4

Linear Regression

Post-menopausal women who exercise less tend to have lower bone mineral density (BMD), putting them at increased risk for fractures. But they also tend to be older, frailer, and heavier, which may explain the association between exercise and BMD. People whose diet is high in fat on average have higher low-density lipoprotein (LDL) cholesterol, a risk factor for coronary heart disease (CHD). But they are also more likely to smoke and be overweight, factors which are also strongly associated with CHD risk. Increasing body mass index (BMI) predicts higher levels of hemoglobin Hb_{a1c}, a marker for poor control of glucose levels; however, older age and ethnic background also predict higher Hb_{a1c}.

These are all examples of potentially complex relationships in observational data where a continuous outcome of interest, such as BMD, SBP, and Hb_{a1c}, is related to a risk factor in analyses that do not take account of other factors. But in each case the risk factor of interest is associated with a number of other factors, or potential *confounders*, which also predict the outcome. So the simple association we observe between the factor of interest and the outcome may be explained by the other factors.

Similarly, in experiments, including clinical trials, factors other than treatment may need to be taken into account. If the randomization is properly implemented, treatment assignment is on average not associated with any prognostic variable, so confounding is usually not an issue. However, in stratified and other complex study designs, multipredictor analysis is used to ensure that confidence intervals, hypothesis tests, and P-values are valid. For example, it is now standard practice to account for clinical center in the analysis of multi-site clinical trials, often using the random effects methodology to be introduced in Chapter 8. And with continuous outcomes, stratifying on a strong predictor in both design and analysis can account for a substantial proportion of outcome variability, increasing the efficiency of the study. Multipredictor analysis may also be used when baseline differences are apparent between the randomized groups, to account for potential confounding of treatment assignment.

Another way the predictor–outcome relationship can depend on other factors is that an association may not be the same in all parts of the population. For example, the association of lipoprotein(a) levels with risk of CHD events appears to vary by ethnicity. Hormone therapy has a smaller beneficial effect on LDL levels among post-menopausal women who are also taking statins, and its effect on BMD may be greater in younger post-menopausal women. These are examples of *interaction*, where the association of a factor of primary interest with a continuous outcome is modified by another factor.

The problem of sorting out complex relationships is not restricted to continuous outcomes; the same issues arise with the binary outcomes covered in Chapter 6, survival times in Chapter 7, and repeated measures in Chapter 8. A general statistical approach to these problems is needed.

The topic of this chapter is the multipredictor linear regression model, a flexible and widely used tool for assessing the joint relationships of multiple predictors with a continuous outcome variable. We begin by illustrating some basic ideas in a simple example (Sect. 4.1). Then in Sect. 4.2 we present the assumptions of the multipredictor linear regression model and show how the simple linear model reviewed in Chapter 3 is extended to accommodate multiple predictors. Sect. 4.3 shows how categorical predictors with multiple levels are coded and interpreted. Sect. 4.4 describes how multipredictor regression models deal with confounding; in particular Sect. 4.4.1 uses a *counterfactual* view of *causal effects* to show how and under what conditions multipredictor regression models might be used to estimate them. These themes recur in Sects. 4.5 and 4.6 on mediation and interaction, respectively. Sect. 4.7 introduces some simple methods for assessing the fit of the model to the data and how well the data conform to the underlying assumptions of the model. In Chapter 5 we discuss the difficult problem of which variables and how many to include in a multipredictor model.

4.1 Example: Exercise and Glucose

Glucose levels above 125 mg/dL are diagnostic of diabetes, while levels in the range from 100 to 125 mg/dL signal increased risk of progressing to this serious and increasingly widespread condition. So it is of interest to determine whether exercise, a modifiable lifestyle factor, would help people reduce their glucose levels and thus avoid diabetes.

To answer this question definitively would require a randomized clinical trial, a difficult and expensive undertaking. As a result, research questions like this are often initially looked at using observational data. But this is complicated by the fact that people who exercise differ in many ways from those who do not, and some of the other differences might explain any unadjusted association between exercise and glucose level.

Table 4.1 shows a simple linear model using a measure of exercise to predict baseline glucose levels among 2,032 participants without diabetes in the HERS

Table 4.1. Unadjusted Regression of Glucose on Exercise

```
. reg glucose exercise if diabetes == 0

      Source |       SS       df       MS              Number of obs =    2032
-------------+------------------------------           F(  1,  2030) =   14.97
       Model |  1412.50418     1  1412.50418            Prob > F      =  0.0001
    Residual |  191605.195  2030  94.3867954            R-squared     =  0.0073
-------------+------------------------------           Adj R-squared =  0.0068
       Total |  193017.699  2031  95.0357946           Root MSE      =  9.7153

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |  -1.692789   .4375862    -3.87   0.000    -2.550954   -.8346243
       _cons |   97.36104   .2815138   345.85   0.000     96.80896    97.91313
------------------------------------------------------------------------------
```

clinical trial of hormone therapy (Hulley *et al.*, 1998). Women with diabetes are excluded because the research question is whether exercise might help to prevent progression to diabetes among women at risk, and because the causal determinants of glucose may be different in that group. Furthermore, glucose levels are far more variable among diabetics, a violation of the assumption of homoscedasticity, as we show in Sect. 4.7.3 below. The coefficient estimate (`Coef.`) for `exercise` shows that average baseline glucose levels were about 1.7 mg/dL lower among women who exercised at least three times a week than among women who exercised less. This difference is statistically significant ($t = -3.87, P < 0.0005$).

However, women who exercise are slightly younger, a little more likely to use alcohol, and in particular have lower average body mass index (BMI), all factors associated with glucose levels. This implies that the lower average glucose we observe among women who exercise could be due at least in part to differences in these other predictors. Under these conditions, it is important that our estimate of the difference in average glucose levels associated with exercise be "adjusted" for the effects of these potential confounders of the unadjusted association. Ideally, adjustment using a multipredictor regression model provides an estimate of the causal effect of exercise on average glucose levels, by *holding the other variables constant*. In Sect. 4.4 below, the rationale for estimation of causal effects using multipredictor regression models is explained in more detail.

From Table 4.2 we see that in a multiple regression model that also includes – that is, adjusts for – age, alcohol use (`drinkany`), and BMI, average glucose is estimated to be only about 1 mg/dL lower among women who exercise (95% CI 0.1–1.8, $P = 0.027$), holding the other three factors constant. The multipredictor model also shows that average glucose levels are about 0.7 mg/dL higher among alcohol users than among non-users. Average levels also increase by about 0.5 mg/dL per unit increase in BMI, and by 0.06 mg/dL for each additional year of age. Each of these associations is statistically significant after adjustment for the other predictors in the model. Furthermore, the

Table 4.2. Adjusted Regression of Glucose on Exercise

```
. reg glucose exercise age drinkany BMI if diabetes == 0;

      Source |       SS       df       MS              Number of obs =    2028
-------------+------------------------------           F(  4,  2023) =   39.22
       Model |  13828.8486     4  3457.21214           Prob > F      =  0.0000
    Residual |  178319.973  2023  88.1463042           R-squared     =  0.0720
-------------+------------------------------           Adj R-squared =  0.0701
       Total |  192148.822  2027  94.7946828           Root MSE      =  9.3886

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |   -.950441     .42873    -2.22   0.027    -1.791239   -.1096426
         age |   .0635495   .0313911     2.02   0.043     .0019872    .1251118
    drinkany |   .6802641   .4219569     1.61   0.107    -.1472513     1.50778
         BMI |    .489242   .0415528    11.77   0.000     .4077512    .5707328
       _cons |   78.96239   2.592844    30.45   0.000     73.87747    84.04732
------------------------------------------------------------------------------
```

association of each of the four predictors with glucose levels is adjusted for the effects of the other three, in the sense of taking account of its correlation with the other predictors and their adjusted associations with glucose levels. In summary, the multipredictor model for glucose levels shows that the unadjusted association between exercise and glucose is partly but not completely explained by BMI, age, and alcohol use, and that exercise remains a statistically significant predictor of glucose levels after adjustment for these three other factors – that is, when they are held constant by the multipredictor regression model.

Still, we have been careful to retain the language of association rather than cause and effect, and in Sect. 4.4 and Chapter 5 will suggest that adjustment for additional potential confounders would be needed before we could consider a causal interpretation of the result.

4.2 Multiple Linear Regression Model

Confounding thus motivates models in which the average value of the outcome is allowed to depend on multiple predictors instead of just one. Many basic elements of the multiple linear model carry over from the simple linear model, which was reviewed in Sect. 3.3. In Sects. 4.4.1–4.4.9 below, we show how this model is potentially suited to estimating causal relationships between predictors and outcomes.

4.2.1 Systematic Part of the Model

For the simple linear model with a single predictor, the regression line is defined by

$$\begin{aligned} \mathrm{E}[y|x] &= \text{ average value of outcome } y \text{ given predictor value } x \\ &= \beta_0 + \beta_1 x. \end{aligned} \tag{4.1}$$

In the multiple regression model, this generalizes to

$$\mathrm{E}[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{4.2}$$

where $\mathbf{x}$ represents the collection of p predictors $x_1, x_2, \ldots x_p$ in the model, and $\beta_1, \beta_2, \ldots \beta_p$ are the corresponding regression coefficients.

The right-hand side of model (4.2) has a relatively simple form, a *linear combination* of the predictors and coefficients. Analogous linear combinations of predictors and coefficients, often referred to as the *linear predictor*, are used in all the other regression models covered in this book. Despite the simple form of (4.2), the multipredictor linear regression model is a flexible tool, and with the elaborations to be introduced later in this chapter, usually allows us to represent with considerable realism how the average value of the outcome varies systematically with the predictors. In Sect. 4.7, we will consider methods for examining the adequacy of this part of the model and for improving it.

Interpretation of Adjusted Regression Coefficients

In (4.2), the coefficient $\beta_j, j = 1, \cdots, p$ gives the change in $\mathrm{E}[y|\mathbf{x}]$ for an increase of one unit in predictor x_j, holding other factors in the model constant; each of the estimates is adjusted for the effects of all the other predictors. As in the simple linear model, the intercept β_0 gives the value of $\mathrm{E}[y|\mathbf{x}]$ when all the predictors are equal to zero; "centering" of the continuous predictors can make the intercept interpretable. If confounding has been persuasively ruled out, we may be willing to interpret the adjusted coefficient estimates as representing causal effects.

4.2.2 Random Part of the Model

As before, individual observations of the outcome y_i are modeled as varying by an error term ε_i about an average determined by their predictor values $\mathbf{x}_i$:

$$\begin{aligned} y_i &= \mathrm{E}[y_i|\mathbf{x}_i] + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \end{aligned} \tag{4.3}$$

where x_{ji} is the value of predictor variable x_j for observation i. We again assume that $\varepsilon_i \sim$ i.i.d $\mathcal{N}(0, \sigma_\varepsilon^2)$; that is, ε is normally distributed with mean zero and the same standard deviation σ_ε at every value of $\mathbf{x}$, and that its values are statistically independent.

Fitted Values, Sums of Squares, and Variance Estimators

From (4.2) it is clear that the fitted values $\hat{y}_i$, defined for the simple linear model in Equation (3.4), now depend on all p predictors and the corresponding regression coefficient estimates, rather than just one predictor and two coefficients. The resulting sums of squares and variance estimators introduced in Sect. 3.3 are otherwise unchanged in the multipredictor model.

In the glucose example, the residual standard deviation, shown as `Root MSE`, declines from 9.7 in the unadjusted model (Table 4.1) to 9.4 in the model adjusting for age, alcohol use, and BMI (Table 4.2).

Variance of Adjusted Regression Coefficients

Including multiple predictors does affect the variance of $\hat{\beta}_j$, which now depends on an additional factor r_j, the multiple correlation of x_j with the other predictors in the model. Specifically,

$$\hat{\mathrm{Var}}(\hat{\beta}_j) = \frac{s^2_{y|\mathbf{x}}}{(n-1)s^2_{x_j}(1-r_j^2)}. \tag{4.4}$$

where, as before, $s^2_{y|\mathbf{x}}$ is the residual variance of the outcome and $s^2_{x_j}$ is the variance of x_j; r_j is equivalent to $r = \sqrt{R^2}$ from a multiple linear model in which x_j is regressed on all the other predictors. The term $1/(1-r_j^2)$ is known as the *variance inflation factor*, since $\mathrm{Var}(\hat{\beta}_j)$ is increased to the extent that x_j is correlated with other predictors in the model.

However, inclusion of other predictors, especially powerful ones, also tends to decrease $s^2_{y|\mathbf{x}}$, the residual or unexplained variance of the outcome. Thus the overall impact of including other predictors on $\mathrm{Var}(\hat{\beta}_j)$ depends on both the correlation of x_j with the other predictors and how much additional variability they explain. In the glucose example, the standard error of the coefficient estimate for exercise declines slightly, from 0.44 to 0.43, after adjustment for age, alcohol use, and BMI. This reflects the reduction in residual standard deviation previously described, as well as a variance inflation factor in the adjusted model of only 1.03.

t-Tests and Confidence Intervals

The t-tests of the null hypothesis H_0: $\beta_j = 0$ and confidence intervals for β_j carry over almost unchanged for each of the βs estimated by the model, only using (4.4) rather than (3.11) to compute the standard error of the regression coefficient, and comparing the t-statistic to a t-distribution with $n-(p+1)$ degrees of freedom (p is the number of predictors in the model, and an extra degree of freedom is used in estimation of the intercept β_0).

However, there is a substantial difference in interpretation, since the results are now adjusted for other predictors. Thus in rejecting the null hypothesis

H_0: $\beta_j = 0$ we would be making the stronger claim that, in the population, x_j predicts y, holding the other factors in the model constant. Similarly, the confidence interval for β_j refers to the parameter which takes account of the other $p - 1$ predictors in the model.

We have just seen that $\text{Var}(\hat{\beta}_j)$ may not be increased by adjustment. However, in Sect. 4.4 we will see that including other predictors in order to control confounding commonly has the effect of attenuating the unadjusted estimate of the association of x_j with y. This reflects the fact that the population parameter being estimated in the adjusted model is often closer to zero than the parameter estimated in the unadjusted model, since some of the unadjusted association is explained by other predictors. If this is the case, then even if $\text{Var}(\hat{\beta}_j)$ is unchanged, it may be more difficult to reject H_0: $\beta_j = 0$ in the adjusted model. In the glucose example, the adjusted coefficient estimate for exercise is considerably smaller than the unadjusted estimate. As a result the t-statistic is reduced in magnitude from –3.87 to –2.22 – still statistically significant, but less highly so.

4.2.3 Generalization of R^2 and r

The coefficient of determination $R^2 = \text{MSS} / \text{TSS}$ retains its interpretation as the proportion of the total variability of the outcome that can be accounted for by the predictor variables. Under the model, the fitted values summarize all the information that the predictors supply about the outcome. Thus the multiple correlation coefficient $r = \sqrt{R^2}$ now represents the correlation between the outcome y and the fitted values $\hat{y}$. It is easy to confirm this identity by extracting the fitted values from a regression model and computing their correlation with the outcome (Problem 4.3). In the glucose example, R^2 increases from less than 1% in the unadjusted model to 7% after inclusion of age, alcohol use, and BMI, a substantial increase in relative if not absolute terms.

4.2.4 Standardized Regression Coefficients

In Sect. 3.3.9 we saw that the slope coefficient β_1 in a simple linear model is systematically related to the Pearson correlation coefficient (3.12); specifically, $r = \beta_1 \sigma_x / \sigma_y$, where σ_x and σ_y are the standard deviations of the predictor and outcome. Moreover, we pointed out that the scale-free correlation coefficient makes it easier to compare the strength of association between the outcome and various predictors across single-predictor models. In the context of a multipredictor model, *standardized regression coefficients* play this role. Obtained using the `beta` option to the `regress` command in Stata, the standardized regression coefficient $\hat{\beta}_j^s$ for predictor x_j is defined in analogy to (3.12) as

$$\hat{\beta}_j^s = \hat{\beta}_j \text{SD}(x_j)/\text{SD}(y), \tag{4.5}$$

where $\text{SD}(x_j)$ and $\text{SD}(y)$ are the sample standard deviations of predictor x_j and the outcome y. These standardized coefficient estimates are what would be obtained from the regression if the outcome and all the predictors were first rescaled to have standard deviation 1. Thus they give the change in standard deviation units in the average value of y per standard deviation increase in the predictor. Standardized coefficients make it easy to compare the strength of association of different continuous predictors with the outcome within the same model.

For binary predictors, however, the unstandardized regression coefficients may be more directly interpretable than the standardized estimates, since the unstandardized coefficients for such predictors simply estimate the differences in the average value of the outcome between the two groups defined by the predictor, holding the other predictors in the model constant.

4.3 Categorical Predictors

In Chapter 3 the simple regression model was introduced with a single continuous predictor. However, predictors in both simple and multipredictor regression models can be binary, categorical, or discrete numeric, as well as continuous numeric.

4.3.1 Binary Predictors

The exercise variable in the model for LDL levels shown in Table 4.1 is an example of a binary predictor. A good way to code such a variable is as an *indicator* or *dummy* variable, taking the value 1 for the group with the characteristic of interest, and 0 for the group without the characteristic. With this coding, the regression coefficient corresponding to this variable has a straightforward interpretation as the increase or decrease in average outcome levels in the group with the characteristic, with respect to the reference group.

