CHAPTER 11

Case Study: An Experimental Evaluation of a Labor Market Program

11.1 INTRODUCTION

In this chapter we illustrate some of the methods discussed in the previous chapters in an application. The application involves a social program designed to improve labor market outcomes for individuals with relatively poor skills and labor market histories: the Saturation Work Initiative Model (SWIM) program in San Diego, evaluated during the period 1985–1987. As is typical, a substantial amount of background information on the individuals in the program was collected, including demographics and recent labor market histories, allowing us to investigate heterogeneity in the effects of the program. The outcomes of interest, post-program earnings and employment records, are either discrete or mixed discrete-continuous, suggesting that constant additive treatment-effect assumptions are typically not plausible.

Using these data we will calculate Fisher exact p-values for sharp null hypotheses and construct Neyman large-sample confidence intervals. We will also discuss, in detail, regression and model-based inferences for various average treatment effects, using the covariates to increase precision as well as to estimate treatment effects for subpopulations. We emphasize the model selection choices and the various other decisions faced by researchers.

11.2 THE SAN DIEGO SWIM PROGRAM DATA

SWIM primarily targeted women who were eligible for Aid to Families with Dependent Children (AFDC), with children at least six years old (although, as the summary statistics show, there was a substantial proportion of women with younger children, a small number of men, and some individuals with no children). It was a mandatory program, with fairly strong participation enforcement, and provided a sequence of group job search, unpaid work experience, education, and job skills training. Compared to similar programs in other locations, it had broad coverage, with the intention to reach a wide range of individuals eligible for AFDC, including those who may not have participated in such assistance programs. The average cost of participating in this program was $919 per trainee, paid for by the local authorities. The participants faced no direct

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.1. *Summary Statistics San Diego SWIM Data*

Variable		All ($N = 3211$)		Controls ($N_c = 1607$)		Treated ($N_t = 1604$)	
		Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
Pre-treatment variables							
female	female	0.91	(0.28)	0.92	(0.28)	0.91	(0.28)
agege35	(age $\geq$ 35)	0.46	(0.50)	0.46	(0.50)	0.46	(0.50)
hsdip	(high school diploma)	0.56	(0.50)	0.56	(0.50)	0.56	(0.50)
nevmar	(never married)	0.30	(0.46)	0.30	(0.46)	0.30	(0.46)
divwid	(divorced or widowed)	0.37	(0.48)	0.37	(0.48)	0.36	(0.48)
numchild	(number of children)	1.76	(1.08)	1.76	(1.07)	1.76	(1.10)
chldlt6	(children younger than 6)	0.10	(0.30)	0.10	(0.31)	0.10	(0.29)
af-amer	(African-American)	0.42	(0.49)	0.43	(0.49)	0.42	(0.49)
hisp	(Hispanic)	0.25	(0.44)	0.25	(0.43)	0.26	(0.44)
earnyrm1	(earnings year minus 1)	1.57	(3.54)	1.60	(3.56)	1.53	(3.51)
empyrm1	(positive earnings year minus 1)	0.39	(0.49)	0.40	(0.49)	0.39	(0.49)
Outcomes variables							
earnyr1	(earnings year 1)	1.85	(3.78)	1.69	(3.76)	2.02	(3.80)
empyr1	(positive earnings year 1)	0.46	(0.50)	0.40	(0.49)	0.52	(0.50)
earnyr2	(earnings year 2)	2.57	(5.08)	2.26	(4.68)	2.89	(5.44)
empyr2	(positive earnings year 2)	0.45	(0.50)	0.40	(0.49)	0.49	(0.50)

expenses for the program, although there are likely to have been indirect costs, such as child care and travel expenses. The evaluation started in 1985. Eligible individuals enrolled in the study were randomized to receive training or not. The randomization did use demographics and labor market histories. This program is typical of many labor market programs in the 1980s and 1990s, a substantial number of which were evaluated using randomized experiments. The general emphasis on experimental evaluations around this time was motivated by research (most notably a paper by Lalonde published in 1986, whose data we use in other chapters) that had concluded that non-experimental evaluations (in practice with analyses limited to linear covariance adjustment or regression methods) were often unable to replicate experimental results, and therefore claimed that non-experimental evaluations were not credible in these settings. See the notes at the end of this chapter for more discussion on this topic.

Table 11.1 presents some summary statistics for this data set. We have information on $N = 3{,}211$ individuals, with $N_t = 1{,}604$ randomly assigned to receive the training, and the remaining $N_c = 1{,}607$ assigned to the control group, which was not to receive any training as part of the SWIM program. Individuals in the control group had no access to SWIM program services but may have had access to other, possibly similar, services outside of the SWIM program. This is a common problem with social programs, where individuals assigned to the control group often have access to related programs. This feature implies that the effects should be interpreted as the effect of participating in the program versus being denied access to this particular program, rather than as the effect of participating versus not participating in any job-training program.

