to assess whether the entire distribution is now exactly the same for the groups. Unlike with the ES, there is no well-accepted cutoff point to use for the KS statistic. Instead, we use the KS statistic $p$-values, which test for statistical significance, to determine whether the KS statistic is significantly different from 0 and thus provides evidence of a meaningful difference between the two curves. **We will aim to have all KS statistic $p$-values be greater than 0.05.** Anything less than 0.05 will be indicative of a lack of balance between the two groups in question with regard to the overall distribution function of the measure.

The TWANG package includes four built-in stop methods based on the ES and the KS. The four stop rules are defined by two components: a balance metric for covariates and a rule for summarizing across covariates. The stop rules are “es.mean,” “es.max,” “ks.mean,” and “ks.max.” The first piece of the stop rule name identifies the balance metric (“es” for the ES or SB, 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 ES and summarizes across all the individual covariate ES values by taking the mean; “ks.max” uses the KS statistic to assess balance and summarizes across all the individual covariate KS statistics using the maximum. For this example, we will select two: es.max and ks.max.

There is no research on which stop rule is best and the choice is likely to depend on the application. McCaffrey, Ridgeway, and Morral (2004) essentially used “es.mean” for their analyses, but our more recent work has sometimes used “ks.max.” (For greater details on stop rules, see McCaffrey et al., 2013.) In general, we recommend that you choose more than one stop rule because it is often the case that the two rules can yield different optimal iterations of the GBM. From there, you should select the rule for use in the outcome analyses that produces the best balance, or when balance is equal, the rule that maximizes the effective sample size (ESS) of the propensity score–weighted samples (see Chapter 4 for more details on ESS).

- **Sampling Weights:** This 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. There are no sampling weights in our LaLonde example.

After setting all of the options above, your screen should now look like Figure 3.3.
Next, hit the run button. You will get a window similar to that in Figure 3.4.
This next step can take around two to three minutes. Once the TWANG package is finished estimating the propensity score weights, you will see a summary table on your screen that shows you the key balance information for your data (Figure 3.5). You are now ready to explore the performance of the GBM and the quality of the estimated propensity score weights, which is a critical step prior to implementing outcome analyses and treatment effect estimation.
Figure 3.5. Propensity Score Model—Summary Table
4. Evaluating the Quality of the Propensity Score Weights

After running the analyses, the “Propensity Score Model” tab will show the “Propensity Score Model Summary Table” (as shown in Figure 3.5, Chapter 3). The TWANG Shiny app provides several detailed summaries on the quality of the estimated propensity score weights in both table and graph form. Prior to any outcome analyses, the analyst should perform several diagnostic checks to assess how well the propensity score weights did at balancing the two groups being compared.

We will return to explaining the output in the “Propensity Score Model Summary Table” after walking through the summaries available under the “Model Evaluation” tab.

Click on the “Model Evaluation” tab. This tab has several model diagnostics and summaries that are useful to review in detail prior to outcome analyses.

First, the page prepopulates with the first option, the “Convergence” plot (Figure 4.1).

![Figure 4.1. Model Evaluation—Convergence Plot](image)
This plot shows diagnostic checks that make sure that the specified value of “GBM iterations” allows the GBM to explore sufficiently complicated models. The convergence plot shows the balance measures (ES or KS) as a function of the number of iterations in the GBM algorithm, with higher iterations corresponding to more-complicated fitted models. In this example, 2,753 iterations minimize the maximum ES difference (es.max) and 1,554 iterations minimize the largest of the eight KS statistics computed for the covariates (ks.max). This can be observed clearly in Figure 4.1. The maximum of ES and KS statistics starts large, decreases, and then increases again somewhere between 1,000 and 3,000 iterations. The plot suggests 5,000 GBM iterations is sufficient because the optimal iteration is clearly occurring well before the 5,000th iteration.

Note that if it seems likely that additional iterations would result in lower values of the balance statistic—for instance, if the balance statistic is still declining without appearing to have attained a minimum by the end of the plot—GBM iterations should be increased and the propensity score model rerun. You want the statistics to get very low before achieving the maximum number of GBM iterations specified in the TWANG Shiny app. As shown in Figure 4.1, additional complexity typically makes the balance worse after a point. Figure 4.1 also gives information on how compatible two or more stop rules are: If the minima for multiple stop rules under consideration are near one another, the results should not be sensitive to which stop rule is used for the final analysis.

Note there is the option to save the graphic below the graph. All graphics in the TWANG Shiny app allow you to download and save diagnostic graphics as .png images. Additionally, it is possible to show findings from only one of the stop rules (e.g., es.max) so that results for specific stop rules can be used easily in presentations and publications.