To see this, consider the simple regression model for average glucose values:

$$\text{E}[\texttt{glucose}|x] = \beta_0 + \beta_1 \texttt{exercise} \tag{4.6}$$

With the indicator coding of `exercise` (1 = yes, 0 = no), the average value of glucose is $\beta_0 + \beta_1$ among women who do exercise, and β_0 among the rest. It follows directly that β_1 is the difference in average glucose levels between the two groups. This is consistent with our more general definition of β_j as the change in $\text{E}[y|\mathbf{x}]$ for a one-unit increase in x_j. Furthermore, the t-test of the null hypothesis H_0: $\beta_1 = 0$ is a test of whether the between-group difference in average glucose levels is statistically significant. In fact this unadjusted model is equivalent to a t-test comparing glucose levels in women who do and do not exercise. A final point: when coded this way, the average value of the exercise variable gives the proportion of women who exercise.

A commonly used alternative coding for binary variables is (1 = yes, 2 = no). With this coding, the coefficient β_1 retains its interpretation as the between-group difference in average glucose levels, but now among women who do not exercise as compared to those who do, a less intuitive way to think of the difference. Furthermore, with this coding the coefficient β_0 has no straightforward interpretation, and the average value of the binary variable is not equal to the proportion of the sample in either group. However, overall model fit, including fitted values of the outcome, standard errors, and *P*-values, are the same with either coding (Problem 4.1).

4.3.2 Multilevel Categorical Predictors

The 2,763 women in the HERS cohort also responded to a question about how physically active they considered themselves compared to other women their age. The five-level response, designated `physact`, ranged from "much less active" to "much more active," and was coded in order from 1 to 5. This is an example of an *ordinal* variable, as described in Chapter 2, with categories that are meaningfully ordered, but separated by increments that may not be accurately reflected in the numerical codes used to represent them. For example, responses "much less active" and "somewhat less active" may represent a larger difference in physical activity than "somewhat less active" and "about as active."

Multilevel categorical variables can also be *nominal*, in the sense that there is no intrinsic ordering in the categories. Examples include ethnicity, marital status, occupation, and geographic region. With nominal variables it is even clearer that the numeric codes often used to represent the variable in the database cannot be treated like the values of a numeric variable such as glucose.

Categories are usually set up to be mutually exclusive and exhaustive, so that every member of the population falls into one and only one category. In that case both ordinal and nominal categories define subgroups of the population.

Both types of categorical variables are easily accommodated in multipredictor linear and other regression models, using indicator or dummy variables. As with binary variables, where two categories are represented in the model by a single indicator variable, categorical variables with $K \geq 2$ levels are represented by $K-1$ indicators, one for each of level of the variable except a baseline or reference level. Suppose level 1 is chosen as the baseline level. Then for $k = 2, 3, \ldots, K$, indicator variable k has value 1 for observations belonging to the category k, and 0 for observations belonging to any of the other categories. Note that for $K = 2$ this also describes the binary case, in which the "no" response defines the baseline or reference group and the indicator variable takes on value 1 only for the "yes" group.

Stata automatically defines indicator variables using the `xi:` command prefix in conjunction with the `i.` variable prefix. By default it uses the level

Table 4.3. Coding of Indicators for a Multilevel Categorical Variable

	Indicator variables			
physact	_Iphysact_2	_Iphysact_3	_Iphysact_4	_Iphysact_5
Much less active	0	0	0	0
Somewhat less active	1	0	0	0
About as active	0	1	0	0
Somewhat more active	0	0	1	0
Much more active	0	0	0	1

with the lowest value as the reference group; for text variables this means using the first in alphabetic order. Following the Stata convention for the naming of the four indicator variables, Table 4.3 shows the values of the four indicator variables corresponding to the five response levels of `physact`. Each level of `physact` is defined by a unique pattern in the four indicator variables.

Furthermore, the corresponding βs have a straightforward interpretation For the moment, consider a simple regression model in which the five levels of `physact` are the only predictors. Then

$$\mathrm{E}[\texttt{glucose}|\mathbf{x}] = \beta_0 + \beta_2 \texttt{\_Iphysact\_2} + \cdots + \beta_5 \texttt{\_Iphysact\_5}. \tag{4.7}$$

For clarity, the βs in (4.7) are indexed in accord with the levels of `physact`, so β_1 does not appear in the model. Letting the four indicators take on values of 0 or 1 as appropriate for the five groups defined by `physact`, we obtain

$$\mathrm{E}[\texttt{glucose}|\mathbf{x}] = \begin{cases} \beta_0 & \texttt{physact = 1} \\ \beta_0 + \beta_2 & \texttt{physact = 2} \\ \beta_0 + \beta_3 & \texttt{physact = 3} \\ \beta_0 + \beta_4 & \texttt{physact = 4} \\ \beta_0 + \beta_5 & \texttt{physact = 5.} \end{cases} \tag{4.8}$$

From (4.8) it is clear that the intercept β_0 gives the value of $\mathrm{E}[\texttt{glucose}|\mathbf{x}]$ in the reference or much less active group (`physact` = 1). Then it is just a matter of subtracting the first line of (4.8) from the second to see that β_2 gives the difference in the average glucose in the somewhat less active group (`physact` = 2) as compared to the much less active group. Accordingly the t-test of H_0: $\beta_2 = 0$ is a test of whether average glucose levels are the same in the much less and somewhat less active groups (`physact` = 1 and 2). And similarly for β_3, β_4, and β_5.

Four other points are to be made from (4.8).

- Without other predictors, or covariates, the model is equivalent to a one-way ANOVA (Problem 4.10). Also the model is said to be *saturated* and the population group means would be estimated under model (4.8) by the sample averages. With covariates, the estimated means for each group would be adjusted for between-group differences in the covariates included in the model.

- The parameters of the model can be manipulated to give the estimated mean in any group, using (4.8), or to give the estimated differences between any two groups. For instance, the difference in average outcome levels between the much more and somewhat more active groups is equal to $\beta_5 - \beta_4$ (why?). All regression packages make it straightforward to estimate and test hypotheses about these *linear contrasts*. This implies that choice of reference group is in some sense arbitrary. While a particular choice may be best for ease of presentation, possibly because contrasts with the selected reference group are of primary interest, alternative reference groups result in essentially the same model (Problem 4.4).
- The five estimated group means can take on almost any pattern with respect to each other, in either the adjusted or unadjusted model. In contrast, if `physact` were treated as a score with integer values 1 through 5, the estimated means would be constrained to lie on a straight regression line.

Table 4.4 shows results for the model with `physact` treated as a categorical variable, again using data for women without diabetes in HERS. In the regression output, $\hat{\beta}_0$ is found in the column and row labeled `Coef.` and `_cons`; we see that average glucose in the much less active group is approximately 98.4 mg/dL. The differences between the reference group and the two most active groups are statistically significant; for instance, the average glucose level in the much more active group (`_Iphysact_5`) is 3.3 mg/dL lower than in the much less active group ($t = -2.92$, $P = 0.003$).

Using (4.8), the first `lincom` command after the regression computes the estimated mean in the somewhat less active group, equal to the sum of $\hat{\beta}_0$ (`_cons`) and $\hat{\beta}_2$ (`_Iphysact_2`), or 97.6 mg/dL (95% CI 96.5–98.6 mg/dL). We can also use the `lincom` command to assess pairwise differences between two groups when neither is the referent. For example, the second `lincom` result in Table 4.4 shows that average glucose is 2.1 mg/dL lower in among women in the much more active (`physact` = 5) group as compared to those who are about as active (`physact` = 3), and that this difference is statistically significant ($t = -2.86$, $P = 0.004$). The last two results in the table are explained below.

4.3.3 The *F*-Test

Although every pairwise contrast between levels of a categorical predictor is readily available, the t-tests for these multiple comparisons provide no overall evaluation of the importance of the categorical variable, or more precisely a single test of the null hypothesis that the mean level of the outcome is the same at all levels of this predictor. In the example, this is equivalent to a test of whether any of the four coefficients corresponding to `physact` differ from zero. The `testparm` result in Table 4.4 ($F(4, 2027) = 4.43, P = 0.0014$) shows that glucose levels clearly differ among the groups defined by `physact`.

Table 4.4. Regression of Glucose on Physical Activity

```
. xi: reg glucose i.physact if diabetes == 0;
i.physact         _Iphysact_1-5       (naturally coded; _Iphysact_1 omitted)

      Source |       SS       df       MS              Number of obs =    2032
-------------+------------------------------           F(  4,  2027) =    4.43
       Model |  1673.09022     4  418.272554           Prob > F      =  0.0014
    Residual |  191344.609  2027  94.3979322           R-squared     =  0.0087
-------------+------------------------------           Adj R-squared =  0.0067
       Total |  193017.699  2031  95.0357946           Root MSE      =  9.7159

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iphysact_2 |  -.8584489   1.084152    -0.79   0.429    -2.984617    1.267719
 _Iphysact_3 |  -1.226199   1.011079    -1.21   0.225     -3.20906    .7566629
 _Iphysact_4 |  -2.433855   1.010772    -2.41   0.016    -4.416114   -.4515951
 _Iphysact_5 |  -3.277704   1.121079    -2.92   0.003    -5.476291   -1.079116
       _cons |   98.42056   .9392676   104.78   0.000     96.57853    100.2626
------------------------------------------------------------------------------
. lincom _cons + _Iphysact_2;
 ( 1)  _Iphysact_2 + _cons = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   97.56211   .5414437   180.19   0.000     96.50027    98.62396
------------------------------------------------------------------------------

. lincom _Iphysact_5 - _Iphysact_3;
 ( 1) - _Iphysact_3 + _Iphysact_5 = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.051505    .717392    -2.86   0.004    -3.458407   -.6446024
------------------------------------------------------------------------------

. testparm _I*;
 ( 1)  _Iphysact_2 = 0
 ( 2)  _Iphysact_3 = 0
 ( 3)  _Iphysact_4 = 0
 ( 4)  _Iphysact_5 = 0
       F(  4,  2027) =    4.43
            Prob > F =    0.0014

. test - _Iphysact_2 + _Iphysact_4 + 2 * _Iphysact_5 = 0;
 ( 1) - _Iphysact_2 + _Iphysact_4 + 2 _Iphysact_5 = 0
       F(  1,  2027) =   12.11
            Prob > F =    0.0005
```

4.3.4 Multiple Pairwise Comparisons Between Categories

When the focus is on the difference between a single pre-specified pair of subgroups, the overall F-test is of limited interest and the t-test for the single contrast between those subgroups can be used without inflation of the type-I error rate. All levels of the categorical predictor should still be retained in the analysis, however, because residual variance can be reduced, sometimes substantially, by splitting out the remaining groups. Furthermore, this avoids combining the remaining subgroups with either of the pre-specified groups, focusing the contrast on the comparison of interest.

However, it is frequently of interest to examine multiple pairwise differences between levels of a categorical predictor, especially when the overall F-test is statistically significant, and in some cases even when it is not. Examples include comparisons between treatments in a clinical trial with more than one active treatment arm, or in longitudinal data, to be discussed in Chapter 8, when between-treatment differences are evaluated at multiple points in time.

For this case, various methods are available for controlling the experiment-wise type-I error rate (EER) for the wider set of comparisons. These methods differ in the trade-off made between power and the breadth of the circumstances under which the type-I error rate is protected. One of the most straightforward is Fisher's *least significant difference* (LSD) procedure, in which the pairwise-comparisons are carried out using t-tests at the nominal type-I error rate, but only if the overall F-test is statistically significant; otherwise the null hypothesis is accepted for all the pairwise comparisons. This protects the EER under the *complete null hypothesis* that all the group-specific population means are the same. However, it is subject to inflation of the EER under *partial null hypotheses* – that is, when there are some real population differences between subgroups.

More conservative procedures that protect the EER under partial null hypotheses include setting the level of the pairwise tests required to declare statistical significance equal to α/k (Bonferroni) or $1-(1-\alpha)^{1/k}$ (Sidak), where α is the desired EER and k is the number of pre-planned comparisons to be made. The Sidak correction is slightly more liberal for small values of k, but otherwise equivalent. The Scheffé method is another, though very conservative, method in which differences can be declared statistically significant only when the overall F-test is also statistically significant. The Tukey *honestly significant difference* (HSD) and Tukey-Kramer methods are more powerful than the Bonferroni, Sidak, or Scheffé approaches and also perform well under partial null hypotheses.

A special case arises when only comparisons with a single reference group are of interest, as might arise in a clinical trial with multiple treatments and a single placebo control. In this situation, Dunnett's test achieves better power than alternatives designed for all pairwise comparisons, while still protecting the EER under partial null hypotheses. It also illustrates the general principle that controlling the EER for a smaller number of contrasts is less costly in terms of power, so that it makes sense to control only for the contrasts of interest. Compare this approach to Scheffé's, which controls the EER for all possible linear contrasts but at a considerable expense in power.

The previous alternatives provide simultaneous inference on all the pairwise comparisons considered. Various *step-down* and *step-up* multiple-stage testing procedures attempt to improve power using testing of cleverly sequenced hypotheses that only continues as long as the test results are statistically significant. The Duncan and Student-Newman-Keuls procedures fall in this class. However, neither protects the EER under partial null hypotheses.

As noted in Sect. 3.1.5, the Bonferroni, Sidak, and Scheffé procedures are available with the `oneway` ANOVA in Stata, but not in the regression `regress` command used for linear regression. Thus using these methods in examining estimates provided by a multipredictor linear model may require help from a statistician.

4.3.5 Testing for Trend Across Categories

The coefficient estimates for the categories of `physact` shown in Table 4.4 decrease in order, so a linear trend in `physact` might be an adequate representation of the association with glucose. Tests for linear trend across the values of `physact` are best performed using a *linear contrast* in the coefficients corresponding to the various levels of the categorical predictor. As compared to a simpler approach in which the numeric values of the categorical variable are treated as a score, this approach is more efficient, in that the model captures both trend and departures from it, reducing the residual variance that makes regression effects harder to detect.

Table 4.5. Linear Contrasts Used for Testing Trend

Number of levels	Linear contrast
3	$\beta_3 = 0$
4	$-\beta_2 + \beta_3 + 3\beta_4 = 0$
5	$-\beta_2 + \beta_4 + 2\beta_5 = 0$
6	$-3\beta_2 - \beta_3 + \beta_4 + 3\beta_5 + 5\beta_6 = 0$

Table 4.5 summarizes linear contrasts that would be used for testing trend when the categorical variable has 3–6 levels with evenly spaced numeric codes (e.g., 1, 2, 3, 4, 5), and the category with the lowest numeric code is treated as the reference. As in the `physact` example, β_k is the coefficient corresponding to the indicator for category k. These contrasts can be motivated as the slope coefficients from a regression in which the group means are modeled as linear in the sequential numeric codes for the categorical variable. Note that for a categorical variable with only three levels, the t-test for β_3, the coefficient for the category with the largest numeric code, provides the test for trend. These formulas are valid for all the other models in this book.

In the `physact` example, shown in Table 4.4, we tested the hypothesis H_0: $-\beta_2 + \beta_4 + 2\beta_5 = 0$. The result ($F(1, 2027) = 12.11, P = 0.0005$) leaves little doubt that there is a declining trend in glucose levels with increasing values of `physact`.

The pattern in average glucose across the levels of a categorical variable could be characterized by both a linear trend and a departure from trend. Given a statistically significant trend according to the test of the linear contrast, it is easy to test for such a departure. This test uses a model in which

Table 4.6. Model Assessing Departures from Linear Trend

```
. xi: reg glucose physact i.physact if diabetes == 0;
i.physact         _Iphysact_1-5       (naturally coded; _Iphysact_1 omitted)

      Source |       SS       df       MS              Number of obs =    2032
-------------+------------------------------           F(  4,  2027) =    4.43
       Model |  1673.09022     4  418.272554           Prob > F      =  0.0014
    Residual |  191344.609  2027  94.3979322           R-squared     =  0.0087
-------------+------------------------------           Adj R-squared =  0.0067
       Total |  193017.699  2031  95.0357946           Root MSE      =  9.7159

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     physact |  -.8194259   .2802698    -2.92   0.003    -1.369073    -.269779
 _Iphysact_2 |   -.039023   .9015677    -0.04   0.965    -1.807119    1.729073
 _Iphysact_3 |   .4126531   .6739888     0.61   0.540      -.90913    1.734436
 _Iphysact_4 |    .024423   .6366194     0.04   0.969    -1.224074     1.27292
 _Iphysact_5 |  (dropped)
       _cons |   99.23999   1.184013    83.82   0.000     96.91798     101.562
------------------------------------------------------------------------------
. testparm _I*;
 ( 1)  _Iphysact_2 = 0
 ( 2)  _Iphysact_3 = 0
 ( 3)  _Iphysact_4 = 0
 ( 4)  _Iphysact_5 = 0
       Constraint 4 dropped
       F(  3,  2027) =    0.26
            Prob > F =    0.8511
```

the categorical variable appears both as a *score* (i.e., is treated as a continuous predictor) and as a set of indicators. In Table 4.6 the F-test for the joint effect of `physact` as a categorical variable ($F(3, 2027) = 0.26, P = 0.85$) shows that there is little evidence for departures from a linear trend in this case.

It is important to note that in Table 4.6, both the coefficient and the t-test for the effect of `physact` as a score ($\hat{\beta} = -0.82, t = -2.92, P = 0.003$) are not easily interpretable, because their values depend on which additional indicator is dropped from the model. The test for trend must be carried out using the linear contrast described earlier.

