Chapter 7

Regression Methods for Completely Randomized Experiments

7.1 Introduction

One of the more common ways of analyzing both experimental and observational data in many disciplines is based on regression methods. Typically an additive linear regression function is specified for the observed outcome in both treatment groups 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 pretreatment variables. The parameters of the regression equation are estimated by least squares, with the 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. In this chapter we discuss the rationale for, and implementation of, these models in the context of 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 models for imputation of the unobserved potential outcomes.

There are three key features of the models considered in this chapter. First, we consider models of the observed outcome, rather than a model of the potential outcomes. Second, we consider only models of the conditional mean, rather than of 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 simply involve questions of inference for parameters of a statistical model. An important implication of this approach is that the
validity of these models, that is, whether the models provide accurate descriptions of the conditional means, is immaterial for the consistency of the least squares estimator of the average treatment effect.

The conventional justification for linear regression models, that the regression function represents the conditional expectation of the outcome given the predictor variables, does not follow from the randomization. Nevertheless, in the setting of a completely randomized experiment, the point estimates and associated inferences can be justified using large sample results. There are two important differences with the previous chapters, however. First, with the exception of the setting without covariates, where all 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 because of the need to estimate the associated regression coefficients. Second, the often implicit focus in this literature is on a large (super-)population average treatment effect, rather than on the finite sample average treatment effect. The potential benefit of the regression methods over the exact methods from the previous chapter is that they provide a straightforward way to incorporate covariates. If these covariates are predictive of the potential outcomes, their inclusion in the regression model often results in inferences that are more precise. The disadvantage of regression models relative to the fully model-based methods that will be discussed in the next chapter, is that the restriction to 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 potentially a somewhat unnatural tension between (a), models that provide a good statistical fit and have good statistical properties and (b), models that answer the substantive question of interest. This tension is not present in the full, model-based, methods discussed in the next chapter.

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), EF from hereon. The data set analyzed here contains
information on $N = 337$ individuals. Of these, $N_t = 165$ were randomly assigned to receive cholestyramine and the remaining $N_c = 172$ 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 first, $\text{chol}_1$, was taken prior to the suggestion, made to all 337 individuals in the population, for a low-cholesterol diet, and the second, $\text{chol}_2$ was taken after this suggestion, but prior to the random assignment to cholestyramine or placebo. We observe two outcomes. The main outcome is an average of post-randomization cholesterol readings, $\text{cholf}$, averaged over two-month readings for a period of time averaging 7.3 years for all the individuals in the study. EF focus on the change in cholesterol level, relative to a weighted average of the two pre-treatment cholesterol levels, $\text{chol}_p = 0.25 \cdot \text{chol}_1 + 0.75 \cdot \text{chol}_2$, as the primary outcome. We denote this change in cholesterol levels by $\text{chold} = \text{cholf} - \text{chol}_p$.

We also observe a compliance measure, denoted by $\text{comp}$, 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, differences in adverse side effects between the active drug and the placebo apparently 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 the actual taking the drug and the effect of being assigned to the taking of cholestyramine. We discuss some of these issues in detail in later chapters on noncompliance and instrumental variables. 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 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 pretreatment 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, and appears to be, in the current study. In fact, with the current data, we can reject the assumption that observed
compliance is a proper covariate.

In Table 7.1 we present summary statistics for the EF data. For the two initial cholesterol levels (cho11 and cho12), as well as the composite pretreatment cholesterol level (cho1p), the averages do not vary much by treatment status, consistent with the randomized assignment. For the subsequent cholesterol level measures (cho1f and cho1d) 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 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 super-population, rather than the sample, average treatment effect. We assume that the population of size $N$ subject to the completely randomized experiment is a simple random sample drawn from a larger, essentially infinite, super-population. Considering the $N$ units in our sample as a random sample of the super population induces a distribution on the pair of potential outcomes. The observed potential outcome and covariate values for a drawn unit $i$ are simply one draw from the distributions in the super population and are therefore themselves stochastic. We assume that we do not know anything about this distribution other than 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 effect and thus on the average of the unit-level treatment effect within the experimental sample. To be clear about this superpopulation perspective, let us index the average treatment effect by FS to denote the finite sample average treatment effect and by SP to denote the super-population average treatment effect. Thus

$$
\tau_{FS} = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))
$$

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

$$
\tau_{SP} = \mathbb{E} [Y_i(1) - Y_i(0)]
$$
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. For the discussion in this chapter, it is useful to introduce some additional notation. Define also the super-population average and variance of the two potential outcomes conditional on the covariates,

\[ \mu_c(x) = \mathbb{E}[Y_i(0)|X_i = x], \quad \mu_t(x) = \mathbb{E}[Y_i(1)|X_i = x], \]

\[ \sigma^2_c(x) = \mathbb{V}(Y_i(0)|X_i = x), \quad \text{and} \quad \sigma^2_t = \mathbb{V}(Y_i(1)|X_i = x), \]

and let the variance of the unit-level treatment effect be denoted by

\[ \sigma^2_{ct}(x) = \mathbb{V}(Y_i(1) - Y_i(0)|X_i = x). \]

In addition, denote the marginal means and variances by

\[ \mu_c = \mathbb{E}[Y_i(0)], \quad \mu_t = \mathbb{E}[Y_i(1)], \]

\[ \sigma^2_c = \mathbb{V}(Y_i(0)), \quad \text{and} \quad \sigma^2_t = \mathbb{V}(Y_i(1)). \]

Note that for the two means, the marginal mean is equal to the expectation of the conditional mean:

\[ \mu_c = \mathbb{E}[\mu_c(X_i)], \quad \text{and} \quad \mu_t = \mathbb{E}[\mu_t(X_i)], \]

but for the variances, the marginal variance differs from the average of the conditional variance by the variance of the conditional mean:

\[ \sigma^2_c = \mathbb{E}[\sigma^2_c(X_i)] + \mathbb{V}(\mu_c(X_i)), \quad \text{and} \quad \sigma^2_t = \mathbb{E}[\sigma^2_t(X_i)] + \mathbb{V}(\mu_t(X_i)). \]

Finally, let \( p \) denote the population fraction of treated units, \( p = N_t/N \), and let \( \mu_X = \mathbb{E}[X_i] \) and \( \Omega_X = \mathbb{E}[(X_i - \mu_X)(X_i - \mu_X)'] \) be the population mean and covariance matrix of the vector of covariates, respectively.

