CHAPTER 7

Regression Methods for Completely Randomized Experiments

7.1 INTRODUCTION

One of the more common ways of estimating causal effects with experimental, as well as observational, data in many disciplines is based on regression methods. Typically an additive linear regression function is specified for the observed outcome as a function of a set of predictor variables. This set of predictor variables includes the indicator variable for the receipt of treatment and usually additional pre-treatment variables. The parameters of the regression equation are estimated by least squares, with the primary focus on the coefficient for the treatment indicator. Inferences, including point estimates, standard errors, tests, and confidence intervals, are based on standard least squares methods. Although popular, the use of these methods in this context is not without controversy, with some researchers arguing that experimental data should be analyzed based on randomization inference. As Freedman writes bluntly, “Experiments should be analyzed as experiments, not as observational studies” (Freedman, 2006, p. 691). It has also been pointed out that the justification for least squares methods does not follow from randomization. Again Freedman: “randomization does not justify the assumptions behind the ols [ordinary least squares] model” (Freedman, 2008a, p. 181). In this chapter we discuss in some detail the rationale for, and the interpretation and implementation of, regression methods in the setting with completely randomized experiments. This chapter can be viewed as providing a bridge between the previous chapter, which was largely focused on exact finite-sample results based on randomization, and the next chapter, which is based on fully parametric models for imputation of the unobserved potential outcomes.

The most important difference between the methods discussed in Chapters 5 and 6 and the ones discussed here is that they rely on different sampling perspectives. Both the Fisher approach discussed in Chapter 5 and the Neyman methods discussed in Chapter 6 view the potential outcomes as fixed and the treatment assignments as the sole source of randomness. In the regression analysis discussed in this chapter, the starting point is an infinite super-population of units. Properties of the estimators are assessed by resampling from that population, sometimes conditional on the predictor variables including the treatment indicator. From that perspective, the potential outcomes in the sample are random, and we can derive the bias and sampling variance of estimators over the distribution induced by this random sampling. The sampling variance of estimators derived in

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

this approach will be seen to be very similar to the Neyman sampling variance for $\hat{\tau}^{\text{dif}}$ derived in Chapter 6, although its interpretation will be different.

There are four key features of the models considered in this chapter. First, we consider models for the observed outcomes rather than for the potential outcomes. Second, we consider models only for the conditional mean rather than for the full distribution. Third, the estimand, here always an average treatment effect, is a parameter of the statistical model. The latter implies that inferential questions can be viewed as questions of inference for parameters of a statistical model. Fourth, in the current context of completely randomized experiments, the validity of these models, that is, whether the models provide accurate descriptions of the conditional mean, is immaterial for the large-sample unbiasedness of the least squares estimator of the average treatment effect.

As the Freedman quote illustrates, the conventional justification for linear regression models, that the regression function represents the conditional expectation of the observed outcome given the predictor variables, does not follow from the randomization if there are predictors beyond the treatment indicator. Nevertheless, in the setting of a completely randomized experiment, the least squares point estimates and associated inferences can be given a causal interpretation. There is an important difference with the causal interpretation in the previous chapter, however. With the exception of the setting without additional covariates beyond the treatment indicator, where the main results are essentially identical to those discussed in the previous chapter from the Neyman approach, all results are now asymptotic (large sample) results. Specifically, exact unbiasedness no longer holds in finite samples with covariates beyond the treatment indicator because of the need to estimate additional nuisance parameters, that is, the associated regression coefficients. The possible benefit of the regression methods over the exact methods from the previous chapter is that they provide a straightforward and, for many researchers, familiar way to incorporate covariates. If these covariates are predictive of the potential outcomes, their inclusion in the regression model can result in causal inferences that are more precise than differences in observed means. This gain in precision can be substantial if the covariates are highly predictive of the potential outcomes, although in practice the gains are often modest. The disadvantage of regression models relative to the fully model-based methods that will be discussed in the next chapter is that the use of standard linear regression models often restricts the set of models considerably, and thereby restricts the set of questions that can be addressed. Thus, when using these regression models, there is often a somewhat unnatural tension between, on the one hand, models that provide a good statistical fit and have good statistical properties and, on the other hand, models that answer the substantive question of interest. This tension is not present in the full, model-based methods discussed in the next chapter.

This chapter is organized as follows. In the next section, Section 7.2, we describe the data that will be used to illustrate the techniques discussed in this chapter. The data come from a completely randomized experiment previously analyzed by Efron and Feldman (1991). Section 7.3 reviews and adds notation regarding the super-population perspective. In Section 7.4 we discuss the case with no predictor variables beyond the treatment indicator. In that case, most of the results are closely related to those from the previous chapter. In Section 7.5 we generalize the results to allow for the presence of additional predictor variables. Next, in Section 7.6, we include interactions between the predictor variables and the treatment indicator. In Section 7.7 we discuss the role of

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

transformations of the outcome variable. The following section, Section 7.8, discusses the limits on the increases in precision that can be obtained by including covariates. In Section 7.9 we discuss testing for the presence of treatment effects. Then, in Section 7.10, we apply the methods to the Efron-Feldman data. Section 7.11 concludes.

7.2 THE LRC-CPPT CHOLESTEROL DATA

We illustrate the concepts discussed in this chapter using data from a randomized experiment, the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT), designed to evaluate the effect of the drug cholestyramine on cholesterol levels. The data were previously analyzed in Efron and Feldman (1991). The data set analyzed here contains information on $N = 337$ individuals. Of these 337 individuals, $N_\mathrm{t} = 165$ were randomly assigned to receive cholestyramine and the remaining $N_\mathrm{c} = 172$ were assigned to the control group, which received a placebo.

For each individual, we observe two cholesterol measures recorded prior to the random assignment. The two measures differ in their timing. The first, `chol1`, was taken prior to a communication, sent to all 337 individuals in the study, about the benefits of a low-cholesterol diet, and the second, `chol2`, was taken after this suggestion, but prior to the random assignment to cholestyramine or placebo. We observe two outcomes. The primary outcome is an average of post-randomization cholesterol readings, `cholf`, averaged over two-month readings for a period of time averaging 7.3 years for all the individuals in the study. Efron and Feldman's primary outcome is the change in cholesterol level, relative to a weighted average of the two pre-treatment cholesterol levels, $\texttt{cholp}= 0.25 \cdot \texttt{chol1} + 0.75 \cdot \texttt{chol2}$. We denote this change in cholesterol levels by $\texttt{chold}=\texttt{cholf}\texttt{-}\texttt{cholp}$. The secondary outcome is a compliance measure, denoted by `comp`, the percentage of the nominally assigned dose of either cholestyramine or placebo that the individual actually took. Although individuals did not know whether they were assigned to cholestyramine or to the placebo, later we shall see that differences in side effects between the active drug and the placebo induced systematic differences in compliance behavior by treatment status. Note that all individuals, whether assigned to the treatment or the control group, were assigned the same nominal dose of the drug or placebo, for the same time period.

The availability of compliance data raises many interesting issues regarding differences between the effect of *being assigned* to the taking of cholestyramine and the effect of actually *taking* cholestyramine. We discuss some of these issues in detail in later chapters on noncompliance and instrumental variables (Chapters 23–25). Here we analyze the compliance measure solely as a secondary outcome. Note, however, that in general it is *not* appropriate to interpret either the difference in final cholesterol levels by assignment, conditional on observed compliance levels, or the difference in final cholesterol levels by actual dosage taken, as estimates of average causal effects. Such causal interpretations would require strong additional assumptions beyond randomization. For example, to validate conditioning on observed compliance levels would require that observed compliance is a proper pre-treatment variable unaffected by the assignment to treatment versus placebo. Because observed compliance reflects behavior subsequent to the assignment, it may be affected by the treatment assigned, which is an assumption. This is an assumption

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

Table 7.1. *Summary Statistics for PRC-CPPT Cholesterol Data*

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	`chol1`	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	`chol2`	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	`cholp`	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	`cholf`	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	`chold`	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	`comp`	74.5	(21.0)	59.9	(24.4)	0	101.0

that can be assessed, and in the current study we can reject, at conventional significance levels, the assumption that observed compliance is a proper pretreatment variable.

In Table 7.1 we present summary statistics for the Efron-Feldman data. For the two initial cholesterol levels (`chol1` and `chol2`), as well as the composite pre-treatment cholesterol level (`cholp`), the averages do not vary much by treatment status, consistent with the randomized assignment. We do see that the second pre-treatment cholesterol-level measurement, `chol2`, is, on average, lower than the first one, `chol1`. This is consistent with the fact that in between the two measurements, the individuals in the study received information about the benefits of a low cholesterol diet that may have induced them to improve their diets. For the subsequent cholesterol-level measures (`cholf` and `chold`), the averages do vary considerably by treatment status. In addition, the average level of compliance (`comp`) is much higher in the control group than in the treatment group. Later in this chapter we investigate the statistical precision of this difference, but here we just comment that this is consistent with relatively severe side effects of the actual drug, which are not present in the placebo. This difference signals the potential dangers of using a post-treatment variable, such as observed compliance, as a covariate.

7.3 THE SUPER-POPULATION AVERAGE TREATMENT EFFECTS

As in Section 6.7 in the previous chapter, we focus in this chapter on the average effect in the super-population, rather than in the sample. We assume that the sample of size N for which we have information can be considered a simple random sample drawn from an infinite super-population. Considering the N units in our sample as a random sample from the super-population induces a distribution on the pair of potential outcomes. The observed potential outcome and covariate values for a drawn unit are simply one draw from the joint distribution in the population and are therefore themselves stochastic. We assume that we have no information about this distribution beyond the values of the observed outcomes and covariates in our sample.