4.4 Confounding

In Table 4.1, the unadjusted coefficient for `exercise` estimates the difference in mean glucose levels between two subgroups of the population of women with heart disease. But this contrast ignores other ways in which those subgroups may differ. In other words, the analysis does not take account of confounding of the association we see. Although the unadjusted contrast may be useful for describing subgroups, it would be risky to infer any causal connection between exercise and glucose on this basis. In contrast, the adjusted coefficient for `exercise` in Table 4.2 takes account of the fact that women who exercise also have lower BMI and are slightly younger and more likely to report alcohol use, all factors which are associated with differences in glucose levels. While

this adjusted model is clearly rudimentary, the underlying premise of multipredictor regression analysis of observational data is that with a sufficiently refined model (and good enough data), we can estimate causal effects, free or almost free of confounding.

To understand what confounding means, and to see how and under what conditions a multipredictor regression model might be able to overcome it, requires that we first state more clearly what we mean by the causal effect of a predictor variable. What would it mean, in more precise terms, for exercise to have a causal effect on glucose?

4.4.1 Causal Effects and Counterfactuals

To simplify the discussion, we focus on a binary predictor, a generic "exposure." Now suppose that we could run an experiment in which every member of a population is exposed and the value of the outcome observed; then, turning back the clock, we observe the outcome in the absence of exposure for every member of the population. Because we can never really turn back the clock, one of the two experimental outcomes for every individual is an unobservable *counterfactual*. However, this counterfactual experiment is central to our definition of the causal effect of the exposure.

> Definition: The *causal effect of an exposure on a continuous outcome* is the difference in population mean values of the outcome in the presence as compared to the absence of exposure, when the actual and counterfactual outcomes are observed for every member of the population as if by experiment, holding all other variables constant. If the means differ, then the exposure is a *causal determinant* of the outcome.

Three comments:

- The causal effect is defined as a *difference in population means*. This does not rule out variation in the causal effects of exposure at the individual level, possibly depending on the values of other variables. It might even be the case that exposure increases outcome levels for some members of the population and decreases them for others, yet the population means under the two conditions are equal. That is, we could have individual causal effects in the absence of an overall population causal effect.
- In our counterfactual experiment, turning back the clock to observe the outcome for each individual under both conditions means that the individual characteristics and experimental conditions that help determine the outcome are held constant, except for exposure. Thus the exposed and unexposed population means represent averaging over the same distribution of individual characteristics

and experimental conditions. In other words, all other causal determinants of outcome levels are perfectly balanced in the exposed and unexposed populations.
- Holding other variables constant does not imply that other causal effects of exposure are held constant after the experiment is initiated. These other effects may include *mediators* of the causal effect of exposure on the outcome (Sect. 4.5).

4.4.2 A Linear Model for the Counterfactual Experiment

To gain insight into our counterfactual experiment, we can write down expressions for Y_1 and Y_0, the outcomes under exposure and in its absence, using notation introduced earlier. In the following, X_1 is the indicator of exposure, with 0 = unexposed and 1 = exposed. For simplicity, we also assume that all the other determinants of the outcome – that is, the personal characteristics and experimental conditions held constant within individuals when we turn back the clock in the counterfactual experiment – are captured by another binary variable, X_2, which also has a causal effect on the outcome in the sense of our definition. Thus, for individual i the outcome under exposure is

$$y_{1i} = \beta_0 + \beta_1^c + \beta_2^c x_{2i} + \varepsilon_{1i}. \tag{4.9}$$

In (4.9)

- β_0 represents the mean of the outcome when $X_1 = X_2 = 0$.
- β_1^c is the *causal effect* of X_1: that is, the difference in population mean values of the outcome in the counterfactual experiment where X_1 is varied and X_2 is held constant.
- β_2^c is the causal effect of X_2, defined analogously as the difference in population means in a second counterfactual experiment in which X_1 is held constant and X_2 is varied.
- Variable x_{2i} is the observed value of X_2 for individual i.
- The error term ε_1 has mean zero and is assumed not to depend on X_1 or X_2. It captures variation in the causal effects across individuals as well as error in the measurement of the outcome.

Thus the population mean value of the outcome under exposure is

$$\begin{aligned} \mathrm{E}[Y_1] &= \mathrm{E}[\beta_0 + \beta_1^c + \beta_2^c X_2 + \varepsilon_1] \\ &= \beta_0 + \beta_1^c + \beta_2^c \mathrm{E}[X_2], \end{aligned} \tag{4.10}$$

where $\mathrm{E}[X_2]$ is the mean value of X_2 across all members of the population. Similarly, the outcome for individual i in the absence of exposure is

$$y_{0i} = \beta_0 + \beta_2^c x_{2i} + \varepsilon_{0i}, \tag{4.11}$$

and the population mean outcome under this experimental condition is

$$\begin{aligned} \mathrm{E}[Y_0] &= \mathrm{E}[\beta_0 + \beta_2^c X_2 + \varepsilon_0] \\ &= \beta_0 + \beta_2^c \mathrm{E}[X_2]. \end{aligned} \tag{4.12}$$

Crucially, in the counterfactual experiment, X_2 has the same mean $\mathrm{E}[X_2]$ under both the exposed and unexposed conditions, because it is held constant within individuals, each of whom contributes both an actual and counterfactual outcome. Subtracting (4.12) from (4.10), the difference in population means is

$$\begin{aligned} \mathrm{E}[Y_1] - \mathrm{E}[Y_0] &= \beta_0 + \beta_1^c + \beta_2^c \mathrm{E}[X_2] - \beta_0 - \beta_2^c \mathrm{E}[X_2] \\ &= \beta_1^c. \end{aligned} \tag{4.13}$$

Thus the linear model reflects the fact that in the counterfactual experiment, the difference in population means is equal to β_1^c, the causal effect of X_1, even in the presence of the other causal effects represented by X_2.

To illustrate using our first example, suppose that β_0, the mean glucose value when $X_1 = X_2 = 0$, is 100 mg/dL; η_1^c, the causal effect of exercise is to lower glucose levels an average of 2 mg/dL; that β_2^c, the causal effect of X_2 (which may represent younger age, lower BMI, alcohol use, as well as other factors) is to lower glucose 4 mg/dL; and that $\mathrm{E}[X_2]$, in this case the proportion of women with $X_2 = 1$, is 0.5. Now consider comparing the counterfactual population means. Using (4.10), mean glucose under the exercise condition would be

$$\beta_0 + \beta_1^c + (\beta_2^c \times 0.5) = 100 - 2 - (4 \times 0.5) = 96 \text{ mg/dL}. \tag{4.14}$$

In the absence of exercise, mean glucose would be

$$\beta_0 + (\beta_2^c \times 0.5) = 100 - (4 \times 0.5) = 98 \text{ mg/dL}. \tag{4.15}$$

Thus, using (4.13), or subtracting (4.15) from (4.14), the difference in the counterfactual means would be

$$\beta_1^c = -2 \text{ mg/dL}. \tag{4.16}$$

Now suppose we could sample randomly from this population of individuals and observe both actual and counterfactual outcomes for each, and that we used the simple linear model

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x_1 \tag{4.17}$$

to estimate the causal effect of exposure. Equation (4.13) implies that fitting the simple linear model (4.17) would result in an unbiased estimate of the causal effect β_1^c. By unbiased we mean that that over many repeated samples drawn from the population, the average or expected value of the estimates based on each sample would equal the population causal effect. Equivalently, using our notation for expected values,

$$\mathrm{E}[\hat{\beta}_1] = \beta_1^c. \tag{4.18}$$

Thus if we could sample from the counterfactual experiment the difference in sample averages under the exposed and unexposed conditions would provide an unbiased estimate of the causal effect of exercise on glucose.

4.4.3 Confounding of Causal Effects

In reality, of course, causal effects cannot be estimated in counterfactual experiments. The outcome is generally observable for each individual under only one of the two conditions. In place of a counterfactual experiment, we usually have to compare mean values of the outcome in two distinct populations, one composed of exposed individuals and the other of unexposed. In doing so, there is no longer any guarantee that the mean values of X_2 would be equal in the exposed ($X_1 = 1$) and unexposed ($X_1 = 0$) populations. Note that this inequality would mean that X_1 and X_2 are correlated.

However, since both β_1^c and β_2^c represent causal effects, we can still use (4.10) and (4.12) to express the two population means. Letting $\mathrm{E}_1[X_2]$ denote the mean of X_2 in the exposed, the mean outcome value in that population is

$$\mathrm{E}[Y_1] = \beta_0 + \beta_1^c + \beta_2^c \mathrm{E}_1[X_2]. \tag{4.19}$$

Similarly, with $\mathrm{E}_0[X_2]$ denoting the mean of X_2 among the unexposed, the mean of the outcome in that population is

$$\mathrm{E}[Y_0] = \beta_0 + \beta_2^c \mathrm{E}_0[X_2]. \tag{4.20}$$

This implies that

$$\begin{aligned} \mathrm{E}[Y_1] - \mathrm{E}[Y_0] &= \beta_0 + \beta_1^c + \beta_2^c \mathrm{E}_1[X_2] - \beta_0 - \beta_2^c \mathrm{E}_0[X_2] \\ &= \beta_1^c + \beta_2^c (\mathrm{E}_1[X_2] - \mathrm{E}_0[X_2]). \end{aligned} \tag{4.21}$$

Thus the difference in population means is now arbitrarily different from the causal effect, depending on the difference between $\mathrm{E}_1[X_2]$ and $\mathrm{E}_0[X_2]$ and on the magnitude of β_2^c, the population causal effect of X_2. From this it follows that if we sampled randomly from the combined exposed and unexposed populations, an estimate of β_1 found using the simple linear model (4.17) ignoring X_2 would usually be biased for β_1^c, the causal effect of X_1 on Y. In short, our estimate of the causal effect of X_1 would be confounded by the causal effect of X_2.

> Definition: *Confounding* is present when the difference in mean values of the outcome between populations defined by a potentially causal variable of interest is not equal to its causal effect on that outcome. In terms of our binary exposure, this can be expressed as $\mathrm{E}[Y_1] - \mathrm{E}[Y_0] \neq \beta_1^c$. As a consequence, the regression coefficient estimate for the causal variable given by fitting a simple linear model to a random sample of data from the combined population will be biased for the causal effect.

The effect of such confounding can be large and go in either direction. Returning to our first example, we again suppose that β_0, the mean glucose value in the population with $X_1 = X_2 = 0$ is 100 mg/dL; β_1^c, the causal effect of exercise is to lower glucose levels an average of 2 mg/dL; and that β_2^c, the causal effect of the potential confounder X_2 is to lower glucose 4 mg/dL. Now consider comparing populations where $\mathrm{E}_1[X_2]$, the proportion with $X_2 = 1$ among women who exercise, is 0.8; but $\mathrm{E}_0[X_2]$, the corresponding proportion among women who do not, is only 0.2. Then, using (4.19), mean glucose in the population of women who exercise would be

$$\beta_0 + \beta_1^c + (\beta_2^c \times 0.8) = 100 - 2 - (4 \times 0.8) = 94.8 \text{ mg/dL}. \tag{4.22}$$

In the population of women who do not exercise, mean glucose would be

$$\beta_0 + (\beta_2^c \times 0.2) = 100 - (4 \times 0.2) = 99.2 \text{ mg/dL}. \tag{4.23}$$

Thus, using (4.21), or subtracting (4.23) from (4.22), the difference in population means would be

$$\beta_1^c + \beta_2^c \times (0.8 - 0.2) = -2 - (4 \times 0.6) = -4.4 \text{ mg/dL}. \tag{4.24}$$

So the difference in population means would be considerably larger than the population causal effect of exercise. It follows that an unadjusted estimate of the causal effect using the simple linear model (4.17) would on average be substantially too large. In sum, under the plausible assumption that the other determinants of glucose have a real causal effect, (that is, $\beta_2^c \neq 0$), then only if the mean of X_2 were the same in both the exposed and unexposed populations – that is, $\mathrm{E}_1[X_2] = \mathrm{E}_0[X_2]$ – would the simple unadjusted comparison of sample averages – or population means – be free of confounding.

4.4.4 Randomization Assumption

The condition under which the difference in population means is equal to the causal effect can now be stated in terms of counterfactual outcomes: this equality will hold if the process determining whether individuals belong to the exposed or unexposed population is independent of their actual and counterfactual outcomes under those two conditions (exposure and its absence). In the glucose example, this would imply that exercising (or not) does not depend in any way on what glucose levels would be under either condition. This is known as the *randomization assumption*.

In general this assumption is met in randomized experiments, since in that setting, exposure – that is, treatment – is determined by a random process and does not depend on future outcomes. But in the setting of observational data where multipredictor regression models are most useful, this assumption clearly cannot be assumed to hold. In the HERS cohort, the randomization assumption holds for assignment to hormone therapy. However, in the glucose

example, the randomization assumption is violated when the co-determinants of glucose differ according to exercise. Essentially this is because the other factors captured by X_2 are causal determinants of glucose levels (or proxies for such determinants) *and* correlated with exercise.

4.4.5 Conditions for Confounding of Causal Effects

There are two conditions under which a covariate X_2 may confound the difference in mean values of an outcome Y in populations defined by the primary causal variable X_1:

- *X_2 is a causal determinant of Y, or a proxy for such determinants.*
- *X_2 is a causal determinant of X_1, or they share a common causal determinant.*

We note that age is one commonly used proxy for underlying causal effects. Further, if X_1 is a causal determinant of X_2, rather than the opposite, then X_2 would *mediate* rather than confound the causal effects of X_1. Mediation is discussed in more detail below in Sect. 4.5. Finally, bi-directional causal pathways between X_1 and X_2 would require more complex methods beyond the scope of this book.

4.4.6 Control of Confounding

The key to understanding how a multiple regression model can control for confounding when the randomization assumption does not hold is the concept of *holding other causal determinants of the outcome constant*. This is easiest to see in our example where all the causal determinants of the outcome Y other than X_1 are captured by the binary covariate X_2. The underlying argument is that within levels of X_2, we should be able to determine the causal effect of X_1, since within those strata X_2 is the same for all individuals and thus cannot explain differences in mean outcome levels according to X_1. Under the two-predictor linear model

$$\mathrm{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{4.25}$$

it is straightforward to write down the population mean value of the outcome for the four groups defined by X_1 and X_2. For purposes of illustration, we assume as in the previous example that $\beta_0 = 100$ mg/dL, $\beta_1^c = -2$ mg/dL, and $\beta_2^c = -4$ mg/dL. The results are shown in Table 4.7.

Examining the effect of X_1 while holding X_2 constant thus means comparing groups 1 and 2 as well as groups 3 and 4. It is easy to see that in both cases the between-group difference in $\mathrm{E}[y|\mathbf{x}]$ is simply β_1^c, or –2 mg/dL. We have made it possible to hold X_2 constant by modeling its effect, β_2^2. Furthermore, under our assumption that all causal determinants of Y other than X_1 are captured by X_2, the randomization assumption holds within the

Table 4.7. Linear Model for Causal Effects of X_1 and X_2

Group	X_1	X_2	E[y|**x**]	Population mean
1	0	0	β_0	100 mg/dL
2	1	0	$\beta_0 + \beta_1^c$	98 mg/dL
3	0	1	$\beta_0 + \beta_2^c$	96 mg/dL
4	1	1	$\beta_0 + \beta_1^c + \beta_2^c$	94 mg/dL

strata defined by X_2. As a result, the regression parameters β_1^2 and β_2^2 are interpretable as causal effects.

By extension from this simple example, the rationale for using multiple regression to control for confounding is the prospect of obtaining unbiased estimates of the causal effects of predictors of interest by modeling the effects of confounding variables. Furthermore, these arguments for the potential to control confounding using the multipredictor linear model can be extended, with messier algebra, to settings where there is more than one causal co-determinant of the outcome, where any or all of the predictor variables are continuous, counts, or multi-level categories, rather than binary, and where the outcome is binary or a survival time, as discussed in later chapters.

4.4.7 Range of Confounding Patterns

In our hypothetical causal example comparing distinct rather than counterfactual populations, the causal effect of X_1 is smaller than the simple difference between population means. We also saw this pattern in the estimate for the effect of exercise on glucose levels after adjustment for age, alcohol use, and BMI.

However, qualitatively different patterns can arise. We now consider a small hypothetical example where x_1, the predictor of primary interest, is binary and coded 0 and 1, and the potential confounder, x_2, is continuous. At one extreme, the effect of a factor of interest may be completely confounded by a second variable. In the upper left panel of Fig. 4.1, x_1 is shown to be strongly associated with y in unadjusted analysis, as represented in the scatterplot. However, the upper right panel shows that the unadjusted difference in y can be entirely explained by the continuous covariate x_2. The regression lines for x_2 are the same for both groups defined by x_1; in other words, there is no association with x_1 after adjustment for x_2.

At the other extreme, we may find little or no association in unadjusted analysis, because it is *masked* or *negatively confounded* by another predictor. The lower panels of Fig. 4.1 show this pattern. On the left, there is clearly no association between the binary predictor x_1 and y, but on the right the regression lines for x_2 are very distinct for the groups defined by x_1. In short, the association between x_1 and y is unmasked by adjustment for x_2. Negative confounding can occur under the following circumstances:

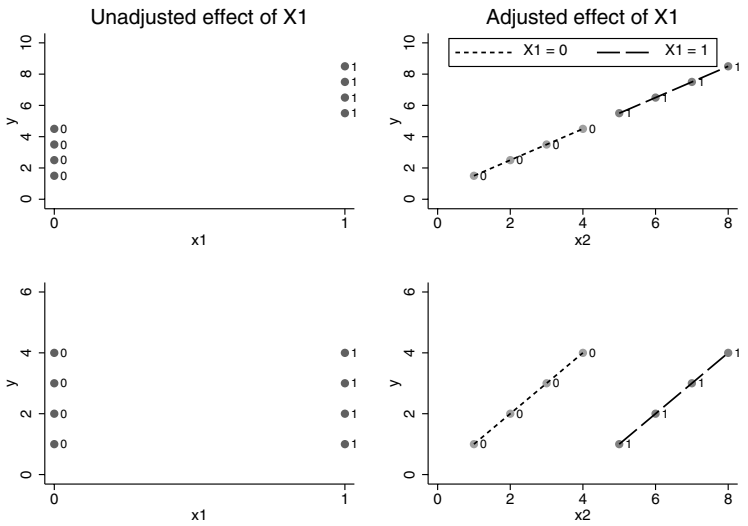

Fig. 4.1. Complete and Negative Confounding Patterns

- the predictors are inversely correlated, but have regression coefficients with the same sign.
- the two predictors are positively correlated, but have regression coefficients with the opposite sign.