There are two sets of pre-treatment variables. First there are some covariates measuring individual-level background characteristics. These pre-treatment variables include whether the individual had a high school diploma, was female (`female`), was at least 35 years old (`agege35`), had a high school diploma (`hsdip`), had never married (`nevmar`), and was divorced or widowed (`divwid`), the number of children (`numchild`); whether any children were present in the household who were younger than six (`chldlt6`); and whether the individual was African-American (`af-amer`) or Hispanic (`hisp`). Second, there are records for earnings for the year prior to the randomization. We use both the actual earnings measure (`earnyrm1`) and an indicator for positive earnings in this pre-randomization year (`empyrm1`). The outcome variables of interest are total earnings in the first and second year post-randomization (`earnyr1` and `earnyr2`) and indicators for these earnings being positive. For these covariates and the outcome variables, means and standard deviations for the entire sample, as well as means and standard deviations by treatment status, are displayed in Table 11.1. Notice that approximately 60% of the participants have no earning the year prior to the assignment, suggesting that simple gain scores may not be particularly helpful. All earnings variables are yearly earnings, measured in thousands of dollars.

11.3 FISHER'S EXACT P-VALUES

First we analyze the experimental data using Fisher's exact p-value approach discussed in Chapter 5. We focus on tests of the null hypothesis that there is no effect of the program for any individual:

$$H_0: \ Y_i(0) = Y_i(1), \quad \text{for } i = 1, \ldots, N.$$

We calculate the p-values for tests of this null hypothesis for a variety of test statistics using the first and second year post-program earnings (`empyr1` and `empyr2`) as the outcomes. We analyze the full sample and, separately, the subsamples created by whether individuals had graduated from high school. Table 11.2 contains all the p-values discussed in the text. Although for illustrative purposes we calculate a large number of p-values, we should note that the formal interpretation of each holds for one p-value at a time.

Our primary p-value is based on the difference in ranks in first year post-program earnings. As before, we define the normalized rank as:

$$R_i = \sum_{i'=1}^{N} \mathbf{1}_{Y_{i'}^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{i'=1}^{N} \mathbf{1}_{Y_{i'}^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N+1}{2}.$$

Then the rank-based test statistic is

$$T^{\text{rank}} = \left|\overline{R}_{\text{t}} - \overline{R}_{\text{c}}\right|,$$

where $\overline{R}_{\text{t}}$ and $\overline{R}_{\text{c}}$ are the average ranks in the treatment and control groups respectively. The average rank is higher for individuals in the treatment group than for individuals in

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.2. ***P-Values for Fisher Exact Tests on San Diego SWIM Data (based on 1,000,000 draws from randomization distribution)***

Post-Program Earnings	Statistic	All (3,211)	No High School (1,409)	High School (1,802)
Year 1	T^{rank}	< 0.0001	< 0.0001	0.0014
	$T^{\text{rank}-\text{gain}}$	< 0.0001	< 0.0001	0.0001
	T^{dif}	0.0131	0.0051	0.1967
Year 2	T^{rank}	< 0.0001	0.0017	< 0.0001
	$T^{\text{rank}-\text{gain}}$	< 0.0001	0.0020	0.0002
	T^{dif}	0.0004	0.0980	0.0018

the control group, leading to a p-value less than 0.0001, strong evidence against the null hypothesis of no effect of the treatment.

For comparison purposes, we report p-values for two other statistics. The first of these exploits the additional information in the form of the covariates. Specifically, because we have values for earnings prior to the program, we may wish to base the test statistic on the rank of the gains, rather than the rank of the level of earnings. Let X_i denote the level of prior earnings. Then the rank of the gains is defined as

$$R'_i = \sum_{i'=1}^{N} \mathbf{1}_{Y_{i'}^{\text{obs}} - X_{i'} < Y_i^{\text{obs}} - X_i} + \frac{1}{2}\left(1 + \sum_{i'=1}^{N} \mathbf{1}_{Y_{i'}^{\text{obs}} - X_{i'} = Y_i^{\text{obs}} - X_i}\right) - \frac{N+1}{2}.$$

Then the rank-based test statistic is

$$T^{\text{rank,gain}} = \left|\overline{R'}_{\text{t}} - \overline{R'}_{\text{c}}\right|,$$

where $\overline{R'}_{\text{t}}$ and $\overline{R'}_{\text{c}}$ are the average ranks of the gain in the treatment and control groups respectively. The p-values based on this statistic are similar to those based on the simple rank statistic. In both cases the evidence against the null is strong for the full sample and for the subsamples based on whether the individuals have a high school degree or not.

The third statistic is the widely (perhaps too widely) used difference in means of the observed outcomes:

$$T^{\text{dif}} = \left|\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right|.$$

Here the evidence against the null hypothesis is statistically significant at conventional levels in most cases, although not quite as strong as for the rank-based tests. The reason appears to be that the distribution of the outcome is heavily skewed. About 50% of the individuals have positive earnings in either Year 1 or Year 2 post-treatment. Figures 11.1 and 11.2 present histograms of the level of earnings and its logarithm, for those with positive earnings. For such distributions, rank-based tests tend to be more sensitive to violations of the null hypothesis of no effect of the treatment than tests based on averages of the levels.