The distribution of the two potential outcomes in turn induces a distribution on the unit-level treatment effects, and thereby on the average of the unit-level treatment effect within the experimental sample. To be clear about this super-population perspective, let us, as we did in the previous chapter, index the average treatment effect τ by fs to denote

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

the finite-sample average treatment effect and by sp to denote the super-population average treatment effect. Thus

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)$$

is the average effect of the treatment in the finite sample, and

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0)\right]$$

is the expected value of the unit-level treatment effect under the distribution induced by sampling from the super-population, or, equivalently, the average treatment effect in the super-population. (We index the expectations operator by sp to make explicit that the expectation is taken over the random sampling, not over the randomization distribution, as in the previous chapter.) For the discussion in this chapter, it is useful to introduce some additional notation. Define the super-population average and variance of the two potential outcomes conditional on the covariates or pre-treatment variables, e.g., $X_i = x$,

$$\mu_{\text{c}}(x) = \mathbb{E}_{\text{sp}}\left[Y_i(0)|X_i = x\right], \quad \mu_{\text{t}}(x) = \mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i = x\right],$$
$$\sigma^2_{\text{c}}(x) = \mathbb{V}_{\text{sp}}\left(Y_i(0)|X_i = x\right), \quad \text{and} \quad \sigma^2_{\text{t}} = \mathbb{V}_{\text{sp}}\left(Y_i(1)|X_i = x\right),$$

and let the mean and variance of the unit-level treatment effects at $X_i = x$ be denoted by

$$\tau(x) = \mathbb{E}_{\text{sp}}(Y_i(1) - Y_i(0)|X_i = x], \quad \text{and} \quad \sigma^2_{\text{ct}}(x) = \mathbb{V}_{\text{sp}}\left(Y_i(1) - Y_i(0)|X_i = x\right),$$

respectively. In addition, denote the marginal means and variances

$$\mu_{\text{c}} = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right], \quad \mu_{\text{t}} = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right],$$
$$\sigma^2_{\text{c}} = \mathbb{V}_{\text{sp}}\left(Y_i(0)\right), \quad \text{and} \quad \sigma^2_{\text{t}} = \mathbb{V}_{\text{sp}}\left(Y_i(1)\right).$$

Note that the two marginal means are equal to the expectation of the corresponding conditional means:

$$\mu_{\text{c}} = \mathbb{E}_{\text{sp}}\left[\mu_{\text{c}}(X_i)\right], \quad \text{and} \quad \mu_{\text{t}} = \mathbb{E}_{\text{sp}}\left[\mu_{\text{t}}(X_i)\right],$$

but, by the law of iterated expectations, the marginal variance differs from the average of the conditional variance by the variance of the conditional mean:

$$\sigma^2_{\text{c}} = \mathbb{E}_{\text{sp}}\left[\sigma^2_{\text{c}}(X_i)\right] + \mathbb{V}_{\text{sp}}\left(\mu_{\text{c}}(X_i)\right), \quad \text{and} \quad \sigma^2_{\text{t}} = \mathbb{E}_{\text{sp}}\left[\sigma^2_{\text{t}}(X_i)\right] + \mathbb{V}_{\text{sp}}\left(\mu_{\text{t}}(X_i)\right).$$

Finally, let

$$\mu_X = \mathbb{E}_{\text{sp}}\left[X_i\right], \quad \text{and} \quad \Omega_X = \mathbb{V}_{\text{sp}}(X_i) = \mathbb{E}_{\text{sp}}\left[(X_i - \mu_X)^T(X_i - \mu_X)\right],$$

denote the super-population mean and covariance matrix of the row vector of covariates X_i, respectively.

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

7.4 LINEAR REGRESSION WITH NO COVARIATES

In this section we focus on the case without covariates, that is, no predictor variables beyond the indicator W_i for the receipt of treatment. We maintain the assumption of a completely randomized experiment. We specify a linear regression function for the observed outcome Y_i^{obs} as

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

where the unobserved residual ε_i captures unobserved determinants of the outcome. The ordinary least squares (or ols for short) estimator for τ is based on minimizing the sum of squared residuals over α and τ,

$$(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i \right)^2,$$

with solutions

$$\hat{\tau}^{\text{ols}} = \frac{\sum_{i=1}^{N} \left(W_i - \overline{W} \right) \cdot (Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})}{\sum_{i=1}^{N} \left(W_i - \overline{W} \right)^2}, \quad \text{and} \quad \hat{\alpha}^{\text{ols}} = \overline{Y}^{\text{obs}} - \hat{\tau}^{\text{ols}} \cdot \overline{W},$$

where

$$\overline{Y}^{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} Y_i^{\text{obs}} \quad \text{and} \quad \overline{W} = \frac{1}{N} \sum_{i=1}^{N} W_i = \frac{N_{\text{t}}}{N}.$$

Simple algebra shows that in this case the ols estimator $\hat{\tau}^{\text{ols}}$ is identical to the difference in average outcomes by treatment status:

$$\hat{\tau}^{\text{ols}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = \hat{\tau}^{\text{dif}},$$

where, as before, $\overline{Y}_{\text{t}}^{\text{obs}} = \sum_{i:W_i=1} Y_i^{\text{obs}}/N_{\text{t}}$ and $\overline{Y}_{\text{c}}^{\text{obs}} = \sum_{i:W_i=0} Y_i^{\text{obs}}/N_{\text{c}}$ are the averages of the observed outcomes in the treatment and control groups respectively.

The least squares estimate of τ is often interpreted as an estimate of the causal effect of the treatment, explicitly in randomized experiments, and sometimes implicitly in observational studies. The assumptions traditionally used in the least squares approach are that the residuals ε_i are independent of, or at least uncorrelated with, the treatment indicator W_i. This assumption is difficult to evaluate directly, as the interpretation of these residuals is rarely made explicit beyond a somewhat vague notion of capturing unobserved factors affecting the outcomes of interest. Statistical textbooks, therefore, often stress that in observational studies the regression estimate $\hat{\tau}^{\text{ols}}$ measures only the association between the two random variables W_i and Y_i^{obs} and that a causal interpretation is generally not warranted. In the current context, however, we already have a formal justification for the causal interpretation of $\hat{\tau}^{\text{ols}}$ because it is identical to $\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, which itself was shown in Chapter 6 to be unbiased for the finite-sample average treatment effect, τ_{fs}, as well as for the super-population average treatment effect, τ_{sp}. Nevertheless, it is useful to

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

justify the causal interpretation of $\hat{\tau}^{\text{ols}}$ more directly in terms of the standard justification for regression methods, using the assumptions that random sampling created the sample and a completely randomized experiment generated the observed data from that sample.

Let α be the population average outcome under the control, $\alpha = \mu_{\text{c}} = \mathbb{E}_{\text{sp}}[Y_i(0)]$, and recall that τ_{sp} is the super-population average treatment effect, $\tau_{\text{sp}} = \mu_{\text{t}} - \mu_{\text{c}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$. Now *define* the residual ε_i in terms of the population parameters, treatment indicator, and the potential outcomes as

$$\varepsilon_i = Y_i(0) - \alpha + W_i \cdot \big(Y_i(1) - Y_i(0) - \tau_{\text{sp}}\big) = \begin{cases} Y_i^{\text{obs}} - \alpha & \text{if } W_i = 0, \\ Y_i^{\text{obs}} - \alpha - \tau_{\text{sp}} & \text{if } W_i = 1. \end{cases}$$

Then we can write

$$\varepsilon_i = Y_i^{\text{obs}} - (\alpha + \tau_{\text{sp}} \cdot W_i),$$

and thus we can write the observed outcome as

$$Y_i^{\text{obs}} = \alpha + \tau_{\text{sp}} \cdot W_i + \varepsilon_i.$$

Random sampling allows us to view the potential outcomes as random variables. In combination with random assignment this implies that assignment is independent of the potential outcomes,

$$\Pr(W_i = 1 | Y_i(0), Y_i(1)) = \Pr(W_i = 1),$$

or in Dawid's (1979) "$\perp\!\!\!\perp$" independence notation,

$$W_i \;\perp\!\!\!\perp\; (Y_i(0), Y_i(1)).$$

The definition of the residual, in combination with random assignment and random sampling from a super-population, implies that the residual has mean zero conditional on the treatment indicator in the population:

$$\mathbb{E}_{\text{sp}}[\varepsilon_i | W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(0) - \alpha | W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(0)] - \alpha] = 0,$$

and

$$\begin{aligned} \mathbb{E}_{\text{sp}}[\varepsilon_i | W_i = 1] &= \mathbb{E}_{\text{sp}}\left[Y_i(1) - \alpha - \tau_{\text{sp}} | W_i = 1\right] \\ &= \mathbb{E}_{\text{sp}}\left[Y_i(1) - \alpha - \tau_{\text{sp}} | W_i = 1\right] = 0, \end{aligned}$$

so that

$$\mathbb{E}_{\text{sp}}[\varepsilon_i | W_i = w] = 0, \qquad \text{for } w = 0, 1.$$

The fact that the conditional mean of ε_i given W_i is zero in turn implies unbiasedness of the least squares estimator, $\hat{\tau}^{\text{ols}}$ for $\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$, over the distribution induced by random sampling. The above derivation shows how properties of residuals commonly asserted as assumptions in least squares analyses actually follow from random

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

sampling and random assignment, and thus have a scientific basis in the context of a completely randomized experiment.

Another way of deriving this result, which is closer to the way we will do this for the general case with pre-treatment variables, is to consider the super-population limits of the estimators. The estimators are defined as

$$(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}) = \arg\min_{\alpha,\tau} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i \right)^2 .$$

Under some regularity conditions, these estimators converge, as the sample size goes to infinity, to the population limits (α^*, τ^*) that minimize the expected value of the sum of squares:

$$\begin{aligned}(\alpha^*, \tau^*) &= \arg\min_{\alpha,\tau} \mathbb{E}_{\text{sp}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i \right)^2 \right] \\ &= \arg\min_{\alpha,\tau} \mathbb{E}_{\text{sp}} \left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i \right)^2 \right] .\end{aligned}$$

This implies that the population limit is $\tau^* = \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 0]$. Random assignment of W_i implies $\mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)] = \tau_{\text{sp}}$, so that the population limit of the least squares estimator is equal to the population average treatment effect, $\tau^* = \tau_{\text{sp}}$.

