

Week 2: Matching

Making an apples-to-apples comparison

[Class slides](#) · [Download slides as PDF](#) · [Problem Set 1](#) · [Download notes as PDF](#) · [Download Quarto source](#)

Discussion assignments are available on Moodle. Prepare your assigned item individually; one person from each group will be called on to discuss it.

Learning Objectives

By the end of this topic, you should be able to:

- Explain how LaLonde uses the experimental NSW result to assess estimates obtained with observational comparison groups.
- State the conditional independence and common support assumptions, and explain how matching aims to create an apples-to-apples comparison.
- Explain what a propensity score measures and use a logit model to estimate it.
- Estimate and interpret the ATT using inverse-probability-weighted regression, assess covariate balance, and compare the result with the experimental benchmark.

Before class

Read the notes and empirical readings below. We will start with the experimental evaluation of the National Supported Work demonstration by MDRC. The Week 2 practice is **Problem Set 1**, due **Friday 2 October 2026 at 12 noon**. [Instructions and Quarto answer template](#).

Empirical readings

- **MDRC, *Summary and Findings of the National Supported Work Demonstration***. Read the programme overview and research-design discussion, and examine Tables 4-3 (p. 61), 5-3 (p. 84), 6-3 (p. 106), and 7-3 (p. 124). Page numbers refer to the printed pages. [Report and download](#).
- **LaLonde (1986), “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.”** Focus on the research question, experimental and comparison samples, and Tables 3 and 5 for men. [Published paper](#).
- **Dehejia and Wahba (1999), “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.”** Focus on the sample, Figures 1 and 2, and Table 3. [Paper from the author’s website](#).

The quotations used in class come from **LaLonde (1984), Chapter 2, footnote 5, printed pp. 81–82**. The working paper gives a fuller account of why his male-sample estimate differs from the report’s findings. [Princeton catalogue record](#).

Textbook references

- **The Effect:** Chapter 14, *Matching*. [Online book](#).
- **Causal Inference: The Remix:** Chapter 5, *Unconfoundedness*, particularly matching and propensity scores. [Online book](#).
- **Mastering Metrics:** Chapter 2, *Regression*, for selection on observables and comparisons conditional on covariates. [Companion resources](#) and [Mastering Econometrics videos at Marginal Revolution University](#).
- **Mostly Harmless Econometrics, optional advanced reading:** Chapter 3, particularly regression, matching, and the propensity score. [Book contents](#).

Use the textbook references for complementary explanations of the material.

1. From selection bias to a credible comparison

Let $D_i = 1$ indicate programme participation, and let Y_i^1 and Y_i^0 be earnings with and without treatment. The observed outcome is

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0.$$

The average treatment effect on the treated is

$$\tau_{\text{ATT}} = E[Y_i^1 - Y_i^0 \mid D_i = 1].$$

We observe treated individuals' outcomes with treatment, but not what those same people would have earned without it. Recall the Week 1 decomposition:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \tau_{\text{ATT}} + E[Y_i^0 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0].$$

The second line is selection bias. Participants and nonparticipants may have had different earnings even without the programme.

Matching tries to make the comparison “**apples to apples**” using observed pre-treatment characteristics. Regression adjustment also addresses differences in observed characteristics, but a fitted regression can extrapolate into combinations of characteristics with little or no comparison data. Making the comparisons explicit helps us see where the evidence comes from.

2. Supported Work: first understand the experiment

The National Supported Work Demonstration offered temporary paid employment in a supportive, performance-oriented environment. Its policy objective was to help people with substantial employment difficulties move into regular, unsubsidised work. Its target groups included long-term recipients of Aid to Families with Dependent Children (AFDC), former drug addicts, former offenders, and young school dropouts. Eligible applicants in the experiment were randomly assigned to the treatment or control group.

Distinguish **earnings while in subsidised employment** from earnings after leaving the programme. The later outcomes tell us about the persistence of the earnings gains.

The report's earnings tables

These tables organise outcomes by **months since enrolment**. The first three columns are regression-adjusted experimental and control means and their difference. The amounts are monthly earnings in nominal dollars. The final columns distinguish earnings received directly from Supported Work. Read the table notes as part of the evidence.

The AFDC sample supplies the discussion of women. The other three target groups were predominantly male, but **their report tables include women as well as men**. LaLonde later restricts his pooled analysis to men.

Table 4-3: AFDC sample

Table 4-3. Average Earnings per Month, by Length of Follow-Up: AFDC Sample

Month After Enrollment in the Sample	Experimental Group Mean	Control Group Mean	Experimental-Control Differential	Mean Earnings From Supported Work ^a	
				Earnings	As Percent of Total Earnings of Experimentals
1-3	\$409.44	\$49.28	\$360.16†	\$401.27	98.0
4-6	412.30	90.84	321.46†	389.01	94.4
7-9	387.20	99.28	287.92†	348.51	90.0
10-12	371.25	117.85	253.40†	275.45	74.2
13-15	237.11	140.12	96.99†	65.76	27.7
16-18	213.30	159.33	53.97†	13.99	6.6
19-21	230.92	161.79	69.13†	10.97	4.8
22-24	244.64	166.54	78.10†	2.93	1.2
25-27	248.47	167.86	80.61†	0.00	0.0

Source: See Table 4-1.

Notes: The data presented in the first three columns are regression-adjusted estimates that control for differences of age, sex, race, education, prior work experience, household composition, site, and length of site operation. Averages are calculated for all members of the sample, including those with no employment in the covered period. Dollar amounts reported are gross earnings unadjusted for inflation.

^aThese data are not regression adjusted. No experimentals should have been in the program beyond month 21. The small number indicated as participating after that time reflects either data errors or failures by program operators to terminate individuals on schedule.

†Statistically significant at the 5 percent level.

