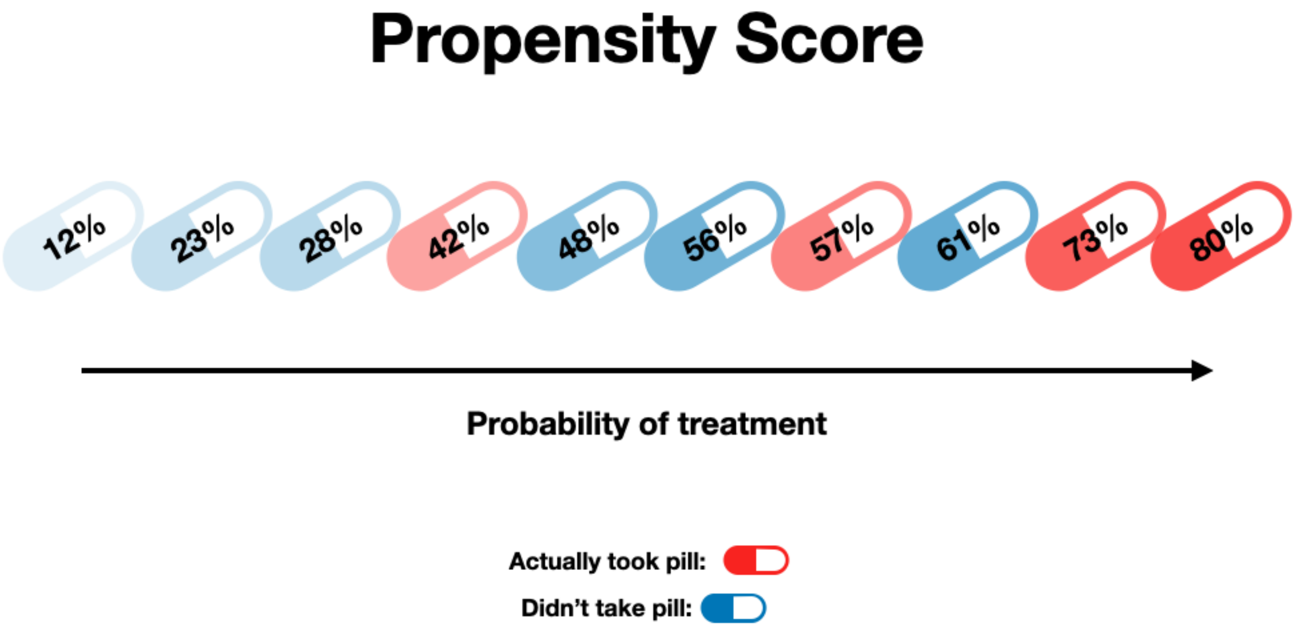
Causal Effects via Propensity Scores

How to estimate effects from observational data

This article is the 2nd post in a series on causal effects. In the previous post, we laid a theoretical foundation for causal effects, but there were some lingering practical concerns. Namely, *how can we compute causal effects from observational data?* Here, I will discuss a set of techniques that do exactly this using something called a propensity score. The discussion will be followed up with example Python code of using these techniques with real-world data.

Key points:

1. Propensity scores estimate the probability of treatment
2. Subjects with similar propensity scores have similar baseline covariates
3. 3 popular propensity score techniques are: matching, stratification, and inverse probability of treatment weighting



Propensity scores estimate the probability of treatment. Image by author.

. . .

To estimate causal effects we need data. More specifically, we need data that contains **outcomes**, **treatments**, and **covariates** (defined previously). However, not all data sources are equal. Before discussing the propensity score, we must draw a clear distinction between two ways data are obtained.

Observational vs Interventional Studies

The **first** we can call an **observational study**. This consists of passively measuring a system without any intervention in the data generating process. Most data collection falls under this category.

The **second** way to get data we can call an **interventional study**. This is when we purposely influence the data generating process for a particular goal. This includes things like **randomized controlled trials (RCTs)**, where a *random* sample of a population is split into 2 groups (i.e. control and treatment groups) and compared. The key benefit of randomization is it will tend to balance the treated and control populations in terms of both **measured and unmeasured** covariates [1].