Now let us analyze the least squares approach to inference (i.e., sampling variance and confidence intervals) applied to the setting of a completely randomized experiment. Let us initially assume homoskedasticity ($\sigma^2_{Y|W} = \sigma^2_{\text{c}} = \sigma^2_{\text{t}}$). Using least squares methods, the variance of the residuals would be estimated as

$$\hat{\sigma}^2_{Y|W} = \frac{1}{N-2} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \frac{1}{N-2} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}} \right)^2 ,$$

where the estimated residual is $\hat{\varepsilon}_i = Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}}$, and the predicted value $\hat{Y}_i^{\text{obs}}$ is

$$\hat{Y}_i^{\text{obs}} = \begin{cases} \hat{\alpha}^{\text{ols}} & \text{if } W_i = 0, \\ \hat{\alpha}^{\text{ols}} + \hat{\tau}^{\text{ols}} & \text{if } W_i = 1. \end{cases}$$

The ols variance estimate can be rewritten as

$$\hat{\sigma}^2_{Y|W} = \frac{1}{N-2} \left(\sum_{i:W_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right)^2 + \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}} \right)^2 \right) ,$$

which is equivalent to our calculation of s^2, the common variance across the two potential outcome distributions, as seen in Equation (6.11) in Chapter 6. The conventional

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

estimator for the sampling variance of $\hat{\tau}_{\text{ols}}$ is then

$$\hat{\mathbb{V}}^{\text{homosk}} = \frac{\hat{\sigma}^2_{Y|W}}{\sum_{i=1}^{N}\left(W_i - \overline{W}\right)^2} = s^2 \cdot \left(\frac{1}{N_{\text{c}}} + \frac{1}{N_{\text{t}}}\right).$$

This expression is equal to $\hat{\mathbb{V}}^{\text{const}}$ in Equation (6.12) in Chapter 6. This result is not surprising, because the assumption of homoskedasticity in the linear model setting is implied by the assumption of a constant treatment effect.

For comparison with subsequent results, it is also useful to have the limit of the estimated sampling variance, normalized by the sample size N. Let p be the probability limit of the ratio of the number of treated units to the total number of units, $p = \text{plim}(N_{\text{t}}/N)$. Then, as the sample size increases, the normalized sampling variance estimator converges in probability to

$$N \cdot \hat{\mathbb{V}}^{\text{homosk}} \xrightarrow{p} \frac{\sigma^2_{Y|W}}{p \cdot (1-p)}. \tag{7.1}$$

Note, however, that the random assignment assumption we used for the causal interpretation of $\hat{\tau}^{\text{ols}}$, although it implies independence between assignments and potential outcomes, implies only zero correlation between the assignment and the residual, not necessarily full independence. Yet we rely on this independence to conclude that the variance is homoskedastic. In many cases, the homoskedasticity assumption will not be warranted, and one may wish to use an estimator for the sampling variance of $\hat{\tau}^{\text{ols}}$ that allows for heteroskedasticity. The standard robust sampling variance estimator for least squares estimators is

$$\hat{\mathbb{V}}^{\text{hetero}} = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2 \cdot \left(W_i - \overline{W}\right)^2}{\left(\sum_{i=1}^{N}\left(W_i - \overline{W}\right)^2\right)^2}.$$

Defining, as the previous chapter,

$$s_{\text{c}}^2 = \frac{1}{N_{\text{c}} - 1} \sum_{i:W_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right)^2, \quad \text{and} \quad s_{\text{t}}^2 = \frac{1}{N_{\text{t}} - 1} \sum_{i:W_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}}\right)^2,$$

we can write the variance estimator under heteroskedasticity as

$$\hat{\mathbb{V}}^{\text{hetero}} = \frac{s_{\text{c}}^2}{N_{\text{c}}} + \frac{s_{\text{t}}^2}{N_{\text{t}}}.$$

This is exactly the same estimator for the sampling variance derived from Neyman's perspective in Chapter 6 ($\hat{\mathbb{V}}^{\text{neyman}}$ in Equation (6.8)). So, in the case without additional predictors, the regression approach leads to sampling variance estimators that are familiar from the discussion in the previous chapter. It does, however, provide a different perspective on these results. First of all, it is based on a random sampling perspective. Second, this perspective allows for a natural and simple extension to the case with additional predictors.

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

7.5 LINEAR REGRESSION WITH ADDITIONAL COVARIATES

Now let us consider the case with additional covariates. In this section these additional covariates are included in the regression function additively. The regression function is specified as:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i, \tag{7.2}$$

where X_i is a row vector of covariates (i.e., pre-treatment variables). We estimate the regression coefficients again using least squares:

$$(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\alpha,\tau,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta \right)^2 .$$

The first question we address in this section concerns the causal interpretation of the least squares estimate $\hat{\tau}^{\text{ols}}$ in the presence of these covariates and the associated parameters. We are not interested per se in the value of the "nuisance" parameters, β and α. In particular, we are not interested in a causal interpretation of those parameters. Moreover, we will *not* make the assumption that the regression function in (7.2) is correctly specified or that the conditional expectation of Y_i^{obs} is actually linear in X_i and W_i. However, in order to be precise about the causal interpretation of $\hat{\tau}^{\text{ols}}$, it is useful, as in Section 7.4, to define the limiting values to which the least squares estimators converge as the sample gets large. We will refer to these limiting values as the super-population values corresponding to the estimators and denote them with a superscript $*$, as in Section 7.4. Using this notation, under some regularity conditions, $(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}})$ converge to $(\alpha^*, \tau^*, \beta^*)$, defined as

$$(\alpha^*, \tau^*, \beta^*) = \arg\min_{\alpha,\tau,\beta} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta \right)^2 \right] .$$

These population values are generally well defined (subject, essentially, only to finite-moment conditions and positive definiteness of Ω_X, the population covariance matrix of X_i), even if the conditional expectation of the observed outcome given covariates is not linear in the covariates.

In this case with additional predictors, it is no longer true that $\hat{\tau}^{\text{ols}}$ is unbiased for τ_{sp} in finite samples. However, irrespective of whether the regression function is truly linear in the covariates in the population, the least squares estimate $\hat{\tau}^{\text{ols}}$ is unbiased in large samples for the population average treatment effect, τ_{sp}. Moreover, τ^*, the probability limit of the estimator, is equal to the population average treatment effect τ_{sp}. Finally, in large samples $\hat{\tau}^{\text{ols}}$ will be distributed approximately normally around τ_{sp}. To be precise, we state the result formally.

Theorem 7.1 *Suppose we conduct a completely randomized experiment in a sample drawn at random from an infinite population. Then,* (*i*)

$$\tau^* = \tau_{\text{sp}},$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

and (*ii*),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[(W_i - p)^2 \cdot \left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} \cdot W_i - X_i\beta^*\right)^2\right]}{p^2 \cdot (1-p)^2}\right).$$

We will prove the first part of the result here in the body of the text. The proof of the second part, and of subsequent results, is given in the Appendix to this chapter.

Proof of Theorem 7.1(i). Consider the limiting objective function:

$$\begin{aligned} Q(\alpha, \tau, \beta) &= \mathbb{E}[(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta)^2] \\ &= \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i - (X_i - \mu_X)\beta\right)^2\right], \end{aligned}$$

where $\tilde{\alpha} = \alpha + \mu_X\beta$, with $\mu_X = \mathbb{E}[X_i]$. Minimizing the right-hand side over $\tilde{\alpha}$, τ, and β leads to the same values for τ and β as minimizing the left-hand side over α, τ, and β, with the least squares estimate of $\tilde{\alpha}$ equal $\hat{\alpha} + \hat{\beta}'\mu_X$. Next,

$$\begin{aligned} Q(\tilde{\alpha}, \tau, \beta) &= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i - (X_i - \mu_X)\beta\right)^2\right] \\ &= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right)^2\right] + \mathbb{E}_{\text{sp}}\left[((X_i - \mu_X)\beta)^2\right] \\ &\quad - 2 \cdot \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right) \cdot (X_i - \mu_X)\beta\right] \\ &= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right)^2\right] + \mathbb{E}_{\text{sp}}\left[((X_i - \mu_X)\beta)^2\right] \\ &\quad - 2 \cdot \mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} \cdot (X_i - \mu_X)\beta\right], \end{aligned} \tag{7.3}$$

because

$$\mathbb{E}_{\text{sp}}\left[(X_i - \mu_X)\beta\right] = 0, \quad \text{and} \quad \mathbb{E}_{\text{sp}}\left[\tau \cdot W_i \cdot (X_i - \mu_X)\beta\right] = 0,$$

the first by definition, and the second because of the random sampling and the random assignment. Because the last two terms in (7.3) do not depend on $\tilde{\alpha}$ or τ, minimizing (7.3) over τ and α is equivalent to minimizing the objective function without the additional covariates,

$$\mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right)^2\right],$$

which leads to the solutions

$$\tilde{\alpha}^* = \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 0] = \mathbb{E}_{\text{sp}}\left[Y_i(0)|W_i = 0\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right] = \mu_c,$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

and

$$\begin{aligned}\tau^* &= \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}}|W_i = 0]\\ &= \mathbb{E}_{\text{sp}}[Y_i(1)|W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i(0)|W_i = 0] = \tau_{\text{sp}}.\end{aligned}$$

Thus, the least squares estimator is consistent for the population average treatment effect τ_{sp}. □

What is important in the first part of the result is that the consistency (large-sample unbiasedness) of the least squares estimator for τ_{sp} does *not* depend on the correctness of the specification of the regression function in a completely randomized experiment. No matter how non-linear the conditional expectations of the potential outcomes given the covariates are in the super-population, simple least square regression is consistent for estimating the population average treatment effect. The key insight into this result is that, by randomizing treatment assignment, the super-population correlation between the treatment indicator and the covariates is zero. Even though in finite samples the actual correlation may differ from zero, in large samples this correlation will vanish, and as a result the inclusion of the covariates does not matter for the limiting values of the estimator. The fact that in finite samples the correlation may differ from zero is what leads to the possibility of finite-sample bias.