Table 5-3: ex-addict sample

Table 5-3. Average Earnings Per Month, by Length of Follow-Up: Ex-Addict Sample

Month After Enrollment in the Sample	Experimental Group Mean	Control Group Mean	Experimental-Control Differential	Mean Earnings from Supported Work ^a	
				Earnings	As Percent of Total Earnings of Experimentals
1-3	\$395.31	\$122.30	\$273.01†	\$355.31	89.9
4-6	348.00	184.58	163.42†	284.68	81.8
7-9	306.08	166.99	139.09†	224.83	73.5
10-12	280.71	205.37	75.34†	154.02	54.9
13-15	251.65	211.81	39.84*	66.17	26.3
16-18	215.51	222.22	-6.71	17.27	8.0
19-21	243.78	250.25	-6.47	5.20	2.1
22-24	281.02	270.36	10.66	3.81	1.4
25-27	287.05	259.88	27.17	1.78	0.6
28-30	304.09	237.45	66.64	0.00	0.0
31-33	332.18	221.85	110.33†	0.00	0.0
34-36	318.60	218.76	99.84†	0.00	0.0

Source: See Table 5-1.

Notes: The data presented in the first three columns are regression-adjusted estimates that control for differences of age, sex, race, education, prior work experience, household composition, site, and length of site operation. Averages are calculated for a members of the sample, including those with no employment in the covered period. Dollar amounts reported are gross earnings unadjusted for inflation.

^aThese data are not regression adjusted. No experimentals should have been in the program beyond month 21. The small number indicated as participating after that time reflects either data errors or failures by program operators to terminate individuals on schedule.

†Statistically significant at the 5 percent level.

*Statistically significant at the 10 percent level.

Table 6-3: youth sample

Table 6-3. Average Earnings per Month, by Length of Follow-Up: Youth Sample

Month After Enrollment in the Sample	Experimental Group Mean	Control Group Mean	Experimental-Control Differential	Mean Earnings From Supported Work ^a	
				Earnings	As Percent of Total Earnings of Experimentals
1-3	\$389.52	\$100.15	\$289.37†	\$353.90	90.9
4-6	340.76	140.51	200.25†	266.80	78.3
7-9	284.40	138.24	146.16†	197.20	69.3
10-12	255.24	163.51	91.73†	134.10	52.5
13-15	218.79	211.16	7.63	63.33	28.9
16-18	208.78	213.04	-4.26	28.23	13.5
19-21	246.22	220.85	25.37	9.33	3.8
22-24	270.77	256.01	14.76	7.75	2.9
25-27	265.98	268.05	-2.07	1.63	0.6
28-30	300.95	323.53	-22.58	0.00	0.0
31-33	323.60	347.73	-24.13	0.00	0.0
34-36	287.13	331.59	-44.46	0.00	0.0

Source: See Table 6-1.

Notes: The data presented in the first three columns are regression-adjusted estimates that control for differences of age, sex, race, education, prior work experience, household composition, site, and length of site operation. Averages are calculated for all members of the sample, including those with no employment in the covered period. Dollar amounts reported are gross earnings unadjusted for inflation.

^aThese data are not regression adjusted. No experimentals should have been in the program beyond month 21. The small number indicated as participating after that time reflects either data errors or failures by program operators to terminate individuals on schedule.

†Statistically significant at the 5 percent level.

Table 7-3: ex-offender sample

Table 7-3. Average Earnings per Month, by Length of Follow-Up: Ex-Offender Sample

Month After Enrollment in the Sample	Experimental Group Mean	Control Group Mean	Experimental-Control Differential	Mean Earnings From Supported Work ^a	
				Earnings	As Percent of Total Earnings of Experimentals
1-3	\$446.91	\$138.74	\$308.17†	\$383.82	85.9
4-6	385.74	185.15	200.59†	265.53	68.8
7-9	323.94	173.74	150.20†	191.17	59.0
10-12	276.56	221.37	55.19†	107.24	38.8
13-15	274.88	246.15	28.73	40.98	14.9
16-18	269.11	250.10	19.01	13.62	5.1
19-21	265.73	252.92	12.81	11.17	4.2
22-24	273.81	265.63	8.18	6.27	2.3
25-27	266.80	263.72	3.08	2.69	1.0
28-30	353.14	285.48	67.66	0.00	0.0
31-33	375.77	319.99	55.78	0.00	0.0
34-36	361.28	295.35	65.93	0.00	0.0

Source: See Table 7-1.

Notes: The data presented in the first three columns are regression-adjusted estimates that control for differences of age, sex, race, education, prior work experience, household composition, site, and length of site operation. Averages are calculated for all members of the sample, including those with no employment in the covered period. Dollar amounts reported are gross earnings unadjusted for inflation.

^aThese data are not regression adjusted. No experimentals should have been in the program beyond month 21. The small number indicated as participating after that time reflects either data errors or failures by program operators to terminate individuals on schedule.

†Statistically significant at the 5 percent level.

Your turn

Explain the outcome and units, the treatment-control comparison, the magnitude and persistence of the estimated impacts, and their uncertainty. Compare the results across target groups, considering their different needs, labour-market opportunities, and follow-up periods.

3. LaLonde: evaluating the evaluation methods

LaLonde wants to evaluate whether substituting a control group from survey or administrative data can yield the same estimate of the NSW impact as in the experiment.

The comparison groups come from the **Panel Study of Income Dynamics (PSID)** and the **Current Population Survey linked to Social Security earnings records (CPS-SSA)**. Different restrictions produce different comparison samples.

Table 3: the data

TABLE 3—ANNUAL EARNINGS OF NSW MALE TREATMENTS, CONTROLS, AND SIX CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND *CPS-SSA*

Year	Treatments	Controls	Comparison Group ^{a,b}					
			<i>PSID</i> -1	<i>PSID</i> -2	<i>PSID</i> -3	<i>CPS-SSA</i> -1	<i>CPS-SSA</i> -2	<i>CPS-SSA</i> -3
1975	\$3,066 (283)	\$3,027 (252)	19,056 ^a (272)	7,569 (568)	2,611 (492)	13,650 (73)	7,387 (206)	2,729 (197)
1976	\$4,035 (215)	\$2,121 (163)	20,267 (296)	6,152 (601)	3,191 (609)	14,579 (75)	6,390 (187)	3,863 (267)
1977	\$6,335 (376)	\$3,403 (228)	20,898 (296)	7,985 (621)	3,981 (594)	15,046 (76)	9,305 (225)	6,399 (398)
1978	\$5,976 (402)	\$5,090 (227)	21,542 (311)	9,996 (703)	5,279 (686)	14,846 (76)	10,071 (241)	7,277 (431)
Number of Observations	297	425	2,493	253	128	15,992	1,283	305

^aThe Comparison Groups are defined as follows: *PSID*-1: All male household heads continuously from 1975 through 1978, who were less than 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all men who were not working when surveyed in the spring of 1976; *PSID*-3: Selects from the *PSID*-1 group all men who were not working when surveyed in either spring of 1975 or 1976; *CPS-SSA*-1: All males based on Westat's criteria, except those over 55-years-old; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all males who were not working when surveyed in March 1976; *CPS-SSA*-3: Selects from the *CPS-SSA*-1 unemployed males in 1976 whose income in 1975 was below the poverty level.