There are **two reasons** for distinguishing these types of studies. **One**, observational data, generally, do not control sub-population sizes (e.g. the control group could be 10x bigger than treated group), and **two**, the control and treatment populations may have systematic differences. For example, if trying to estimate the causal effect of smoking on liver health, comparing smokers and non-smokers (in the wild) might produce biased estimates if smokers tend to drink more alcohol than non-smokers.

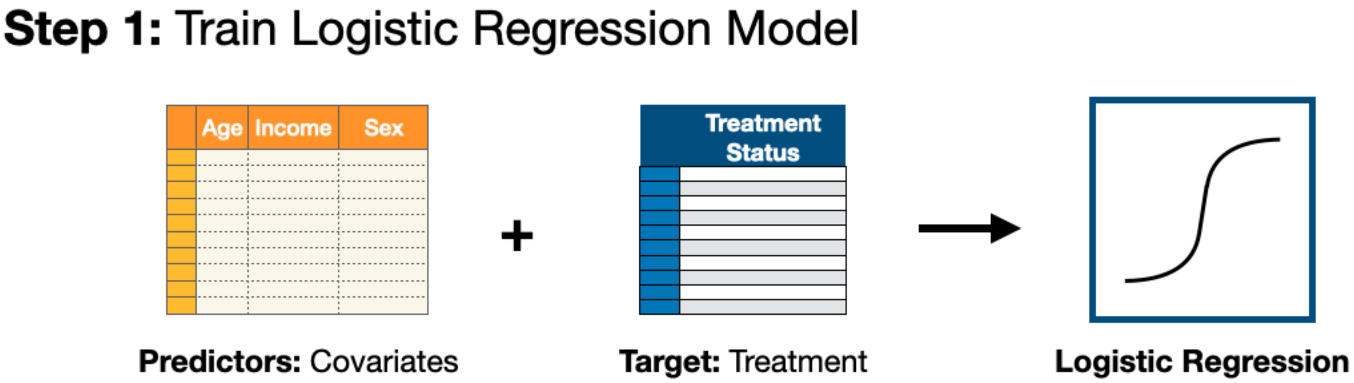
With that, we may feel hopeless if trying to compute unbiased causal effects from observational data. However, as you may have guessed from the subtitle, there is a solution: propensity scores.

Propensity Score

A **propensity score** estimates the **probability a subject receives a particular treatment based on other characteristics** (i.e. covariates). This score approximates a balancing score, meaning when conditioned on, the treatment and control group covariates look similar. This enables comparison of control and treated populations reminiscent of interventional studies, but from observational data [1].

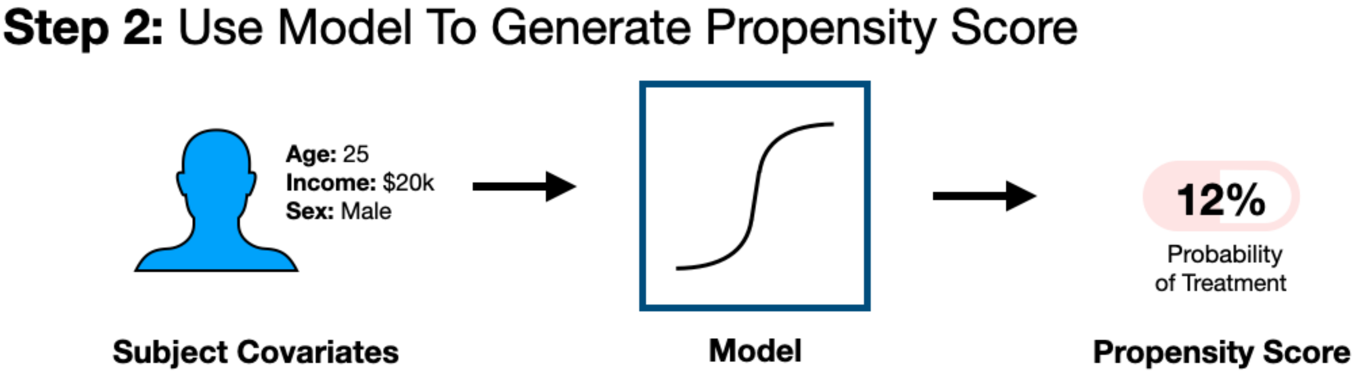
The most popular way to generate a propensity score is via logistic regression [1]. Consider the Tylenol example from the previous post. There we had a treatment (i.e. a pill) and we wanted to measure its effect on an outcome (i.e. headache status). But suppose now we don't have access to data from RCTs, only observational data containing the following information: age, income, sex, took pill or not, and change in headache status.