### 7.4 Linear Regression With No Additional Covariates

In this section we focus on the case with no covariates and thus no predictor variables beyond the dummy for the receipt of treatment. We maintain the assumption of a completely
randomized experiment. We specify a linear regression function for the observed outcome \( Y_i^{\text{obs}} \) as

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

where the unobserved residual \( \varepsilon_i \) is supposed to capture unobserved determinants of the outcome. The ordinary least squares (or ols) estimator for \( \tau \) is based on minimizing the sum of squared residuals over \( \alpha \) and \( \tau \),

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

with solution

\[
\hat{\tau}_{\text{ols}} = \frac{\sum_{i=1}^{N} (W_i - \bar{W}) \cdot (Y_i^{\text{obs}} - \bar{Y}^{\text{obs}})}{\sum_{i=1}^{N} (W_i - \bar{W})^2}, \quad \text{and} \quad \hat{\alpha}_{\text{ols}} = \bar{Y}^{\text{obs}} - \hat{\tau}_{\text{ols}} \cdot \bar{W},
\]

where

\[
\bar{Y}^{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} Y_i^{\text{obs}} \quad \text{and} \quad \bar{W} = \frac{1}{N} \sum_{i=1}^{N} W_i.
\]

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:

\[
\hat{\tau}_{\text{ols}} = \bar{Y}^{\text{obs}}_t - \bar{Y}^{\text{obs}}_c,
\]

where, as before, \( \bar{Y}^{\text{obs}}_c = \sum_{i:W_i=0} Y_i^{\text{obs}}/N_c \) and \( \bar{Y}^{\text{obs}}_t = \sum_{i:W_i=1} Y_i^{\text{obs}}/N_t \) are the average of the observed outcomes in the control and treatment group respectively.

The least squares estimate of \( \tau \) is often interpreted as an estimate of the causal effect of the treatment, explicitly in randomized experiments, and typically implicitly in observational studies. The assumptions traditionally used in this context are that the residuals \( \varepsilon_i \) 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 \( Y_i^{\text{obs}} \), 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 \( \bar{Y}_{\text{obs}} - \bar{Y}_{c} \), which was shown to be unbiased for the finite sample average treatment effect \( \tau_{FS} \) in the previous chapter. However, it is useful to justify the causal interpretation of \( \hat{\tau}_{\text{ols}} \) more directly in terms of the standard justification for regression methods, using the assumption that a completely randomized experiment generated the data. We also maintain the assumption, made in Section 7.3, of simple random sampling from a super-population.

Let \( \alpha \) be the super-population average outcome under the control, \( \alpha = \mu_{c} = \mathbb{E}[Y_{i}(0)] \), and recall that \( \tau_{\text{SP}} \) is the super-population average treatment effect, \( \tau_{\text{SP}} = \mu_{t} - \mu_{c} = \mathbb{E}[Y_{i}(1) - Y_{i}(0)] \). Now define the residual \( \varepsilon_{i} \) in terms of the super-population parameters and the potential outcomes as

\[
\varepsilon_{i} = Y_{i}(0) - \alpha + W_{i} \cdot (Y_{i}(1) - Y_{i}(0) - \tau_{\text{SP}}) = \begin{cases} Y_{i}^{\text{obs}} - \alpha & \text{if } W_{i} = 0, \\ Y_{i}^{\text{obs}} - \alpha - \tau_{\text{SP}} & \text{if } W_{i} = 1. \end{cases}
\]

Then we can write

\[
\varepsilon_{i} = Y_{i}^{\text{obs}} - (\alpha + \tau_{\text{SP}} \cdot W_{i}),
\]

and thus can write the observed outcome as

\[
Y_{i}^{\text{obs}} = \alpha + \tau_{\text{SP}} \cdot W_{i} + \varepsilon_{i}.
\]

Random assignment guarantees 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 “\( \perp \)” notation,

\[
W_{i} \perp \perp (Y_{i}(0), Y_{i}(1)).
\]

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

\[
\mathbb{E}[\varepsilon_{i}|W_{i} = 0] = \mathbb{E}[Y_{i}(0) - \alpha|W_{i} = 0] = \mathbb{E}[Y_{i}(0)] - \alpha = 0,
\]

and

\[
\mathbb{E}[\varepsilon_{i}|W_{i} = 1] = \mathbb{E}[Y_{i}(1) - \alpha - \tau_{\text{SP}}|W_{i} = 1] = \mathbb{E}[Y_{i}(1) - \alpha - \tau_{\text{SP}}|W_{i} = 1] = 0,
\]
so that
\[ \mathbb{E}[\varepsilon_i | W_i] = 0. \]

Thus the least squares estimator, \( \hat{\tau}_{\text{ols}} \), is unbiased for \( \tau_{\text{SP}} = \mathbb{E}[Y_i(1) - Y_i(0)] \). The above derivation shows how properties of residuals commonly asserted as assumptions in least squares analyses actually follow from the randomization, 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 case with covariates, is to consider the population limits of the estimators. The estimators are defined as
\[
(\hat{\alpha}_{\text{ols}}, \hat{\tau}_{\text{ols}}) = \arg \min_{\alpha, \tau} \sum_{i=1}^{N} (Y_i^{\text{obs}} - \alpha - \tau \cdot W_i)^2.
\]

Under some regularity conditions, these estimators converge, as the sample size goes to infinity, to the population limits \((\alpha^*, \tau^*)\) that minimize the expected value of the sum of squares:
\[
(\alpha^*, \tau^*) = \arg \min_{\alpha, \tau} \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^{N} (Y_i^{\text{obs}} - \alpha - \tau \cdot W_i)^2 \right] = \arg \min_{\alpha, \tau} \mathbb{E} \left[ (Y_i^{\text{obs}} - \alpha - \tau \cdot W_i)^2 \right].
\]

The preceding argument now translates as stating that \( \tau^* = \mathbb{E}[Y_i^{\text{obs}} | W_i = 1] - \mathbb{E}[Y_i^{\text{obs}} | W_i = 0] \) Random assignment implies \( \mathbb{E}[Y_i^{\text{obs}} | W_i = 1] - \mathbb{E}[Y_i^{\text{obs}} | W_i = 0] = \mathbb{E}[Y_i(1) - Y_i(0)] = \tau_{\text{SP}}, \) so that \( \tau^* = \tau_{\text{SP}}. \)