Click on the option “Propensity Score” (Figure 4.2).
This option produces boxplots illustrating the spread of the estimated propensity score in the treatment and control groups. This is a useful way to examine whether the groups have overlap in the propensity score. Figure 4.2 shows a reasonable amount of overlap (e.g., shared values of the propensity score between the treatment and control groups). We note that excellent covariate balance can often be achieved with propensity score weights, even when the propensity scores estimated for the treatment and control groups show little overlap. However, it is important to carefully consider whether there is sufficient overlap in the treatment and control groups on all the pretreatment covariates. While one cannot formally test for overlap, it is possible to check the univariate distributions of each covariate by treatment group to at least eyeball areas where overlap might not exist. For example, if the minimum age of the treatment group is clearly much lower than that of the control group, then overlap does not exist between the two groups in the lowest age categories. If this is the case, you can subset the sample down to where you have overlap (e.g., the age range within which you have individuals in both groups) and refit the propensity score model on this subset. When creating a subset, it is important to adjust the population to which you are generalizing the findings to ensure you do not extrapolate findings beyond the sample used in the final analysis.

Assessing “Balance” Using Graphics

The TWANG Shiny app will generate several useful diagnostic plots to evaluate the propensity scores. The full set of plots are given on the “Model Evaluation” tab. These include a plot showing balance before and after weighting as measured by ES differences, and plots
showing the ES and KS \( p \)-values before and after weighting. We now discuss each of these plots in turn.

Select the “Balance Plot” option (Figure 4.3).

**Figure 4.3. Model Evaluation—Balance Plot**

This plot illustrates the effect of the propensity score weights on the magnitude of differences between groups on each pretreatment covariate. These magnitudes are standardized using the standardized ES described earlier. In these plots, substantial reductions in ESs are observed for most variables (blue lines), with two variables showing an increase in ES (red lines), but only a seemingly trivial increase. Closed red circles indicate a statistically significant difference that can occur on both sides. There are typically many more closed circles on the unweighted side because after weighting, differences between the groups are reduced on many of the control covariates included in the propensity score model. 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, you are warned that some ESs are too large to plot. Here, the ESs are still pretty large even after propensity score weights; this is because of the difficulty in getting high-quality ATE weights on this data set. Obtaining good balance in practice is generally a function of several things: the size of the imbalance prior to weighting, the sample sizes in each group, and the number of covariates on which the propensity score weights are trying to balance the sample. In this case, it is the very large race difference between the two groups that is making it difficult to obtain good balance even after weighting. The treatment group that participated in the work
demonstration program was 84 percent black while only 20 percent of the control group was black. After weighting, these percentages change to 58 percent and 36 percent, respectively (see detailed balance tables in the section titled Assessing “Balance” Using Balance Tables).

It is generally recommended that effective sizes below 0.1 are indicative of good balance and high-quality propensity score weights.

The last two plot options examine the \( p \)-values for the ES and KS statistics. 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 has been achieved. We use a quantile-quantile (QQ) plot comparing the quantiles of the observed \( p \)-values with 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 and fall along the 45-degree line. 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 allows you to visually inspect the \( p \)-values of the \( t \)-tests for group differences in the covariate means and the \( p \)-values from testing the KS statistics for each pretreatment covariate.

Select the “ES \( p \)-values” option (Figure 4.4). This plot 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 moving toward the 45-degree line, which represents the cumulative distribution of a uniform variable on \([0,1]\). This indicates that although balance was improved for some covariates, there are still a few for which imbalances exist. The detailed balance tables described in the next section will help us understand which variables these imbalances correspond to. Note that the ES \( p \)-values are shown for the optimal iteration chosen from each stop method used in this example. Even when one is using a stop method defined by the KS (ks.max and ks.mean), one can compute an ES and ES \( p \)-values for each covariate used in the model. The stop methods use the same set of balance criteria even though they select the optimal iteration differently, which will be used to produce the propensity score weights that go into computing the various balancing metrics.
Select the “KS p-values” option (Figure 4.5). We see a similar story here; however, several of the covariates show $p$-values above the 45-degree line for the KS statistic. This suggests that we were able to obtain good balance in terms of the KS for several of the pretreatment covariates.
As noted before, there is the option to save these graphics as .png images by clicking on the “save” button at the bottom of the screen.

All graphics in the TWANG Shiny app allow you to download and save diagnostic graphics. Additionally, it is possible to show findings from only one of the stop rules (e.g., es.max) so that results for specific stop rules can be used easily in presentations and publications.