Armed with this dataset, we can obtain a propensity score via a logistic regression model in a straightforward way. We simply set our covariates (i.e. age, income, and sex) as the **predictors** and the "took pill or not" variable as the **target**. This is illustrated in the image below.



Step 1 in generating propensity score. Pass covariates and treatment variables into logistics regression model. Image by author.

Once we have a logistic regression model, we can obtain a probability of treatment (i.e. propensity score) for each subject in the dataset.



Step 2 in generating propensity score. Input each subject's covariate data into logistic regression model to obtain score. Image by author.

3 Propensity Score Methods

Now that we know what a propensity score is and how to compute it, we turn to **how to use a propensity score to estimate causal effects**. Here, I describe 3 popular ways to do this.

1) Matching

Matching consists of creating **treated-untreated pairs with similar propensity scores**. One simple way to do this is by matching each treated subject with the untreated subject with the closest propensity score in a *greedy* way *(greedy search described previously).*

In other words, pick a treated subject and match them to an untreated subject with the closest propensity score one by one. If we ensure that no untreated subject is matched to more than one treated subject, it is called **matching without replacement**.

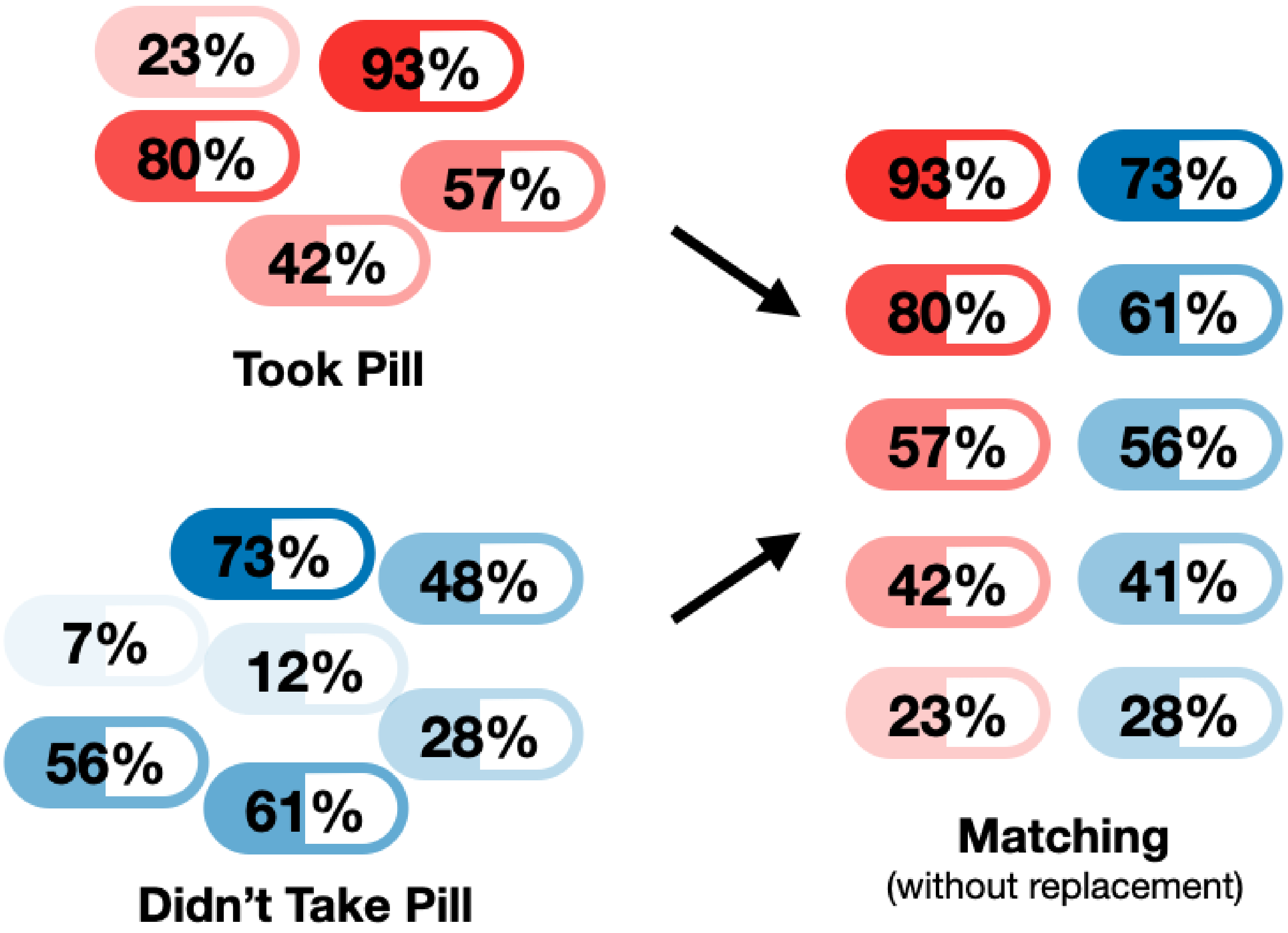


Illustration of how matching on propensity score (without replacement) works. Image by author.

This process will create a so-called **matched sample**, where **every treated subject is matched with an untreated subject**. Notice, in the figure above the subjects with propensity scores of 0.07 and 0.12, did not get matched to a treated subject, thus these subjects did not make it into the matched sample.

With our matched sample, treatment effects can then computed by directly comparing treated and untreated populations just like we did in the previous blog. For more matching strategies see the paper by Austin [1].

2) Stratification

Stratification on the propensity score **splits subjects into groups with similar propensity scores** in which causal effects can be computed. This is typically done as a **2-step process**. **First**, subjects are rank ordered based on their propensity score (i.e. ordered from smallest to largest PS). **Then**, the ordered subjects are split into equal sized groups (i.e. quantiles).

A good rule of thumb based on work by Cochran (1968) is to split subjects into 5 groups (i.e. quintiles), however other choices can be made. For instance, in the figure below, this 2-step process is illustrated for subjects split into 4 equal sized groups (i.e. quartiles).