Now let us analyze the least squares approach to inference (i.e., sampling variance and confidence intervals) applied to the completely randomized experiment setting. Let us first assume homoskedasticity. \( (\sigma_{Y|W}^2 = \sigma_e^2 = \sigma_\tau^2) \). Using least squares methods, the variance of the residuals would be estimated as
\[
\hat{\sigma}_{Y|W}^2 = \frac{1}{N - 2} \sum_{i=1}^{N} \tilde{\varepsilon}_i^2 = \frac{1}{N - 2} \sum_{i=1}^{N} \left( Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}} \right)^2,
\]
where the 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

\[ \hat{\sigma}^2_{Y|W} = \frac{1}{N - 2} \left( \sum_{i:W_i=0} \left( Y_{i}^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right)^2 + \sum_{i:W_i=1} \left( Y_{i}^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2 \right), \]

which is equivalent to our calculation of \( s^2 \), the common variance across the two potential outcome distributions, as seen in equation 6.7 in Chapter 6. The conventional estimator for the sampling variance of ols estimator \( \hat{\tau}_{\text{ols}} \) is

\[ \hat{V}(\hat{\tau}_{\text{ols}})_{\text{homosk}} = \frac{\hat{\sigma}^2_{Y|W}}{\sum_{i=1}^N (W_i - \overline{W})^2} = s^2 \cdot \left( \frac{1}{N_c} + \frac{1}{N_t} \right). \]

Normalized by the sample size \( N \), the sampling variance estimator converges to

\[ N \cdot \hat{V}(\hat{\tau}_{\text{ols}})_{\text{homosk}} \xrightarrow{p} \frac{\sigma^2_{Y|W}}{p \cdot (1 - p)}. \quad (7.1) \]

The estimator \( \hat{V}(\hat{\tau}_{\text{ols}})_{\text{homosk}} \) is identical to the estimator for the sampling variance justified under constant treatment effects (\( \hat{V}_{\text{const}} \) in equation (6.8) 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.

Note, however, that the random assignment assumption we used for the causal interpretation of \( \hat{\tau}_{\text{ols}} \) 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 variance estimator for least squares estimators is

\[ \hat{V}(\hat{\tau}_{\text{ols}})_{\text{hetero}} = \frac{\sum_{i=1}^N \xi_i^2 \cdot (W_i - \overline{W})^2}{\left( \sum_{i=1}^N (W_i - \overline{W})^2 \right)^2}. \]

Defining, as the previous chapter,

\[ s_c^2 = \frac{1}{N_c - 1} \sum_{i|W_i=0} \left( Y_{i}^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right)^2, \quad \text{and} \quad s_t^2 = \frac{1}{N_t - 1} \sum_{i|W_i=1} \left( Y_{i}^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2, \]

we can write the variance estimator under heteroskedasticity as

\[ \hat{V}(\hat{\tau}_{\text{ols}})_{\text{hetero}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}. \]
This is exactly the same estimator for the variance derived from Neyman’s perspective in Chapter 6 (\( \hat{V}_{\text{neyman}} \) in equation (6.6)). So, in the case without additional predictors, the regression approach does not contribute anything not already known from the previous chapter. It does, however, provide a different perspective on these results, and one that allows for a natural extension to the case with additional predictors.

### 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. We specify the regression function as:

\[
Y_{\text{obs}}^i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i,
\]

where \( X_i \) is a vector of covariates. We estimate the regression coefficients again using least squares:

\[
(\hat{\tau}_{\text{ols}}, \hat{\alpha}_{\text{ols}}, \hat{\beta}_{\text{ols}}) = \arg \min_{\tau, \alpha, \beta} \sum_{i=1}^{N} (Y_{i}^{\text{obs}} - \alpha - \tau \cdot W_i - \beta' X_i)^2.
\]

The first question we address in this section concerns the causal interpretation of the least squares estimate \( \hat{\tau}_{\text{ols}} \). We are not interested per se in the coefficients on the additional covariates \( \beta \) or in the intercept \( \alpha \). In particular we are not interested in a causal interpretation of those parameters. Moreover, we will not make the assumption that the conditional expectation of \( Y_{i}^{\text{obs}} \) is 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 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

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

These population values are generally well-defined (subject essentially only to finite moment conditions and positive definiteness of \( \Omega_X \), 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 $\tau_{SP}$ 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 consistent for the population average treatment effect $\tau_{SP}$. In other words, $\tau^*$, the probability limit of the estimator, is equal to the population average treatment effect $\tau_{SP}$. In addition, in large samples $\hat{\tau}_{\text{ols}}$ will be distributed normally around $\tau_{SP}$. To be precise we state the result formally.

**Theorem 1** Suppose we conduct a completely randomized experiment in a sample drawn at random from an infinite population. Then, (i)

$$\tau^* = \tau_{SP},$$

and (ii),

$$\sqrt{N} \cdot (\hat{\tau}_{\text{ols}} - \tau_{SP}) \overset{d}{\rightarrow} N\left(0, \frac{E\left[(W_i - p)^2 \cdot (Y_{\text{obs}}i - \alpha^* - \tau_{SP} \cdot W_i - \beta^* W_i)^2\right]}{p^2 \cdot (1 - p)^2}\right).$$

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, are given in the Appendix to this chapter.

**Proof of Theorem 1(i):** Consider the limiting objective function:

$$Q(\alpha, \tau, \beta) = E[(Y_{\text{obs}}i - \alpha - \tau \cdot W_i - \beta' X_i)^2]$$

$$= E\left[(Y_{\text{obs}}i - \alpha - \tau \cdot W_i - \beta' (X_i - \mu_X))^2\right],$$

where $\alpha = \alpha + \beta' \mu_X$, with $\mu_X = E[X_i]$. Minizing the righthand side over $\alpha$, $\tau$ and $\beta$ leads to the same values for $\tau$ and $\beta$ as minizing the left hand side over $\alpha$, $\tau$, and $\beta$, with the least squares estimate of $\alpha$ equal to the least squares estimate of $\alpha$ plus $\hat{\beta}' \mu_X$. Next,

$$Q(\tilde{\alpha}, \tau, \beta) = E\left[(Y_{\text{obs}}i - \tilde{\alpha} - \tau \cdot W_i - \beta' (X_i - \mu_X))^2\right]$$

$$= E\left[(Y_{\text{obs}}i - \tilde{\alpha} - \tau \cdot W_i)^2\right] + E\left[(\beta' (X_i - \mu_X))^2\right]$$

$$- 2 \cdot E\left[(Y_{\text{obs}}i - \tilde{\alpha} - \tau \cdot W_i) \cdot \beta' (X_i - \mu_X)\right]$$
$$= \mathbb{E} \left[ \left( Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i \right)^2 \right] + \mathbb{E} \left[ \left( \beta'(X_i - \mu_X) \right)^2 \right] - 2 \cdot \mathbb{E} \left[ Y_i^{\text{obs}} \cdot \beta'(X_i - \mu_X) \right], \quad (7.2)$$

because by definition
$$\mathbb{E} \left[ \tilde{\alpha} \cdot \beta'(X_i - \mu_X) \right] = 0,$$

and by the randomization
$$\mathbb{E} \left[ \tau \cdot W_i \cdot \beta'(X_i - \mu_X) \right] = 0.$$