In principle, we can also use sequences of Fisher tests to create Fisher intervals as described in Chapter 5. Such Fisher intervals require specification of the treatment effect for each unit. In most cases we would implement this by considering the set of values c

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

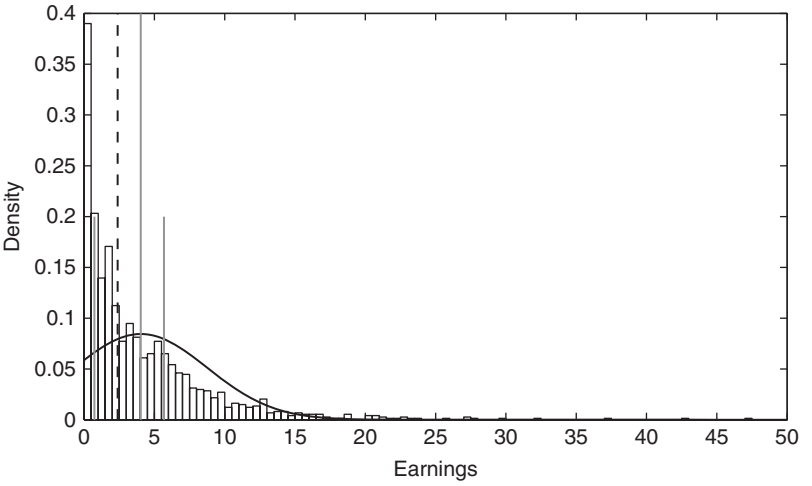

Figure 11.1. Histogram-based estimate of the distribution of Year 1 earnings, for those with positive earnings, San Diego SWIM program data

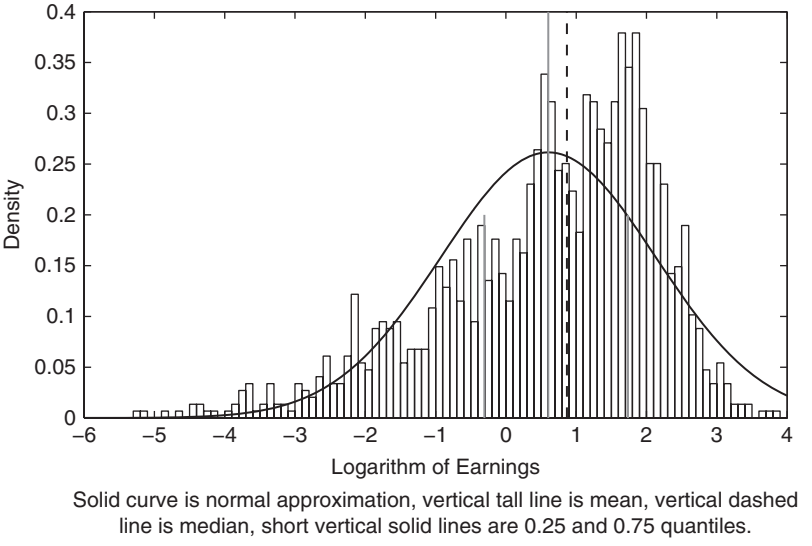

Figure 11.2. Histogram-based estimate of the distribution of the logarithm of year 1 earnings, for those with positive earnings, San Diego SWIM program data

such that we cannot reject the null hypothesis of a constant treatment effect equal to c. In this data set, such an approach is possible, but it is not attractive. Many individuals have earnings equal to zero in some year, because they do not have a job in that year. It is difficult to imagine that the training program would move all these individuals to some positive amount of earnings. On substantive grounds it is therefore extremely unlikely that there is a constant treatment effect, even after considering transformations of the outcome. We will therefore not pursue this strategy.

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

11.4 NEYMAN'S REPEATED SAMPLING-BASED POINT ESTIMATES AND LARGE-SAMPLE CONFIDENCE INTERVALS

In this section we apply Neyman's repeated sampling approach. For the full sample, as well as various subsamples, we estimate the average treatment effect on earnings in the first year after the program, and construct confidence intervals for this average effect. The results for these analyses are displayed in Table 11.3.

First we consider the full sample. The simple difference in average treatment and control outcomes is

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 2.02 - 1.69 = 0.33, \tag{11.1}$$

with sampling variance

$$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \mathbb{E}\left[\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - \tau_{\text{fs}}\right)^2\right] = \frac{S_{\text{c}}^2}{N_{\text{c}}} + \frac{S_{\text{t}}^2}{N_{\text{t}}} - \frac{S_{\text{ct}}^2}{N}.$$

Using the standard estimator for this sampling variance discussed in Chapter 6, we find

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_{\text{c}}^2}{N_{\text{c}}} + \frac{s_{\text{t}}^2}{N_{\text{t}}} = \frac{3.76^2}{1607} + \frac{3.80^2}{1604} = 0.13^2.$$

The implied large sample 95% confidence interval is

$$\text{CI}^{0.95}(\tau_{\text{fs}}) = \left(\hat{\tau}^{\text{dif}} - 1.96 \cdot \sqrt{\frac{s_{\text{c}}^2}{N_{\text{c}}} + \frac{s_{\text{t}}^2}{N_{\text{t}}}}, \hat{\tau}^{\text{dif}} + 1.96 \cdot \sqrt{\frac{s_{\text{c}}^2}{N_{\text{c}}} + \frac{s_{\text{t}}^2}{N_{\text{t}}}}\right) = (0.07, 0.59). \tag{11.2}$$