^bAll earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. The number of observations refer only to 1975 and 1978. In the other years there are fewer observations. The sample of treatments is smaller than the sample of controls because treatments still in Supported Work as of January 1978 are excluded from the sample, and in the young high school target group there were by design more controls than treatments.

For 1978, treatment and control means are \$5,976 and \$5,090: a difference of **\$886**, in 1982 dollars. There are **297 treated men and 425 experimental controls**. The remaining columns show how different observational comparison groups can be, including in pre-programme earnings.

Table 5: the estimates

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		Controlling for All Observed Variables and Pre-Training Earnings (10)
		Pre-Training Year, 1975		Post-Training Year, 1978		Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)	
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)					
Controls	\$2,063 (325)	\$39 (383)	\$ – 21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)
<i>PSID</i> -1	\$2,043 (237)	– \$15,997 (795)	– \$7,624 (851)	– \$15,578 (913)	– \$8,067 (990)	\$425 (650)	– \$749 (692)	– \$2,380 (680)	– \$2,119 (746)	– \$1,228 (896)
<i>PSID</i> -2	\$6,071 (637)	– \$4,503 (608)	– \$3,669 (757)	– \$4,020 (781)	– \$3,482 (935)	\$484 (738)	– \$650 (850)	– \$1,364 (729)	– \$1,694 (878)	– \$792 (1024)
<i>PSID</i> -3	(\$3,322 (780)	(\$455 (539)	\$455 (704)	\$697 (760)	– \$509 (967)	\$242 (884)	– \$1,325 (1078)	\$629 (757)	– \$552 (967)	\$397 (1103)
<i>CPS-SSA</i> -1	\$1,196 (61)	– \$10,585 (539)	– \$4,654 (509)	– \$8,870 (562)	– \$4,416 (557)	\$1,714 (452)	\$195 (441)	– \$1,543 (426)	– \$1,102 (450)	– \$805 (484)
<i>CPS-SSA</i> -2	\$2,684 (229)	– \$4,321 (450)	– \$1,824 (535)	– \$4,095 (537)	– \$1,675 (672)	\$226 (539)	– \$488 (530)	– \$1,850 (497)	– \$782 (621)	– \$319 (761)
<i>CPS-SSA</i> -3	\$4,548 (409)	\$337 (343)	\$878 (447)	– \$1,300 (590)	\$224 (766)	– \$1,637 (631)	– \$1,388 (655)	– \$1,396 (582)	\$17 (761)	\$1,466 (984)

^a The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^b Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^c The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^d See Table 3 for definitions of the comparison groups.

Focus on **column 4 (unadjusted post-training differences)** and **column 5 (regression-adjusted differences)**, comparing the controls row with the *PSID* and *CPS-SSA* rows. The standard error of the unadjusted experimental estimate is \$476.

Changing the comparison group or adjustment can change the answer substantially. The question is whether the observed characteristics and adjustment adequately remove selection bias. We return to difference-in-differences later in the module.

Your turn

Explain why LaLonde might have considered the samples from the *PSID* and *CPS* to be good potential sources for counterfactuals for the treated groups in the NSW experiment.

Why the experimental estimates differ

The report's target-group estimates and LaLonde's pooled male estimates differ in time units, prices, and sample construction. His annual-earnings sample has a larger share observed at the later follow-up, when the full-sample effects were larger:

“First, the partial sample contains a much larger proportion of males who completed a 36 month interview than does the full sample. Second, the training effects for the full sample are much larger for the 36 month interview than for the 27 month interview.”

—LaLonde (1984), Chapter 2, footnote 5, pp. 81–82.

He also acknowledges that constructing the sample selected treated people with higher subsequent earnings:

“Thus the process of selecting individuals with complete earnings in 1975 and 1978 caused us to select treatments, but not controls, with higher earnings after the 27th and 36th month interviews. Despite this potential selection bias, this paper treats the partial sample as if it was an experimental data set.”

—LaLonde (1984), Chapter 2, footnote 5, p. 82.

His Table 2.4 compares the full male sample with the annual-earnings subsample. The discussion on pp. 41–42 also recognises the problem of excluding treated people still in Supported Work in January 1978. **Random assignment in the original experiment does not guarantee that a subsequently selected sample preserves that comparison.**

4. Exact matching and the assumptions

Let X_i collect characteristics measured before treatment, such as schooling, age, and earlier earnings. For the ATT, the central assumption is

$$Y_i^0 \perp D_i \mid X_i.$$

Under **unconfoundedness**, or **selection on observables**, participation provides no further information about untreated potential earnings once we condition on X_i . Unobserved determinants of earnings can exist, but they must not remain associated with participation conditional on X_i in a way that violates this assumption.

We also need **common support**:

$$P(D_i = 0 \mid X_i = x) > 0 \quad \text{for } x \text{ occurring among the treated.}$$

We need to estimate what would have happened to the treated individuals if they had not received treatment. Retain the Week 1 assumptions about well-defined treatment and interference.

Under these assumptions,

$$E[Y_i^0 \mid D_i = 1, X_i = x] = E[Y_i \mid D_i = 0, X_i = x].$$

This step lets observed controls supply the missing counterfactual.

Within-cell comparisons

With discrete covariates, form cells of people with exactly the same characteristics. Compare treated and control means within each cell:

$$\hat{\tau}(x) = \bar{Y}_{D=1, X=x} - \bar{Y}_{D=0, X=x}.$$

For the ATT, average these differences using the treated sample’s cell shares. If N_{1x} is the treated count in cell x and N_1 the total treated count,

$$\hat{\tau}_{\text{ATT}} = \sum_x \frac{N_{1x}}{N_1} \hat{\tau}(x).$$

For example, if 75% of the treated are in a cell with an effect of \$1,000 and 25% in a cell with an effect of \$2,000, the ATT estimate is \$1,250. Equal weighting of the two cells would answer a different question.