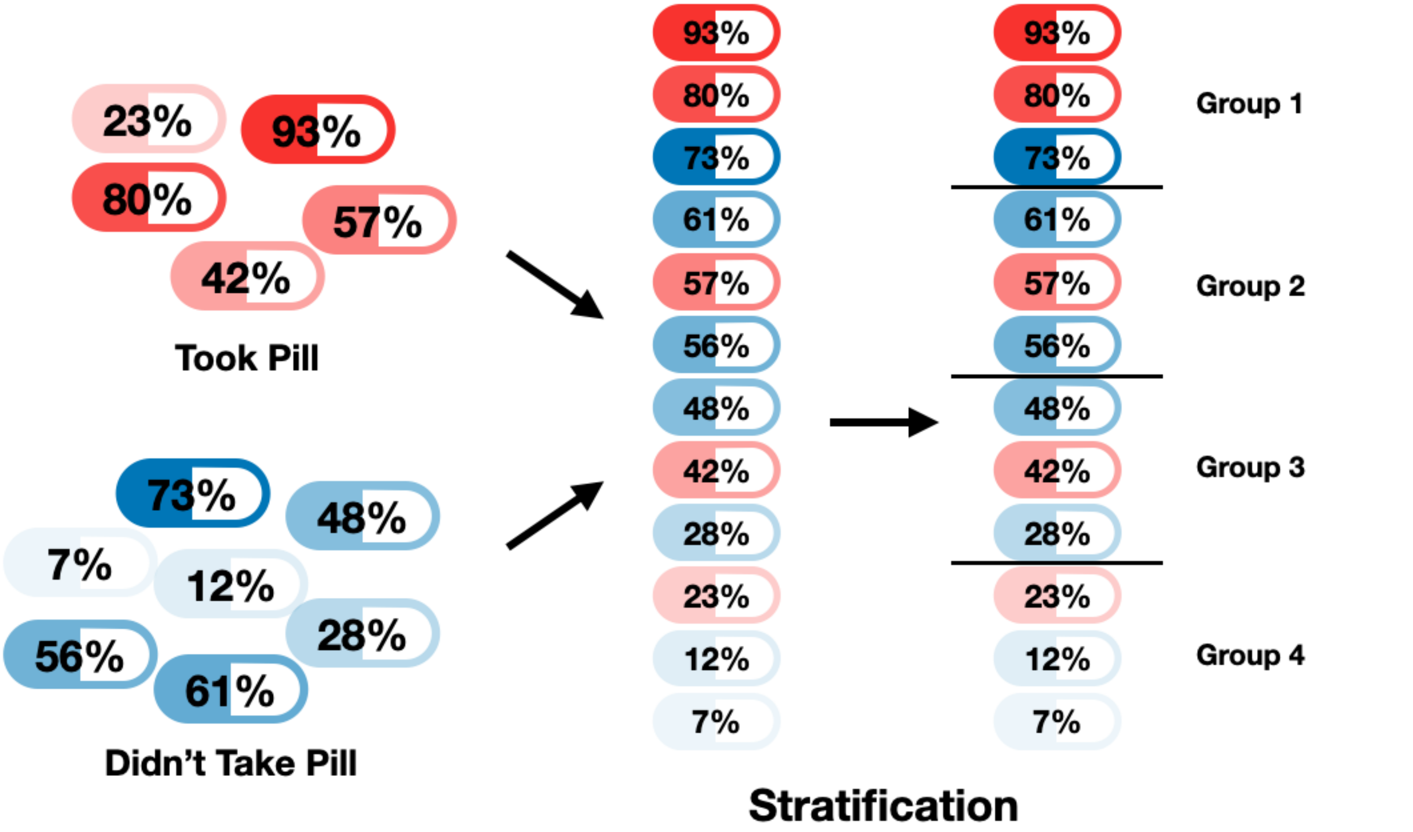


Illustration of 2-step process to do stratification on propensity score. Image by author.

Once we have stratified the subject based on the PS, we can compute treatment effects by **directly comparing treated and untreated populations within each group** as before. We then go one step further and pool the group-level treatment effects via a weighted average, to obtain an overall treatment effect.

3) Inverse probability of treatment weighting (IPTW)

Inverse probability of treatment weighting has a fundamental difference compared to the methods we have discussed so far. In matching and stratification, we use the propensity score to derive *subject groups* which can be used to compute treatment effects. With IPTW, however, **we use the propensity scores to derive *weights for each subject* which are then used to compute treatment effects directly**.