Next, we carry out the same calculations for some subpopulations. This serves two purposes. First, we may be interested in average treatment effects by subpopulations. Second, it may lead to more precise estimates of the overall average treatment effect. We begin by partitioning the sample into those at least thirty-five years old and those younger than thirty-five. The subsample of older individuals consists of 1,473 individuals, and the younger subsample consists of 1,738 individuals. For the older group we find

$$\hat{\tau}^{\text{dif}}(\text{old}) = 0.50 \ \ (\widehat{\text{s.e.}}\ 0.21), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{old})) = (0.09, 0.91).$$

For the younger group the estimated average treatment effect is

$$\hat{\tau}^{\text{dif}}(\text{young}) = 0.19 \ \ (\widehat{\text{s.e.}}\ 0.17), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{young})) = (-0.14, 0.51).$$

Next we partition the sample into those with no employment experience during the pre-program period, as indicated by zero earnings in the pre-program year (`empyrm1` equal to zero, which holds for 1,949 individuals) versus those with positive experience (1,262 individuals with `empyrm1` equal to one). For the first group, the estimated effect and associated estimated standard error are

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.3. ***Estimates for Average Treatment Effects on Year 1 Earnings Based on Neyman's Repeated Sampling Approach, San Diego SWIM Program Data***

Post-Program Earnings		All (3,211)	Young (1,738)	Old (1,473)	Unemployed (1,949)	Employed (1,262)	No HS (1,409)	HS (1,802)
Year 1	Est	0.33	0.19	0.50	0.34	0.38	0.41	0.27
	$\widehat{(\text{s.e.})}$	(0.13)	(0.17)	(0.21)	(0.13)	(0.25)	(0.15)	(0.21)
Year 2	Est	0.63	0.52	0.76	0.58	0.77	0.31	0.87
	$\widehat{(\text{s.e.})}$	(0.18)	(0.24)	(0.27)	(0.19)	(0.33)	(0.19)	(0.28)

$$\hat{\tau}^{\text{dif}}(\text{unempl}) = 0.34 \ \ (\widehat{\text{s.e.}}\ 0.13), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{unempl})) = (0.08, 0.601).$$

For the second group, the estimated average treatment effect is

$$\hat{\tau}^{\text{dif}}(\text{empl}) = 0.38 \ \ (\widehat{\text{s.e.}}\ 0.25), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{emp})) = (-0.12, 0.87).$$

We can also combine these to obtain an estimate of the overall average treatment effect τ_{fs} that is possibly more precise than $\hat{\tau}^{\text{dif}}$. We implement this by weighting the two estimates, $\hat{\tau}^{\text{dif}}(\text{empl})$ for the employed and $\hat{\tau}^{\text{dif}}(\text{unempl})$ for the unemployed, by their shares in the full sample. These shares are $1{,}262/(1{,}262 + 1{,}949) = 0.39$ for those with positive earnings and 0.61 for those with zero earnings in the year prior to the program. The weighted estimated average treatment effect, or employment-adjusted estimate is

$$\begin{aligned}\hat{\tau}^{\text{strat}} &= \frac{N(\text{empl})}{N(\text{empl}) + N(\text{unempl})} \cdot \hat{\tau}^{\text{dif}}(\text{empl}) + \frac{N(\text{unempl})}{N(\text{empl}) + N(\text{unempl})} \cdot \hat{\tau}^{\text{dif}}(\text{unempl}) \\ &= \frac{1262}{1262 + 1949} \cdot 0.38 + \frac{1949}{1262 + 1949} \cdot 0.34 = 0.36 \ \ (\widehat{\text{s.e.}}\ 0.15),\end{aligned}$$

with the large sample 95% confidence interval equal to

$$\text{CI}^{0.95}_{\text{combined}}(\tau_{\text{fs}}) = (0.11, 0.61).$$

Note that this point estimate differs slightly from $\hat{\tau}$ in (11.1) where we took the simple difference in average outcomes by treatment status, which reflects a small imbalance in the proportion of treated and control units among those with positive and zero earnings. More specifically, among those with positive earnings, 49.2% were assigned to the active treatment and 50.8% were assigned to the control treatment; and among those with zero earnings, 50.4% were assigned to the active treatment and 49.6% were assigned to the control treatment. This does not mean the randomization was compromised, merely that there is some random variation in these proportions because the randomization was not stratified on initial employment status.

The estimated sampling variance of the average treatment effect is also affected by the post-stratification on prior employment. If the treatment effect varies by covariates, then estimating the average effects within relatively homogeneous subpopulations, and then averaging over them will often reduce the sampling variance and lead to more precise inferences. Here, the change in estimated precision is fairly small.