Although the inclusion of the additional covariates does not matter for the limit of the corresponding estimator, it does matter for the sampling variance of the estimators. Let us interpret the sampling variance in some special cases. Suppose that, in fact, the conditional expectation of the two potential outcomes is linear in the covariates, with the same slope coefficients but different intercepts in the two treatment arms, or

$$\mathbb{E}_{\text{sp}}[Y_i(0)|X_i = x] = \alpha_{\text{c}} + x\beta, \quad \text{and} \quad \mathbb{E}_{\text{sp}}[Y_i(1)|X_i = x] = \alpha_{\text{t}} + x\beta,$$

so that, in combination with random assignment, we have

$$\mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} \middle| X_i = x, W_i = 1\right] = \alpha_{\text{c}} + \tau_{\text{sp}} \cdot t + x\beta,$$

where $\tau_{\text{sp}} = \alpha_{\text{t}} - \alpha_{\text{c}}$. Suppose that, in addition, the variance of the two potential outcomes does not vary by treatment or covariates:

$$\mathbb{V}_{\text{sp}}(Y_i(w)|X_i = x) = \sigma^2_{Y|W,X},$$

for $w = 0, 1$, and all x. Then the normalized sampling variance for the least squares estimator for τ_{sp}, given for the general case in Theorem 7.1, simplifies to

$$N \cdot \mathbb{V}_{\text{sp}}^{\text{homosk}} = \frac{\sigma^2_{Y|W,X}}{p \cdot (1-p)}. \tag{7.4}$$

This expression reveals the gain in precision from including the covariates. Instead of the unconditional variance of the potential outcomes, as in the expression for the sampling variance in the case without covariates in (7.1), we now have the conditional variance of the outcome given the covariates. If the covariates explain much of the variation in the potential outcomes, so that the conditional variance $\sigma^2_{Y|W,X}$ is substantially smaller than

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

the marginal variance $\sigma^2_{Y|W}$, then including the covariates in the regression model will lead to a considerable increase in precision. The price paid for the increase in precision from including covariates is relatively minor. Instead of having (exact) unbiasedness of the estimator in finite samples, unbiasedness now only holds approximately, that is, in large samples.

The sampling variance for the average treatment effect can be estimated easily using standard least squares methods. Substituting averages for the expectations, and least squares estimates for the unknown parameters, we estimate the sampling variance as

$$\hat{\mathbb{V}}^{\text{hetero}}_{\text{sp}} = \frac{1}{N\,(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^N \left(W_i - \overline{W}\right)^2 \cdot \left(Y_i^{\text{obs}} - \hat{\alpha}^{\text{ols}} - \hat{\tau}^{\text{ols}} - X_i\hat{\beta}^{\text{ols}}\right)^2}{\left(\overline{W}\cdot(1-\overline{W})\right)^2}.$$

If one wishes to impose homoskedasticity, one can still use the heteroskedasticity-consistent sampling variance estimator, but a more precise estimator of the sampling variance imposes homoskedasticity, leading to the form:

$$\hat{\mathbb{V}}^{\text{homo}}_{\text{sp}} = \frac{1}{N\,(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^N \left(Y_i^{\text{obs}} - \hat{\alpha}^{\text{ols}} - \hat{\tau}^{\text{ols}} - X_i\hat{\beta}^{\text{ols}}\right)^2}{\overline{W}\cdot(1-\overline{W})}.$$

7.6 LINEAR REGRESSION WITH COVARIATES AND INTERACTIONS

In this section we take the analysis of Section 7.5 one step further. In addition to including the covariates linearly, one may wish to interact the covariates with the indicator for the receipt of treatment if we expect that the association between the covariates and the outcome varies by treatment status. The motivation for this is twofold. First, adding additional covariates of any form, including those based on interactions, may further improve the precision of the estimator. Second, by interacting all such predictors with the treatment indicators, we achieve a particular form of robustness to model misspecification that we discuss in more detail later. This robustness is not particularly important in the current setting of a completely randomized experiment, but it will be important in observational studies discussed in Parts III and IV of this text. We specify the regression function as

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

We include the interaction of the treatment indicator with the covariates in deviations from their sample means to simplify the relationship between the population limits of the estimators for the parameters of the regression function and τ_{sp}.

Let $\hat{\alpha}^{\text{ols}}$, $\hat{\tau}^{\text{ols}}$, $\hat{\beta}^{\text{ols}}$, and $\hat{\gamma}^{\text{ols}}$ denote the least squares estimates,

$$(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}) = \arg\min_{\alpha,\tau,\beta,\gamma} \sum_{i=1}^N \left(Y_i^{\text{obs}} - \alpha - \tau\cdot W_i - X_i\beta - W_i\cdot(X_i-\overline{X})\gamma\right)^2,$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

and let α^*, τ^*, β^*, and γ^* denote the corresponding population values:

$$(\alpha^*, \tau^*, \beta^*, \gamma^*) = \arg \min_{\alpha,\tau,\beta,\gamma} \mathbb{E}_{\text{sp}} \left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta - W_i \cdot (X_i - \mu_X)\gamma \right)^2 \right].$$

Results similar to Theorem 7.1 can be obtained for this case. The least squares estimator $\hat{\tau}^{\text{ols}}$ is consistent for the average treatment effect τ_{sp}, and inference can be based on least squares methods.

Theorem 7.2 *Suppose we conduct a completely randomized experiment in a random sample from a super-population. Then* (*i*)

$$\tau^* = \tau_{\text{sp}},$$

and (*ii*),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}} \right) \xrightarrow{d} \mathcal{N}$$
$$\left(0, \frac{\mathbb{E}_{\text{sp}} \left[(W_i - p)^2 \cdot \left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} \cdot W_i - X_i\beta^* - W_i \cdot (X_i - \mu_X)\gamma^* \right)^2 \right]}{p^2 \cdot (1-p)^2} \right).$$

The proof for this theorem is provided in the Appendix.

A slightly different interpretation of this result connects it to the imputation-based methods that are the topic of the next chapter. Suppose we take the model at face value and assume that the regression function represents the conditional expectation:

$$\mathbb{E}_{\text{sp}} \left[Y_i^{\text{obs}} \middle| X_i = x, W_i = w \right] = \alpha + \tau \cdot t + x\beta + w \cdot (x - \mu_X)\gamma. \tag{7.5}$$

In combination with the random assignment, this implies that

$$\begin{aligned} \mathbb{E}_{\text{sp}} [Y_i(0) | X_i = x] &= \mathbb{E}_{\text{sp}} [Y_i(0) | X_i = x, W_i = 0] \\ &= \mathbb{E}_{\text{sp}} \left[Y_i^{\text{obs}} \middle| X_i = x, W_i = 0 \right] = \alpha + x\beta, \end{aligned}$$

and

$$\mathbb{E}_{\text{sp}} [Y_i(1) | X_i = x] = \alpha + \tau + x\beta + (x - \mu_X)\gamma.$$

Suppose that unit i was exposed to the treatment ($W_i = 1$), so $Y_i(1)$ is observed and $Y_i(0)$ is missing. Under the model in (7.5), the predicted value for the missing potential outcome $Y_i(0)$ is

$$\hat{Y}_i(0) = \hat{\alpha}^{\text{ols}} + X_i\hat{\beta}^{\text{ols}},$$

so that for this treated unit the predicted value for the unit-level causal effect is

$$\hat{\tau}_i = Y_i(1) - \hat{Y}_i(0) = Y_i^{\text{obs}} - \left(\hat{\alpha}^{\text{ols}} + X_i\hat{\beta}^{\text{ols}} \right).$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

For a control unit i (with $W_i = 0$) the predicted value for the missing potential outcome $Y_i(1)$ is

$$\hat{Y}_i(1) = \hat{\alpha}^{\text{ols}} + \hat{\tau}^{\text{ols}} + X_i\hat{\beta}^{\text{ols}} + (X_i - \overline{X})\hat{\gamma}^{\text{ols}},$$

and the predicted value for the unit-level causal effect for this control unit i is

$$\hat{\tau}_i = \hat{Y}_i(1) - Y_i(0) = \hat{\alpha}^{\text{ols}} + \hat{\tau}^{\text{ols}} + X_i\hat{\beta}^{\text{ols}} + (X_i - \overline{X})\hat{\gamma}^{\text{ols}} - Y_i^{\text{obs}}.$$

Now we can estimate the overall average treatment effect τ_{fs} by averaging the estimates of the unit-level causal effects $\hat{\tau}_i$. Simple algebra shows that this leads to the ols estimator:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i=1}^{N}\left\{W_i \cdot \left(Y_i(1) - \hat{Y}_i(0)\right) + (1 - W_i) \cdot \left(\hat{Y}_i(1) - Y_i(0)\right)\right\} = \hat{\tau}^{\text{ols}}.$$

Thus, the least squares estimator $\hat{\tau}^{\text{ols}}$ can be interpreted as averaging estimated unit-level causal effects in the sample, based on imputing the missing potential outcomes through a linear regression model. However, as has been stressed repeatedly, thanks to the randomization, the consistency of the ols estimator does not rely on the validity of the regression model as an approximation to the conditional expectation.

There is another important feature of the estimator based on linear regression with a full set of interactions that was alluded to at the beginning of this chapter. As the above derivation shows, the estimator essentially imputes the missing potential outcomes. The regression model with a full set of interactions does so separately for the treated and control units. When imputing the value of $Y_i(0)$ for the treated units, this procedure uses only the observed outcomes, Y_i^{obs}, for control units, without any dependence on observations on $Y_i(1)$ (and vice versa). This gives the estimator attractive robustness properties, clearly separating imputation of control and treated outcomes. This will be important in the context of observational studies.

7.7 TRANSFORMATIONS OF THE OUTCOME VARIABLE

If one is interested in the average effect of the treatment on a transformation of the outcome, one can first transform the outcome and then apply the methods discussed so far. For example, in order to estimate the average effect on the logarithm of the outcome, we can first take logarithms and then estimate the regression function

$$\ln\left(Y_i^{\text{obs}}\right) = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i.$$

Irrespective of the form of the association between outcomes and covariates, in a completely randomized experiment, least squares estimates of τ are consistent for the average effect $\mathbb{E}[\ln(Y_i(1)) - \ln(Y_i(0))]$. This follows directly from the previous discussion. There is an important issue, though, involving such transformations that relates to the correctness of the specification of the regression function. Suppose one is interested in the average effect $\mathbb{E}[Y_i(1) - Y_i(0)]$, but suppose that one actually suspects that a model

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

linear in logarithms provides a better fit to the distribution of Y_i^{obs} given X_i and W_i. Estimating a model linear in logarithms and transforming the estimates back to an estimate of the average effect in levels requires assumptions beyond those on the conditional expectation of the logarithm of the potential outcomes: one needs to make distributional assumptions on the unobserved component. We discuss such modeling strategies in the next chapter.