Exact matching becomes data intensive: ten binary covariates generate 1,024 possible cells. Continuous earnings make exact matches rarer still. Empty control cells cannot supply a counterfactual.

5. The propensity score

The propensity score is

$$p(X_i) = P(D_i = 1 | X_i).$$

It summarises a vector of characteristics in one number. Rosenbaum and Rubin (1983) establish the **balancing property**

$$D_i \perp X_i | p(X_i),$$

and show that

$$(Y_i^0, Y_i^1) \perp D_i | X_i \implies (Y_i^0, Y_i^1) \perp D_i | p(X_i).$$

If conditioning on the characteristics makes treatment unconfounded, conditioning on the true score does too. For our ATT argument, the corresponding result for Y_i^0 is sufficient.

The score therefore lets us condition on a single number. The balancing result concerns the distribution of observed characteristics within groups with the same true score.

We will estimate the score using logit, introduced below. The theorem applies to the true score; with $\hat{p}(X_i)$, examine whether the resulting adjustment actually balances covariates. Our aim is to balance the observed characteristics. Almost perfectly predicted participation may signal a lack of comparable controls.

Use characteristics measured before treatment. Subsequent employment may itself be affected by the programme.

Estimating a probability: the LPM and logit

Consider a simulated example of university attendance. Let $D_i = 1$ if individual i attends university and $D_i = 0$ otherwise. Let X_i be a family-income index, with larger values representing higher incomes. Write the probability of attendance as

$$p_i = P(D_i = 1 | X_i).$$

The **linear probability model (LPM)** specifies this probability as a straight-line relationship:

$$p_i = \beta_0 + \beta_1 X_i.$$

We estimate its coefficients using OLS, as in Week 1. A fitted straight line can extend below zero or above one. Those predictions cannot be interpreted as probabilities.

A **logit model**, also called logistic regression, instead specifies

$$p_i = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 X_i))}.$$

Here $\exp(z) = e^z$. The expression $\gamma_0 + \gamma_1 X_i$ is linear, but the transformation makes the probability curved and keeps it between zero and one. When the expression is zero, the probability is 0.5; when it is 2, the probability is about 0.88; when it is -2 , the probability is about 0.12.

A latent-variable interpretation

We can obtain this probability model from an underlying, unobserved variable:

$$D_i^* = \gamma_0 + \gamma_1 X_i + \epsilon_i,$$

with the rule that $D_i = 1$ when $D_i^* > 0$ and $D_i = 0$ otherwise. The word **latent** means unobserved. Think of D_i^* as an underlying inclination to attend university. We observe whether the threshold is crossed. The disturbance ϵ_i represents other influences on the decision.

Assume that ϵ_i is independent of X_i and has a **standard logistic distribution**. Its cumulative distribution function gives the probability that it is below a number t :

$$F(t) = P(\epsilon_i \leq t) = \frac{1}{1 + \exp(-t)}.$$

Set $z_i = \gamma_0 + \gamma_1 X_i$. Then

$$\begin{aligned} P(D_i = 1 \mid X_i) &= P(D_i^* > 0 \mid X_i) \\ &= P(\epsilon_i > -z_i \mid X_i) \\ &= 1 - F(-z_i) \\ &= \frac{1}{1 + \exp(-z_i)}. \end{aligned}$$

The final equality follows from the symmetry of the logistic distribution. Its scale is set to one: multiplying the whole latent relationship by a positive constant would leave the zero-threshold decision unchanged.

Generate data, estimate, and plot

For the simulation, take $\gamma_0 = 0$, $\gamma_1 = 1.5$, and draw X_i uniformly between -3 and 3 . Thus the true probability is

$$p_i = \frac{1}{1 + \exp(-1.5X_i)}.$$

Conditional on income, attendance is a zero-or-one draw with probability p_i of being one. We write this as $D_i \mid X_i \sim \text{Bernoulli}(p_i)$. For someone with $p_i = 0.8$, attendance occurs in about 80% of repeated draws. Drawing a uniform random number U_i between zero and one and setting $D_i = 1$ when $U_i < p_i$ would generate the same conditional distribution.

The code below uses the latent-variable construction. `rng.logistic()` draws the disturbances, and `(d_star > 0).astype(int)` converts the threshold condition to zeros and ones. We fit both models to exactly the same observations.

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

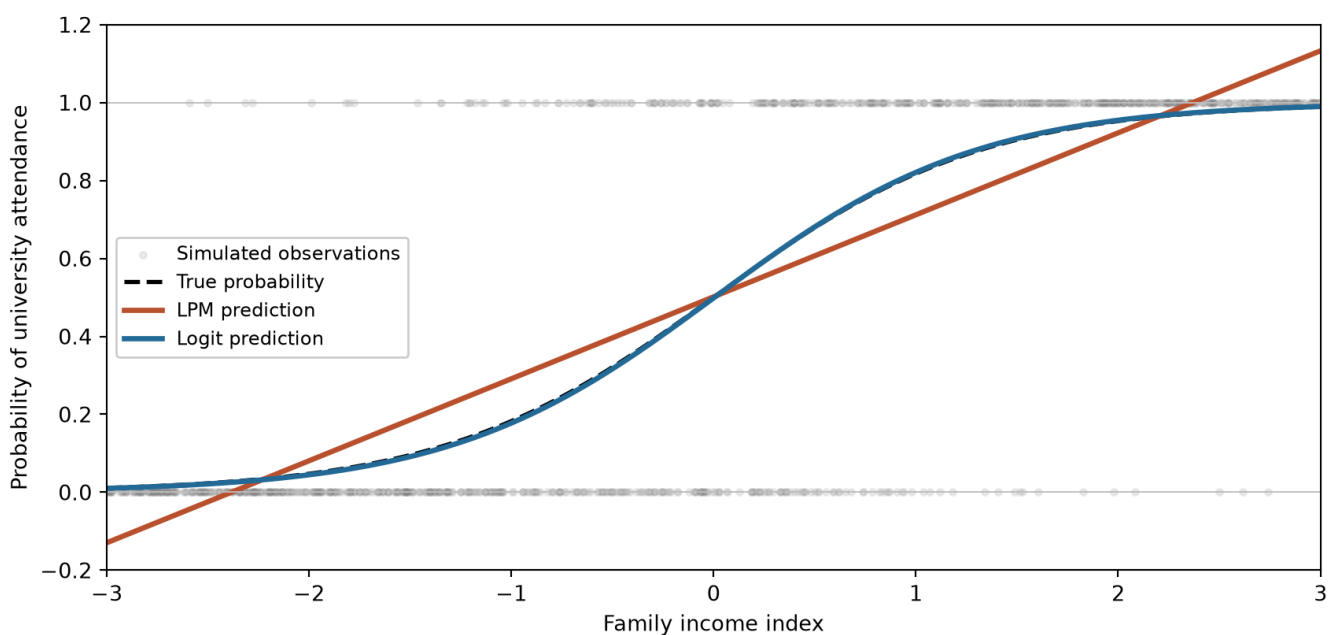
rng = np.random.default_rng(5001)
n = 1000
x = rng.uniform(-3, 3, n)
epsilon = rng.logistic(loc=0, scale=1, size=n)
d_star = 1.5 * x + epsilon
d = (d_star > 0).astype(int)
simulated = pd.DataFrame({"attend": d, "income": x})