The example shown in the lower panels of Fig. 4.1 is of the second kind.

4.4.8 Diagnostics for Confounding in a Sample

In Sects. 4.4.3 and 4.4.5 a definition and conditions for confounding were stated in terms of causal relationships defined by counterfactual differences in population means, which are clearly not verifiable. Randomized experiments provide the best approximation to these conditions, since the randomization assumption holds in that context. However, many epidemiologic questions about the causes of disease cannot be answered by experiments. In an observational sample, we do our best to control confounding by modeling the effects of potential confounders in multipredictor regression models.

In this context, we have to assess the potential for confounding in terms of associations between predictors and outcomes, an assessment best carried out within a hypothesized causal framework to help us distinguish potential confounders from mediators, defined below in Sect. 4.5. There are four useful diagnostics for potential confounding of the effect of a predictor of interest:

- The potential confounder must be associated with the outcome.

- The potential confounder must be associated with the predictor of interest.
- Adjustment for the potential confounder must affect the magnitude of the coefficient estimate for the predictor of interest. Note that this change could be in either direction, and may even involve change in sign; attenuation is the most common pattern, but increases in the absolute value of the coefficient are consistent with negative confounding.
- The potential confounder must make sense in terms of the hypothetical causal framework. In particular it should be plausible as a causal determinant of the predictor of interest, or as a proxy for such a determinant, and at the same time, it should clearly not represent a causal effect of the predictor of interest.

The first two diagnostics are the sample analogs of the conditions for confounding of causal effects given in Sect. 4.4.5. The third condition is the sample analog of a discrepancy between the causal effect of exposure, defined as the difference in mean values of the outcome between counterfactual populations, and the simple but potentially confounded difference between outcome means in distinct populations defined by exposure. If the fourth condition does not hold, we might see a similar pattern of associations and change in coefficients, but a different analysis is appropriate, as explained below in Sect. 4.5 on mediation.

4.4.9 Confounding Is Difficult To Rule Out

The problem of confounding is more resistant to multipredictor regression modeling than the simple two-predictor causal model in Sect. 4.4.6 might suggest. We assumed in that case that all causal determinants of Y other than X_1 were completely captured in the binary covariate X_2 – a substantial idealization. Of course, the multipredictor linear model (4.2) can (within limits imposed by sample size) include many more than two predictors, giving us considerable freedom to model the effects of other causal determinants. Nonetheless, for the multipredictor linear model to control confounding successfully and estimate causal effects without bias, all potential confounders must have been–

- recognized and assessed by design in the study,
- measured without error, and
- accurately represented in the systematic part of the model.

Logically, of course, it is not possible to show that all confounders have been measured, and in some cases it may be clear that they have not. Furthermore, the hypothetical causal framework may be uncertain, especially in the early stages of an investigating a research question. Also, measurement error in predictors is common; this may arise in some some cases because the study

has only measured proxies for the causal variables which actually confound a predictor of interest. Finally, Sect. 4.7 will show that accurate modeling of systematic relationships cannot be taken for granted.

4.4.10 Adjusted vs. Unadjusted $\hat{\beta}$s

In Sect. 4.4.3 we emphasized that confounding induces bias in unadjusted (or inadequately adjusted) estimates of the causal effects that are commonly the focus of our attention. This implies that unadjusted parameter estimates are always biased and adjusted estimates less so. But there is a sense in which this is misleading. In fact the two estimate different population quantities. The observed difference in average glucose levels between women who do and do not exercise is clearly interpretable, though it almost surely does not have a causal interpretation. Thus it should not be expected to have the same value as the causal parameter.

4.4.11 Example: BMI and LDL

With a more formal definition of confounding now in hand, we turn to a relatively simple example, again using data from the HERS cohort. Body mass index (BMI) and LDL cholesterol are both established heart disease risk factors. It reasonable to hypothesize that BMI is a causal determinant of LDL. An unadjusted model for BMI and LDL is shown in Table 4.8. The unadjusted estimate shows that average LDL increases .42 mg/dL per unit increase in BMI (95% CI: 0.16–0.67 mg/dL, $P = 0.001$). However, age, ethnicity (`nonwhite`), smoking, and alcohol use (`drinkany`) may confound this unadjusted association. These covariates may either represent causal determinants of LDL or be proxies for such determinants, and are correlated with but almost surely not caused by BMI, and so may confound the BMI–LDL relationship. After adjustment for these four demographic and lifestyle factors, the estimated increase in average LDL is 0.36 mg/dL per unit increase in BMI, an association that remains highly statistically significant ($P = 0.007$). In addition, average LDL is estimated to be 5.2 mg/dL higher among nonwhite women, after adjustment for between-group differences in BMI, age, smoking, and alcohol use. The association of smoking with higher LDL is also statistically significant, and there is some evidence for lower LDL among older women and those who use alcohol.

In this example, smoking is a negative confounder, because women with higher BMI are less likely to smoke, but both are associated with higher LDL. Negative confounding is further evidenced by the fact that the adjusted coefficient for BMI is *larger* (0.36 vs. 0.32 mg/dL) in the fully adjusted model shown in Table 4.8 than in a model adjusted for `age`, `nonwhite`, and `drinkany` but not for `smoking` (reduced model not shown).

The covariates in the adjusted model shown in Table 4.8 can all be shown to meet sample diagnostic criteria for potential confounding of the effect of

Table 4.8. Unadjusted and Adjusted Regressions of LDL on BMI

```
. reg LDL bmi

      Source |       SS       df       MS              Number of obs =    2747
-------------+------------------------------           F(  1,  2745) =   10.14
       Model |  14446.0223     1  14446.0223           Prob > F      =  0.0015
    Residual |  3910928.63  2745  1424.74631           R-squared     =  0.0037
-------------+------------------------------           Adj R-squared =  0.0033
       Total |  3925374.66  2746  1429.48822           Root MSE      =  37.746

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .4151123   .1303648     3.18   0.001     .1594894    .6707353
       _cons |   133.1913     3.7939    35.11   0.000     125.7521    140.6305
------------------------------------------------------------------------------

. reg LDL bmi age nonwhite smoking drinkany

      Source |       SS       df       MS              Number of obs =    2745
-------------+------------------------------           F(  5,  2739) =    5.97
       Model |  42279.1877     5  8455.83753           Prob > F      =  0.0000
    Residual |  3881903.3  2739  1417.27028           R-squared     =  0.0108
-------------+------------------------------           Adj R-squared =  0.0090
       Total |  3924182.49  2744  1430.09566           Root MSE      =  37.647

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .3591038   .1341047     2.68   0.007     .0961472    .6220605
         age |  -.1897166   .1130776    -1.68   0.094    -.4114426    .0320095
    nonwhite |   5.219436   2.323673     2.25   0.025     .6631081    9.775764
     smoking |   4.750738   2.210391     2.15   0.032     .4165363     9.08494
    drinkany |  -2.722354   1.498854    -1.82   0.069    -5.661351    .2166444
       _cons |   147.3153   9.256449    15.91   0.000      129.165    165.4656
------------------------------------------------------------------------------
```

BMI. For example, LDL is 5.2 mg/dL higher and average BMI 1.7 kg/m^2 higher among nonwhite women, and the adjusted effect of BMI is 13% smaller than the unadjusted estimate. Note that while the associations of ethnicity with both BMI and LDL are statistically significant in this example, ethnicity might still meaningfully confound BMI even if the differences were not nominally signficant. Evidence for this would still be provided by the substantial ($\geq$ 10%) change in the coefficient for BMI after adjustment for ethnicity, according to a useful (albeit ultimately arbitrary) rule of thumb (Greenland, 1989). Recommendations for inclusion of potential confounders in multipredictor regression models are given in Chapter 5.

Fig. 4.2 shows the unadjusted regression line for LDL and BMI, together with the adjusted lines specific to the white and nonwhite women, holding the other variables constant at their respective means. Two comments about Fig. 4.2:

- Some of the upward slope of the unadjusted regression line reflects the fact that women with higher BMI are more likely to be nonwhite, younger, and not to use alcohol – all factors associated with

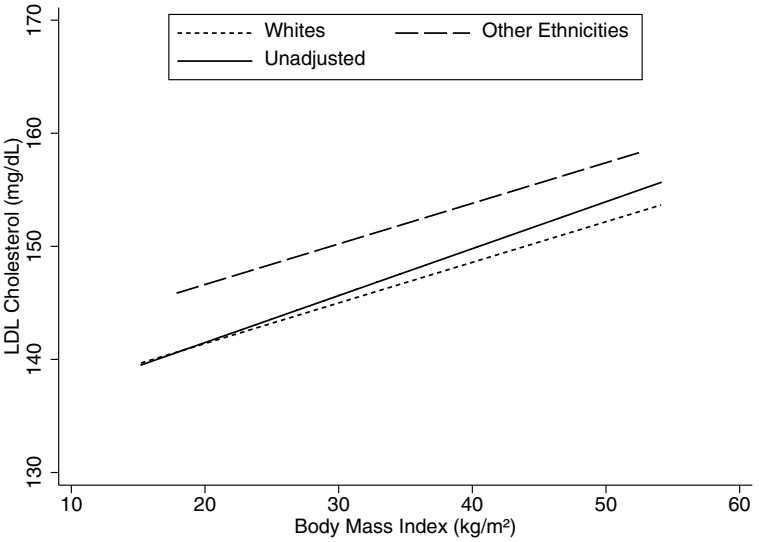

Fig. 4.2. Unadjusted and Adjusted Regression Lines

higher LDL. Despite the negative confounding by smoking, when these all these effects are accounted for using the multipredictor regression model, the slope for BMI is attenuated.

- The adjusted regression lines for white and nonwhite women are parallel, both with the same slope of 0.36 mg/dL per unit increase in BMI. Similar patterns are assumed to hold for adjusted regression lines specific to subgroups defined by smoking and alcohol use. Accordingly, the lines are separated by a vertical distance of 5.2 mg/dL at every value of BMI – the adjusted difference in average LDL by ethnicity. This pattern reflects the fact that the model does not allow for interaction between BMI and ethnicity. We assume that the slope for BMI is the same in both ethnic groups, and, equivalently, that the difference in LDL due to ethnicity is the same at every value of BMI. Testing the no-interaction assumption will be examined in Sect. 4.6 below.

4.5 Mediation

In Sect. 4.4.5 we presented conditions under which a covariate X_2 may confound the difference in mean values of an outcome Y in populations defined by the primary causal variable X_1:

- *X_2 is a causal determinant of Y, or a proxy for such determinants.*

- *X_2 is a causal determinant of X_1, or they share a common causal determinant.*

However, if X_1 is a causal determinant of X_2, then X_2 would not confound X_1 even if the first condition held; rather this would be an instance of *mediation* of the causal effects of X_1 on Y via its causal effects on X_2. That is, X_2 is affected by X_1 and in turn affects Y. For example, statin drugs reduce low-density LDL cholesterol levels, which in turn appear to reduce risk of heart attack; in this model, reductions in LDL mediate the protective effect of statins. The causal pathway from increased abdominal fat to development of diabetes and heart disease may operate through – that is, be mediated by – chemical messengers made by fat cells. The protective effect of bisphosponate drugs against fracture is mediated in part by the increases in bone mineral density (BMD) achieved by the drugs.

> Definition: A *mediating variable* is a predictor hypothesized to lie on the causal pathway between a predictor of interest and the outcome, and thus to mediate the predictor's effects.

With both mediation and confounding, the mediator/confounder is associated with both the predictor of interest and the outcome, and adjustment for it typically attenuates the estimated association of the primary predictor with the outcome. At the extreme, a mediating variable based on continuous monitoring of a common pathway at a point near the final outcome may almost completely remove the effects of antecedent predictors; an example is heart failure and death from coronary heart disease.

However, in contrast to confounding, the coefficient for the primary predictor before adjustment for the proposed mediator has the more direct interpretion as the overall causal effect, while the coefficient adjusted for the mediator represents its direct causal effect via other pathways that do not involve the mediator. In this instance, the adjusted analysis is used to estimate the direct effect of the primary predictor via other pathways, its indirect effect via the mediator, and the degree of mediation. In the context of clinical trials, the relative change in the coefficient for treatment after adjustment for a mediator is sometimes referred to as the *proportion of treatment effect explained*, or PTE (Freedman *et al.*, 1992). A new approach to estimation of PTE has been developed by Li *et al.* (2001).

4.5.1 Modeling Mediation

If a potential mediator is identified on *a priori* grounds, then a series of models can be used to examine whether–

- the predictor of interest also predicts the mediator;
- the mediator predicts the outcome in a model controlling for the predictor of interest;

- addition of the mediator to a multipredictor model for the outcome attenuates the estimated coefficient for the predictor of interest.

If all three elements of this pattern are present, then the data are consistent with the mediation hypothesis. However, because this pattern also reflects what is typically seen with confounding, the two causal models must be distinguished on non-statistical grounds.

Estimation of the overall and direct effects of the predictor of interest, as well as its indirect effects via the proposed mediator, has many potential difficulties. For example, longitudinal data would clearly provide stronger support the hypothesized causal model by potentially showing that changes or differences in the predictor of interest are associated with subsequent changes in the mediator, which in turn predict the outcome still later in time. However, as discussed in Sect. 7.3.1, longitudinal analyses set up to examine such temporal patterns can be misleading if the mediator also potentially confounds the association between the primary predictor and outcome (Hernan *et al.*, 2001). Furthermore, bias in estimation of the direct effects of the primary predictor can arise from uncontrolled confounding of the association between the mediator and the outcome (Robins and Greenland, 1992; Cole and Hernan, 2002) – even in clinical trials where the primary predictor is randomized treatment assignment.

4.5.2 Confidence Intervals for Measures of Mediation

In principle, confidence intervals for PTE or for the difference in coefficient estimates for the same predictor before and after adjustment for a mediator are straightforward to compute, particularly for linear models; they have also been developed for evaluating mediation using logistic (Freedman *et al.*, 1992; Li *et al.*, 2001) and Cox proportional hazards models (Lin *et al.*, 1997). However, unlike simple comparisons of coefficients estimated in the same model, assessing mediation involves comparing coefficient estimates from two *different* models estimated using the *same* data. As a result, the two estimates are correlated, making confidence intervals more difficult to compute. Since standard statistical packages generally do not provide them, this would require the analyst to carry out computations requiring moderately advanced programming skills. An alternative is provided by *bootstrap* procedures, which were introduced in Sect. 3.6.

4.5.3 Example: BMI, Exercise, and Glucose

In Sect. 4.1 we saw that the association of exercise and glucose levels among women at risk for diabetes was substantially confounded by age, alcohol use, and BMI. In that model, BMI was shown to be a powerful predictor of glucose levels, with each kg/m^2 increase in BMI associated with a 0.49 mg/dL increase in average glucose (95% CI 0.41–0.57, $P < 0.0005$). In fact, most of the

attenuation of the coefficient for exercise in the adjusted model was due to controlling for BMI, as is easily demonstrated by re-fitting the adjusted model omitting BMI.

In treating BMI as a confounder of exercise, we implicitly assumed that higher BMI makes women less likely to exercise: in short, BMI is a causal determinant of exercise. Of course, exercise might also be a determinant of BMI, which would considerably complicate the picture. However, exercise vigorous enough to result in weight loss was very uncommon in this cohort of older post-menopausal women with heart disease; furthermore, exercise was weakly associated ($P = 0.06$) with a small *in*crease in BMI of 0.12 kg/m^2 over the first year of the study, after adjusting for age, ethnicity, smoking, and self-report of poor or fair health. Thus the potential causal pathway from exercise to decreased BMI appears negligible in this population.

Accordingly, we examined the extent to which the effects of BMI on glucose levels might be mediated through its effects on likelihood of exercise. In implementing the series of models set out in Sect. 4.5.1, we first used a multipredictor logistic regression model (Chap. 6) to show that each kg/m^2 increase in BMI is associated with an 8% decrease in the odds of exercise (95% CI 4–10%, $P < 0.0005$). We have already observed that exercise is associated with a decrease in average glucose of about 1 mg/dL (95% CI 0.1–1.9, $P = 0.027$), after adjusting for BMI as well as age and alcohol use. However, the coefficient for BMI is only slightly attenuated when exercise is added to the model, from 0.50 to 0.49 mg/dL per kg/m^2 increase in BMI, a decrease of only 2.9%. As shown in Table 4.9, a bias-corrected bootstrap confidence interval for the percentage decrease in the BMI coefficient due to mediation by exercise, the equivalent of PTE, was 0.3–6.0%, showing that the attenuation was not just due to chance. Nonetheless, this analysis suggests that only a very small part of the effect of BMI on glucose levels is mediated by its effects on likelihood of exercising. Note that a short program had to be "defined" in order to fit the nested models before and after adjustment for exercise in each bootstrap sample and then compute PTE.