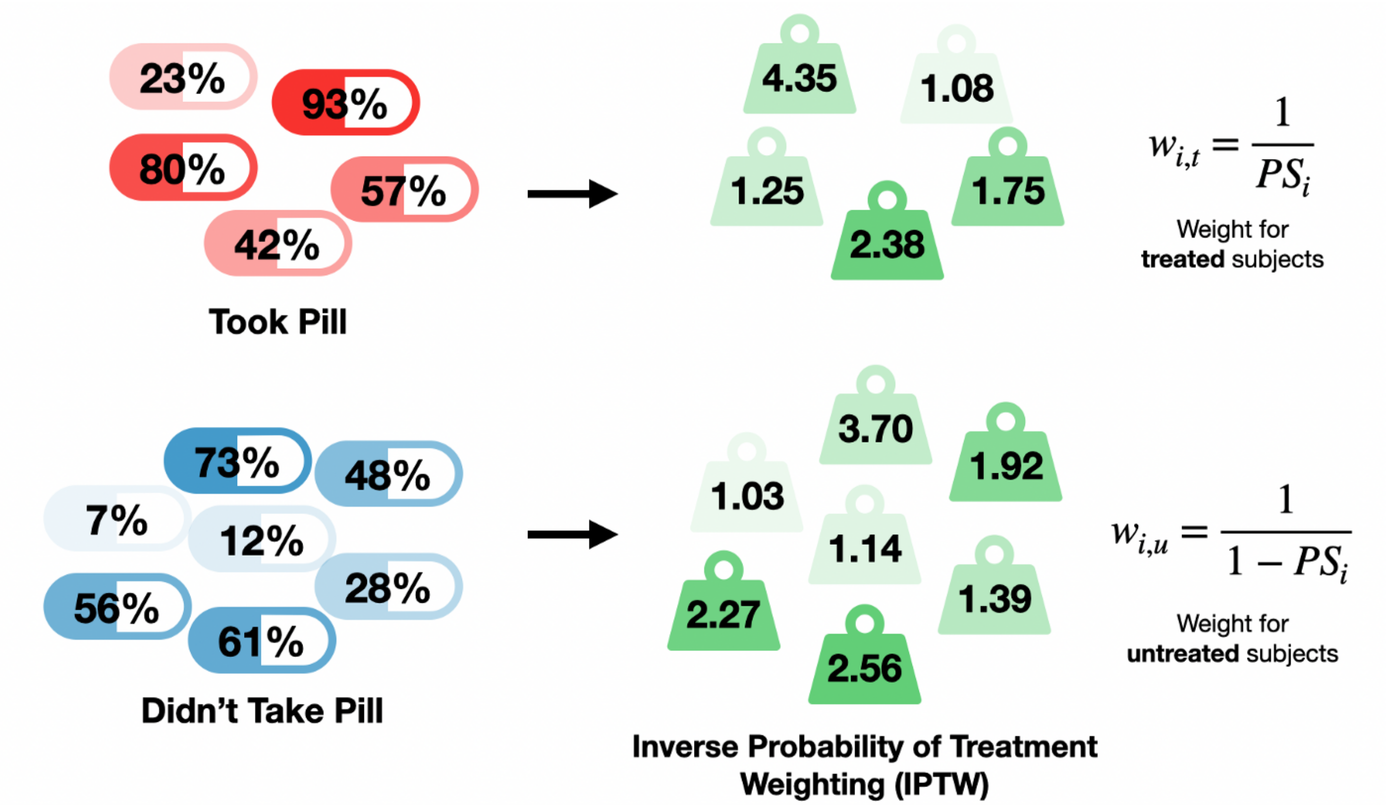