Because the last two terms in (7.2) do not depend on $\tilde{\alpha}$ or $\tau$, minimizing (7.2) over $\tau$ and $\alpha$ is equivalent to minizing the objective function without the additional covariates,
$$\mathbb{E} \left[ \left( Y_i^{\text{obs}} - \bar{\alpha} - \tau \cdot W_i \right)^2 \right],$$

which leads to the solutions
$$\bar{\alpha}^* = \mathbb{E}[Y_i^{\text{obs}}|W_i = 0] = \mathbb{E}[Y_i(0)|W_i = 0] = \mathbb{E}[Y_i(0)],$$

and
$$\tau^* = \mathbb{E}[Y_i^{\text{obs}}|W_i = 1] - \mathbb{E}[Y_i^{\text{obs}}|W_i = 0] = \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] = \tau_{SP}.$$

Thus the least squares estimator is consistent for the population average treatment effect $\tau_{SP}$.

What is important in the first part of the result is that the consistency of the least squares estimator for $\tau_{SP}$ does not depend on the correctness of the specification of the regression function in a completely randomized experiment. No matter how nonlinear the conditional expectations of the potential outcomes given the covariates is in the 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 population correlation between the treatment indicator and the covariates is zero. Even though in finite samples the 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 for $\tau$. 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 estimator, it does matter for the sampling variance of 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

\[ E[Y_i(0) | X_i = x] = \alpha_c + \beta'x, \quad \text{and} \quad E[Y_i(1) | X_i = x] = \alpha_t + \beta'x, \]

so that, in combination with random assignment, we have

\[ E[Y_{i,\text{obs}} | X_i = x, W_i = t] = \alpha_c + \tau_{SP} \cdot t + \beta'x, \]

where \( \tau_{SP} = \alpha_t - \alpha_c \). Suppose that, in addition, the variance of the two potential outcomes does not vary by treatment or covariates:

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

for \( w = 0, 1 \), and all \( x \). Then the normalized sampling variance for the least squares estimator for \( \tau_{SP} \), given for the general case in Theorem 1, simplifies to

\[ N \cdot \mathbb{V}(\hat{\tau}_{\text{ols}})_{\text{homosk}} = \frac{\sigma^2_{Y|W,X}}{p \cdot (1 - p)}. \tag{7.3} \]

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 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 the marginal variance \( \sigma^2_{Y|W} \), then including the covariates in the regression model will lead a considerable gain in precision. The price paid for the gain 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 variance as

\[ \hat{\mathbb{V}}(\hat{\tau})_{\text{hetero}} = \frac{1}{N (N - 1 - \text{dim}(X_i))} \cdot \frac{\sum_{i=1}^{N} (W_i - \bar{W})^2 \cdot (Y_{i,\text{obs}} - \hat{\alpha}_{\text{ols}} - \hat{\tau}_{\text{ols}} - \hat{\beta}_{\text{ols}}X_i)^2}{(\bar{W} \cdot (1 - \bar{W}))^2}. \]
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:

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

### 7.6 Linear Regression With Additional 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 predictors with the treatment indicators, we achieve a particular form of robustness that we discuss in more detail later. This robustness is not particularly important in the current setting of a randomized experiment, but it will be important in observational studies discussed in Part II of this book. We specify the regression function as

$$
Y_{i,\text{obs}} = \alpha + \tau \cdot W_i + \beta'X_i + \gamma'(X_i - \bar{X}) \cdot W_i + \varepsilon_i.
$$

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 $$\tau_{\text{SP}}$$.

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

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

and let $$\alpha^*, \tau^*, \beta^*$$, and $$\gamma^*$$ denote the corresponding population values:

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

Results similar to Theorem 1 can be obtained for this case. The least squares estimator $$\hat{\tau}_{\text{ols}}$$ is consistent for the average treatment effect $$\tau_{\text{SP}}$$, and inference can be based on least squares methods.
Theorem 2 Suppose we conduct a completely randomized experiment in a random sample from a large population. Then (i)

$$\tau^* = \tau_{SP},$$

and (ii),

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

The proof for this theorem is 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 it represents the conditional expectation:

$$\mathbb{E} \left[ Y_{i \text{obs}} | \ X_i = x, W_i = t \right] = \alpha + \tau \cdot t + \beta' x + \gamma' (x - \mu_X) \cdot t. \quad (7.4)$$

In combination with the random assignment, this implies that

$$\mathbb{E} \left[ Y_i(0) | \ X_i = x \right] = \mathbb{E} \left[ Y_i(0) | \ X_i = x, W_i = 0 \right] = \mathbb{E} \left[ Y_{i \text{obs}} | \ X_i = x, W_i = 0 \right] = \alpha + \beta' x,$$

and

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

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 (7.4), the predicted value for the missing potential outcome $Y_i(0)$ is

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

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

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

For a control unit $j$ ($W_j = 0$) the predicted value for the missing potential outcome $Y_j(1)$ is

$$\hat{Y}_j(1) = \hat{\alpha}_{ols} + \hat{\tau}_{ols} + \hat{\beta}_{ols}' X_j + \hat{\gamma}_{ols}' (X_j - \overline{X}),$$
and the predicted value for the unit-level causal effect for this control unit is
\[
\hat{\tau}_j = \hat{Y}_j(1) - Y_j(0) = \hat{\alpha} + \hat{\tau} + \hat{\beta}'X_j + \hat{\gamma}'(X_j - \bar{X}) - Y_{j,\text{obs}}.
\]
Now we can estimate the overall average treatment effect \(\tau_{SP}\) 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, 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 \(Y_i(0)\), this procedure uses only the outcomes \(Y_{i,\text{obs}}\) 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
\[
\ln \left( Y_{i,\text{obs}} \right) = \alpha + \tau \cdot W_i + \beta'X_i + \varepsilon_i.
\]
Irrespective of the form of the association between outcomes and covariates, in a completely randomized experiment, least squares estimates of \(\tau\) are consistent for the average effect
\[ E[\ln(Y_i(1)) - \ln(Y_i(0))] \]. 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 \( E[Y_i(1) - Y_i(0)] \), but suppose that one actually suspects that a model 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 leads to a consistent estimator for the average effect only if the specification of the model is correct. We discuss such modelling 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 to the conditional expectation of the outcome given covariates and treatment indicator. Logit models (where \( \Pr(Y_i^{obs} = 1|W_i = w, X_i = x) \) is modelled as \( \exp(\alpha + \tau \cdot w + \beta'x)/(1 + \exp(\alpha + \tau \cdot w + \beta'x)) \)), or probit models (where \( \Pr(Y_i^{obs} = 1|W_i = w, X_i = x) = \Phi(\alpha + \tau \cdot w + \beta'x) \)), 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 necessarily 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 randomized experiments the bias is unlikely to be substantial with moderate sized samples, as flexible models are likely to have at most minimal bias. Moreover, this consistency property despite potential misspecification of the regression function only holds under 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 Precision Gains