Of note, there was little evidence of bias in the estimate of PTE in this example, since bias correction did not affect the percentile-based confidence interval. However, the interval based on the normal approximation was somewhat different from either percentile-based CI, running from 0.1 to 5.7% and thus indicating some departure from normality in the sampling distribution of PTE; this is not uncommon with ratio estimates. Of course, the qualitative interpretation would be unchanged.

4.6 Interaction

In Sect. 4.4.5 we outlined the conditions under which a two-predictor linear model could be successfully used to eliminate confounding of the effects of a primary predictor X_1 by a confounder X_2. We presented the two-predictor

Table 4.9. Bootstrap Confidence Interval for PTE

```
. program define mediate, rclass
  1. version 8.0
  2. tempname bg0
  3. reg glucose BMI age10 drinkany if diabetes == 0
  4. scalar ‘bg0’ = _b[BMI]
  5. reg glucose exercise BMI age10 drinkany if diabetes == 0
  6. return scalar pte = (‘bg0’ - _b[BMI]) / ‘bg0’ * 100
  7.  end
. bootstrap "mediate" r(pte), reps(1000)
command:      mediate
statistic:    _bs_1      = r(pte)

Bootstrap statistics                              Number of obs    =      2028
                                                  Replications     =      1000

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |  1000  2.900681  .0546537  1.406885   .1398924   5.661469   (N)
             |                                      .3417959   6.031512   (P)
             |                                      .3424993   6.032569  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

model (4.25) under the tacit assumption that causal effect of X_1 on Y was the same within both strata defined by X_2. However, this may not hold. In this section we show how linear models can be used to model the interaction and thereby estimate causal effects that differ according to the level of a covariate.

4.6.1 Causal Effects and Interaction

In Sect. 4.4.2 we recognized that individual causal effects of exposure might vary across individuals from the overall population causal effect. In that context we assumed that the two differ for each individual by an amount that has mean zero and does not depend on either X_1 or X_2. However, suppose that in our counterfactual experiment, the population causal effect of X_1 does differ systematically according to the value of X_2. We could continue to compare the actual and counterfactual outcomes for each individual, thus holding all other variables constant, but in this case find that the causal effects, defined as the difference in population mean values of the outcome under exposure as compared to its absence, now vary across the strata of the population defined by X_2.

> Definition: *Interaction* means that the causal effect of a predictor on the outcome differs according to the level of another predictor.

Interaction is also referred to as *effect modification* or *moderation*, and must be distinguished from mediation (Baron and Kenny, 1986).

4.6.2 Modeling Interaction

Continuing with our example of a primary predictor X_1 and a single covariate X_2, it is straightforward to model the interaction between X_1 and X_2 using a three-predictor linear model. As before, the randomization assumption must hold within the two strata defined by X_2, so that the stratum-specific difference in population means is equal to the causal effect of X_1 within each stratum. But in this case we do not assume that the causal effects of X_1 are the same in both strata. To allow for the interaction, we use the following three-predictor linear model:

$$\mathrm{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \tag{4.26}$$

where $x_1 x_2$ simply denotes the product of the two predictors, equal to one only in the case where $X_1 = X_2 = 1$. It is again straightforward to write down the population mean values of the outcome for the four groups defined by X_1 and X_2. We assume as in the previous example that $\beta_1^c = -2$ mg/dL, $\beta_2^c = -4$ mg/dL, and $\beta_0 = 100$ mg/dL. But now in addition we assume that $\beta_3^c = -2$ mg/dL. The results are shown in Table 4.10.

Table 4.10. Interaction Model for Causal Effects of X_1 and X_2

Group	X_1	X_2	X_1X_2	$\mathrm{E}[y\|\mathbf{x}]$	Population mean
1	0	0	0	β_0	100 mg/dL
2	1	0	0	$\beta_0 + \beta_1$	98 mg/dL
3	0	1	0	$\beta_0 + \beta_2$	96 mg/dL
4	1	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	92 mg/dL

Examining the effect of X_1 while holding X_2 constant again means comparing groups 1 and 2 as well as groups 3 and 4. Now we do so not only to eliminate confounding, but also because the causal effects differ. In this case, when $X_2 = 0$, the between-group difference in $\mathrm{E}[y|\mathbf{x}]$ is simply β_1, or –2 mg/dL. However, when $X_2 = 1$, the difference is $\beta_1 + \beta_3$, or –4 mg/dL. We hold X_2 constant by modeling its effect with the parameter β_2, and allow for the interaction by modeling the difference in the causal effects of X_1 with the parameter β_3. Again, assuming that the causal determinants of Y other than X_1 are captured by X_2, the randomization assumption holds within the strata defined by X_2. As a result, $\beta_1 = \beta_1^c$ and $\beta_3 = \beta_3^c$, so the regression parameters remain interpretable as causal effects.

4.6.3 Overall Causal Effect in the Presence of Interaction

It is important to point out that even when the causal effect of X_1 differs according to the level of X_2, the overall causal effect of X_1 remains well-defined in the counterfactual experiment as the difference in population mean

values of the outcome in the presence as compared to the absence of exposure defined by X_1. In fact this overall causal effect is simply the weighted average of its causal effects within the strata defined by X_2, with weights defined by the proportions of the total population with $X_2 = 0$ and 1. This is not very different from averaging over the individual causal effects, except that in this case X_2 has a systematic effect on them. Furthermore, an estimate of β_1 using the two-predictor linear model (4.25) would be unbiased for this overall causal effect, provided the four groups in Table 4.7 are sampled in proportion to their relative sizes in the population.

However, in settings where an important interaction operates – especially where the causal effects differ in direction across strata – the overall causal effect is sometimes difficult to interpret. In addition, estimation and comparison of the stratum-specific causal effects will usually be of greater interest.

4.6.4 Example: Hormone Therapy and Statin Use

As an example of interaction, we examined whether the effect of hormone therapy (HT) on LDL cholesterol differs according to baseline statin use, using data from HERS. Suppose both assignment to hormone therapy and use of statins at baseline are coded using indicator variables. Then the product term for assessing interaction is also an indicator, in this case with value 1 only for the subgroup of women who reported using statins at baseline and were randomly assigned to hormone therapy. Now consider the regression model

$$\mathrm{E}[\mathtt{LDL}|\mathbf{x}] = \beta_0 + \beta_1 \mathtt{HT} + \beta_2 \mathtt{statins} + \beta_3 \mathtt{HTstatins}, \tag{4.27}$$

where `HT` is the indicator of assignment to hormone therapy, `statins` the indicator of baseline statin use, and `HTstatins` the product term.

Table 4.11. Model for Interaction of HT and Statins

Group	`HT`	`statins`	`HTstatins`	E[`LDL`$|\mathbf{x}$]
1	0	0	0	β_0
2	1	0	0	$\beta_0 + \beta_1$
3	0	1	0	$\beta_0 + \beta_2$
4	1	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Table 4.11 shows the values of (4.27) for each of the four groups of women defined by `HT` and `statins`. The difference in $\mathrm{E}[y|\mathbf{x}]$ between groups 1 and 2 is β_1, the effect of HT among women not using statins. Similarly, the difference in $\mathrm{E}[y|\mathbf{x}]$ between groups 3 and 4 is $\beta_1 + \beta_3$, the effect of HT among statin users. So the interaction term β_3 gives the difference in treatment effects in these two groups. Accordingly, a t-test of H_0: $\beta_3 = 0$ is a test for the equality

of the effects of HT among statin users as compared to non-users. Note that within the strata defined by baseline statin use, the randomization assumption can clearly be assumed to hold for HT, the indicator for random treatment assignment.

Taking analogous differences between groups 1 and 3 or 2 and 4 would show that β_2 gives the difference in average LDL among statin users as compared to non-users among women assigned to placebo, while $\beta_2 + \beta_3$ gives the analogous difference among women assigned to HT. However, in this case the randomization assumption does not hold, implying that that unbiased estimation of the causal effects of statin use would require careful adjustment for *confounding by indication* – that is, for the prognostic factors that lead physicians to prescribe this treatment.

Table 4.12. Interaction of Hormone Therapy and Statin Use

```
. gen HTstatins = HT * statins
. reg LDL1 HT statins HTstatins

      Source |       SS       df       MS              Number of obs =    2608
-------------+------------------------------           F(  3,  2604) =   52.68
       Model |  227141.021     3  75713.6735           Prob > F      =  0.0000
    Residual |  3742707.78  2604  1437.29177           R-squared     =  0.0572
-------------+------------------------------           Adj R-squared =  0.0561
       Total |  3969848.80  2607  1522.76517           Root MSE      =  37.912

------------------------------------------------------------------------------
        LDL1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HT |  -17.72836   1.870629    -9.48   0.000    -21.39643   -14.06029
     statins |  -13.80912    2.15213    -6.42   0.000    -18.02918   -9.589065
   HTstatins |   6.244416   3.076489     2.03   0.042     .2118044    12.27703
       _cons |   145.1567   1.325549   109.51   0.000     142.5575     147.756
------------------------------------------------------------------------------
. lincom HT + HTstatins
 ( 1)  HT + HTstatins = 0.0
------------------------------------------------------------------------------
        LDL1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -11.48394   2.442444    -4.70   0.000    -16.27327   -6.694615
------------------------------------------------------------------------------
```

Table 4.12 shows that there is some evidence for a smaller effect of HT on LDL among women reporting statin use at study baseline. The coefficient for `HT`, or $\hat{\beta}_1$, shows that among women who did not report statin use at baseline, average cholesterol at the first annual HERS visit was almost 18 mg/dL lower in the HT arm than in placebo, a statistically significant subgroup treatment effect. To obtain the estimate of the effect of HT among baseline statin users, we sum the coefficients for `HT` and `HTstatins` (that is, $\hat{\beta}_1 + \hat{\beta}_3$) using the `lincom` command. This shows that the treatment effect among baseline statin users was only –11.5 mg/dL, although this was also statistically significant. The difference ($\hat{\beta}_3$) of 6.2 mg/dL between the two treatment effects was also statistically significant ($t = 2.03, P = .042$). Finally, the results for variable

statins indicate that among women assigned to placebo, baseline statin use is a statistically significant predictor of LDL levels at the first annual visit.

4.6.5 Example: BMI and Statin Use

While it is often hard to obtain unbiased estimates of the causal effects of treatments like statins using observational data, a more tractable question of interest is whether the causal relationships of variables related to statin use may be modified by use of these drugs. Or it may be of interest simply to find out whether other risk factors differentially predict outcomes of interest according to use of related medications.

For example, the association between BMI and baseline LDL cholesterol levels was shown in Sect. 4.4.11 to be statistically significant after adjustment for demographics and lifestyle factors. However, treatment with statins may modify this association, possibly by interrupting the causal pathway between higher BMI and increased LDL. This would imply that BMI is less strongly associated with increased average LDL among statin users than among non-uses.

In examining this interaction, centering the continuous predictor variable BMI about its mean value of 28.6 kg/m^2 makes the parameter estimate for statin use more interpretable, as we show below. Then, to implement the analysis, we would first compute the product term statcBMI = statins $\times$ cBMI, where cBMI is the new centered BMI variable. Note that because statins is an indicator variable coded 1 for users and 0 for non-users, the product variable statcBMI is by definition equal to cBMI in statin users, but equal to zero for non-users. We then fit a multipredictor regression model including all these three predictors, as well as the potential confounders adjusted for previously. The resulting model for baseline LDL is

$$\begin{aligned} \mathrm{E}[\mathtt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_1 \mathtt{statins} + \beta_2 \mathtt{cBMI} + \beta_3 \mathtt{statcBMI} \\ & + \beta_4 \mathtt{age} + \beta_5 \mathtt{nonwhite} + \beta_6 \mathtt{smoking} + \beta_7 \mathtt{drinkany}. \end{aligned} \quad (4.28)$$

Thus among women who do not use statins,

$$\begin{aligned} \mathrm{E}[\mathtt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_2 \mathtt{cBMI} \\ & + \beta_4 \mathtt{age} + \beta_5 \mathtt{nonwhite} + \beta_6 \mathtt{smoking} + \beta_7 \mathtt{drinkany}, \end{aligned} \quad (4.29)$$

and the slope associated with cBMI in this group is β_2. In contrast, among statin users

$$\begin{aligned} \mathrm{E}[\mathtt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_1 \mathtt{statins} + \beta_2 \mathtt{cBMI} + \beta_3 \mathtt{statcBMI} \\ & + \beta_4 \mathtt{age} + \beta_5 \mathtt{nonwhite} + \beta_6 \mathtt{smoking} + \beta_7 \mathtt{drinkany} \\ = {} & \beta_0 + \beta_1 \mathtt{statins} + (\beta_2 + \beta_3) \mathtt{cBMI} \\ & + \beta_4 \mathtt{age} + \beta_5 \mathtt{nonwhite} + \beta_6 \mathtt{smoking} + \beta_7 \mathtt{drinkany}. \end{aligned} \quad (4.30)$$

In this group, the slope associated with BMI is $\beta_2+\beta_3$; so clearly the interaction parameter β_3 gives the difference between the two slopes.

The model also posits that the difference in average LDL between statin users and non-users depends on BMI. Subtracting (4.29) from (4.30), the difference in average LDL in statin users as compared to non-users is $\beta_1+\beta_3$cBMI. However, we may be reluctant to interpret this result as an unbiased estimate of the causal effects of statin use in view of the potential for uncontrolled confounding by indication.

Table 4.13. Interaction Model for BMI and Statin Use

```
. reg LDL cBMI statins statcBMI age nonwhite smoking drinkany

      Source |       SS       df       MS              Number of obs =    2745
-------------+------------------------------           F(  7,  2737) =   22.85
       Model |  216681.484     7  30954.4978           Prob > F      =  0.0000
    Residual |     3707501  2737  1354.58568           R-squared     =  0.0552
-------------+------------------------------           Adj R-squared =  0.0528
       Total |  3924182.49  2744  1430.09566           Root MSE      =  36.805

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     statins |  -16.25301   1.468788   -11.07   0.000    -19.13305   -13.37296
        cBMI |   .5821275    .160095     3.64   0.000     .2682082    .8960468
    statcBMI |   -.701947   .2693752    -2.61   0.009    -1.230146   -.1737478
         age |  -.1728526   .1105696    -1.56   0.118    -.3896608    .0439556
    nonwhite |   4.072767   2.275126     1.79   0.074    -.3883702    8.533903
     smoking |   3.109819    2.16704     1.44   0.151    -1.139381    7.359019
    drinkany |  -2.075282   1.466581    -1.42   0.157    -4.950999    .8004354
       _cons |   162.4052   7.583312    21.42   0.000     147.5356    177.2748
------------------------------------------------------------------------------