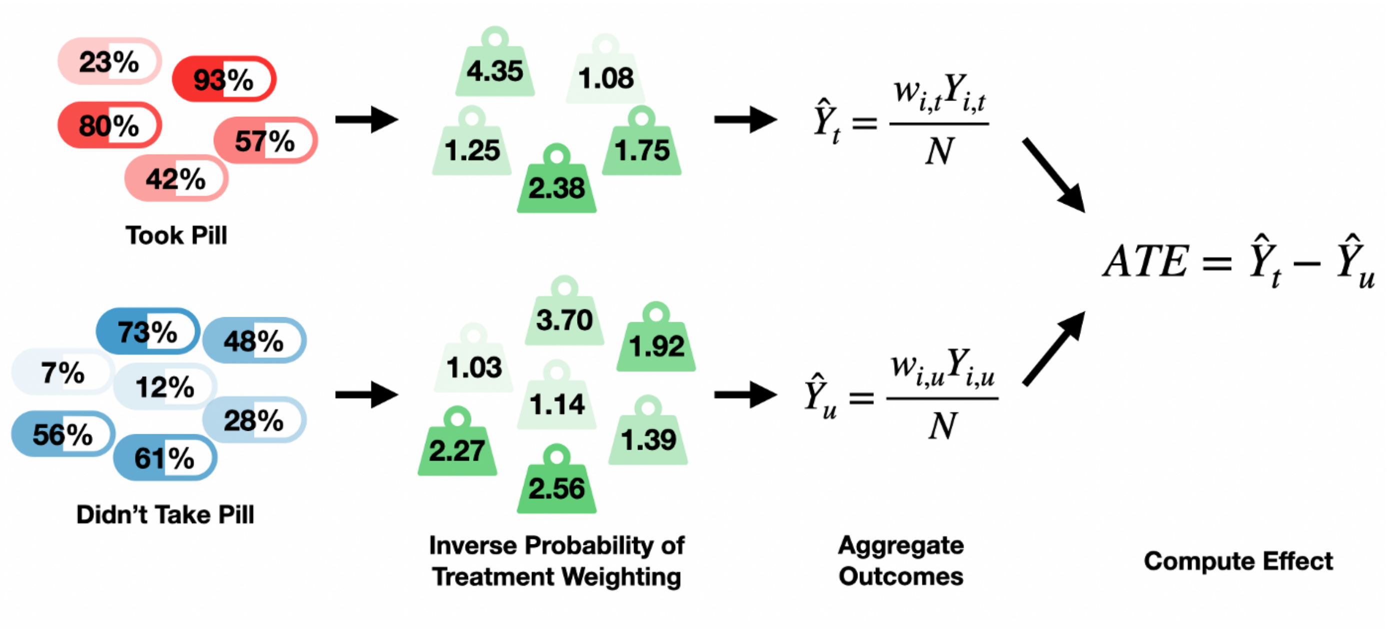