Including an additional covariate in the regression function will not lower, and in generally will increase, the precision of the estimator for the average treatment effect, at least in large samples. However, beyond the first few covariates additional covariates are unlikely to improve the precision substantially in finite 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 dummy variable indicating the treatment. 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

\[ \text{V}_{\text{nocov}} = \frac{\sigma_c^2}{1 - p} + \frac{\sigma_t^2}{p}, \]

familiar in various forms from this and the previous chapter. Now suppose we have a vector of additional covariates, \(X_i\), available. 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

\[ \text{V}_{\text{bound}} = \frac{\mathbb{E}[\sigma_c^2(X_i)]}{1 - p} + \frac{\mathbb{E}[\sigma_t^2(X_i)]}{p}. \]

Instead of the marginal variances \(\sigma_w^2\) in each of the two terms, we now take the expectation of the conditional variance \(\sigma_w^2(X_i)\). 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 \(X_i\):

\[ \text{V}_{\text{nocov}} - \text{V}_{\text{bound}} = \left( \frac{\sigma_c^2}{1 - p} + \frac{\sigma_t^2}{p} \right) - \left( \frac{\mathbb{E}[\sigma_c^2(X_i)]}{1 - p} + \frac{\mathbb{E}[\sigma_t^2(X_i)]}{p} \right) \]

\[ = \frac{\text{V}(\mu_c(X_i))}{1 - p} + \frac{\text{V}(\mu_t(X_i))}{p}. \]

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 \(\mu_c(x)\) and \(\mu_t(x)\) do not vary with the predictor variables, there is no gain, even in large samples from using the covariates, and 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 may be useful for testing 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 important. 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 null hypotheses are not sharp. As in the discussion of estimation, we focus on procedures that are valid, at least 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

\[ Y_{i}^{\text{obs}} = \alpha + \tau \cdot W_{i} + \beta'X_{i} + \gamma'(X_{i} - \bar{X}) \cdot W_{i} + \varepsilon_{i}. \]

In that case we can test the null hypothesis of a zero average treatment effect by testing the null hypothesis that \( \tau_{\text{SP}} = 0 \). However, we can construct a different test by focusing on the deviation of either \( \hat{\tau}_{\text{SP}} \) 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 was equal to

\[ E\left[ Y_{i}^{(1)} - Y_{i}^{(0)} \mid X_{i} = x, W_{i} = w \right] = \alpha + \tau \cdot w + \beta'x + \gamma'(x - \mu_{X}) \cdot w, \]

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

\[ H_{0} : E[Y_{i}(1) - Y_{i}(0) \mid X_{i} = x] = 0, \forall x, \]

against the alternative hypothesis

\[ H_{a} : E[Y_{i}(1) - Y_{i}(0) \mid X_{i} = x] \neq 0, \text{ for some } x. \]

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

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

The precise form of the components of the covariance matrix, as well as consistent estimators for these components, are 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

$$Q_{\text{zero}} = \begin{pmatrix} \hat{\tau}_{SP} \\ \hat{\gamma} \end{pmatrix}^\prime \hat{\mathbb{V}}_{\tau,\gamma}^{-1} \begin{pmatrix} \hat{\tau}_{SP} \\ \hat{\gamma} \end{pmatrix}. \quad (7.5)$$

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:

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

against the alternative hypothesis

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

This null hypothesis is 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 a second order problem.) In order to test this null hypothesis we can use the quadratic form

$$Q_{\text{const}} = \hat{\gamma}^\prime \hat{\mathbb{V}}_\gamma^{-1} \hat{\gamma}. \quad (7.6)$$
Theorem 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 $\tau$ and all units, then

(i): $\gamma^* = 0$,

and (ii)

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

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

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

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 posttreatment cholesterol levels, the primary outcome of interest. Second, partly anticipating some of the analyses in later chapters, we estimate the effect of assignment to treatment on the level of compliance. Because compliance was far from perfect (on average, individuals in the control group took 75% of the nominal dose, and individual in the group assigned to the active treatment, on average, took 60% of the nominal dose), the estimates of the effect on postassignment cholesterol levels should be interpreted as estimates of intention-to-treat (ITT) effects, that is effects of assignment to the drug versus assigned 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 no covariates beyond the indicator for assignment. Second, we include the composite prior cholesterol level $\text{cholp}$ as a linear predictor. Third, we include both prior cholesterol level measurements, $\text{chol1}$ and $\text{chol2}$, 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 to be a reduction of approximately 25-26 units, approximately a 8% reduction. Including additional predictors beyond the treatment indicator improves the precision considerably, reducing the estimated standard error by a third. Including predictors beyond the simple composite prior cholesterol level $\text{cholp}$ 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 part 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 R-squared decreases as a result of this transformation.

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 $Q_{\text{zero}}$ given in equation (7.5), based on the regression with the two prior cholesterol levels and their interactions with the treatment as the additional 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 observations, 172 in the control group and 165 in the treatment group, and the data values, the normal approximations that underly the large sample properties of the tests are accurate.

Next, we test the null hypothesis that the treatment effect is constant, using the statistic
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, 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 much weaker.

Overall, the message from this application is that including covariates can substantially improve the precision of the inferences, although including many covariates is unlikely to be helpful after the inclusion of the most important ones.