. lincom cBMI + statcBMI;
 ( 1)  cBMI + statcBMI = 0
------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1198195   .2206807    -0.54   0.587    -.5525371    .3128981
------------------------------------------------------------------------------
```

Table 4.13 shows the results of the interaction model for statin use and BMI. The estimated coefficients have the following interpretations:

- statins: Among women with cBMI = 0, or equivalently, with BMI = 28.6 kg/m^2, statin use was associated with LDL levels that were more than 16 mg/dL lower on average. Note that if we had not first centered BMI, this coefficient would be an estimate of the statin effect in women with BMI = 0.
- cBMI: Among women who do not use statins, the increase in average LDL is 0.58 mg/dL per unit increase in BMI. The association is statistically signficant (t=3.64, $P < 0.0005$).
- statcBMI: The slopes for the average change in LDL per unit increase in BMI differ by approximately –0.70 mg/dL according to

baseline statin use. That is, the increase in average LDL associated with increases in BMI is much less rapid among women who use statins. Moreover, the interaction is statistically significant ($t = -2.61, P = 0.009$).

- `lincom` is used to estimate the slope for BMI among statin users, equal to the sum of the slope among non-users plus the estimated difference in slopes. The estimate of -.12 mg/dL per unit increase in BMI is not statistically significant ($t = -0.54, P = 0.59$), but the 95% confidence interval (–0.55 to 0.31 mg/dL per unit increase in BMI) is consistent with effects comparable in magnitude to the point estimate for non-users.

Fig. 4.3 shows the estimated regression lines in the two groups, demonstrating that the parallel lines assumption is no longer constrained to hold in the interaction model. In summary, the analysis suggests that any adverse causal effects of higher BMI on LDL may be blocked by statin use.

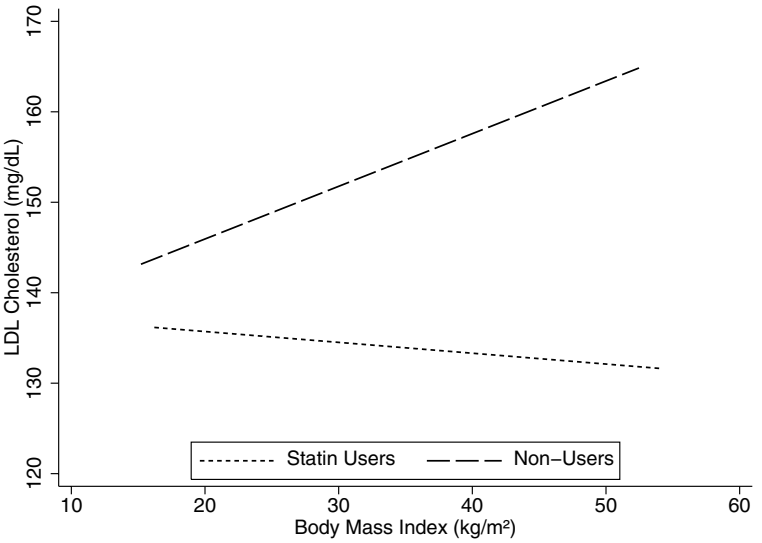

Fig. 4.3. Stratum-Specific Regression Lines

4.6.6 Interaction and Scale

Interaction models like (4.26) are often distinguished from simpler *additive* models typified by (4.25), which do not included product terms such as x_1x_2. Moreover, the simpler additive model is generally treated as the default in

predictor selection, with a product term being added only if there is more-or-less persuasive evidence that it is needed. It is important to recognize, however, that the need for interaction terms is dependent on the scale on which the outcome is measured (or, in the models discussed in later chapters, the scale on which its mean is modeled).

In Sects. 4.7.2 and 4.7.3 below we examine changes of the scale on which the outcome is measured to address violations of the linear model assumptions of normality and constant variance. Log transformation of the outcome, among the most commonly used changes of scale, effectively means modeling the average value of the outcome on a relative rather than absolute scale, as we show in Sect. 4.7.5 below. Similarly, in the analysis of before-and-after measurements of a response to treatment, we have the option of modeling percent rather than absolute change from baseline.

The issue of the dependence of interaction on scale arises in a similar but subtly different way with the other models discussed later in this book. For example, in logistic regression (Chap. 6) the *logit* transformation of $\mathrm{E}[Y|\mathbf{x}]$ is modeled, while in some generalized linear models (GLMs; Chap. 9), including the widely used Poisson model, the log of $\mathrm{E}[Y|\mathbf{x}]$ is modeled. Note that modeling $\mathrm{E}[\log(Y)|\mathbf{x}]$, as we might do in a linear model, is different from modeling $\log(\mathrm{E}[Y|\mathbf{x}])$ in the Poisson model. In these cases, the default model is *additive on a multiplicative scale*, as explained in Chapters 6, 7, and 9.

The need to model interaction depends on outcome scale because the simpler additive model can only hold exactly on one such scale, and may be an acceptable approximation on some scales but not others. This is in contrast to confounding; if X_2 confounds X_1, then it does so on every outcome scale. In the case of the linear model, the dependence of interaction on scale means that transformation of the outcome will sometimes succeed in eliminating an interaction.

4.6.7 Example: Hormone Therapy and Baseline LDL

The effect of hormone therapy on LDL cholesterol in the HERS trial was dependent on baseline values of LDL, with larger reductions seen among women with higher baseline values. An interaction model for absolute change in LDL from baseline to the first annual visit is shown in Table 4.14. Note that baseline LDL is centered in this model in order to make the coefficient for hormone therapy (`HT`) easier to interpret. The coefficients in the model have the following interpretations:

- `HT`: Among women with the average baseline LDL level of 135 mg/dL, the effect of HT is to lower LDL an average of 15.5 mg/dL over the first year of the study.
- `cLDL0`: Among women assigned to placebo, each mg/dL increase in baseline LDL is associated with a 0.35 mg/dL greater decrease in LDL over the first year. That is, women with higher baseline LDL

Table 4.14. Interaction Model for HT Effects on Absolute Change in LDL

```
. reg LDLch HT cLDL0 HTcLDL0

      Source |       SS       df       MS              Number of obs =    2597
-------------+------------------------------           F(  3,  2593) =  258.81
       Model |  721218.969     3  240406.323           Prob > F      =  0.0000
    Residual |  2408575.51  2593  928.876015           R-squared     =  0.2304
-------------+------------------------------           Adj R-squared =  0.2295
       Total |  3129794.48  2596  1205.62191           Root MSE      =  30.477

------------------------------------------------------------------------------
       LDLch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HT |  -15.47703   1.196246   -12.94   0.000    -17.82273   -13.13134
       cLDL0 |  -.3477064   .0225169   -15.44   0.000    -.3918593   -.3035534
     HTcLDL0 |  -.0786871   .0316365    -2.49   0.013    -.1407226   -.0166517
       _cons |  -4.888737   .8408392    -5.81   0.000    -6.537522   -3.239953
------------------------------------------------------------------------------
```

experience greater decreases in the absence of treatment; this is in part due to regression to the mean and in part to greater likelihood of starting use of statins.

- `HTcLDL0`: The effect of HT is to lower LDL an additional 0.08 mg/dL for each additional mg/dL in baseline LDL. In short, larger treatment effects are seen among women with higher baseline values. The interaction is statistically significant ($P = 0.013$).

Table 4.15. Interaction Model for HT Effects on Percent Change in LDL

```
. reg LDLpctch HT cLDL0 HTcLDL0

      Source |       SS       df       MS              Number of obs =    2597
-------------+------------------------------           F(  3,  2593) =  165.33
       Model |  233394.163     3  77798.0542           Prob > F      =  0.0000
    Residual |  1220171.82  2593  470.563756           R-squared     =  0.1606
-------------+------------------------------           Adj R-squared =  0.1596
       Total |  1453565.98  2596  559.925263           Root MSE      =  21.692