lpm = smf.ols("attend ~ income", data=simulated).fit(cov_type="HC1")
logit = smf.logit("attend ~ income", data=simulated).fit(disps=False)

grid = pd.DataFrame({"income": np.linspace(-3, 3, 400)})
grid["true_probability"] = 1 / (1 + np.exp(-1.5 * grid["income"]))
grid["lpm_prediction"] = lpm.predict(grid)
grid["logit_prediction"] = logit.predict(grid)

fig, ax = plt.subplots(figsize=(9, 4.4))
ax.scatter(simulated["income"], simulated["attend"],
           color="0.5", alpha=0.12, s=9, label="Simulated observations")
ax.plot(grid["income"], grid["true_probability"],
        color="black", linestyle="--", linewidth=2, label="True probability")
ax.plot(grid["income"], grid["lpm_prediction"],
        color="#bc542f", linewidth=2.5, label="LPM prediction")
ax.plot(grid["income"], grid["logit_prediction"],
        color="#236b99", linewidth=2.5, label="Logit prediction")
ax.axhline(0, color="0.75", linewidth=0.7)
ax.axhline(1, color="0.75", linewidth=0.7)
ax.set(xlabel="Family income index", ylabel="Probability of university attendance",
       xlim=(-3, 3), ylim=(-0.2, 1.2))
ax.legend(loc="center left", fontsize=9, framealpha=0.95)
fig.tight_layout()
plt.show()

```



In this simulation the true curve is known because we chose it. The estimated logit curve is close to it, while the LPM's straight line gives predictions below zero and above one at the ends of the observed income range. The faint points show the actual zero-or-one observations. This example illustrates the models' different shapes; real data need not follow a logistic curve.

What does the estimation code do?

`smf.ols()` fits the LPM. `smf.logit()` specifies the logit model, and `.fit()` estimates its coefficients by **maximum likelihood**: it chooses coefficients that make the observed pattern of attendance and non-attendance most likely under the model. The computer searches for these coefficients numerically. `disp=False` suppresses its routine progress message. The formula includes an intercept automatically.

`.predict(grid)` uses the estimated coefficients to calculate fitted probabilities at the income values in `grid`. Logit coefficients describe changes in the linear expression inside the transformation; they are not percentage-point effects as in the LPM. For propensity scores, we use the resulting predicted probabilities.

From this example to the propensity score

In PS1, replace attendance with `treat` and income with the listed pre-treatment characteristics. Fit the model to the 185 NSW treated men combined with **one observational comparison sample**. Here the model describes membership in the treated group within that combined sample. Refit it for the other comparison sample.

If `sample` contains that combined sample, the first PS1 model is:

```
import statsmodels.formula.api as smf

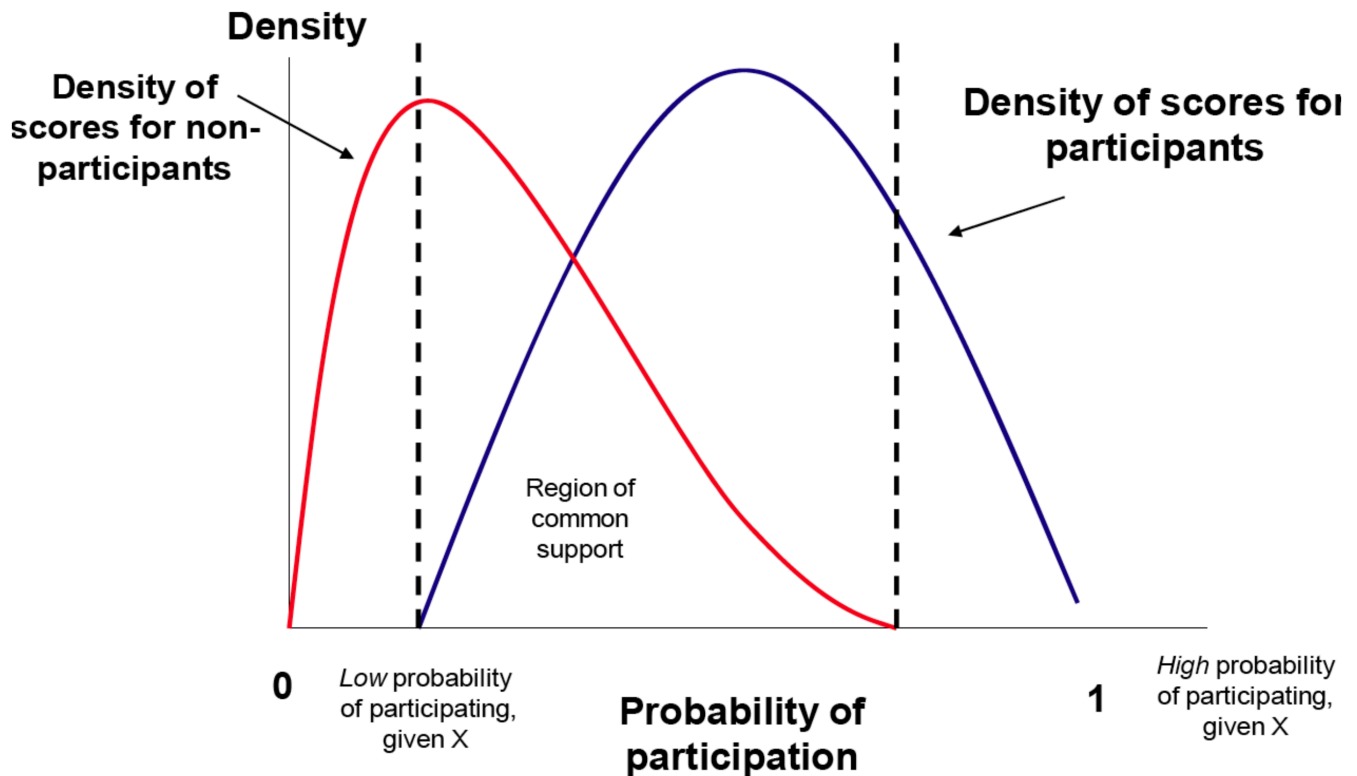
score_model = smf.logit(
    "treat ~ age + I(age ** 2) + education + black + hispanic "
    "+ married + nodegree + I(re75 / 1000)",
    data=sample
).fit(disp=False, maxiter=200)