### 7.11 Conclusion

In this chapter we discussed 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 other than as approximations. In the context of a completely randomized experiment, however, we can use the randomization to justify the key assumptions necessary for consistency of the least squares estimator. In contrast to the methods discussed in previous chapters, 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 generally 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 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 Chapters 12-19, simple regression methods are not recommended.
Notes
Data were also analyzed in Jin and Rubin (2006).
Mitchell Gail (Biometrika, 1985?)
Hahn (1998, efficiency bound)
Freedman article on regression analyses in the context of randomized experiments
**Proof of Theorem 1:**

It is convenient to reparametrize the model differently. Instead of \((\alpha, \tau, \beta)\), we parametrize the model using \((\hat{\alpha}, \hat{r}, \hat{\beta})\), where \(\hat{\alpha} = \alpha - p \cdot \tau - \beta' E[X_i]\). The reparametrization does not change the ols estimates for \(\tau\) and \(\beta\), nor their limiting values. The limiting value of the new parameter is \(\hat{\alpha}^* = \alpha^* - p \cdot \tau^* - \beta'^* E[X_i]\) in terms of these parameters, the objective function is

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

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

The first order conditions for the estimators \((\hat{\alpha}, \hat{r}, \hat{\beta})\) are

\[
\sum_{i=1}^{N} \psi(Y_i^{\text{obs}}, W_i, X_i, \hat{\alpha}, \hat{r}, \hat{\beta}) = 0,
\]

where

\[
\psi(y, w, x, \alpha, \tau, \beta) = \begin{pmatrix} y - \alpha - \tau \cdot (w - p) - \beta' (x - E[X_i]) \\ (w - p) \cdot (y - \alpha - \tau \cdot (w - p) - \beta' (x - E[X_i])) \\ (x - E[X_i]) \cdot (y - \alpha - \tau \cdot (w - p) - \beta' (x - E[X_i])) \end{pmatrix}
\]

Given the population values of the parameters, \(\alpha^*, \tau_{\text{SP}}, \text{ and } \beta^*\), standard M-estimation (or generalized method of moments) results imply that under standard regularity conditions the estimator is consistent and asymptotically normally distributed:

\[
\sqrt{N} \cdot \begin{pmatrix} \hat{\alpha} - \alpha^* \\ \hat{r} - \tau_{\text{SP}} \\ \hat{\beta} - \beta^* \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1} \Delta (\Gamma')^{-1} \right),
\]

where the two components of the covariance matrix are

\[
\Gamma = E \left[ \frac{\partial}{\partial (\alpha, \tau, \beta')} \psi(Y_i^{\text{obs}}, W_i, X_i, \alpha, \tau, \beta) \bigg| (\alpha^*, \tau_{\text{SP}}, \beta^*) \right]
\]

\[
E \begin{pmatrix} -1 & -(W_i - p) & -(W_i - p) \\ -(W_i - p) & -(W_i - p)^2 & -(W_i - p) \cdot (X_i - E[X_i]) \\ -(X_i - E[X_i]) & -(W_i - p) \cdot (X_i - E[X_i]) & -(X_i - E[X_i]) \cdot (X_i - E[X_i])' \end{pmatrix}
\]

\[
= E \left[ \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} -p(1-p) \\ 0 \\ -E[(X_i - E[X_i]) \cdot (X_i - E[X_i])'] \end{pmatrix} \right],
\]

and

\[
\Delta = E \left[ \psi(Y_i^{\text{obs}}, W_i, X_i, \alpha^*, \tau_{\text{SP}}, \beta^*') \cdot \psi(Y_i^{\text{obs}}, W_i, X_i, \alpha^*, \tau_{\text{SP}}, \beta^*)' \right]
\]

\[
= E \left[ \left( Y_i^{\text{obs}} - \alpha^* - \tau_{\text{SP}} - \beta'^* X_i \right)^2 \cdot \begin{pmatrix} W_i - p \\ X_i - E[X_i] \end{pmatrix} \left( W_i - p \\ X_i - E[X_i] \right)' \right].
\]

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

\[
E \left[ \left( Y_i^{\text{obs}} - \alpha^* - \tau_{\text{SP}} - \beta'^* X_i \right)^2 \cdot (W_i - p)^2 \right].
\]
Hence the variance of \( \hat{\tau} \), normalized by the sample size \( N \), is equal to

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

\[\square\]

**Proof of Theorem 2:**
First we show that in this case \( \tau^* \) the population value of \( \hat{\tau} \), equal to

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

is equal to \( \tau_{\text{SP}} \). Again it is useful to reparametrize. The new vector of parameters is

\[
\begin{pmatrix}
\hat{\alpha}_c \\
\beta_c \\
\hat{\alpha}_t \\
\beta_t
\end{pmatrix} = \begin{pmatrix}
\alpha + \beta' \mu_X \\
\beta \\
\alpha + \tau + \beta' \mu_X \\
\gamma + \beta
\end{pmatrix},
\]

with inverse

\[
\begin{pmatrix}
\alpha \\
\beta \\
\tau \\
\gamma
\end{pmatrix} = \begin{pmatrix}
\hat{\alpha}_c - \beta'_c \mu_X \\
\beta_c \\
\hat{\alpha}_t - \alpha_c \\
\beta_t - \beta_c
\end{pmatrix}.
\]

In terms of this parameter vector the minimization problem is

\[
(\hat{\alpha}^*_c, \hat{\alpha}^*_t, \beta^*_c, \beta^*_t) = \arg \min_{\alpha_c, \alpha_t, \beta_c, \beta_t} \mathbb{E} \left[ (Y_i^{\text{obs}} - \alpha_c - (\alpha_t - \alpha_c) \cdot W_i - \beta'_c X_i - (\beta_t - \beta_c) X_i - \beta'_c (X_i - \mu_X) \cdot W_i)^2 \right].
\]

\[
= \arg \min_{\alpha_c, \alpha_t, \beta_c, \beta_t} \mathbb{E} \left[ (1 - W_i) \cdot (Y_i^{\text{obs}} - \alpha_c - \beta'_c (X_i - \mu_X))^2 + W_i \cdot (Y_i^{\text{obs}} - \alpha_c - \beta'_c (X_i - \mu_X))^2 \right].
\]

Hence we can solve separately

\[
(\hat{\alpha}^*_c, \beta^*_c) = \arg \min_{\alpha_c, \beta_c} \mathbb{E} \left[ (1 - W_i) \cdot (Y_i^{\text{obs}} - \alpha_c - \beta'_c (X_i - \mu_X))^2 \right],
\]

and

\[
(\hat{\alpha}^*_t, \beta^*_t) = \arg \min_{\alpha_t, \beta_t} \mathbb{E} \left[ W_i \cdot (Y_i^{\text{obs}} - \alpha_t - \beta'_t (X_i - \mu_X))^2 \right].
\]

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

\[
\hat{\alpha}^*_c = \mathbb{E}[Y_i(0)], \quad \text{and} \quad \hat{\alpha}^*_t = \mathbb{E}[Y_i(1)].
\]

Hence

\[
\tau^* = \hat{\alpha}^*_t - \hat{\alpha}^*_c = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau_{\text{SP}},
\]

proving part (i).