As an extreme example of this issue, consider the case where the researcher is interested in the average effect of the treatment on a binary outcome. Estimating a linear regression function by least squares will lead to a consistent estimator for the average treatment effect. However, such a linear probability model is unlikely to provide an accurate approximation of the conditional expectation of the outcome given covariates and treatment indicator. Logistic models (where $\Pr(Y_i^{\text{obs}} = 1|W_i = w, X_i = x)$ is modeled as $\exp(\alpha+\tau\cdot w+x\beta)/(1+\exp(\alpha+\tau\cdot w+x\beta))$), or probit models (where $\Pr(Y_i^{\text{obs}} = 1|W_i = w, X_i = x) = \Phi(\alpha + \tau \cdot w + x\beta)$, with $\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-1/2} \exp(-z^2/2)$ the normal cumulative distribution function) are more likely to lead to an accurate approximation of the conditional expectation of the outcome given the covariates and the treatment indicator. However, such a model will not generally lead to a consistent estimator for the average effect unless the model is correctly specified. Moreover, the average treatment effect cannot be expressed directly in terms of the parameters of the logistic or probit regression model.

The issue is that in the regression approach, the specification of the statistical model is closely tied to the estimand of interest. In the next chapter we separate these two issues. This separation is attractive for a number of reasons discussed in more detail in the next chapter, but it also carries a price, namely that consistency of the estimators will be tied more closely to the correct specification of the model. We do not view this as a major issue. In the setting of completely randomized experiments, the bias is unlikely to be substantial with moderate-sized samples, as flexible models are likely to have minimal bias. Moreover, this consistency property despite possible misspecification of the regression function holds only with completely randomized experiments. In observational studies, even regression models rely heavily on the correct specification for consistency of the estimator. Furthermore, large-sample results, such as consistency, are only guidelines for finite-sample properties, and as such not always reliable.

7.8 THE LIMITS ON INCREASES IN PRECISION DUE TO COVARIATES

In large samples, including covariates in the regression function will not lower, and generally will increase, the precision of the estimator for the average treatment effect. However, beyond the first few covariates, more covariates are unlikely to improve the precision substantially in modest-sized samples. Here we briefly discuss some limits to the gains in precision from including covariates in settings where the randomized assignment ensures that the covariates are not needed for bias removal.

Suppose we do not include any predictor variables in the regression beyond the indicator variable for the treatment, W_i, that is, we include no covariates. Normalized by the sample size, the sampling variance of the least squares estimator, in this case equal to the simple difference in means, is equal to

$$N \cdot \mathbb{V}_{\text{nocov}} = \frac{\sigma_{\text{c}}^2}{1-p} + \frac{\sigma_{\text{t}}^2}{p},$$

familiar in various forms from this and the previous chapter. Now suppose we have available a vector of covariates, X_i. Including these covariates, their interactions with the treatment indicator, and possibly higher-order moments of these covariates, leads to a normalized sampling variance that is bounded from below by

$$N \cdot \mathbb{V}_{\text{bound}} = \frac{\mathbb{E}_{\text{sp}}[\sigma_{\text{c}}^2(X_i)]}{1-p} + \frac{\mathbb{E}_{\text{sp}}[\sigma_{\text{t}}^2(X_i)]}{p}.$$

Instead of the marginal variances σ_{c}^2 and σ_{t}^2 in the two terms, we now take the expectation of the conditional variances $\sigma_{\text{c}}^2(X_i)$ and $\sigma_{\text{t}}^2(X_i)$. The difference between the two expressions for the sampling variance, and thus the gain from including the covariates in a flexible manner, is the sum of the sampling variances of the conditional means of $Y_i(w)$ given X_i:

$$\begin{aligned}\mathbb{V}_{\text{nocov}} - \mathbb{V}_{\text{bound}} &= \left(\frac{\sigma_{\text{c}}^2}{1-p} + \frac{\sigma_{\text{t}}^2}{p}\right) - \left(\frac{\mathbb{E}_{\text{sp}}[\sigma_{\text{c}}^2(X_i)]}{1-p} + \frac{\mathbb{E}_{\text{sp}}\left[\sigma_{\text{t}}^2(X_i)\right]}{p}\right)\\ &= \frac{\mathbb{V}_{\text{sp}}(\mu_{\text{c}}(X_i))}{1-p} + \frac{\mathbb{V}_{\text{sp}}(\mu_{\text{t}}(X_i))}{p}.\end{aligned}$$

The more the covariates X_i help in explaining the potential outcomes, and thus the bigger the variation in $\mu_w(x)$, the bigger the gain from including them in the specification of the regression function. In the extreme case, where neither $\mu_{\text{c}}(x)$ nor $\mu_{\text{t}}(x)$ varies with the predictor variables, there is no gain from using the covariates, even in large samples. Moreover, in small samples there will actually be a loss of precision due to the estimation of coefficients, that are, in fact, zero.

7.9 TESTING FOR THE PRESENCE OF TREATMENT EFFECTS

In addition to estimating average treatment effects, the regression models discussed in this chapter have been used to test for the presence of treatment effects. In the current setting of completely randomized experiments, tests for the presence of any treatment effects are not necessarily as attractive as the Fisher exact p-value calculations discussed in Chapter 5, but their extensions to observational studies are relevant. In addition, we may be interested in testing hypotheses concerning the heterogeneity in the treatment effects that do not fit into the FEP framework because the associated null hypotheses are not sharp. As in the discussion of estimation, we focus on procedures that are valid in large samples, irrespective of the correctness of the specification of the regression model.

The most interesting setting is the one where we allow for a full set of first-order interactions with the treatment indicator and specify the regression function as

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

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

In that case we can test the null hypothesis of a zero average treatment effect by testing the null hypothesis that $\tau_{\rm sp} = 0$. However, we can construct a different test by focusing on the deviation of either $\hat{\tau}_{\rm sp}$ or $\hat{\gamma}$ from zero. If the regression model were correctly specified, that is, if the conditional expectation of the outcome in the population given covariates and treatment indicator were equal to

$$\mathbb{E}_{\rm sp}\left[Y_i^{\rm obs} \middle| X_i = x, W_i = w \right] = \alpha + \tau \cdot w + x\beta + w \cdot (x - \mu_X)\gamma',$$

this would test the null hypothesis that the average treatment effect conditional on each value of the covariates is equal to zero, or

$$H_0 : \; \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0)|X_i = x] = 0, \quad \forall\, x,$$

against the alternative hypothesis

$$H_a : \; \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0)|X_i = x] \neq 0, \quad \text{for some } x.$$

Without making the assumption that the regression model is correctly specified, it is still true that, if the null hypothesis that $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = 0$ for all x were correct, then the population values $\tau_{\rm sp}$ and γ^* would be equal to zero. However, it is no longer true that for *all* deviations of this null hypothesis the limiting values of either $\tau_{\rm sp}$ or γ^* differ from zero. It is possible that $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ differs from zero for some values of x even though $\tau_{\rm sp}$ and γ^* are both equal to zero.

In order to implement these tests, one can again use standard least squares methods. The normalized covariance matrix of the vector $(\hat{\tau}^{\rm ols}, \hat{\gamma}^{\rm ols})$ is

$$\mathbb{V}_{\tau,\gamma} = \begin{pmatrix} \mathbb{V}_\tau & \mathbb{C}_{\tau,\gamma} \\ \mathbb{C}_{\tau,\gamma}^T & \mathbb{V}_\gamma \end{pmatrix}.$$

The precise form of the components of the covariance matrix, as well as consistent estimators for these components, is given in the Appendix. In order to test the null hypothesis that the average effect of the treatment given the covariates is zero for all values of the covariates, we then use the quadratic form

$$Q_{\rm zero} = \begin{pmatrix} \hat{\tau}^{\rm ols} \\ \hat{\gamma}^{\rm ols} \end{pmatrix}^T \hat{\mathbb{V}}_{\tau,\gamma}^{-1} \begin{pmatrix} \hat{\tau}^{\rm ols} \\ \hat{\gamma}^{\rm ols} \end{pmatrix}. \tag{7.6}$$

Note that this is not a test that fits into the Fisher exact p-value approach because it does not specify all missing potential outcomes under the null hypothesis.

The second null hypothesis we consider is that the average treatment effect is constant as a function of the covariates:

$$H_0' : \; \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0)|X_i = x] = \tau_{\rm sp}, \quad \text{for all } x,$$

against the alternative hypothesis

$$H_a' : \; \exists\, x_0, x_1, \text{ such that } \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0)|X_i = x_0] \neq \mathbb{E}_{\rm sp}[Y_i(1) - Y_i(0)|X_i = x_1].$$

This null hypothesis may be of some importance in practice. If there is evidence of heterogeneity in the effect of the treatment as a function of the covariates, one has to be more careful in extrapolating to different subpopulations. On the other hand, if there is no evidence of heterogeneity by observed characteristics, and if the distribution of these characteristics in the sample is sufficiently varied, it may be more credible to extrapolate estimates to different subpopulations. (Of course, lack of positive evidence for heterogeneity does not imply a constant treatment effect, but in cases with sufficient variation in the covariates, it does suggest that treatment-effect heterogeneity may be a second-order problem.) In order to test this null hypothesis, we can use the quadratic form

$$Q_{\text{const}} = (\hat{\gamma}^{\text{ols}})^T \hat{\mathbb{V}}_{\gamma}^{-1} \hat{\gamma}^{\text{ols}}. \tag{7.7}$$

Theorem 7.3 *Suppose we conduct a completely randomized experiment in a random sample from a large population. If $Y_i(1) - Y_i(0) = \tau$ for some value τ and all units, then* (*i*)*: $\gamma^* = 0$,*
and (*ii*)

$$Q_{\text{const}} \xrightarrow{d} \mathcal{X}(\dim(X_i)).$$

If $Y_i(1) - Y_i(0) = 0$ for all units, then (*iii*)*,*

$$Q_{\text{zero}} \xrightarrow{d} \mathcal{X}(\dim(X_i) + 1).$$

7.10 ESTIMATES FOR LRC-CPPT CHOLESTEROL DATA

Now let us return to the LRC-CPPT cholesterol data. We look at estimates for two average effects. First, the effect on post-treatment cholesterol levels, the primary outcome of interest, denoted by `cholf`. Second, partly anticipating some of the analyses in Chapters 23–25, we estimate the effect of assignment to treatment on the level of compliance, `comp`. Because compliance was far from perfect (on average, individuals assigned to the control group took 75% of the nominal dose, and individuals in the group assigned to the active treatment, on average, took 60% of the nominal dose), the estimates of the effect on post-assignment cholesterol levels should be interpreted as estimates of *intention-to-treat* (ITT) effects, that is, average effects of assignment to the drug versus assignment to the placebo, rather than as estimates of the effects of the efficacy of the drug.