23%
93%
80%
57%
42%
Took Pill
4.35
1.08
1.25
1.75
2.38
$w_{i,t} = \frac{1}{PS_i}$
Weight for treated subjects
73%
48%
7%
12%
28%
56%
61%
Didn't Take Pill
3.70
1.92
1.03
1.14
1.39
2.27
2.56
$w_{i,u} = \frac{1}{1 - PS_i}$
Weight for untreated subjects
Inverse Probability of Treatment Weighting (IPTW)

Illustration of how to compute weights for inverse probability of treatment weighting (IPTW). Image by author.

To compute effects we **first**, we assign a weight to each subject based on the equations on right hand side of the above figure. Notice, the weight equation is different for treated and untreated subject.

Next, we aggregate the outcome variable values for the treated and untreated subjects, respectively, using the weights we just derived. **Finally**, we take the difference of the aggregated outcome values for the treated and untreated populations. This difference is taken as the average treatment effect (ATE). A visual summary of the full process is given below.


23%
93%
80%
57%
42%
Took Pill
4.35
1.08
1.25
1.75
2.38
$\hat{Y}_t = \frac{w_{i,t}Y_{i,t}}{N}$
$ATE = \hat{Y}_t - \hat{Y}_u$
73%
48%
7%
12%
28%
56%
61%
Didn't Take Pill
3.70
1.92
1.03
1.14
1.39
2.27
2.56
$\hat{Y}_u = \frac{w_{i,u}Y_{i,u}}{N}$
Inverse Probability of Treatment Weighting
Aggregate Outcomes
Compute Effect

Visual summary of how to compute average treatment effect using inverse probability of treatment weighting (IPTW). Image by author.

Balance Diagnostics: evaluating propensity score performance

In the ideal case, a propensity score is a **balancing score**, meaning if we look at **2 subjects with the same score, their underlying covariates should be identical.** In practice, however, this is rarely the case. The propensity score serves merely as an approximation of this ideal case [3]**.**

With that being said, we can use this ideal to derive strategies to evaluate the performance of our propensity score model. For example, in the case of matching, we can take the treated and untreated groups in the matched sample and compare their underlying covariate distributions.

For normally distributed variables, this can be done by simply comparing means and standard deviations. For other types of distributions one can use techniques such as: the KS-test, KL-divergence, Wasserstein distance, etc. For further discussion on balance diagnostic I refer the reader to the nice paper by Austin [1]**.**

Regression-based techniques

Before moving on to the example code, it is worth mentioning that propensity score-based methods are not our only option when trying to compute causal effects from observational data. There is another class of techniques that we can call **regression-based**. This includes approaches such as: **linear regression, meta-learners, and double machine learning**. In fact, in a preceding post on causal inference we used a meta-learner to estimate the treatment effect of going to grad school on income. Here, we will repeat this analysis, but now using propensity score-based methods.

Example: Estimating Treatment Effect of Grad School on Income (with propensity scores)

In this example, we will estimate the causal effect of a graduate degree on income using the 3 propensity score methods discussed earlier. The analysis is almost identical to the example from the causal inference blog. We again use the DoWhy Python library and consider only 3 variables in our causal model: age, hasGraduateDegree, and greaterThan50k (income). The example code can be found at the GitHub Repo.

We start by importing the necessary Python modules.

```
import pickle

import dowhy
from dowhy import CausalModel
import numpy as np
```