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

\[
(\tilde{\alpha}, \tilde{\tau}, \tilde{\beta}, \tilde{\gamma}) = \arg \min_{\alpha, \tau, \beta, \gamma} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left[ (Y_i^{\text{obs}} - \alpha - \tau \cdot (W_i - p) - \beta' (X_i - \mu_X) - \gamma' (X_i - \mu_X) \cdot W_i)^2 \right].
\]
The first order conditions for the estimators \((\hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\gamma})\) are

\[
\sum_{i=1}^{N} \psi(Y_{i}^{\text{obs}}, W_{i}, X_{i}, \hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\gamma}) = 0,
\]

where

\[
\psi(y, t, x, \alpha, \tau, \beta, \gamma) = \begin{pmatrix}
y - \alpha - \tau \cdot (t - p) - \beta' (x - E[X_i]) - \gamma' (x - E[X_i]) \cdot t \\
(t - p) \cdot (y - \alpha - \tau \cdot (t - p) - \beta' (x - E[X_i]) - \gamma' (x - E[X_i]) \cdot t) \\
(x - E[X_i]) \cdot (y - \alpha - \tau \cdot (t - p) - \beta' (x - E[X_i]) - \gamma' (x - E[X_i]) \cdot t)
\end{pmatrix}.
\]

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

\[
\sqrt{N} \cdot \begin{pmatrix}
\hat{\alpha} - \alpha^* \\
\hat{\tau} - \tau_{\text{SP}} \\
\hat{\beta} - \beta^* \\
\hat{\gamma} - \gamma^*
\end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1} \Delta (\Gamma')^{-1} \right),
\]

where the two components of the covariance matrix are now

\[
\Gamma = E \left[ \frac{\partial}{\partial (\alpha, \tau, \beta', \gamma)} \psi(Y_{i}^{\text{obs}}, W_{i}, X_{i}, \alpha, \tau, \beta, \gamma) \right]_{(\alpha^*, \tau_{\text{SP}}, \beta^*, \gamma^*)} = E \begin{bmatrix}
-1 & -W_i - p & -(W_i - p)X_i - \mu_X & W_i (X_i - \mu_X) \\
-(W_i - p) & -(W_i - p)^2 & -(W_i - p)(X_i - \mu_X) & (W_i - p)W_i (X_i - \mu_X) \\
-(X_i - \mu_X) & -(W_i - p)(X_i - \mu_X) & -(X_i - \mu_X)W_i (X_i - \mu_X) & W_i (X_i - \mu_X)W_i (X_i - \mu_X) \\
W_i (X_i - \mu_X) & W_i (X_i - \mu_X) & W_i (X_i - \mu_X)W_i (X_i - \mu_X) & W_i^2 (X_i - \mu_X)W_i (X_i - \mu_X)
\end{bmatrix}
\]

and

\[
\Delta = E \left[ \psi(Y_{i}^{\text{obs}}, W_{i}, X_{i}, \alpha^*, \tau_{\text{SP}}, \beta^*, \gamma^*) \cdot \psi(Y_{i}^{\text{obs}}, W_{i}, X_{i}, \alpha^*, \tau_{\text{SP}}, \beta^*, \gamma^*)' \right] = E \begin{pmatrix}
(Y_{i}^{\text{obs}} - \alpha^* - \tau_{\text{SP}} - \beta^* X_i)^2 & \begin{pmatrix} 1 \\ W_i - p \\ W_i - \mu_X \\ W_i \cdot (X_i - \mu_X) \end{pmatrix}
\end{pmatrix} \begin{pmatrix}
(Y_{i}^{\text{obs}} - \alpha^* - \tau_{\text{SP}} - \beta^* X_i) \\
1 \\
W_i - p \\
W_i \cdot (X_i - \mu_X)
\end{pmatrix}'.
\]

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

\[
\frac{E \left[ (Y_{i}^{\text{obs}} - \alpha^* - \tau_{\text{SP}} - \beta^* X_i)^2 \cdot (W_i - p)^2 \right]}{p^2 \cdot (1 - p)^2}.
\]

\[\square\]