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Finally, we partition the sample into those with no high school diploma (1,409 individuals) and those with a high school diploma (1,802 individuals). For the high school dropouts, we find

$$\hat{\tau}^{\text{dif}}(\text{no-hs}) = 0.41 \ \ (\widehat{\text{s.e.}}\ 0.15), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{no-hs})) = (0.12, 0.70).$$

For the high school graduates, the estimated average treatment effect is

$$\hat{\tau}^{\text{dif}}(\text{hs}) = 0.27 \ \ (\widehat{\text{s.e.}}\ 0.21), \qquad \text{CI}^{0.95}(\tau_{\text{fs}}(\text{hs})) = (-0.14, 0.68).$$

11.5 REGRESSION-BASED ESTIMATES

We now consider regression-based estimates of the average effect of the treatment, on the earnings in both the first and the second year after the program started. We consider specifications of the regression function that include the set of eleven pre-treatment variables listed in Table 11.1, indicators for being female (`female`), being at least 35 years old (`agege35`), having a high school diploma (`hsdip`), never having been married (`nevmar`), being divorced or widowed (`divwid`), having children younger than six years (`chldlt6`), being African-American (`af-amer`), being Hispanic (`hisp`), the discrete variable giving the number of children (`numchild`), and the lagged outcome, earnings in the year preceding the training program (`earnyrm1`), and an indicator for earnings being positive in that prior year (`empyrm1`). Denoting the row vector of these eleven pre-treatment variables by X_i, the basic specification of the regression function we estimate includes an intercept, the indicator for the treatment, the vector of pre-treatment variables, and the interaction of the two:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + (X_i - \overline{X})\beta + W_i \cdot (X_i - \overline{X})\gamma \ + \varepsilon_i.$$

The covariates are included in deviations from the sample average, so that the estimated coefficient on the treatment indicator, τ, can be interpreted as an estimator for the average effect of the treatment in the population. Implicitly this specification allows for separate slope coefficients for treated and control regression functions. For comparison, we also include least squares estimates of the regression function without pre-treatment variables:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$$

which gives the least squares estimate for τ equal to the difference in average outcomes by treatment status,

$$\hat{\tau}^{\text{ols}} = \hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 2.02 - 1.69 = 0.33.$$

The estimates of the average effect of the treatment do not change much with the inclusion of the eleven pre-treatment variables. For the first year earnings, the point estimate increases from 0.33 (in thousands of dollars) to 0.36, and in the second year, the estimate increases from 0.63 to 0.66. This is not unexpected: the fact that the randomization was done without regard to the pre-treatment variables implies that, on average, the

Table 11.4. ***Regression Estimates for Average Treatment Effects on Earnings, for the San Diego Swim Data***

Covariates	Earnings Year 1				Earnings Year 2			
	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$
`Treat`	0.33	(0.13)	0.36	(0.12)	0.63	(0.18)	0.66	(0.17)
`Intercept`	1.69	(0.09)	1.68	(0.09)	2.26	(0.12)	2.25	(0.11)
Covariates								
`female`			0.35	(0.29)			−0.03	(0.39)
`agege35`			−0.09	(0.17)			−0.01	(0.23)
`hsdip`			0.79	(0.20)			0.86	(0.25)
`nevmar`			0.38	(0.21)			0.47	(0.29)
`divwid`			0.32	(0.20)			0.41	(0.26)
`numchild`			0.10	(0.08)			0.03	(0.11)
`chldlt6`			−0.46	(0.25)			−0.20	(0.36)
`af-amer`			−0.22	(0.22)			−0.54	(0.28)
`hisp`			0.05	(0.23)			−0.25	(0.30)
`earnyrm1`			0.33	(0.08)			0.33	(0.09)
`empyrm1`			0.75	(0.30)			0.78	(0.34)
Interactions with treatment indicator								
`treat×female`			−0.01	(0.43)			0.48	(0.59)
`treat×age 35`			0.17	(0.25)			0.18	(0.36)
`treat×high school dip`			−0.15	(0.27)			0.54	(0.36)
`treat×never married`			−0.40	(0.29)			−0.33	(0.41)
`treat×divorced/widowed`			0.34	(0.29)			0.36	(0.41)
`treat×number of children`			−0.18	(0.11)			−0.29	(0.15)
`treatchldlt6`			0.42	(0.39)			1.15	(0.60)
`treat×african-american`			−0.29	(0.31)			−0.14	(0.42)
`treat×hispanic`			−0.26	(0.34)			0.31	(0.48)
`treatearnyrm1`			0.09	(0.10)			0.22	(0.13)
`treatempyrm1`			−0.30	(0.40)			−0.72	(0.50)
R-squared	0.002		0.190		0.004		0.151	

pre-treatment variables should be approximately the same in treatment group and control group and that their inclusion or omission usually should not change point estimates of treatment effects as a result of the linear predictive power. The estimated standard error does not change much either. They decrease slightly, as a result of the predictive power of the covariates, but because this predictive power is fairly modest, the reduction in estimated standard error is small.

The main interest in the regression estimates is that they provide some evidence regarding heterogeneity in the effect of the program, which can be seen directly by inspecting the least squares estimates of the coefficients of the interactions of the pre-treatment variables with the treatment indicator, as reported in Table 11.4. In addition to these estimates, we also report tests of hypotheses about the coefficients in the linear

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.5. ***P-Values for Tests of Constant and Zero Treatment Effects Assumptions, for San Diego SWIM Data***