------------------------------------------------------------------------------
    LDLpctch |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HT |  -10.79035   .8514335   -12.67   0.000    -12.45991   -9.120789
       cLDL0 |  -.2162436   .0160265   -13.49   0.000    -.2476697   -.1848176
     HTcLDL0 |   .0218767   .0225175     0.97   0.331    -.0222773    .0660307
       _cons |  -1.284976   .5984713    -2.15   0.032    -2.458506   -.1114456
------------------------------------------------------------------------------
```

Inasmuch as the reduction in LDL caused by HT appears to be greater in proportion to baseline LDL, it is reasonable to ask whether the HT effect on *percent change* in LDL might be constant across baseline LDL levels. In that case, modeling an interaction between HT and the baseline value would

not be necessary. This turns out to be the case, as shown in Table 4.15. In particular, the interaction term `HTcLDL0` is no longer statistically significantly ($P = 0.331$) and could be dropped from the model. Note that the coefficient for HT now estimates the average *percent* change in LDL due to treatment, among women at the average baseline level. In summary, analyzing percent rather than absolute change in LDL eliminates the interaction between HT and baseline LDL.

4.6.8 Details

There are several other more general points to be made about dealing with interaction in multipredictor regression models.

- Interactions between two multilevel categorical predictors require extra care in coding and interpretation. Simple computation of product terms involving a categorical predictor will almost always give mistaken results. The `xi:` command prefix and `i.` variable prefix in Stata handle this situation, but must be used with care. Furthermore, if one of the predictors has R levels and the other S levels, then the F-test for interaction would have $(R-1)(S-1)$ degrees of freedom. Many different patterns are subsumed by the alternative hypothesis of interaction, only a few of which may be of interest or biologically plausible.
- Interactions between two continuous variables are also tricky, especially if the two predictors are highly correlated. Both main effects in this case are hard to interpret. "Centering" of both variables on their respective sample means (Problem 4.7) resolves the interpretative problem only in part, since the coefficient for each predictor still refers only to the case where the value of other predictor is at its sample mean. Both the linearity of the interaction effect and the need for higher order interactions would need to be checked.
- In examining interactions, it is not enough to show that the predictor of primary interest has a statistically significant association with the outcome in a subgroup, especially when it is not a statistically significant predictor overall. So-called subgroup analysis of this kind can severely inflate the type-I error rate, and has a justifiably bad reputation in the analysis of clinical trials. Showing that the subgroup-specific regression coefficients are statistically different by testing for interaction sets the bar higher, is less prone to type-I error, and thus more persuasive (Brookes *et al.*, 2001).
- Methods have been developed (Gail and Simon, 1985) for assessing *qualitative interaction*, in which the sign of the coefficient for the predictor of interest differs across subgroups. This was nearly the case in the interaction of BMI and statin use. A more specific alternative of this kind is often easier to detect.

- Interaction can be hard to detect if the interacting variables are highly correlated. For example, it would be difficult to assess the interaction between two types of exposure if they occurred together either little or most of the time. This was not the case in the second HERS example, because statin use was reported by 36% of the cohort at baseline, and was uncorrelated with assignment to HT by virtue of randomization. However, in an observational cohort it might be much less common for women to report use of both medications. In that case, oversampling of dual users might be used if the interaction were of sufficient interest.

4.7 Checking Model Assumptions and Fit

In the simple linear model (4.1) as well as the multipredictor linear model (4.2), it has been assumed so far that $\mathrm{E}[y|\mathbf{x}]$ changes linearly with each continuous predictor, and that the error term ε has a normal distribution with mean zero and constant variance for every value of the predictors. We have also implicitly assumed that model results are not unduly driven by any small subset of observations. Violations of these assumptions have the potential to bias regression coefficient estimates and undermine the validity of confidence intervals and P-values.

In this section, we show how to assess the validity of the linearity assumption for continuous predictors and suggest modifications to the model which can make it more reasonable. We also discuss assessments of normality, how to transform the outcome in order to make this assumption approximately hold, and discuss conditions under which it may be relaxed. We then discuss departures from the assumption of constant variance and methods for addressing them. All these procedures rely heavily on the transformations of both predictor and outcome that were introduced in Chapter 2. Finally, we show how to deal with *influential points*. Throughout, we emphasize the *severity* of departures, since model assumptions rarely hold exactly, and small departures are often benign, especially in large data sets. Nonetheless, careful attention to meeting model assumptions can prevent us from being seriously misled, and sometimes increase the efficiency of our analysis into the bargain.

4.7.1 Linearity

In modeling the effect of BMI on LDL, we have assumed that the regression is a straight line. However, this may not be an adequate representation of the true relationship. For example, we might find that average LDL stops increasing, or increases more slowly, among women with BMI in the upper reaches of its range – a *ceiling effect*. Analogously, the inverse relationship between BMI and HDL ("good") cholesterol may depart from linearity, with floor effects among very heavy women.

Component-Plus-Residual (CPR) Plots

In unadjusted analysis, checks for departures from linearity could be carried out using LOWESS, the nonparametric scatterplot smoother introduced in Chapter 2. This smoother approximates the regression line under the weaker assumption that it is smooth but not necessarily linear, with the degree of smoothness under our control, via the bandwidth. If the linear fit were satisfactory, the LOWESS curve would be close to the model regression line; that is, the nonparametric estimate found under the weaker assumption of smoothness would agree with the estimate found when linearity is assumed.

However, the direct approach of adding a LOWESS smooth to a scatterplot of predictor versus outcome is only effective for simple linear models with a single continuous predictor. For multipredictor regression models the analogous plot would have to accommodate $p + 1$ dimensions, where p is the number of predictors in the model – hard to imagine even for $p = 2$. Moreover, nonparametric smoothers work less well in higher dimensions.

Fortunately, the residuals from a regression model make it possible to examine the linearity of the adjusted association between a given predictor and the outcome, after taking account of the other predictors in the model. The basic idea is to plot the residuals versus each continuous predictor in the model; then a nonparametric smoother is used to detect departures from a linear trend in the average value of the residuals across the values of the predictor. This is a *residual versus predictor* (RVP) plot, obtained in Stata using the `rvpplot` command. However, for doing this check in Stata, we recommend the closely related *component plus residual* (CPR) plot, mainly because the `cprplot` command allows LOWESS smooths, which we find more informative and easier to control than the smooths available with `rvpplot`.

Fig. 4.4 shows CPR plots for multipredictor regression models for LDL and HDL, each adjusting the estimated effect of BMI for age, ethnicity, smoking, and alcohol use. If the linear fits for BMI were satisfactory, then there would be no nonlinear pattern across values of BMI in the component-plus-residuals. For LDL, shown on the left, the linear and LOWESS fits agree quite well, but for HDL, there is a substantial divergence. Thus the linearity assumption is rather clearly met by BMI in the model for LDL, but not in the model for HDL. The curvature in the relationship between BMI and HDL can be approximated by adding a quadratic term in BMI to the multipredictor linear model. The augmented model is then

$$\begin{aligned} \mathrm{E}[\mathtt{HDL}|\mathbf{x}] = \beta_0 + \beta_1 \mathtt{BMI} + \beta_2 \mathtt{BMI2} \\ + \beta_3 \mathtt{age} + \beta_4 \mathtt{nonwhite} + \beta_5 \mathtt{smoking} + \beta_6 \mathtt{drinkany}, \end{aligned} \quad (4.31)$$

where `BMI2` is the square of `BMI`. A CPR plot for the relationship between BMI and HDL in this model is shown in Fig. 4.5. Except at the extremes of the range of `BMI`, where the LOWESS smooth would usually be unreliable, the quadratic fit is clearly an improvement on the simpler model. Moreover,

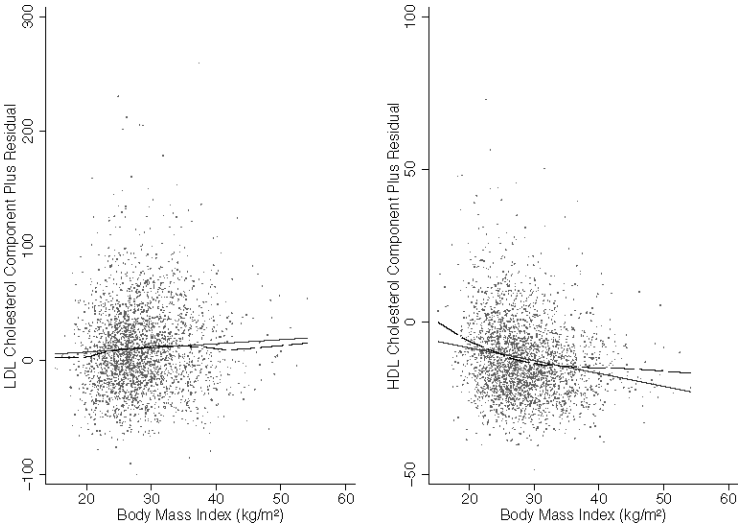

Fig. 4.4. CPR Plots for Multiple Regressions of LDL and HDL on BMI

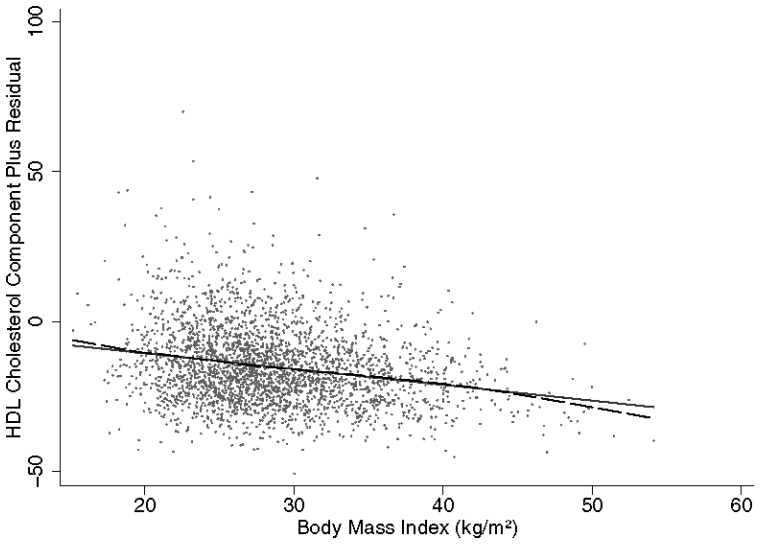

Fig. 4.5. CPR Plot for HDL Model with Quadratic Term in BMI

both the linear and quadratic terms in `BMI` are statistically significant (both $P < 0.0005$), and R^2 increases from 0.074 to 0.081, a gain of 9%.

Smooth Transformations of the Predictors

In the example of HDL and BMI, the departure from linearity was approximately addressed by adding a quadratic term in `BMI` to the model. This solution is often useful when the regression line estimated by the LOWESS smooth is convex or concave, and especially if the line becomes steeper at either side of the CPR plot.

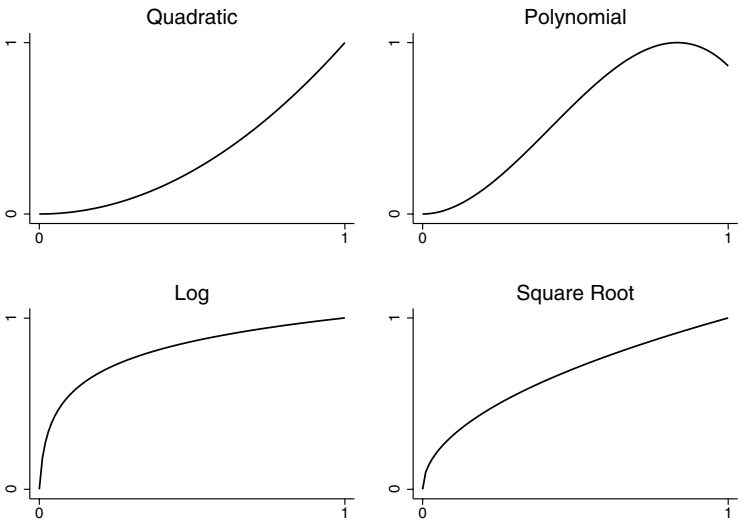

Fig. 4.6. Linearizing Predictor Transformations

However, other transformations of the predictor may sometimes be more successful and should be considered. Fig. 4.6 shows some of the predictor transformations commonly used to linearize the association between the predictor and the outcome. The upper left panel shows the typical curvature captured by adding a quadratic term in the predictor to the model. On the upper right, both quadratic and cubic terms have been included; in general such higher order polynomial tranformations are useful for S-shapes. A drawback is that these lines often fit badly in the tails of the predictor distribution if the data there are sparse. The lower panels show the log and square root transformations, which are useful in situations where the regression line increases more slowly with increasing values of the predictor, as we might expect in cases of floor or ceiling effects, and more generally where the slope becomes

less steep. In Sect. 4.7.5 below, we discuss interpretation of the regression coefficients for a log-transformed predictor. Each of these transformations would work just as well for modeling the mirror image of the nonlinear shape, reversed top-to-bottom.

Categorizing the Predictor

Another transformation useful in exploratory analysis is to categorize the continuous predictor, either at cutpoints selected *a priori* or at percentiles that ensure adequate representation in each category. Then the model is estimated using indicators for all but the reference category of the transformed predictor, as in the `physact` example in Sect. 4.3. Clearly the transformed variable is ordinal in this case. This method models the association between the ordinal categories and the outcome as a *step function* (Fig. 4.7). Although this approach is unrealistic in not providing a smooth estimate of the regression line, and also less efficient, it has the advantage of flexibility, in that each step can be of any height. Such transformations are also easy to understand, especially when the categories are defined by familiar clinical cutpoints. In contrast, smooth transformations, in particular polynomials, are harder to motivate, present, and interpret.

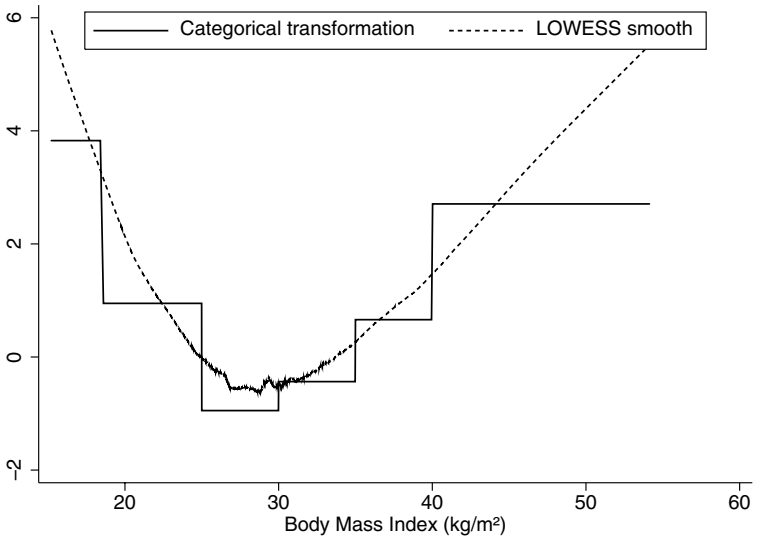

Fig. 4.7. Categorical Transformation of BMI

A final note: while diagnostics for nonlinearity using RVP and CPR plots do not carry over to the logistic, Cox, and generalized linear models presented

in later chapters, departures from linearity can be addressed using quadratic terms as well as smooth and categorical transformations in all of these settings.

Evaluation

The choice of transformation will in a few cases be suggested by an understanding of mechanism or to make results more interpretable, but more often it will be made on the basis of what appears to fit best. Comparison of the LOWESS smooth in CPR plots with the transformations in Fig. 4.6 can help identify the best candidate transformations. After the revised model is estimated, repeating the diagnostic using a new CPR plot then provides an initial check on the adequacy of the transformation: there should be no remaining pattern in the residuals, and the smooth should be close to the linear fit. In cases where a quadratic or quadratic plus cubic term is added to the model, we can use t- or F-tests to evaluate the statistical significance of the addition to the model. This works because the original model is "nested" in the final model, in the sense that the predictors in the smaller model are a subset of those in the larger model. In other cases, for example, when we substitute the log-transformed for the untransformed predictor, the original and final models are not nested, so this testing procedure does not apply, although alternatives are available (Vuong, 1989). In both cases, however, we can check whether R^2 improves substantially with the transformation.

4.7.2 Normality

In Sect. 4.1 we stated that in the multipredictor linear model, the error term ε is assumed to have a normal distribution. Confidence intervals for regression coefficients and related hypothesis tests are based on the assumption that the coefficient estimates have a normal distribution. If ε has a normal distribution, and other assumptions of the multipredictor linear model are met, then ordinary least squares estimates of the regression coefficients can be shown to have a normal distribution, as required.

However, it can be shown that the regression coefficients are approximately normal in larger samples even if ε does not have a normal distribution. In that case, characterizing the distribution of the residuals is helpful for assessing whether the sample is large enough to trust the confidence intervals and hypothesis tests, since larger samples are required for this approximation to hold when departures from the normality of the errors are relatively serious. As with the t-test reviewed in Sect. 3.1, outliers are the principal worry with such departures, with the potential to erode the power of the model to detect real effects.

Residual Plots

Various graphical methods introduced in Chapter 2 are useful for assessing the normality of ε. In using these tools, it is important to distinguish between the distribution of the outcome y and the distribution of the residuals, which are the sample analogue of ε. The point here is that the residuals may be normally distributed when y is not, and conversely. Since our assumptions concern the distribution of ε, it is important to apply the diagnostic tools to the residuals rather than to the outcome variable itself.

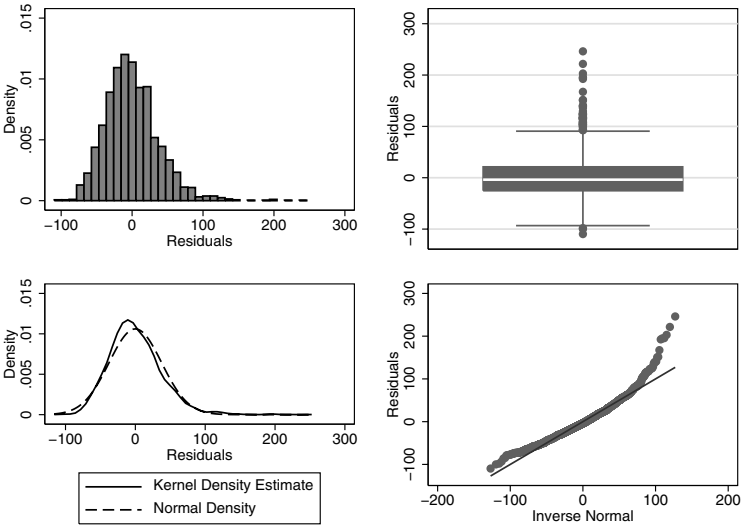

Fig. 4.8. Residuals With Untransformed LDL

Fig. 4.8 shows four useful graphical tools for assessing the normality of the residuals, in this case from our multipredictor regression model for LDL. In the upper panels the histogram and boxplot both suggest a somewhat long tail on the right. The lower left panel presents a nonparametric estimate of the distribution of the residuals obtained using the `kdensity, normal` command in Stata. For comparison, the solid line in that panel shows the normal distribution with the same mean and standard deviation. Comparing these two curves suggests some skewing to the right, with a long right and short left tail; but overall the shapes are quite close. Finally, as explained in Chapter 2, the upward curvature of the normal quantile-quantile (Q-Q) plot on the lower right is also diagnostic of right-skewness.

Interpretation of the results shown in Fig. 4.8 depends on the sample size. With 2,763 observations, there is little reason for concern about the moderate

right-skewness. Given such a large data set, the distribution of the parameter estimates is likely to be well approximated by the normal, despite the mild departure from normality in the residuals. However, in a small data set, say, with 50 or fewer observations, the long right tail might be reason for concern, in part because it could make parameter estimates less precise and tests less powerful.

Testing for Departures From Normality

Various statistical tests are available for assessing the normality of the residuals, but have the drawback of being sensitive to sample size, often failing to reject the null hypothesis of normality in small samples where meeting this assumption is most important, and conversely rejecting it even for small violations in large data sets where inferences are relatively robust to departures from normality. For this reason, we do not recommend use of these tests; instead, the graphical methods just described should be used to judge the potential seriousness of the violation in the light of the sample size.

Log, Power, and Other Transformations of the Outcome

Transforming the outcome is often successful for reducing the skewness of residuals. The rationale is that the more extreme values of the outcome are usually the ones with large residuals (defined as $r_i = y_i - \hat{y}_i$); if we can "pull in" the outcome values in the tail of the distribution toward the center, then the corresponding residuals are likely to be smaller too.

One such transformation is to replace the outcome y with $\log(y)$. A constant can be added to an outcome variable with negative or zero values, so that all values are positive, though this may complicate interpretation. The log tranformation is now conventionally used to analyze viral load in studies of HIV and hepatitis infections, triglyceride levels in studies of cardiovascular disease, and in many other contexts. Fig. 4.9 shows that after log transformation of LDL, there is no more evidence of right-skewness; in fact there is slight evidence of too long a tail on the left. It should also be noted that there is no qualitative change in inferences for BMI. In Sect. 4.7.5 below, we discuss interpretation of regression coefficients in models where the outcome is log-transformed.

Power transformations are a flexible alternative to the log transformation. In this case, y is replaced by y^k. Smaller values of k "pull in" the right tail more strongly. As an example, square ($k = 1/2$) and cube ($k = 1/3$) root transformations were commonly used in analyzing CD4 lymphocyte counts in studies of HIV infection, since the distribution is very long-tailed on the right. Adding a constant so that all values of the outcome are non-negative will sometimes be necessary in this case too. The `ladder` command in Stata systematically searches for the power transformation of the outcome which is closest to normality.

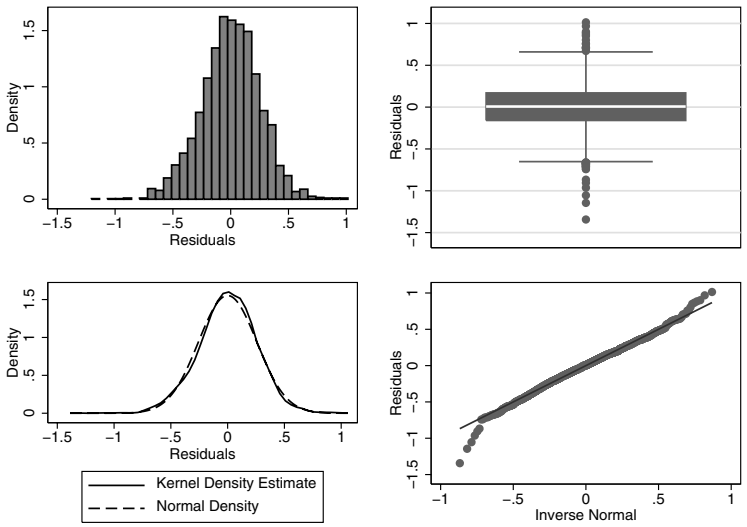

Fig. 4.9. Residuals With Log-Transformed LDL

A more difficult problem arises if both tails of the distribution of the residuals are too long, since neither log nor fractional power transformations will fix both tails. In this case one solution is the rank transformation, in which each outcome is replaced by its rank in the ordering of all the outcomes, as in the computation of the Spearman correlation coefficient (Sect. 3.2); this does not achieve normality but may reduce the loss of power. Another possibility is trimming the tails; for example, "Winsorizing" the outcome involves replacing outcome values more than 2 or 3 standard deviations from the average by that limiting value.

Generalized Linear Models (GLMs)

Some outcome variables cannot be satisfactorily transformed, or there may be compelling reasons to analyze them on the original scale. A good alternative is provided by the generalized linear models (GLMs) discussed in Chapter 9. A less efficient alternative is to dichotomize the outcome and analyze it using logistic models; alternatively, more than two outcome categories can be analyzed using proportional-odds or continuation-ratio models (Ananth and Kleinbaum, 1997; Greenland, 1994), as briefly described in Chapter 6.

4.7.3 Constant Variance

An additional assumption concerning ε is *homoscedasticity*, meaning that its variance σ_ε^2 is constant across observations. When this assumption is violated,