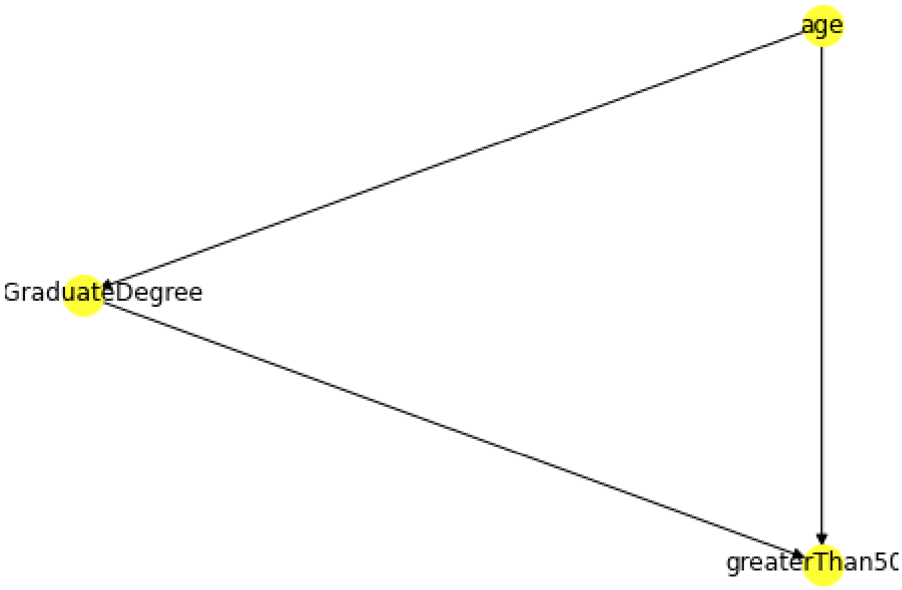
Next, we load in our dataset. Here we use the same US census data from the causal discovery example [3].

```
df = pickle.load( open( "df_propensity_score.p", "rb" ) )
```

The first step in estimating causal effects using the Python DoWhy library is explicitly defining the **causal model** i.e. stating the **causal connections between variables**. Here, we will assume the same causal model as before.

```
model=CausalModel(
        data = df,
        treatment= "hasGraduateDegree",
        outcome= "greaterThan50k",
        common_causes="age",
        )

model.view_model()
```



Output of view_model() method is graphical representation of our causal model.

Next, we generate an estimand which is essentially recipe for computing our causal effect. This ensures the measured confounder (i.e age) is properly adjusted for.

```
identified_estimand_experiment =
model.identify_effect(proceed_when_unidentifiable=True)
```

Then, we can finally use each propensity score method to estimate the causal effect of hasGraduateDegree on greaterThan50k.

```
# create list of names of the propensity score methods
ps_method_name_list = ["matching", "stratification", "weighting"]
```

```
# initialize dictionary to store estimates from each method and list to
store ATEs
ps_estimate_dict = {}
ps_estimate_val_list = []

# estimate effect for each method
for ps_method_name in ps_method_name_list:
    ps_estimate =  model.estimate_effect(identified_estimand_experiment,

method_name="backdoor.propensity_score_" + ps_method_name,
                                    confidence_intervals=False,
                                    method_params={})
    # add estimate to dict and ATE to list
    ps_estimate_dict[ps_method_name] = ps_estimate
    ps_estimate_val_list.append(ps_estimate.value)

    print(ps_estimate)
    print("\n")
```

The average treatment effect estimations from each method are as follows: **matching** — 0.136, **stratification** — 0.25, and **IPTW** — 0.331. Giving an overall mean value of about 0.24.

How do I interpret this?

We can interpret this overall ATE as: a unit change in the expectation value of having a grad degree given age i.e. **going to grad school, will result in approximately a 24% increase in the probability someone makes more than 50k a year**. These results are in the ballpark of what we saw previously using the meta-learner approach (there we saw ATE = 0.20).

Word of caution

As a final note, I would caution practitioners to take these methods with a grain of salt. Unlike RCTs where randomization can handle biases from both **measured and unmeasured** confounders, propensity score methods only help us against **measured** confounders.

For example, there are conceivably other factors that impact both having grad degree and income that are not captured by our dataset such as: the occupation/income of parents, field of study, and work ethic.

In the next blog of this series, we will discuss **how to cope with unmeasured confounders** and bring the do-operator back into our discussion of causal effects.

. . .

Resources

More on Causality: Causality: Intro| Causal Inference | Causal Discovery | Causal Effects | Do-operator

Connect: My website | Book a call | Message me

Socials: YouTube | LinkedIn | Twitter | Tik Tok | Instagram| Pinterest

Support: Buy me a coffee ☕

Join Medium with my referral link — Shawhin Talebi

As a Medium member, a portion of your membership fee goes to writers you read, and you get full access to every story…

shawhin.medium.com

. . .

[1] An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies by Peter C. Austin

[2] The central role of the propensity score in observational studies for causal effects by Paul R. Rosenbaum & Donald B. Rubin

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.