Null Hypothesis		Earnings Year 1	Earnings Year 2
Zero effect	$\mathcal{X}^2(12)$ approximation	0.018	<0.001
	Fisher exact p-value	0.157	0.014
Constant effect	$\mathcal{X}^2(11)$ approximation	0.122	0.002

regression model. Specifically we consider two null hypotheses. First, consider the null hypotheses that all least squares coefficients involving the treatment indicator are equal to zero. Formally,

$$H_0: \quad \tau = 0 \text{ and } \gamma = 0,$$

against the alternative that either τ or some components of γ differ from zero,

$$H_a: \quad \tau \neq 0 \text{ or } \gamma \neq 0,$$

where 0 denotes a vector of zeros. The results from an F-test on the least squares coefficients are reported in Table 11.5. The value of the F-statistic using the first-year earnings as the outcome variable is 2.11, leading to a p-value of 0.018 based on the asymptotic approximation using the F-distribution with 12 degrees of freedom. We also carried out a different version of this test, where we used the F-statistic in a Fisher-exact-p-value calculation, under the null of no effect of the treatment whatsoever. This led to a considerably less significant p-value of 0.157. The results for the p-value are also reported in Table 11.5. For the second-year earnings outcome, the F-statistic is 3.78, leading to a p-value based on the F-distribution less than 0.001, and a p-value based on the randomization distribution equal to 0.014. Next, we considered the null hypothesis of no treatment effect heterogeneity by pre-treatment variables. In terms of the least squares coefficients, this corresponds to testing the null hypothesis

$$H_0: \quad \gamma = 0,$$

against the alternative that some components of γ differ from zero,

$$H_a: \quad \gamma \neq 0.$$

We find somewhat different results for the first- and second-year earnings. For the first year we find an F-statistic equal to 1.50, leading to a p-value of 0.122. This suggests little evidence for heterogeneity of the treatment effect. The F-statistic for second-year earnings is 2.68, leading to a p-value of 0.002, suggesting clear evidence that the treatment effect on second-year earnings varies by the values of the pre-treatment variables.

11.6 MODEL-BASED POINT ESTIMATES

Now let us consider the model-based approach. To avoid reporting a large number of estimates, we focus first on estimating the average treatment effect for earnings in the second year.

A simple strategy is to specify a joint normal distribution for the two potential outcomes with unit correlation. If we use a normal prior distribution for the mean parameters and inverse Chi-squared distributions for the two variance parameters, we return to the case analyzed in Chapter 8. With the number of observations as large as in the SWIM program, the choice of prior distribution is unlikely to matter much. We estimate two versions of the normal model. First, a model with no covariates; for the mean parameters, we use normal prior distributions centered at zero with prior variances equal to 100^2. For the variance parameters, we use inverse Chi-squared distributions with parameters equal to 1/2 and 0.0005. The posterior mean for τ_{fs} is 0.33, and the posterior standard deviation is equal to 0.09. Next we include the eleven covariates in the model, assuming they enter linearly for the mean. Now the posterior mean for τ_{fs} is 0.36 and the posterior standard deviation is 0.08. Although the covariates are moderately strongly associated with the potential outcomes, including the covariates does not affect the posterior distribution for the average effect of interest very much. These results are very similar to those obtained through the Neyman approach, which is not surprising because the sample size implies that, using versions of the central limit theorem, normal distributions are likely to give accurate approximations to both the sampling and the posterior distributions.

It is clear, however, that the model used in this first attempt is not an appropriate one. The distributions are far from normal, with 54% of individuals having zero earnings one year after the program started, as the summary statistics in Table 11.1 show. A more plausible approximation to the distribution of earnings in each treatment regime is therefore a mixed discrete-continuous distribution. We use the following model with one parameter governing the probability of the point mass at zero and a normal distribution for the continuous component (which led to a better fit than a log normal distribution for the continuous part),

$$\Pr(Y_i(0) > 0|X_i,\theta) = \frac{\exp(\gamma_{\text{c}})}{1+\exp(\gamma_{\text{c}})}, \qquad \big(Y_i(0)|Y_i(0) > 0, X_i, \theta\big) \sim \mathcal{N}(\mu_{\text{c}}, \sigma_{\text{c}}^2),$$

$$\Pr(Y_i(1) > 0|X_i,\theta) = \frac{\exp(\gamma_{\text{t}})}{1+\exp(\gamma_{\text{t}})}, \qquad \big(Y_i(1)|Y_i(1) > 0, X_i, \theta\big) \sim \mathcal{N}(\mu_{\text{t}}, \sigma_{\text{t}}^2),$$

and assume independence between the potential outcomes. For this specification, it is difficult to derive an analytic expression for the posterior distribution of the average treatment effect in terms of the observed data for most prior distributions. We focus, therefore, on simulation methods.

We use independent prior distributions for the six elements of the parameter vector $\theta = (\gamma_{\text{c}}, \gamma_{\text{t}}, \mu_{\text{c}}, \mu_{\text{t}}, \sigma_{\text{c}}^2, \sigma_{\text{t}}^2)$. For γ_{c}, γ_{t}, μ_{c}, and μ_{t}, we use normal prior distributions centered at zero and with variance equal to 100^2. The prior distributions for the variance parameters are inverse Chi-squared, with parameters 1/2 and $\sigma_{\text{c}}^2/2$ and $\sigma_{\text{t}}^2/2$, respectively. The mean and standard deviation of the posterior distribution for τ are 0.33 and

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.6. ***Posterior Means and Standard Deviations for Model-Based Imputation Estimates, Year 1 Earnings, for San Diego SWIM Data***