Assessing “Balance” Using Balance Tables

The TWANG Shiny app generates a series of “Balance Tables” that provide a tabular summary of the balance between the covariate distributions for the treatment and control groups before and after propensity score weights and for each stop method. The tables can be found by clicking on the “Balance Tables” option on the “Model Evaluation” tab (Figure 4.6).
The unweighted table provides the same information without weighting to give a sense of the initial general imbalance in the groups. For both the weighted and unweighted tables, balance is assessed separately for each covariate by each stop method. A different weighted balance table is available for each stop method used. Here we used two (es.max and ks.max), so we have the option to look at the weighted balance table for both by selecting the desired stop method in the “Stop Method” box. Figure 4.6 shows the table for the es.max stop method. Figure 4.7 illustrates what you will see if you select ks.max in the “Stop Method” box.
Recall that we have selected to estimate ATE weights for our analysis for this illustrative example.

The columns in the tables consist of the following items:

- **Treatment and Control Means**: The unweighted table shows the unweighted means. The weighted table shows the propensity score-weighted means for each stop rule. Note that when the 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 treatment means.

- **Treatment and Control Standard Deviations**: These columns show the propensity score-weighted treatment and control groups' standard deviations for each of the
pretreatment covariates. The unweighted table shows the unweighted standard deviations.

- **Standardized difference**: This is the standardized ES, defined as the treatment group mean minus the control group mean divided by the pooled sample (treatment and control) standard deviation if the estimand = “ATE,” or the treatment group standard deviation if the estimand = “ATT.” Occasionally, lack of treatment group or pooled sample variance on a covariate results in very large (or infinite) standardized ESs. For purposes of analyzing mean ESs across multiple covariates, we set all standardized ESs larger than 500 to NA (missing values).

- **Kolmogorov-Smirnov**: The KS test statistic and its associated p-value directly test for differences in the distributions of the given covariates between the treatment and control groups. They are derived from analytic approximations. If we have good balance, we expect no evidence of a difference between the two groups’ covariate distribution for a given covariate (e.g., $p > 0.05$).

The weighted tables show how well the resulting weights succeed in balancing the treatment and control group by showing how the propensity score weights adjust the control and treatment group summary statistics. For each stop rule, the means are weighted using propensity score weights corresponding to the optimal iteration of the GBM model selected for the given stop rule (e.g., here the one that yielded the smallest maximum ES or KS statistic). The hope is that, after using the propensity score weights, the weighted pretreatment covariates match, or balance, between the treatment and control groups. In practice, this can be difficult to achieve for all pretreatment covariates. Note that if the estimand = “ATT,” then only the control group is weighted and the weighted table shows how well the propensity score weights the control group to match the treated group.

Also note that all tables have “copy,” “csv,” and “excel” buttons so that you can save the output into formats that are easily used for future publications and project meetings.

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.

Now, let’s return to the first table shown on the “Propensity Score Model” tab (Figure 4.8).
The summary table includes one row for the results using the propensity score weights produced by each stop rule specified in the “Stop Method” option and one row for the unweighted data. The treatment and control group sample sizes are presented in the columns, along with the ESSs of the treatment and control groups. Note when the estimand = “ATT,” the ESS for the treatment group equals the treatment group sample size because the weights equal 1 (additional details on the ESS follow). The maximum and mean or average of the ES or KS statistics for the covariates and the optimal number of iterations or trees in the GBM that minimizes the stop rule will be missing for the unweighted row. In this case, the optimal iteration for both stop methods is 1,175.
In general, weighted means can have greater sampling variance than unweighted means from a sample of equal size. The ESS of the weighted comparison group captures this increase in variance as

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

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 control observations. Therefore, the ESS will give an estimate of the number of control participants that are comparable with the number of participants in the treatment group when the 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 1 for that group. The ESS formula above is an accurate measure of the relative size of the variance of means when the weights are fixed or they are uncorrelated with outcomes, but otherwise underestimates the ESS (Little and Vartivarian, 2004). With propensity score weights, it is rare that weights are uncorrelated with outcomes. Therefore, the ESS formula above really represents a lower bound on the ESS, 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 because of weighting. When proceeding to the outcome analysis phase, we recommend that only one set of propensity score weights is used, ideally the version that produces the best balance across multiple balance criteria and that also yields optimal ESS values.

Return now to the Model Evaluation tab and click on the “Relative Influence” tab, which displays the relative influence that each pretreatment covariate has on the propensity score model predicting treatment (Figure 4.9). Here, we see that being black, wages from 1974, and age are the three most influential factors in the propensity score model.
Figure 4.9. Model Evaluation—Relative Influence

The `gwe` package has various tools for exploring the relationship between the covariates and the treatment assignment indicator if these are of interest. This plot shows the relative influence of each variable for estimating the probability of treatment assignment.
5. Analysis of Outcomes

The aim of the NSWD analysis is to determine whether the program was effective at increasing earnings in 1978. The TWANG Shiny app provides the ability for you to estimate treatment effect estimates for either binary (using logistic regression) or continuous outcomes (using linear regression and assuming the outcome is normally distributed).