sample["pscore"] = score_model.predict(sample)
```

Each row now has an estimated probability between zero and one. `I(age ** 2)` includes age squared and `I(re75 / 1000)` measures earnings in thousands of dollars. `maxiter=200` allows up to 200 estimation steps. If estimation has not converged, the numerical search has not found a satisfactory solution; check the data and model before using its predictions. In Question 5, add `I(re74 / 1000)` to the formula, then estimate and predict again.

Python reference: [statsmodels Logit documentation](#).

6. Common support and Dehejia–Wahba



Overlap means that comparable observations exist. It can be weak even when the formal probability condition holds. If many treated observations have scores near one and few controls do, estimates may depend on a handful of controls.

Excluding controls far outside treated support can focus the comparison. Excluding **treated** observations changes the population for which the ATT is estimated. Explain how exclusions change the group whose effect is being estimated.

Dehejia and Wahba's sample and experimental benchmark

Dehejia and Wahba use a smaller male subsample with earnings observed in 1974 and 1975: **185 treated men and 260 experimental controls**. Their unadjusted benchmark is **\$1,794 (SE \$633)**. It differs from \$886 because it uses a different sample. **PS1 uses this 185/260 subsample throughout**. You will reproduce the DW experimental benchmark, then hold the sample fixed while comparing IPW with 1975 earnings alone and with both 1974 and 1975 earnings.

Figure 1

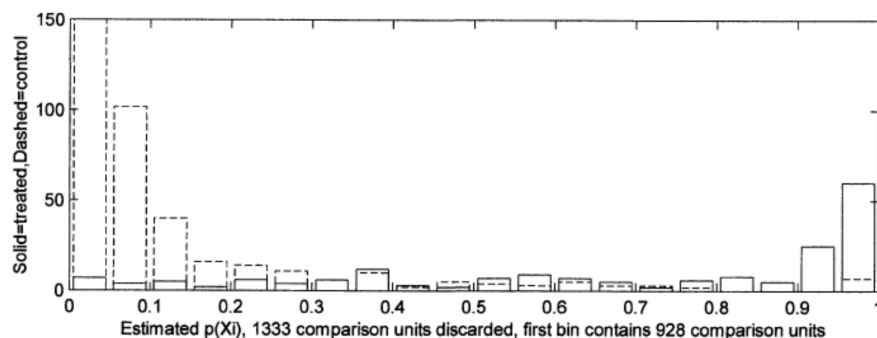


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8–.85, .85–.9, and .9–.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

Figure 2

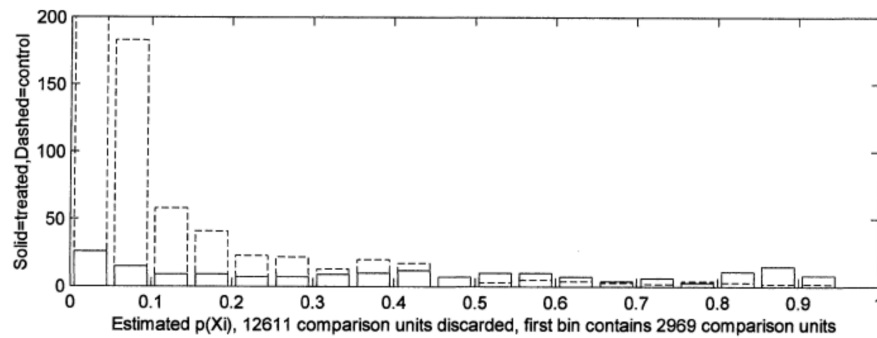


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45–.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Your turn

Identify which distribution represents the treated, where comparison observations are concentrated, and where treated people have few potential comparisons. Read the captions about discarded observations. Explain how the different sample sizes affect your interpretation of the histograms.

Table 3

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
	(1) Unadjusted	(2) Adjusted ^a	Quadratic in score ^b (3)	Stratifying on the score			Matching on the score	
				(4) Unadjusted	(5) Adjusted	(6) Observations ^c	(7) Unadjusted	(8) Adjusted ^d
NSW	1,794 (633)	1,672 (638)						
PSID-1 ^e	−15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 ^f	−3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 ^f	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 ^g	−8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 ^g	−3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 ^g	−635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

^a Least squares regression: RE78 on a constant, a treatment indicator, age, age², education, no degree, black, Hispanic, RE74, RE75.

^b Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

^c Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

^d Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)]. Propensity scores are estimated using the logistic model, with specifications as follows:

^e PSID-1: $\text{Prob}(T_i = 1) = F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{married}, \text{no degree}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE75}, \text{RE74}^2, \text{RE75}^2, \text{u74} * \text{black})$.

^f PSID-2 and PSID-3: $\text{Prob}(T_i = 1) = F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE74}^2, \text{RE75}, \text{RE75}^2, \text{u74}, \text{u75})$.

^g CPS-1, CPS-2, and CPS-3: $\text{Prob}(T_i = 1) = F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE75}, \text{u74}, \text{u75}, \text{education} * \text{RE74}, \text{age}^3)$.

Your turn

Concentrate on the **NSW benchmark, column 1, and columns 7–8**. Explain how the estimates change and how uncertain they remain.

The table's note (d) describes weights reflecting how often a control is matched. Our problem set uses the propensity-score weights introduced below.

Investigate the assignment process and data comparability. A comparison group from a different labour market, measured with different questions or lacking important work-history information, may remain problematic after adjustment. Assess the selection-on-observables assumption in the setting of each study.

Why not simply match on the estimated score?

The propensity score's balancing property does not guarantee that matching people with similar **estimated** scores produces well-balanced samples. Similar scores can conceal substantial differences in individual characteristics. Discarding observations to improve score matches can sometimes worsen covariate balance: this is a central criticism in [King and Nielsen \(2019\)](#). Their critique concerns propensity-score matching; it does not invalidate the score's use for weighting or the population balancing result discussed above.