	Linear Model No Covariates		Linear Model Covariates		Two-Part Model No Covariates				Two-Part Model Covariates			
					Logit		Normal		Logit		Normal	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
Control Outcome												
Intercept	1.69	(0.09)	−0.13	(0.40)	−0.39	(0.05)	4.17	(0.20)	−1.56	(0.27)	1.77	(0.86)
female			0.35	(0.32)					0.04	(0.21)	0.55	(0.68)
agege35			−0.09	(0.19)					−0.28	(0.13)	0.19	(0.41)
hsdip			0.78	(0.19)					0.46	(0.12)	1.39	(0.42)
nevmar			0.38	(0.24)					0.26	(0.16)	0.52	(0.52)
divwid			0.32	(0.21)					0.13	(0.14)	0.68	(0.46)
numchild			0.10	(0.09)					0.06	(0.06)	0.14	(0.19)
chldlt6			−0.47	(0.29)					−0.13	(0.19)	−0.89	(0.63)
af-amer			−0.22	(0.21)					−0.05	(0.14)	−0.58	(0.45)
hisp			0.05	(0.24)					0.04	(0.16)	0.13	(0.53)
earnyrm1			0.33	(0.03)					0.10	(0.02)	0.32	(0.05)
empyrm1			0.75	(0.21)					1.49	(0.14)	−0.63	(0.44)
σ_c	3.76	(0.07)	3.45	(0.06)			4.97	(0.14)			4.72	(0.13)
Treated Outcome												
Intercept	2.02	(0.09)	0.69	(0.38)	0.06	(0.05)	3.92	(0.16)	−0.62	(0.24)	2.67	(0.67)
female			0.34	(0.30)					0.09	(0.20)	0.08	(0.51)
agege35			0.08	(0.18)					−0.17	(0.12)	0.31	(0.32)
hsdip			0.64	(0.18)					0.22	(0.11)	0.91	(0.32)
nevmar			−0.02	(0.23)					0.23	(0.15)	−0.39	(0.41)
divwid			0.66	(0.21)					0.51	(0.13)	0.59	(0.35)
numchild			−0.08	(0.09)					−0.08	(0.05)	−0.07	(0.15)
chldlt6			−0.04	(0.30)					−0.23	(0.18)	0.40	(0.53)
af-amer			−0.51	(0.21)					−0.14	(0.13)	−0.80	(0.34)
hisp			−0.21	(0.24)					−0.09	(0.15)	−0.29	(0.40)
earnyrm1			0.42	(0.03)					0.09	(0.03)	0.43	(0.04)
empyrm1			0.45	(0.21)					1.13	(0.14)	−0.37	(0.34)
σ_t	3.80	(0.07)	3.38	(0.06)			4.53	(0.11)			4.10	(0.10)
τ_{fs}	0.33	(0.09)	0.36	(0.08)			0.33	(0.09)			0.36	(0.08)

0.09, respectively. The posterior means and standard deviations for all elements of θ are presented in Tables 11.6 (year 1 earnings) and 11.7 (year 2 earnings).

Next, we consider a similar mixed discrete-continuous model with covariates, often called a "two-part" model. Let X_i denote the vector of covariates reported in Table 11.1. The model is now

$$\Pr(Y_i(0) > 0|X_i = x, \theta) = \frac{\exp(x\gamma_c)}{1 + \exp(x\gamma_c)}, \quad \Big(Y_i(0)|X_i = x, Y_i(0) > 0, \theta\Big) \sim \mathcal{N}(x\beta_c, \sigma_c^2),$$

$$\Pr(Y_i(1) > 0|X_i = x, \theta) = \frac{\exp(x\gamma_t)}{1 + \exp(x\gamma_t)} \text{ and } \Big(Y_i(1)|X_i = x, Y_i(1) > 0, \theta\Big) \sim \mathcal{N}(x\beta_t, \sigma_t^2).$$

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.7. ***Posterior Means and Standard Deviations for Model-Based Imputation Estimates, Year 2 Earnings, for San Diego SWIM Data***