For each outcome, we present four regression estimates of the average effects. First, we use a simple linear regression with only the indicator for assignment. Second, we include the composite prior cholesterol level `cholp` as a linear predictor. Third, we include both prior cholesterol-level measurements, `chol1` and `chol2`, as linear predictors. Fourth, we add interactions of the two prior cholesterol-level measurements with the assignment indicator.

Table 7.2 presents the results for these regressions. For the cholesterol-level outcome, the average effect is estimated in all cases reported to be a reduction of approximately 25–26 units, approximately an 8% reduction. Including predictors beyond the treatment

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

Table 7.2. ***Regression Estimates for Average Treatment Effects for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Effect of Assignment to Treatment on			
	Post-Cholesterol Level		Compliance	
	$\hat{\tau}$	$\widehat{\text{(s. e.)}}$	$\hat{\tau}$	$\widehat{\text{(s. e.)}}$
No covariates	−26.22	(3.93)	−14.64	(3.51)
`cholp`	−25.01	(2.60)	−14.68	(3.51)
`chol1, chol2`	−25.02	(2.59)	−14.95	(3.50)
`chol1, chol2`, interacted with *W*	−25.04	(2.56)	−14.94	(3.49)

Table 7.3. ***Regression Estimates for Average Treatment Effects on Post-Cholesterol Levels for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Model for Levels		Model for Logs	
	Est	$\widehat{\text{(s. e.)}}$	Est	$\widehat{\text{(s. e.)}}$
Assignment	−25.04	(2.56)	−0.098	(0.010)
Intercept	−3.28	(12.05)	−0.133	(0.233)
`chol1`	0.98	(0.04)	−0.133	(0.233)
`chol2-chol1`	0.61	(0.08)	0.602	(0.073)
`chol1` × Assignment	−0.22	(0.09)	−0.154	(0.107)
(`chol2-chol1`) × Assignment	0.07	(0.14)	0.184	(0.159)
R-squared	0.63		0.57	

indicator improves the precision considerably, reducing the estimated standard error by a third. Including predictors beyond the simple composite prior cholesterol level `cholp` does not affect the estimated precision appreciably. For the effect of the assignment on receipt of the drug, the estimated effect is also stable across the different specifications of the regression function. For this outcome the estimated precision does not change with the inclusion of additional predictors.

The left panel of Table 7.3 presents more detailed results for the regression of the outcome on the covariates and the interaction of covariates with the treatment indicator. Although substantively the coefficients of the covariates are not of interest in the current setting, we can see from these results that the covariates do add considerable predictive power to the regression function. This predictive power is what leads to the increased precision of the estimator for the average treatment effect based on the regression with covariates relative to the regression without covariates. For the purpose of assessing the relative predictive power of different specifications, we also report, in the right panel of Table 7.3, the results for a regression after transforming all cholesterol levels to logarithms. As stressed before, this changes the estimand, and so the results are not directly comparable. It is useful to note, though, that in this case the transformation does not improve the predictive power, in the sense that the squared correlation between the observed outcomes and the covariates decreases as a result of this transformation.

Table 7.4. ***P-Values for Tests for Constant and Zero Treatment Effects, Using*** `chol1` ***and*** `chol2-chol1` ***as Covariates for the PRC-CPPT Cholesterol Data from Table 7.1***

		Post-Cholesterol Level	Compliance
Zero treatment effect	$\mathcal{X}^2(3)$ approximation	<0.001	<0.001
	Fisher exact p-value	<0.001	0.001
Constant treatment effect	$\mathcal{X}^2(2)$ approximation	0.029	0.270

In Table 7.4 we report p-values for some of the tests discussed in Section 7.9. First we consider the null hypothesis that the effect of the treatment on the final cholesterol level is zero. We use the statistic Q_{zero} given in Equation (7.6), based on the regression with the two prior cholesterol levels and their interactions with the treatment as covariates. Under this null hypothesis, this statistic has, in large samples, a chi-squared distribution with three degrees of freedom. The value of the statistic in the sample is 100.48, which leads to an approximate p-value based on the chi-squared distribution with three degrees of freedom less than 0.001. We perform the same calculations using the compliance variable as the outcome of interest. Now the value of the test statistic is 19.27, again leading to an approximate p-value less than 0.001. Because under the null hypothesis of no effect whatsoever, we can apply the FEP approach, we also calculate the exact p-values. For the post-cholesterol level, the FEP calculations lead to a p-value less than 0.001. For the compliance outcome, the p-value based on the FEP approach is 0.001. The p-values under the FEP approach are similar to those based on large-sample approximations because, with the sample size used in this example, a total of 337 units, 172 in the control group and 165 in the treatment group, and the data values, the normal approximations that underlie the large-sample properties of the tests are accurate.

Next, we test the null hypothesis that the treatment effect is constant against the alternative that it varies between units, using the statistic Q_{const} given in (7.7). For the final cholesterol-level outcome, the value of the test statistic is 7.05, leading to a p-value based on the chi-squared approximation with two degrees of freedom equal to 0.029. For the compliance outcome, the value of the statistic is 2.62, leading to an approximate p-value of 0.269. Note that in this case, because of the presence of nuisance parameters (we do not restrict the level of the treatment effect, only its variance), the FEP approach is not applicable. Together the tests suggest that the evidence for the presence of treatment effects is very strong but that the evidence for heterogeneity in the treatment effect is weak.

Overall, with the caveat of the multiple testing, the message from this application supports the conclusion that including some covariates can substantially improve the estimated precision of the inferences, although including many covariates is unlikely to be helpful beyond the inclusion of the most important ones.

7.11 CONCLUSION

In this chapter we discuss regression methods for estimating causal effects in the context of a completely randomized experiment. Regression models are typically motivated by assumptions on conditional mean functions. Such assumptions are difficult to justify

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

other than as approximations. In the context of a completely randomized experiment, however, we can use the randomization to help justify the key assumptions necessary for consistency of the least squares estimator. In contrast to the methods discussed in previous chapters, most of these results are only approximate, relying on large samples. In that sense, the regression methods can be viewed as providing a bridge from the exact results based on randomization inference to the model-based methods that will be discussed in the next chapter.

Regression methods can easily incorporate covariates into estimands and, in that sense lead to an attractive extension of Neyman's basic approach discussed in Chapter 6. In settings with completely randomized experiments, they offer a simple and widely used framework for estimating and constructing confidence intervals for average treatment effects. The main disadvantage is that they are closely tied to linearity. In completely randomized experiments, this linearity is not a particularly important concern, because the methods still lead to consistent estimators for average treatment effects. In observational studies, however, this reliance on linearity can make regression methods sensitive to minor changes in specification. In those settings, discussed in detail in Parts III and IV of this text, simple regression methods are not recommended.

NOTES

The Efron-Feldman data were also analyzed in Jin and Rubin (2008) using a principal stratification approach. In their analysis, the focus is on the causal effect of the actual dose of the drug taken, rather than on the (intention-to-treat) effect of the assignment to the drug.

Cochran (1977) and Goldberger (1991) have extensive discussions on the properties of least squares estimators in settings where the conditional expectation is not necessarily linear, and on the notion of the "best linear predictor" (Goldberger, 1991, p. 52). Gail, Wieand, and Piantadosi (1984) discuss biases in estimated treatment effects in the context of non-linear regression models with experimental data. See also Lin (2012) and Miratrix, Sekhon, and Yu (2013). Lesaffre and Senn (2003) discuss the properties of alternative covariance adjustment methods. Koch, Tangen, Jung, and Amara (1998) discuss regression methods in settings with binary and ordered discrete outcome data. Victora, Habicht, and Bryce (2004) discuss regression methods in health applications.

The discussion in Section 7.8 on the limits of the gains in precision from incorporating pre-treatment variables draws on the results in Hahn (1998). See also Robins and Rotnitzky (1995) and Hirano, Imbens, and Ridder (2003).

Freedman (2008ab) discusses the role of regression analyses in the context of randomized experiments. He suggests, as evidenced by the quotes in the introduction to this chapter, that the use of regression analysis is not always warranted, a view to which we also subscribe. Angrist and Pischke (2008) and Lin (2012) present a less critical view of the use of regression methods for causal inference.

Senn (1994) and Imai, King, and Stuart (2008) discuss the motivation for testing or not testing for baseline balance in randomized experiments.

APPENDIX

Proof of Theorem 7.1

It is convenient to reparametrize the model. Instead of (α, τ, β), we parametrize the model using $(\tilde{\alpha}, \tau, \beta)$, where $\tilde{\alpha} = \alpha - p \cdot \tau - \mathbb{E}_{\text{sp}}[X_i]\beta$. The reparametrization does not change the ols estimates for τ and β, nor their limiting values. The limiting value of the new parameter is $\tilde{\alpha}^* = \alpha^* - p \cdot \tau_{\text{sp}} - \mathbb{E}_{\text{sp}}[X_i]\beta^*$. In terms of these parameters, the objective function is

$$\sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \left(\tilde{\alpha} - p \cdot \tau - \mathbb{E}_{\text{sp}}[X_i]\beta \right) - \tau \cdot W_i - X_i\beta \right)^2$$

$$= \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot (W_i - p) - \left(X_i - \mathbb{E}_{\text{sp}}[X_i] \right) \beta \right)^2 .$$

The first-order conditions for the estimators $(\hat{\tilde{\alpha}}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}})$ are

$$\sum_{i=1}^{N} \psi(Y_i^{\text{obs}}, W_i, X_i, \hat{\tilde{\alpha}}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = 0,$$

where $\psi(\,\cdot\,)$ is a three-component column vector:

$$\psi(y, w, x, \alpha, \tau, \beta) = \begin{pmatrix} y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i] \right) \beta \\ (w - p) \cdot \left(y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i] \right) \beta \right) \\ \left(x - \mathbb{E}_{\text{sp}}[X_i] \right) \cdot \left(y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i] \right) \beta \right) \end{pmatrix}.$$