To estimate treatment effects, go to the “Treatment Effect Estimation” tab (Figure 5.1).

**Figure 5.1. Treatment Effect Estimation**

The different boxes on this tab require some instructions.

- **Outcome**: This is where you select the outcome to be analyzed. You can only examine one outcome at a time. Select “re78,” wages from 1978 for the LaLonde example.

- **Outcome Type**: Two types of models can be used in the TWANG Shiny app, linear regression models for continuous outcomes and logistic regression models for binary outcomes. Because “re78” is continuous, select the “Continuous” option for this illustration. If your outcome is binary, please ensure it is coded as a 0/1 variable to avoid warning messages.

- **Covariates**: This option allows for additional pretreatment covariates to be used in the analysis. For this first run, we will not select any additional covariates to be included in the model run. However, it is generally recommended to include the same pretreatment covariates used in the propensity score model in the outcome model, assuming the sample size in the data allow for such an outcome model to be fit. This type of model
can be referred to as a *doubly robust* model. Doubly robust models guard against bias that could result if either the outcome model (with covariates) or the propensity model is incorrect. As long as one is right, our causal effect estimates should be unbiased, so it is best practice to use doubly robust models to the extent possible in a propensity score analysis (Kang and Schafer, 2007). When the sample size is not sufficient to fit a fully doubly robust model, it can also suffice to control for the pretreatment covariates that have lingering imbalances (e.g., ES > 0.10 or KS-statistic $p$-values < 0.05) in the outcome model. We will explore additional doubly robust models below for the LaLonde data.

- **Stop Method**: You should select which stop method will be used to provide the propensity score weights. This should be selected based on which method produces optimal balance. If the stop rules perform similarly, it can be selected based on which method yields the largest ESS for the treatment and control groups. For now, select “es.max.”

When you are done, the screen should look like Figure 5.2 and you can hit the run button. After doing so, you will see two tables that show the same findings because no additional control covariates were used in the regression model.

**Figure 5.2. Treatment Effect Estimation—Results**

The first set of results shows the estimated ATE (or ATT, depending on the selection during the propensity score model fitting stage). These results are estimated using predictive margins, also known as recycled predictions, which convert the model coefficients to the scale of interest (Graubard and Korn, 1999; Kleinman and Norton, 2009). The analysis estimates a decrease in earnings of $889 for those who participated in the NSWD compared with similarly observed
people in the CPS. The effect, however, does not appear to be statistically significant ($p = 0.342$).

The second set of results shows the estimated model coefficients. For linear regression, as shown in Figure 5.2, the coefficient on the treatment variable will match the estimated ATE. However, for other outcome types, these will not necessarily agree. For example, for a binary outcome, logistic regression will be used to fit the model. The coefficients from a logistic regression are log-odds ratios, which do not directly correspond to the ATE or ATT.

Now, let’s go back and run the model with covariates that had lingering imbalances (race, married, and re74). To run adjusted analysis using these covariates, you must select them in the covariates box and then hit run for updated results (Figure 5.3).

**Figure 5.3. Treatment Effect Estimation—Adjusted Results**

The analysis estimates a decrease in earnings of $439 for those who participated in the NSWD compared with similarly observed people in the CPS. However, this effect is still not statistically significant ($p = 0.747$).

As noted earlier, 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 (Bang and Robins, 2005; Hullsiek and Louis, 2002). These estimators are consistent if the propensity scores are estimated correctly or if the regression model is specified correctly. For example, note that in the balance table for es.max.ate, the two groups still have lingering imbalances on age, black, married, and re74. Although linear regression is sensitive to model misspecification when the treatment and control 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.

Let’s go back and run one final model that includes all the pretreatment covariates used in the propensity score model. To run this fully adjusted analyses using these covariates, you must select them in the covariates box and then hit run to generate updated results (Figure 5.4).

**Figure 5.4. Treatment Effect Estimation—Doubly Robust**

Notably, in the fully doubly robust model, the estimated impact of NSWD has still remained around $500 with a nonsignificant $p$-value. The overarching lesson here is that it is important to use additional covariate control in these models. The covariates should be the same ones that are included in the propensity score model.
Now that we are done with the analysis, it is useful to save the weights into your data set. To do so, click on the last tab of the app, labeled “Weights.” You can specify the label name you like for the weights and click save (Figure 6.1).

**Figure 6.1. Saving Propensity Score Weights**

Returns the original dataset with a new column containing the weights.
- Weights for ATT are 1 for the treatment cases and p(1-q) for the control cases.
- Weights for ATE are 1/p for the treatment cases and 1/(1-p) for the control cases.