	Linear Model No Covariates		Linear Model Covariates		Two-Part Model No Covariates				Two-Part Model Covariates			
					Logit		Normal		Logit		Normal	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
Control Outcome												
Intercept	2.26	(0.12)	0.96	(0.50)	−0.40	(0.05)	5.62	(0.23)	−1.03	(0.25)	4.04	(1.01)
female			−0.06	(0.40)					−0.12	(0.20)	−0.26	(0.82)
agege35				(0.23)					−0.18	(0.12)	0.35	(0.51)
hsdip			0.88	(0.24)					0.08	(0.12)	2.18	(0.49)
nevmar			0.46	(0.31)					0.29	(0.15)	0.70	(0.64)
divwid			0.40	(0.27)					0.30	(0.13)	0.40	(0.56)
numchild			0.03	(0.11)						(0.05)	0.16	(0.24)
chldlt6			−0.22	(0.38)					−0.02	(0.17)	−0.55	(0.77)
af-amer			−0.52	(0.26)					0.05	(0.12)	−1.59	(0.55)
hisp			−0.24	(0.31)					0.06	(0.14)	−0.83	(0.62)
earnyrm1			0.33	(0.04)					0.06	(0.02)	0.38	(0.06)
empyrm1			0.76	(0.27)					1.06	(0.14)	−0.61	(0.51)
σ_c	4.68	(0.08)	4.42	(0.08)			5.97	(0.17)			5.65	(0.16)
Treated Outcome												
Intercept	2.89	(0.13)	1.05	(0.55)	−0.03	(0.05)	5.86	(0.24)	−0.73	(0.24)	4.06	(0.98)
female			0.43	(0.43)					0.10	(0.18)	0.02	(0.75)
agege35			0.18	(0.28)					0.01	(0.11)	0.36	(0.46)
hsdip			1.39	(0.28)					0.36	(0.12)	2.09	(0.49)
nevmar			0.15	(0.34)					0.13	(0.14)	0.10	(0.59)
divwid			0.78	(0.31)					0.33	(0.14)	0.87	(0.51)
numchild			−0.26	(0.12)					−0.12	(0.06)	−0.22	(0.24)
chldlt6			0.96	(0.45)					0.26	(0.18)	1.17	(0.72)
af-amer			−0.65	(0.30)					−0.20	(0.12)	−0.96	(0.51)
hisp			0.06	(0.36)					0.33	(0.14)	−0.61	(0.57)
earnyrm1			0.55	(0.04)					0.09	(0.02)	0.59	(0.06)
empyrm1			0.06	(0.31)					0.77	(0.13)	−1.23	(0.52)
σ_t	5.44	(0.10)	4.97	(0.09)			6.53	(0.16)			5.97	(0.15)
τ_{fs}	0.64	(0.13)	0.66	(0.12)			0.63	(0.13)			0.67	(0.12)

The posterior mean for τ_{fs} given this model is 0.36 with a posterior standard deviation equal to 0.08 (Table 11.8) The posterior means and standard deviations for all other elements of θ are again presented in Tables 11.6 and 11.7.

One major advantage of the model-based imputation approach is that we can easily accommodate different estimands. Suppose that instead of focusing on the average effect of the treatment, we are interested in the effect of the training program on the probability that individuals who were not working before now have jobs paying more than $5,000. Within the context of the imputations, this is a straightforward calculation. The imputation procedure is exactly as before. Now to calculate the posterior distribution of the

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Table 11.8. ***Summary Statistics Posterior Distribution for Finite-Sample Average Treatment Effect, for San Diego SWIM Data***

Post-Program Earnings	Model	Covariates	Mean	(S. D.)	Posterior Quantiles				
					0.025	0.25	0.5	0.75	0.975
Year 1	Linear	No	0.33	(0.09)	0.14	0.27	0.33	0.40	0.51
Year 1	Linear	Yes	0.36	(0.08)	0.19	0.30	0.36	0.41	0.52
Year 1	Two-part	No	0.33	(0.09)	0.14	0.27	0.33	0.39	0.51
Year 1	Two-part	Yes	0.37	(0.09)	0.20	0.31	0.37	0.42	0.53
Year 2	Linear	No	0.63	(0.13)	0.38	0.54	0.63	0.71	0.88
Year 2	Linear	Yes	0.66	(0.12)	0.43	0.58	0.66	0.74	0.89
Year 2	Two-part	No	0.63	(0.13)	0.38	0.54	0.63	0.71	0.87
Year 2	Two-part	Yes	0.67	(0.12)	0.44	0.59	0.67	0.75	0.90

estimand, we simply calculate the fraction, among individuals who had zero earnings before, of individuals who now have earnings more than $5,000. Using the two-part model with covariates, the posterior mean and standard deviation for this probability are 0.029 and 0.009. Another advantage is that it is straightforward to report results on the posterior distribution of the estimands beyond moments, for example posterior quantiles.

Table 11.8 reports posterior quantiles for the average effect of the treatment on post-program earnings.

11.7 CONCLUSION

In this chapter we illustrate the four basic methods for analyzing classical randomized experiments discussed in the second part of the text. Taking as the example a randomized experiment of a job-training program, we illustrate the calculation of Fisher exact p-values, the construction of confidence intervals based on Neyman's repeated sampling approach, regression analyses, and model-based analyses. The methods generally agree here: there is strong evidence of an effect of the program, and we can estimate its average effects precisely. Ultimately the choice of methods here is somewhat subtle: the randomization ensures that the point estimates tend to be similar, the estimated precisions are similar because the covariates are only moderately predictive of the potential outcomes, and the methods differ mostly in the precise questions they ask. In the next parts of the book, where we address observational studies, these differences are often amplified, and the choices become more consequential.

NOTES

For more detail on the San Diego SWIM program and similar labor market training programs, see Friedlander and Robbins (1995), Friedlander and Gueron (1995), Hotz, Imbens, and Mortimer (2005), and Hotz, Imbens, and Klerman (2001).

https://doi.org/10.1017/CBO9781139025751.012 Published online by Cambridge University Press

Research that concluded that non-experimental evaluations were not credible in social sciences led to a renewed interest in experimental evaluations. Important papers in this literature are Lalonde (1986), Fraker and Maynard (1987), and Friedlander and Robbins (1995). The central thesis in this literature was the claim that non-experimental methods led to a wide range of results, with no reliable methods for choosing among these results. Later research cast some doubt on these claims. Dehejia and Wahba (1999) showed that methods based on the propensity score were considerably more successful in replicating experimental results than the regression-based methods considered by Lalonde (1986).