**Proof of Theorem 3:**
We use the same reparametrization as in the first part of the proof of Theorem 2:

\[
\begin{pmatrix}
\hat{\alpha}_c \\
\hat{\beta}_c \\
\hat{\alpha}_t \\
\hat{\beta}_t
\end{pmatrix} = \begin{pmatrix}
\alpha + \beta \mu_X \\
\beta \\
\alpha + \tau + \beta' \mu_X \\
\gamma + \beta
\end{pmatrix}.
\]
In terms of the new parameters \( \gamma^* = \beta_i^* - \beta_c^* \). In the proof of Theorem 2 it was shown that the population values for \((\hat{\alpha}_c, \beta_c)\) solve
\[
(\hat{\alpha}_c, \beta_c^*) = \arg\min_{\alpha_c, \beta_c} \mathbb{E} \left[ (1 - W_i) \cdot (Y_{i}^{\text{obs}} - \alpha_c - \beta_c^* (X_i - \mu_X))^2 \right]
\]
\[
= \arg\min_{\alpha_c, \beta_c} \mathbb{E} \left[ (1 - W_i) \cdot (Y_{i}(0) - \alpha_c - \beta_c^* (X_i - \mu_X))^2 \right].
\]
Because of the randomization \( W_i \) is independent of \( Y_i(0) \) and \( X_i \), and so
\[
(\hat{\alpha}_c, \beta_c^*) = \arg\min_{\alpha_c, \beta_c} (1 - p) \cdot \mathbb{E} \left[ (Y_{i}(0) - \alpha_c - \beta_c^* (X_i - \mu_X))^2 \right].
\]
A similar argument shows that \((\hat{\alpha}_i^*, \beta_i^*)\) solve the same optimization problem:
\[
(\hat{\alpha}_i^*, \beta_i^*) = \arg\min_{\alpha_i, \beta_i} \mathbb{E} \left[ (Y_{i}(1) - \alpha_i - \beta_i^* (X_i - \mu_X))^2 \right]
\]
\[
= \arg\min_{\alpha_i, \beta_i} (1 - p) \cdot \mathbb{E} \left[ (Y_{i}(0) + \tau - \alpha_i - \beta_i^* (X_i - \mu_X))^2 \right].
\]
(because by the null hypothesis of zero effects \( Y_i(1) = Y_i(0) + \tau \) and so \( \gamma^* = \beta_i^* - \beta_c^* = 0 \). This finishes part \( (i) \).
Under the null hypothesis \( Y_i(1) = Y_i(0) + \tau \), \( \gamma^* = 0 \). Then \( \sqrt{N} \hat{\gamma} \) will in large samples have a normal distribution with variance \( V_{\gamma} \), and the quadratic form \( Q_{\text{const}} \) will have a Chi-squared distribution with degrees of freedom equal to the dimension of \( X_i \). This concludes part \( (ii) \).
Under the null hypothesis \( Y_i(1) = Y_i(0) \) for all units) it also follows that \( \tau_{SP} = 0 \). In that case \( \sqrt{N}(\hat{\tau}, \hat{\gamma}) \) are in large samples normally distributed with covariance matrix \( V_{\tau,\gamma} \). Hence the quadratic form \( Q_{\text{zero}} \) will in large samples have a Chi-squared distribution with degrees of freedom equal to the dimension of \( \tau \) and \( \gamma \), which is equal to the dimension of \( X_i \) plus one.
The covariance matrix for \((\hat{\tau}, \hat{\gamma})\) is most easily obtained from the parametrization in part \( (ii) \) of the proof of Theorem 2, in terms of \((\hat{\alpha}, \tau, \beta, \gamma)\). The point estimates for \( \tau \) and \( \gamma \) under this parametrization are identical to those under the parametrization \((\alpha, \tau, \beta, \gamma)\). Under the parametrization in terms of \((\hat{\alpha}, \tau, \beta, \gamma)\) the full covariance matrix of \( \sqrt{N}(\hat{\alpha} - \hat{\alpha}, \hat{\tau} - \tau, \hat{\beta} - \beta, \hat{\gamma} - \gamma) \) is given by \( \Gamma^{-1} \Delta(\Gamma')^{-1} \) as given in (A.1). To obtain the covariance matrix for \( \sqrt{N}(\hat{\tau} - \tau, \hat{\gamma} - \gamma) \) partition \( \Gamma^{-1} \Delta(\Gamma')^{-1} \) as
\[
V = \Gamma^{-1} \Delta(\Gamma')^{-1} = \begin{pmatrix}
V_{\hat{\alpha}, \hat{\alpha}} & V_{\hat{\alpha}, \tau} & V_{\hat{\alpha}, \beta} & V_{\hat{\alpha}, \gamma} \\
V_{\hat{\tau}, \hat{\alpha}} & V_{\hat{\tau}, \tau} & V_{\hat{\tau}, \beta} & V_{\hat{\tau}, \gamma} \\
V_{\hat{\beta}, \hat{\alpha}} & V_{\hat{\beta}, \tau} & V_{\hat{\beta}, \beta} & V_{\hat{\beta}, \gamma} \\
V_{\hat{\gamma}, \hat{\alpha}} & V_{\hat{\gamma}, \tau} & V_{\hat{\gamma}, \beta} & V_{\gamma, \gamma}
\end{pmatrix}.
\]
The covariance matrix for \( \sqrt{N}(\hat{\tau} - \tau, \hat{\gamma} - \gamma) \) is then
\[
V_{\tau, \gamma} = \begin{pmatrix}
V_{\tau, \tau} & V_{\tau, \gamma} \\
V_{\gamma, \tau} & V_{\gamma, \gamma}
\end{pmatrix}.
\]
The covariance matrix for \( \sqrt{N}(\hat{\gamma} - \gamma) \) is simply \( V_{\gamma, \gamma} \).
### Table 7.1: Summary Statistics for PRC-CPPT Cholesterol Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control (N = 172)</th>
<th>Treatment (N = 165)</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ave</td>
<td>Sample S.D</td>
<td>Ave</td>
<td>Sample S.D</td>
</tr>
<tr>
<td>pretreatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chol1</td>
<td>297.1</td>
<td>23.1</td>
<td>297.0</td>
<td>20.4</td>
</tr>
<tr>
<td>chol2</td>
<td>289.2</td>
<td>24.1</td>
<td>287.4</td>
<td>21.4</td>
</tr>
<tr>
<td>cholp</td>
<td>291.2</td>
<td>23.2</td>
<td>289.9</td>
<td>20.4</td>
</tr>
<tr>
<td>posttreatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cholf</td>
<td>282.7</td>
<td>24.9</td>
<td>256.5</td>
<td>26.2</td>
</tr>
<tr>
<td>chold</td>
<td>-8.5</td>
<td>10.8</td>
<td>-33.4</td>
<td>21.3</td>
</tr>
<tr>
<td>comp</td>
<td>74.5</td>
<td>21.0</td>
<td>59.9</td>
<td>24.4</td>
</tr>
</tbody>
</table>

### Table 7.2: Regression Estimates for Average Treatment Effects

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Effect of Assignment to Treatment On:</th>
<th>Post Cholesterol Level (s.e.)</th>
<th>Compliance (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Cov</td>
<td></td>
<td>-26.22 (3.93)</td>
<td>-14.64 (3.51)</td>
</tr>
<tr>
<td>cholp</td>
<td></td>
<td>-25.01 (2.60)</td>
<td>-14.68 (3.51)</td>
</tr>
<tr>
<td>chol1, chol2</td>
<td></td>
<td>-25.02 (2.59)</td>
<td>-14.95 (3.50)</td>
</tr>
<tr>
<td>chol1, chol2, interacted with W</td>
<td></td>
<td>-25.04 (2.56)</td>
<td>-14.94 (3.49)</td>
</tr>
</tbody>
</table>
### Table 7.3: Regression Estimates for Average Treatment Effects on Post Cholesterol Levels

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Model for Levels</th>
<th>Model for Logs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>est (s.e.)</td>
<td>est (s.e.)</td>
</tr>
<tr>
<td>Assignment</td>
<td>-25.04 (2.56)</td>
<td>-0.098 (0.010)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.28 (12.05)</td>
<td>-0.133 (0.233)</td>
</tr>
<tr>
<td>chol1</td>
<td>0.98 (0.04)</td>
<td>-0.133 (0.233)</td>
</tr>
<tr>
<td>chol2-chol1</td>
<td>0.61 (0.08)</td>
<td>0.602 (0.073)</td>
</tr>
<tr>
<td>chol1 × Assignment</td>
<td>-0.22 (0.09)</td>
<td>-0.154 (0.107)</td>
</tr>
<tr>
<td>(chol2-chol1) × Assignment</td>
<td>0.07 (0.14)</td>
<td>0.184 (0.159)</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.63</td>
<td>0.57</td>
</tr>
</tbody>
</table>

### Table 7.4: P-values for Tests for Constant and Zero Treatment Effects, Using chol1 and chol2-chol1 as Covariates

<table>
<thead>
<tr>
<th></th>
<th>post chol level</th>
<th>compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero treatment effect</td>
<td>( \chi^2(3) ) approximation</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Fisher Exact P-value</td>
<td>0.000</td>
</tr>
<tr>
<td>constant treatment effect</td>
<td>( \chi^2(2) ) approximation</td>
<td>0.029</td>
</tr>
</tbody>
</table>