For our weighting exercise, **check balance in the actual covariates and inspect the weights**. Large weights can make estimates unstable.

7. Inverse-probability weighting for the ATT

We can use weights to construct a comparison group. **For the ATT**, keep treated observations equally weighted and reweight controls to resemble them:

$$w_i = \begin{cases} 1 & D_i = 1, \\ \frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)} & D_i = 0. \end{cases}$$

These are **ATT weights**, often called odds weights for controls. They estimate the average treatment effect for those who received treatment by making the controls resemble the treated group.

ATT and ATE weights

ATE weights estimate the average treatment effect for the whole population, including people who did not receive treatment. Weight each observation by the inverse of its estimated probability of receiving its observed treatment:

$$w_i^{\text{ATE}} = \begin{cases} \frac{1}{\hat{p}(X_i)} & D_i = 1, \\ \frac{1}{1 - \hat{p}(X_i)} & D_i = 0. \end{cases}$$

Both groups are reweighted to represent the population's distribution of observed characteristics.

For NSW, our question concerns the effect on programme participants. ATT weights make the survey controls resemble those participants and address the effect we compare with the experimental benchmark. Applying ATE weights to the combined NSW-survey sample would also reweight the treated observations towards the survey respondents' characteristics. Some of those characteristics are scarcely represented among NSW participants, so a few treated observations could receive very large weights. The composition of that combined sample also depends on which survey sample we append and its size.

ATE weights are useful when the question concerns a clearly defined broader population and there is adequate overlap between its treated and control groups.

Why these weights?

Within a covariate cell, the treated share is proportional to $p(x)$ and the control share to $1 - p(x)$. To make controls represent the treated, multiply their contribution by the ratio $p(x) / (1 - p(x))$.

More formally, Bayes' rule gives

$$\frac{f(x | D = 1)}{f(x | D = 0)} = \frac{p(x)}{1 - p(x)} \frac{P(D = 0)}{P(D = 1)}.$$

The last factor is constant across observations and cancels when we normalise the weighted control mean.

A control with score 0.2 receives weight $0.2/0.8 = 0.25$. One with score 0.8 receives $0.8/0.2 = 4$, sixteen times as much. At 0.95, the weight is 19. Large weights make the connection to weak common support concrete.

From OLS to weighted least squares

In PP5000, we derived OLS using the **method of moments**. The same coefficient estimates minimise the sum of squared residuals:

$$\min_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 D_i)^2.$$

Every observation receives equal weight. Here b_0 and b_1 are candidate coefficient values; the values that minimise the sum are the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Weighted least squares instead minimises

$$\min_{b_0, b_1} \sum_{i=1}^n w_i (Y_i - b_0 - b_1 D_i)^2.$$

An observation with a larger weight contributes more to this sum.

IPW as a regression

Estimate

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

by **weighted least squares**, choosing coefficients to minimise

$$\min_{b_0, b_1} \sum_i w_i (Y_i - b_0 - b_1 D_i)^2.$$

Here regression calculates two group means with the desired weights. Under the identification assumptions, its treatment coefficient estimates the ATT.

Write the fitted control mean as $\mu_0 = \beta_0$ and treated mean as $\mu_1 = \beta_0 + \beta_1$. The objective becomes

$$\sum_{i:D_i=0} w_i (Y_i - \mu_0)^2 + \sum_{i:D_i=1} w_i (Y_i - \mu_1)^2.$$

Each part is minimised by its group mean:

$$\hat{\beta}_0 = \frac{\sum_{i:D_i=0} w_i Y_i}{\sum_{i:D_i=0} w_i}, \quad \hat{\beta}_0 + \hat{\beta}_1 = \frac{1}{N_1} \sum_{i:D_i=1} Y_i.$$

Therefore,

$$\hat{\beta}_1 = \frac{1}{N_1} \sum_{i:D_i=1} Y_i - \frac{\sum_{i:D_i=0} w_i Y_i}{\sum_{i:D_i=0} w_i}.$$

Normalising by the sum of control weights makes the second term an average. Multiplying all control weights by a common positive constant does not change this mean or treatment coefficient.

The equivalence applies to a regression with **only an intercept and treatment indicator**. Adding regressors changes the coefficient; it is no longer simply this difference in means.

Python template

This template uses a DataFrame called `df` containing the NSW treated observations and one survey comparison group, with the columns listed below.

If `nsw_treated` and `survey_controls` are the DataFrames you imported from the two files, stack their rows as follows. Use the same variable names in both DataFrames.