Given the population values of the parameters, α^*, τ_{sp}, and β^*, standard M-estimation (or generalized method of moments) results imply that under standard regularity conditions the estimator is consistent and asymptotically normally distributed:

$$\sqrt{N} \cdot \begin{pmatrix} \hat{\tilde{\alpha}}^{\text{ols}} - \alpha^* \\ \hat{\tau}^{\text{ols}} - \tau_{\text{sp}} \\ \hat{\beta}^{\text{ols}} - \beta^* \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1} \Delta (\Gamma^T)^{-1} \right),$$

where the two components of the covariance matrix are

$$\Gamma = \mathbb{E}_{\text{sp}} \left[\frac{\partial}{\partial(\alpha, \tau, \beta)} \psi(Y_i^{\text{obs}}, W_i, X_i, \alpha, \tau, \beta) \right] \Bigg|_{(\tilde{\alpha}^*, \tau_{\text{sp}}, \beta^*)}$$

$$= \mathbb{E}_{\text{sp}} \left[\begin{pmatrix} -1 & -(W_i - p) & -(X_i - \mathbb{E}_{\text{sp}}[X_i]) \\ -(W_i - p) & -(W_i - p)^2 & -(W_i - p) \cdot (X_i - \mathbb{E}_{\text{sp}}[X_i]) \\ -(X_i - \mathbb{E}_{\text{sp}}[X_i])^T & -(W_i - p) \cdot (X_i - \mathbb{E}_{\text{sp}}[X_i])^T & -(X_i - \mathbb{E}_{\text{sp}}[X_i])^T \cdot (X_i - \mathbb{E}_{\text{sp}}[X_i]) \end{pmatrix} \right]$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

$$= \mathbb{E}_{\rm sp}\left[\begin{pmatrix} -1 & 0 & 0 \\ 0 & -p(1-p) & 0 \\ 0 & 0 & -\mathbb{E}_{\rm sp}\left[(X_i - \mathbb{E}_{\rm sp}[X_i])^T \cdot (X_i - \mathbb{E}_{\rm sp}[X_i])\right] \end{pmatrix}\right],$$

and

$$\Delta = \mathbb{E}_{\rm sp}\left[\psi(Y_i^{\rm obs}, W_i, X_i, \tilde{\alpha}^*, \tau_{\rm sp}, \beta^*) \cdot \psi(Y_i^{\rm obs}, W_i, X_i, \tilde{\alpha}^*, \tau_{\rm sp}, \beta^*)^T\right]$$
$$= \mathbb{E}_{\rm sp}\left[\left(Y_i^{\rm obs} - \alpha^* - \tau_{\rm sp} - X_i\beta^*\right)^2 \cdot \begin{pmatrix} 1 \\ W_i - p \\ (X_i - \mathbb{E}_{\rm sp}[X_i])^T \end{pmatrix} \begin{pmatrix} 1 \\ W_i - p \\ (X_i - \mathbb{E}_{\rm sp}[X_i])^T \end{pmatrix}^T\right].$$

The variance of $\hat{\tau}$ is the $(2,2)$ element of the covariance matrix. Because Γ is block diagonal, the $(2,2)$ element of $\Gamma^{-1}\Delta(\Gamma^T)^{-1}$ is equal to the $(2,2)$ element of Δ divided by $(p(1-p))^2$, which is equal to

$$\mathbb{E}_{\rm sp}\left[\left(Y_i^{\rm obs} - \alpha^* - \tau_{\rm sp} - X_i\beta^*\right)^2 \cdot (W_i - p)^2\right].$$

Hence the variance of $\hat{\tau}$, normalized by the sample size N, is equal to

$$\frac{\mathbb{E}_{\rm sp}\left[\left(Y_i^{\rm obs} - \alpha^* - \tau_{\rm sp} - X_i\beta^*\right)^2 \cdot (W_i - p)^2\right]}{p^2 \cdot (1-p)^2}.$$

□

Proof of Theorem 7.2

First we show that in this case τ^* the population value of $\hat{\tau}$, equal to

$$(\alpha^*, \tau^*, \beta^*, \gamma^*) = \arg\min_{\alpha,\beta,\tau,\gamma} \mathbb{E}_{\rm sp}\left[\left(Y_i^{\rm obs} - \alpha - \tau \cdot W_i - X_i\beta - W_i \cdot (X_i - \mu_X)\gamma\right)^2\right],$$

is equal to $\tau_{\rm sp}$. Again it is useful to reparametrize. The new vector of parameters is

$$\begin{pmatrix} \tilde{\alpha}_{\rm c} \\ \beta_{\rm c} \\ \tilde{\alpha}_{\rm t} \\ \beta_{\rm t} \end{pmatrix} = \begin{pmatrix} \alpha + \mu_X\beta \\ \beta \\ \alpha + \tau + \mu_X\beta \\ \gamma + \beta \end{pmatrix},$$

with inverse

$$\begin{pmatrix} \alpha \\ \beta \\ \tau \\ \gamma \end{pmatrix} = \begin{pmatrix} \tilde{\alpha}_{\rm c} - \mu_X\beta_{\rm c} \\ \beta_{\rm c} \\ \tilde{\alpha}_{\rm t} - \tilde{\alpha}_{\rm c} \\ \beta_{\rm t} - \beta_{\rm c} \end{pmatrix}.$$

In terms of this parameter vector the minimization problem is

$$\begin{aligned}&(\tilde{\alpha}_{\mathrm{c}}^*, \tilde{\alpha}_{\mathrm{t}}^*, \beta_{\mathrm{c}}^*, \beta_{\mathrm{t}}^*)\\ &= \arg \min_{\alpha_{\mathrm{c}},\alpha_{\mathrm{t}},\beta_{\mathrm{c}},\beta_{\mathrm{t}}} \mathbb{E}_{\mathrm{sp}}\left[\left(Y_i^{\mathrm{obs}} - \alpha_{\mathrm{c}} - (\alpha_{\mathrm{t}} - \alpha_{\mathrm{c}}) \cdot W_i - X_i\beta_{\mathrm{c}}\right.\right.\\ &\quad\left.\left. - W_i \cdot (X_i - \mu_X)(\beta_{\mathrm{t}} - \beta_{\mathrm{c}})\right)^2\right]\\ &= \arg \min_{\alpha_{\mathrm{c}},\alpha_{\mathrm{t}},\beta_{\mathrm{c}},\beta_{\mathrm{t}}} \mathbb{E}_{\mathrm{sp}}\left[(1 - W_i) \cdot \left(Y_i^{\mathrm{obs}} - \alpha_{\mathrm{c}} - (X_i - \mu_X)\beta_{\mathrm{c}}\right)^2\right.\\ &\quad\left. + W_i \cdot \left(Y_i^{\mathrm{obs}} - \alpha_{\mathrm{t}} - (X_i - \mu_X)\beta_{\mathrm{t}}\right)^2\right].\end{aligned}$$

Hence, we can solve separately

$$(\tilde{\alpha}_{\mathrm{c}}^*, \beta_{\mathrm{c}}^*) = \arg \min_{\alpha_{\mathrm{c}},\beta_{\mathrm{c}}} \mathbb{E}_{\mathrm{sp}}\left[(1 - W_i) \cdot \left(Y_i^{\mathrm{obs}} - \alpha_{\mathrm{c}} - (X_i - \mu_X)\beta_{\mathrm{c}}\right)^2\right],$$

and

$$(\tilde{\alpha}_{\mathrm{t}}^*, \beta_{\mathrm{t}}^*) = \arg \min_{\alpha_{\mathrm{t}},\beta_{\mathrm{t}}} \mathbb{E}_{\mathrm{sp}}\left[W_i \cdot \left(Y_i^{\mathrm{obs}} - \alpha_{\mathrm{t}} - (X_i - \mu_X)\beta_{\mathrm{t}}\right)^2\right].$$

Because $\mathbb{E}_{\mathrm{sp}}[X_i|W_i = w] = \mu_X$ for $w = 0, 1$ by the randomization, this leads to the solutions

$$\tilde{\alpha}_{\mathrm{c}}^* = \mathbb{E}_{\mathrm{sp}}[Y_i(0)], \qquad \text{and } \tilde{\alpha}_{\mathrm{t}}^* = \mathbb{E}_{\mathrm{sp}}[Y_i(1)].$$

Hence

$$\tau^* = \tilde{\alpha}_{\mathrm{t}}^* - \tilde{\alpha}_{\mathrm{c}}^* = \mathbb{E}_{\mathrm{sp}}[Y_i(1)] - \mathbb{E}_{\mathrm{sp}}[Y_i(0)] = \tau_{\mathrm{sp}},$$

proving part (*i*).

For part (*ii*) we use a different reparametrization. Let $\tilde{\alpha} = \alpha - \tau \cdot p - \mu_X\beta$, with the other parameters unchanged, so that the minimization problem becomes

$$\begin{aligned}(\hat{\tilde{\alpha}}^{\mathrm{ols}}, \hat{\tau}^{\mathrm{ols}}, \hat{\beta}^{\mathrm{ols}}, \hat{\gamma}^{\mathrm{ols}}) &= \arg \min_{\alpha,\tau,\beta,\gamma} \frac{1}{N}\sum_{i=1}^{N}\\ &\quad\times \left(Y_i^{\mathrm{obs}} - \alpha - \tau \cdot (W_i - p) - \beta'(X_i - \mu_X) - \gamma'(X_i - \mu_X) \cdot W_i\right)^2.\end{aligned}$$

The first-order conditions for the estimators $(\hat{\tilde{\alpha}}^{\mathrm{ols}}, \hat{\tau}^{\mathrm{ols}}, \hat{\beta}^{\mathrm{ols}}, \hat{\gamma}^{\mathrm{ols}})$ are

$$\sum_{i=1}^{N} \psi(Y_i^{\mathrm{obs}}, W_i, X_i, \hat{\tilde{\alpha}}^{\mathrm{ols}}, \hat{\tau}^{\mathrm{ols}}, \hat{\beta}^{\mathrm{ols}}, \hat{\gamma}^{\mathrm{ols}}) = 0,$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

where

$$\psi(y,w,x,\alpha,\tau,\beta,\gamma)=\begin{pmatrix} y-\alpha-\tau\cdot(w-p)-\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\beta-\gamma'\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\cdot\tau \\ (w-p)\cdot\left\{y-\alpha-\tau\cdot(w-p)-\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\beta\right. \\ \qquad \left.-w\cdot\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\gamma\right\} \\ \left(x-\mathbb{E}_{\text{sp}}[X_i]\right)^T\cdot\left\{y-\alpha-\tau\cdot(w-p)-\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\beta\right. \\ \qquad \left.-w\cdot\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\gamma\right\} \\ \left(x-\mathbb{E}_{\text{sp}}[X_i]\right)^T\cdot w\cdot\left\{y-\alpha-\tau\cdot(w-p)-\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\beta\right. \\ \qquad \left.-w\cdot\left(x-\mathbb{E}_{\text{sp}}[X_i]\right)\gamma\right\} \end{pmatrix}.$$

In large samples we have, by standard M-estimation methods,

$$\sqrt{N}\cdot\begin{pmatrix}\hat{\tilde{\alpha}}^{\text{ols}}-\alpha^* \\ \hat{\tau}^{\text{ols}}-\tau_{\text{sp}} \\ \hat{\beta}^{\text{ols}}-\beta^* \\ \hat{\gamma}^{\text{ols}}-\gamma^*\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix},\Gamma^{-1}\Delta(\Gamma^T)^{-1}\right), \tag{A.1}$$

where the two components of the covariance matrix are now

$$\Gamma=\mathbb{E}_{\text{sp}}\left[\frac{\partial}{\partial(\alpha,\tau,\beta^T,\gamma^T)}\psi(Y_i^{\text{obs}},W_i,X_i,\alpha,\tau,\beta,\gamma)\right]\Bigg|_{(\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*,\gamma^*)}$$

$$=\mathbb{E}_{\text{sp}}\left[\begin{pmatrix} -1 & -(W_i-p) & -(X_i-\mu_X) & W_i\,(X_i-\mu_X) \\ -(W_i-p) & -(W_i-p)^2 & -(W_i-p)(X_i-\mu_X) & (W_i-p)W_i\,(X_i-\mu_X) \\ -(X_i-\mu_X)^T & -(W_i-p)(X_i-\mu_X)^T & -(X_i-\mu_X)^T(X_i-\mu_X) & W_i\,(X_i-\mu_X)^T\,(X_i-\mu_X) \\ W_i\,(X_i-\mu_X)^T & (W_i-p)W_i\,(X_i-\mu_X)^T & W_i\,(X_i-\mu_X)^T\,(X_i-\mu_X) & W_i^2\,(X_i-\mu_X)^T\,(X_i-\mu_X) \end{pmatrix}\right]$$

$$=\mathbb{E}_{\text{sp}}\left[\begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -p(1-p) & 0 & 0 \\ 0 & 0 & -\Omega_X & 0 \\ 0 & 00 & & -p\cdot\Omega_X \end{pmatrix}\right],$$

and

$$\Delta=\mathbb{E}_{\text{sp}}\left[\psi(Y_i^{\text{obs}},W_i,X_i,\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*,\gamma^*)\cdot\psi(Y_i^{\text{obs}},W_i,X_i,\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*,\gamma^*)^T\right]$$

$$=\mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}}-\alpha^*-\tau_{\text{sp}}-\beta^{*\prime}X_i\right)^2\cdot\begin{pmatrix}1\\W_i-p\\(X_i-\mu_X)^T\\W_i\cdot(X_i-\mu_X)^T\end{pmatrix}\begin{pmatrix}1\\W_i-p\\(X_i-\mu_X)^T\\W_i\cdot(X_i-\mu_X)^T\end{pmatrix}^T\right].$$

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

The normalized variance of $\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}$ is the $(2,2)$ element of the matrix $\Gamma^{-1}\Delta(\Gamma^{T})^{-1}$, which is equal to

$$\frac{\mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} - X_i\beta^*\right)^2 \cdot (W_i - p)^2\right]}{p^2 \cdot (1-p)^2}.$$

□

Proof of Theorem 7.3

We use the same reparametrization as in the first part of the proof of Theorem 7.2:

$$\begin{pmatrix} \tilde{\alpha}_{\text{c}} \\ \beta_{\text{c}} \\ \tilde{\alpha}_{\text{t}} \\ \beta_{\text{t}} \end{pmatrix} = \begin{pmatrix} \alpha + \mu_X\beta \\ \beta \\ \alpha + \tau + \mu_X\beta \\ \gamma + \beta \end{pmatrix}.$$

In terms of the new parameters, $\gamma^* = \beta_{\text{t}}^* - \beta_{\text{c}}^*$. In the proof of Theorem 7.2 it was shown that the population values for $(\tilde{\alpha}_{\text{c}}, \beta_{\text{c}})$ solve

$$\begin{aligned}(\tilde{\alpha}_{\text{c}}^*, \beta_{\text{c}}^*) &= \arg\min_{\alpha_{\text{c}},\beta_{\text{c}}} \mathbb{E}_{\text{sp}}\left[(1-W_i)\cdot\left(Y_i^{\text{obs}} - \alpha_{\text{t}} - (X_i - \mu_X)\beta_{\text{c}}\right)^2\right] \\ &= \arg\min_{\alpha_{\text{c}},\beta_{\text{c}}} \mathbb{E}_{\text{sp}}\left[(1-W_i)\cdot(Y_i(0) - \alpha_{\text{t}} - (X_i - \mu_X)\beta_{\text{c}})^2\right].\end{aligned}$$

Because of the randomization, W_i is independent of $Y_i(0)$ and X_i, and so

$$(\tilde{\alpha}_{\text{c}}^*, \beta_{\text{c}}^*) = \arg\min_{\alpha_{\text{c}},\beta_{\text{c}}} (1-p)\cdot\mathbb{E}_{\text{sp}}\left[(Y_i(0) - \alpha_{\text{c}} - (X_i - \mu_X)\beta_{\text{c}})^2\right].$$

A similar argument shows that $(\tilde{\alpha}_{\text{t}}^*, \beta_{\text{t}}^*)$ solve the same optimization problem:

$$\begin{aligned}(\tilde{\alpha}_{\text{t}}^*, \beta_{\text{t}}^*) &= \arg\min_{\alpha_{\text{t}},\beta_{\text{t}}} p\cdot\mathbb{E}_{\text{sp}}\left[(Y_i(1) - \alpha_{\text{c}} - (X_i - \mu_X)\beta_{\text{t}})^2\right] \\ &= \arg\min_{\alpha_{\text{t}},\beta_{\text{t}}} (1-p)\cdot\mathbb{E}_{\text{sp}}\left[(Y_i(0) + \tau - \alpha_{\text{c}} - (X_i - \mu_X)\beta_{\text{t}})^2\right]\end{aligned}$$

(because by the null hypothesis of zero effects $Y_i(1) = Y_i(0) + \tau$) and so $\gamma^* = \beta_{\text{t}}^* - \beta_{\text{c}}^* = 0$. This finishes the proof of part (*i*) of the theorem.

Under the null hypothesis $(Y_i(1) = Y_i(0) + \tau)$, $\gamma^* = 0$. Then $\sqrt{N}\hat{\gamma}^{\text{ols}}$ will in large samples have a normal distribution with variance V_γ, and the quadratic form Q_{const} will have a Chi-squared distribution with degrees of freedom equal to the dimension of X_i. This concludes the proof of part (*ii*) of the theorem.

Under the null hypothesis ($Y_i(1) = Y_i(0)$ for all units) it also follows that $\tau_{\text{sp}} = 0$. In that case $\sqrt{N}(\hat{\tau}^{\text{ols}}, \hat{\gamma}^{\text{ols}})$ are in large samples normally distributed with covariance matrix $V_{\tau,\gamma}$. Hence the quadratic form Q_{zero} will in large samples have a chi-squared distribution with degrees of freedom equal to the dimension of τ and γ, which is equal to the dimension of X_i plus one.

The covariance matrix for $(\hat{\tau}^{\text{ols}}, \hat{\gamma}^{\text{ols}})$ is most easily obtained from the parametrization in part (*ii*) of the proof of Theorem 7.2, in terms of $(\tilde{\alpha}, \tau, \beta, \gamma)$. The point estimates

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press

for τ and γ under this parametrization are identical to those under the parametrization $(\alpha, \tau, \beta, \gamma)$. Under the parametrization in terms of $(\tilde{\alpha}, \tau, \beta, \gamma)$ the full covariance matrix of $\sqrt{N}(\hat{\tilde{\alpha}}^{\text{ols}} - \tilde{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}} - \tau, \hat{\beta}^{\text{ols}} - \beta, \hat{\gamma}^{\text{ols}} - \gamma)$ is given by $\Gamma^{-1}\Delta(\Gamma^T)^{-1})$ as given in (A.1). To obtain the covariance matrix for $\sqrt{N}(\hat{\tau}^{\text{ols}} - \tau, \hat{\gamma}^{\text{ols}} - \gamma)$ partition $\Gamma^{-1}\Delta(\Gamma^T)^{-1})$ as

$$\mathbb{V} = \Gamma^{-1}\Delta(\Gamma^T)^{-1}) = \begin{pmatrix} \mathbb{V}_{\tilde{\alpha},\tilde{\alpha}} & \mathbb{V}_{\tilde{\alpha},\tau} & \mathbb{V}_{\tilde{\alpha},\beta^T} & \mathbb{V}_{\tilde{\alpha},\gamma^T} \\ \mathbb{V}_{\tau,\tilde{\alpha}} & \mathbb{V}_{\tau,\tau} & \mathbb{V}_{\tau,\beta^T} & \mathbb{V}_{\tau,\gamma^T} \\ \mathbb{V}_{\beta,\tilde{\alpha}} & \mathbb{V}_{\beta,\tau} & \mathbb{V}_{\beta,\beta^T} & \mathbb{V}_{\beta,\gamma^T} \\ \mathbb{V}_{\gamma,\tilde{\alpha}} & \mathbb{V}_{\gamma,\tau} & \mathbb{V}_{\gamma,\beta^T} & \mathbb{V}_{\gamma,\gamma^T} \end{pmatrix}.$$

The covariance matrix for $\sqrt{N}(\hat{\tau}^{\text{ols}} - \tau, \hat{\gamma}^{\text{ols}} - \gamma)$ is then

$$\mathbb{V}_{\tau,\gamma} = \begin{pmatrix} \mathbb{V}_{\tau,\tau} & \mathbb{V}_{\tau,\gamma^T} \\ \mathbb{V}_{\gamma,\tau} & \mathbb{V}_{\gamma,\gamma^T} \end{pmatrix}.$$

The covariance matrix for $\sqrt{N}(\hat{\gamma}^{\text{ols}} - \gamma)$ is simply $\mathbb{V}_{\gamma,\gamma^T}$. □

https://doi.org/10.1017/CBO9781139025751.008 Published online by Cambridge University Press