```
import pandas as pd

df = pd.concat([nsw_treated, survey_controls], ignore_index=True)
print(df.groupby("treat").size())
```

`pd.concat` stacks the observations and matches columns by name. `ignore_index=True` gives the combined data a new row index. The count checks how many treated and comparison observations it contains.

Use one analysis sample for the comparisons and report observations lost to missing values. Keep `re74` in that sample from the start so Question 5 can add it to the propensity-score model without changing the observations. The initial models below use 1975 earnings only.

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf

columns = [
    "treat", "re78", "re74", "re75", "age", "education",
    "married", "nodegree", "black", "hispanic"
]
sample = df[columns].dropna().copy()
print("Observations omitted:", len(df) - len(sample))

unadjusted = smf.ols(
    "re78 ~ treat", data=sample
).fit(cov_type="HC1", use_t=True)

adjusted = smf.ols(
    "re78 ~ treat + age + I(age ** 2) + education + married +
    nodegree + black + hispanic + I(re75 / 1000)",
    data=sample
).fit(cov_type="HC1", use_t=True)
```

```

# Model treatment using pre-treatment covariates
score_model = smf.logit(
    "treat ~ age + I(age ** 2) + education + married "
    "+ nodegree + black + hispanic + I(re75 / 1000)",
    data=sample
).fit(displ=False, maxiter=200)

sample["pscore"] = score_model.predict(sample)
if not score_model.mle_retvals["converged"]:
    raise ValueError("The propensity-score model did not converge.")
if not sample["pscore"].between(0, 1, inclusive="neither").all():
    raise ValueError("Inspect the score model and common support.")

sample["att_weight"] = 1.0
controls = sample["treat"].eq(0)
sample.loc[controls, "att_weight"] = (
    sample.loc[controls, "pscore"]
    / (1 - sample.loc[controls, "pscore"])
)

# ATT-weighted regression: intercept and treatment only
weighted = smf.wls(
    "re78 ~ treat",
    data=sample,
    weights=sample["att_weight"]
).fit(cov_type="HC1", use_t=True)

print(weighted.summary())

# Check the difference-in-means interpretation
treated_mean = sample.loc[~controls, "re78"].mean()
control_mean = np.average(
    sample.loc[controls, "re78"],
    weights=sample.loc[controls, "att_weight"]
)
print("Weighted difference:", treated_mean - control_mean)
print("WLS coefficient:", weighted.params["treat"])

```

Plot the propensity-score distributions

After estimating `pscore`, use the same bins and axes for both groups. `density=True` makes each histogram have total area one, allowing us to compare distribution shapes when group sizes differ.

```

import matplotlib.pyplot as plt

bins = np.linspace(0, 1, 21)
fig, axes = plt.subplots(1, 2, figsize=(9, 3.5), sharex=True, sharey=True)

axes[0].hist(sample.loc[sample["treat"].eq(1), "pscore"],
             bins=bins, density=True, edgecolor="white")
axes[0].set_title("NSW treated")
axes[0].set_xlabel("Estimated propensity score")
axes[0].set_ylabel("Density")

axes[1].hist(sample.loc[sample["treat"].eq(0), "pscore"],
             bins=bins, density=True, edgecolor="white")
axes[1].set_title("Survey comparison group")
axes[1].set_xlabel("Estimated propensity score")

```

```
fig.tight_layout()
```

In an executable Quarto Python cell, the figure appears below the code. Compare the ranges of scores and where observations in each group are concentrated.

In PS1, add `I(re74 / 1000)` to the score model in Question 5, keeping the observations fixed. Fit the score separately for each observational comparison sample.

Pass the weights in the squared-residual criterion to `smf.wls`. See the [statsmodels WLS documentation](#).

Standard errors

Heteroskedasticity means disturbance variance can differ across observations. Our treatment weights make the groups comparable on observed characteristics. Use **heteroskedasticity-consistent standard errors**, as in `.fit(cov_type="HC1", use_t=True)`, to allow the disturbance variance to differ across observations.

These robust WLS errors treat estimated propensity-score weights as fixed. A full calculation of uncertainty also accounts for estimating the propensity score.

8. Diagnose the comparison

Inspect score distributions and the largest control weights before interpreting the coefficient. If Python reports a warning or cannot estimate the model, check your data import and regression code. You can ask Claude to explain the message and help diagnose the problem. Make sure you understand any suggested changes before using them.

Compare **unweighted and weighted covariate balance**. The weighted control mean of a covariate is

$$\bar{X}_{0,w} = \frac{\sum_{i:D_i=0} w_i X_i}{\sum_{i:D_i=0} w_i}.$$

A standardised mean difference helps compare covariates in different units. One convention is

$$\text{SMD}_w = \frac{\bar{X}_1 - \bar{X}_{0,w}}{\sqrt{(s_1^2 + s_0^2) / 2}},$$

where s_1^2 and s_0^2 are the **unweighted** within-group variances. Keep the denominator fixed before and after weighting, so changes reflect improved mean balance.

Here `sample` continues to contain both groups. `treated` and `control_group` are the two subsets used to calculate balance.

```
covariates = [
    "age", "education", "married", "nodegree",
    "black", "hispanic", "re75"
]
treated = sample.loc[sample["treat"].eq(1)]
control_group = sample.loc[sample["treat"].eq(0)]
rows = []

for variable in covariates:
    mean1 = treated[variable].mean()
    mean0 = control_group[variable].mean()
```

```

mean0w = np.average(
    control_group[variable], weights=control_group["att_weight"]
)
scale = np.sqrt(
    (treated[variable].var() + control_group[variable].var()) / 2
)
rows.append({
    "Covariate": variable,
    "Treated mean": mean1,
    "Control mean": mean0,
    "Weighted control mean": mean0w,
    "SMD before": (mean1 - mean0) / scale if scale > 0 else np.nan,
    "SMD after": (mean1 - mean0w) / scale if scale > 0 else np.nan
})

balance = pd.DataFrame(rows).set_index("Covariate")
print(balance.round(3))
print(control_group["att_weight"].describe(
    percentiles=[0.5, 0.9, 0.95, 0.99]
))

```

Mean balance can conceal distributional differences. Inspect earlier earnings and the frequency of zero earnings. Poor balance means the adjustment has not achieved its observed-data objective; good balance still leaves unconfoundedness to be defended.

9. Preparing for Problem Set 1

Compare OLS and ATT weighting with the experimental benchmark using **two observational comparison samples**. Explain why the estimates differ.

Your Quarto document should combine code, labelled output, and prose explaining the comparison group, sample, units, assumptions, overlap, balance, and results. The [assignment](#) specifies the data and required tasks.

References

- Manpower Demonstration Research Corporation (1980). *Summary and Findings of the National Supported Work Demonstration*. Cambridge, MA: Ballinger. [Report](#).
- LaLonde, Robert J. (1984). *Evaluating the Econometric Evaluations of Training Programs with Experimental Data*. Industrial Relations Section Working Paper 183, Princeton University. [Record](#).
- LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604–620. [Paper](#).
- Rosenbaum, Paul R., and Donald B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55. [DOI](#).
- Dehejia, Rajeev H., and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–1062. [Paper](#).
- King, Gary, and Richard Nielsen (2019). "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27(4): 435–454. [DOI](#).