the validity of confidence intervals and P-values can be affected. In particular, between-group contrasts can be misleading if σ_ε^2 differs substantially across the subgroups being compared, especially if the subgroups differ in size. Furthermore, in contrast to violations of the assumption that the residuals are normally distributed, heteroscedasticity is no less a problem in large samples than in small ones. Finally, while violations do not make the coefficient biased, some precision can be lost.

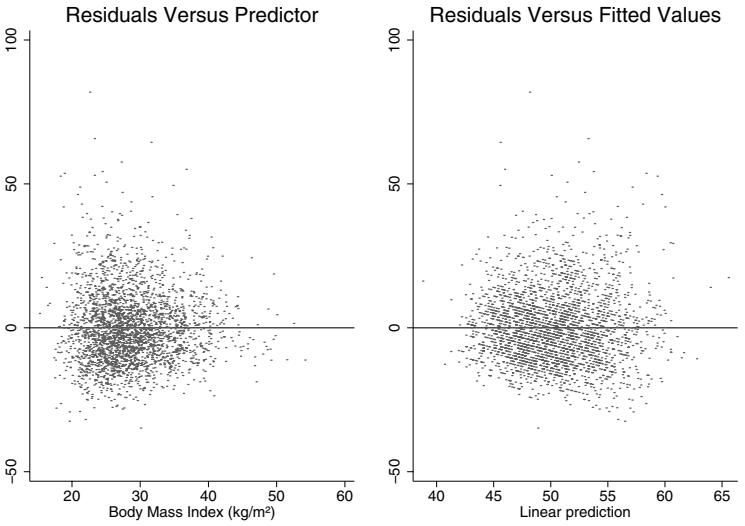

Fig. 4.10. Checking for Constant Residual Variance

Residual Plots

Diagnostics for violations of the constant variance assumption also use the residual versus predictor (RVP) plots used to check linearity of response to continuous predictors, as well as analogously defined residual versus fitted (RVF) plots. If the constant variance assumption is met, then the vertical spread of the residuals should be similar across the ranges of the predictors and fitted values; in contrast, heteroscedasticity is signaled by horizontal funnel shapes. Since the residuals of the LDL analysis gave no evidence of trouble, we examined the residuals from the companion model for HDL, which was shown in Sect. 4.7.1 to need a quadratic term in BMI to meet the linearity assumption.

Fig. 4.10 shows scatterplots of the residuals of the regression of HDL on BMI and its square, as well as age, ethnicity, smoking, and alcohol use. The

plot against BMI shows somewhat wider range on the left, although this may partly be due to the fact that there are more observations on the left, and so more likely a few large residuals purely by chance. This evidence for non-constant variance is mirrored in the slightly wider spread on the right in the facing plot of the residuals against the fitted values.

Sub-Sample Variances

Constancy of variance across levels of categorical predictor can be checked by comparing the sample variance of the residuals for each category. In this example, the variance was essentially identical across groups defined by ethnicity, smoking, and alcohol use.

In contrast, in our analysis of the influence of exercise on glucose levels in Sect. 4.1, violation of the assumption of constant variance was one of several motivations for excluding women with diabetes. If they had been included, the variance of the residuals would have varied between this group of 734 women and the remainder of the HERS cohort by a factor of 26 (2,332 vs. 90). Even after log transformation of glucose, the variance would still have differed by a factor of 10 (0.097 vs. 0.0094). This pattern reflects the fact that diabetes is characterized by loss of control over glucose levels, and also variation in the use of medications that control them. These large differentials in residual variance would call into question inferences drawn from comparisons between women with and without diabetes.

Testing for Departures From Constant Variance

Statistical methods available for testing the assumption of homoscedasticity share the sensitivity to sample size described earlier for tests of normality. The resulting potential for giving false reassurance in small samples leads us to recommend against the use of these formal tests. Instead, we need to examine the severity of the violation.

When Departures May Cause Trouble

Violations of the assumption of constant variance should be addressed in cases where the variance of the residuals–

- changes by a factor of 2 or more across the range of the fitted values or a continuous predictor, judging from the LOWESS smooth of the squared residuals;
- differs by a factor of 2 or more between subgroups that differ in size by a factor of 2 or more;
- differs by a factor of 3 or more between subgroups that differ in size by a factor of less than 2.

Note that smaller differences in the *standard deviation* of the residuals would give reason for transformation.

Variance-Stabilizing Outcome Transformations

In simple cases where multiple predictors do not need to be taken into account, we could use t-tests with the `unequal` option to compare subgroups, allowing for the unequal variances. However, multipredictor modeling is often crucial; furthermore, use of a t-test with unequal variances would not address smooth dependence of σ_ε^2 either on $\mathrm{E}[y|\mathbf{x}]$ or on a continuous predictor. In that case, non-constant variance can sometimes be addressed using a *variance-stabilizing* transformation of the outcome, including the log and square root transformations. As shown in Fig. 4.11, log transformation of HDL reduces, though it does not completely eliminate, the evidence for non-constant variance we found in Fig. 4.10. However, in this case our qualitative conclusions would be unchanged by log transformation of HDL.

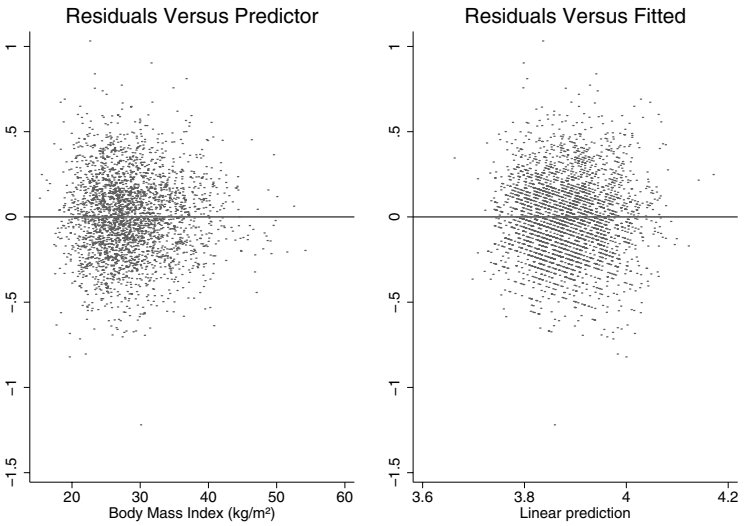

Fig. 4.11. Rechecking Constant Variance After Log-Transforming HDL

GLMs

The square root transformation has been widely used to stabilize the variance of counts. However, this has now been largely supplanted by GLMs such as the Poisson and negative binomial regression models (Chap. 9). As in other GLMs, including the logistic model (Chap. 6), the variance of the Poisson outcome is modeled as a function of its mean. In particular, this would potentially be useful in cases where a LOWESS smooth of the squared residuals, an

alternative diagnostic for heteroscedasticity, increased in proportion to the fitted values. GLMs represent the primary alternative when transformation of the outcome fails to rectify substantial violations of the assumption of constant variance.

4.7.4 Outlying, High Leverage, and Influential Points

We have already pointed out that outlying observations with relatively large residuals can cause trouble, in part by inflating the variance of coefficient estimates, making it harder to detect statistically significant effects. In this section we consider *high-leverage* points, which could be described as x-outliers, since they tend to have extreme values of one or more predictors, or represent an unusual combination of predictor values. The importance of high-leverage points is that they are also potentially *influential*, in the sense that one or more of the coefficient estimates would change by an unduly large amount if the influential points were omitted from the data set. This can happen when a high-leverage point also has a large residual.

> Definitions: *High leverage points* are x-outliers with the potential to exert undue influence on regression coefficient estimates. *Influential points* are points that have exerted undue influence on the regression coefficient estimates.

Ultimately, our concern is that changes in coefficient estimates resulting from the omission of one or a few influential points could qualitatively affect the conclusions drawn from the analysis. This could arise if associations that were clearly statistically significant become clearly non-significant, or vice versa, including interaction and quadratic terms, or if associations change substantially in magnitude or direction. We would have good reason to mistrust substantive conclusions that were dependent on a few observations in this way. Similarly, in regression models oriented to prediction of future outcomes (Sect. 5.2), prediction error might be substantially affected.

Outlying, high leverage, and influential points are illustrated in Fig. 4.12. In all three of these small samples ($n = 26$), a problematic data point, marked with an X, is included. The solid and dashed lines in each plot show the regression lines estimated with and without the point, as a graphical measure of influence. The sample shown on the upper left includes an outlier with a very large positive residual. However, the leverage of the outlier is minimal, because it is in the center of the distribution of x. Accordingly, the slope estimate is unaffected by omission of this data point, Note that the point is influential for the intercept estimate, but this parameter may be of less direct interest.

In the upper right panel, the point at the extreme right has high leverage, but because this data point is fairly consistent with the prediction based on the other 25 data points, its influence is limited, and the estimated slope and

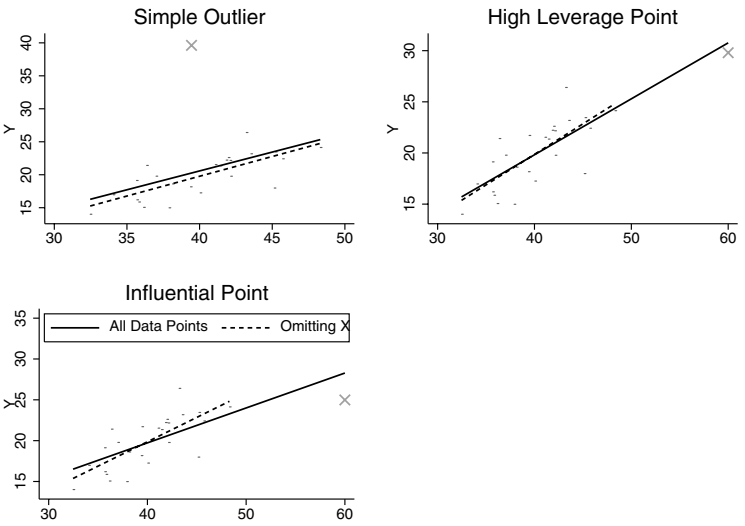

Fig. 4.12. Outlying, High-Leverage, and Influential Points

its statistical significance are almost unchanged by by omission of the high-leverage point. Certainly our qualitative interpretation of the slope would be unaffected.

In contrast, the point at the extreme right in the lower left panel has the same leverage as the point in the upper right panel, but in this case its influence is very strong, moving the slope estimate by more than 2 standard errors. The slope remains positive and statistically significant in this instance, so our qualitative interpretation would be similar, but in some circumstances omission of such a data point could make a non-significant result highly statistically significant, or vice versa. In part this reflects the small sample size, since a high leverage point is has a better chance of outweighing a relatively small number of other observations.

DFBETAs

To check for sensitivity of the conclusions of an analyis to a small number of high-leverage observations, we first need to identify potentially influential points. Of the various statistics for quantifying influence that have been defined, we recommend using DFBETA statistics, which quantify how much each of the coefficients would change if each observation were omitted from the data set. In linear regression, these statistics are exact; for logistic and Cox models, accurate approximations are available. DFBETA statistics are in standard error units – effectively on the same scale as the t-statistic, which is equal to $\hat{\beta}$ divided by its standard error. If the analysis is focused on one

predictor of primary interest, then clearly the DFBETAs for that predictor are of central concern.

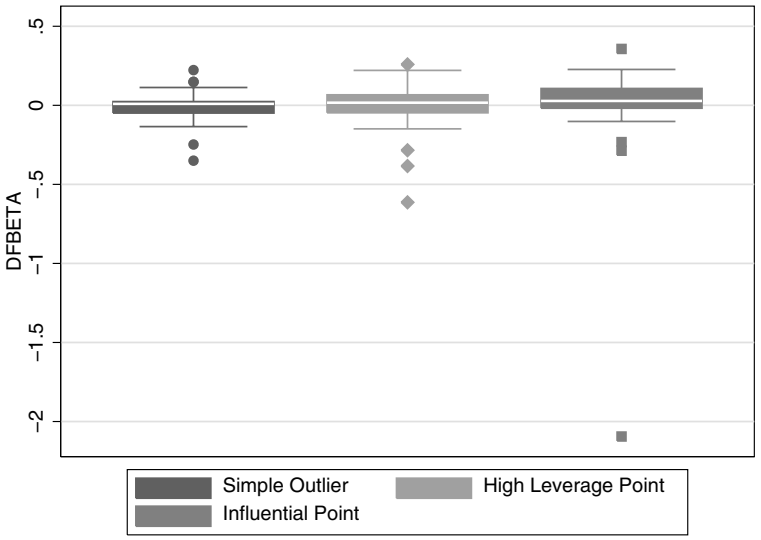

Fig. 4.13. DFBETAs for Data Sets Shown in Fig. 4.12

Boxplots are convenient for identifying a small set of extreme outliers among the DFBETA values for each predictor. DFBETAs often have a very small inter-quartile range, so that a substantial set of observations may lie beyond the whiskers of the plot. Thus we need to look for a small number of extreme values that are set off from the rest. Fig. 4.13 shows boxplots of the DFBETA statistics for the single predictor in the three data sets shown in Fig. 4.12. These plots clearly indicate the single influential point.

If a small set of observations meeting diagnostic criteria for undue influence is identified, the accuracy of those data points should first be checked and clearly erroneous observations corrected, or if this is impossible, deleted. Then if any of the apparently influential points are retained, a final step is sensitivity analyses in which the final model is rerun omitting some or all of the retained influential points. For example, suppose we have identified ten influential points that are not due to data errors, and that these include two observations with absolute DFBETAs greater than 2, three observations with values between 1 and 2, and five more with values between 0.5 and 1. Then a convenient *ad hoc* procedure would be to delete the two worst observations, then the worst five, and finally all ten potentially influential points. In each model, we would check whether the important conclusions of the analysis were affected. In prediction models, sensitivity would be assessed in terms of esti-

mated prediction error (Sect. 5.2). In summary, we emphasize the underlying theme of sensitivity to the omission of a *small* number of points, relative to sample size; if we omit 10% or 20% of the data and the conclusions change, this would probably not indicate undue sensitivity.

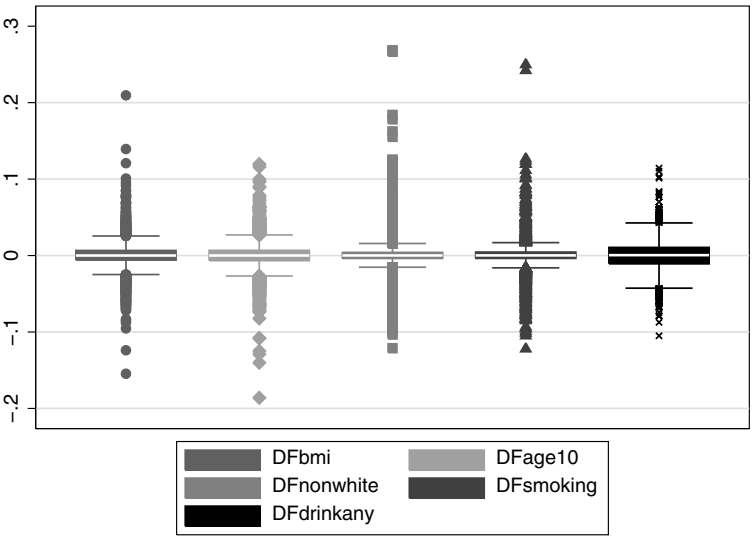

Fig. 4.14. DFBETAs for LDL Model

Fig. 4.14 below shows boxplots of DFBETAs for the multiple regression of LDL on BMI, age, ethnicity, smoking, and alcohol use. As compared to the clearly influential point shown in Fig. 4.13, the largest DFBETAs are much less extreme. Examination of the four observations with DFBETAs > 0.2 identified women with high LDL values (between 346 and 393 mg/dL).

Table 4.16. Sensitivity of LDL Model to Omission of Four Most Influential Points

Predictor	All observations			Omitting four observations		
variable	$\hat{\beta}$	95% CI	P-Value	$\hat{\beta}$	95% CI	P-Value
BMI	0.36	0.10, 0.62	0.007	0.34	0.08, 0.60	0.010
Age	–1.89	–4.11, 0.32	0.090	–1.86	–4.03, 0.31	0.090
Nonwhite	5.22	0.66, 9.78	0.025	4.19	–0.27, 8.66	0.066
Smoking	4.75	0.42, 9.08	0.032	3.78	–0.47, 8.03	0.072
Alcohol Use	–2.72	–5.66, 0.22	0.069	–2.64	–5.51, 0.23	0.072

The sensitivity of model results to the omission of these four points is summarized in Table 4.16.The changes are mostly minor, in particular for BMI, the predictor of primary interest. The P-values for ethnicity and smoking shift from nominally statistically significant to borderline significant, but these are not variables of primary interest and in any case our conclusions should not be unduly influenced by small shifts of this kind.

A potential weakness of these procedures is that DFBETAs capture the influence of omitting one observation at a time, but do not tell us how the omission of various *sets* of points, some of which may have small DFBETAs, will affect our conclusions. Unfortunately, user-friendly diagnostics for checking sensitivity to omission of sets of observations have not been developed, in part because the computational burden is too great.

Addressing Influential Points

If substantive conclusions are qualitatively affected by omission of influential points in the sensitivity analysis, *this should be reported*. In addition, it is often worthwhile to consider in substantive terms why these points have high leverage and are influential. For example, the WCGS data include an influential point with an extreme but accurately recorded cholesterol level of 645 mg/dL, which resulted from familial hypercholesterolemia, a rare condition. For research questions concerning the effects of cholesterol levels in the usual range determined by common risk factors, it would be reasonable to delete this point. But in many circumstances, deletion of influential points is hard to justify persuasively.

In that case, it may also be worth considering a more complex model that better accommodates the influential points. In Fig. 4.12, for example, a quadratic term would almost certainly reduce the influence of the observation causing trouble. Alternatively, interaction terms might accommodate influential data points characterized by an unusual combination of two predictor values. Nonetheless, changing the model in such a substantial way to accommodate one or a few data points should be undertaken with caution, with attention to the plausibility of the modified model, and the results clearly presented as data-driven, sensitive to influential points, and hypothesis-generating.

4.7.5 Interpretation of Results for Log-Transformed Variables

In Sect. 4.7 we discussed log-transforming predictors to achieve linearity, and proposed log transformation of the outcome as a means of normalizing the residuals or stabilizing their variance. Even if substantive interpretation and P-values are often not much changed, these transformations have a substantial effect on the estimated regression coefficients and their literal interpretation.

For both predictors and outcomes, log transformation changes the focus from absolute to relative or percentage change. Recall that for a predictor and outcome on their measured scale, the regression coefficient is interpretable as

the change in the average value of the outcome for every unit increase in the predictor; for both predictor and outcome, we mean change on the measured, or absolute, scale.

Log Transformation of the Predictor

First consider log transformation of the predictor. In this case, the regression coefficient multiplied by log(1.01) can be interpreted as the change in the average value of the outcome for every 1% increase in the predictor. This is valid whether we use the natural log or logarithms with other bases. In a linear model using the natural log (ln) transformation of weight to predict systolic blood pressure (SBP), the estimated coefficient for ln weight is 3.004517. Thus we estimate that average SBP increases $3.004517 \times \ln(1.01) \approx 0.03$ mmHg for each 1% increase in weight. Similarly, if we multiply $\hat{\beta}$ by ln(1.05) or ln(1.1) we obtain the estimates that average SBP increases 0.15 mmHg for each 5% increase in weight and 0.29 mmHg for each 10% increase.

Within limits, we can approximate these results without using a calculator. Specifically, if the predictor is natural log-transformed, we can estimate the increase in the average value of the outcome per 1% increase in the predictor simply by $\hat{\beta}/100$. This follows because $\ln(1.01) \approx 0.01$. But this shortcut is not valid for logarithms with other bases, and analogous calculations for larger percentage increases in the predictor get progressively less accurate and should not be attempted by this means.

Log Transformation of the Outcome

Similarly, with natural log transformation of the outcome, $100(e^{\hat{\beta}} - 1)$ is interpretable as the *percentage* increase in the average value of the outcome per unit increase in the predictor. If base-10 logs were used to transform the outcome, then $100(10^{\hat{\beta}} - 1)$ has this interpetation. The coefficient for BMI in a linear model for the natural log transformation of triglyceride (TGL) is 0.0133487, so the model predicts a $100(e^{0.0133487} - 1) = 1.34\%$ increase in TGL per unit increase in BMI.

Again, we can approximate these results without a calculator under some circumstances. When the outcome is natural log-transformed, we can approximate the percentage change in the average value of the outcome per unit increase in the predictor by $100\hat{\beta}$. But this is acceptably accurate only if $\hat{\beta}$ is smaller than 0.1 in absolute value, and is again not valid using log transformations with other bases.

Log Transformation of Both Predictor and Outcome

If both predictor and outcome are transformed using natural logs, then $100(e^{\hat{\beta}\ln(1.01)} - 1)$ can be interpreted as the percentage increase in the average value of the outcome per 1% increase in the predictor. With the $\log_{10}$

transformation, $100(10^{\hat{\beta}\log_{10}(1.01)} - 1)$ has this interpretation. In this case, the back-of-the-envelope approximation for the percent increase in outcome for each 1% increase in the predictor is simply $\hat{\beta}$; this is accurate if both predictor and outcome are natural log-transformed and $\hat{\beta}$ is smaller than 0.1 in absolute value.

4.7.6 When to Use Transformations

Our graphical diagnostics for linearity, normality, and constant variance do not provide clearcut decision rules analogous to $P < 0.05$, and we do not recommend formal statistical tests in this context. Furthermore, addressing these violations will in many cases involve using transformations of predictors or outcomes that may make the results harder to interpret. A natural criterion for assessing the necessity for transformation is whether important substantive results differ qualitatively before and after transformation. If not, it may be reasonable not to use the transformations. Our example using BMI and diabetes to predict HDL is probably a case in point: while log transformation of HDL corrected departures from both normality and constant variance, the conclusions were unchanged. But if substantial differences do arise, then using transformed variables to meet model assumptions more closely helps us to avoid misleading results.

4.8 Summary

The multipredictor linear model is a straightforward extension of the simple linear model for continuous outcomes. Inclusion of multiple predictors in the model makes it possible to adjust for confounding variables, examine mediation, check for and model interactions, and increase efficiency, especially in experiments, by accounting for design factors. It is important to check the assumptions of the linear model and to use transformations of predictor and outcome variables as necessary to meet them more closely, especially in small samples. It is also important to recognize common data types where linear regression is not appropriate; these include binary, time-to-event, count, and repeated measures or clustered outcomes, and are addressed in subsequent chapters.

4.9 Further Notes and References

For more detailed information on the linear regression model, first-rate books include Weisberg (1985) and Draper and Smith (1981). Jewell (2004), in particular Chapter 8, gives an excellent introduction – to which we are indebted – to issues of causality in observational studies; Rothman and Greenland (1998)

also address these issues in some detail. A cutting-edge book in this area, unfortunately of considerable difficulty, is van der Laan and Robins (2003). A standard book on regression diagnostics is Belsey *et al.* (1980), while Cleveland (1985) covers graphical methods for model checking in detail. See Breiman (2001) for a skeptical view of the sensitivity of the methods presented here for detecting lack of fit.

Splines and Generalized Additive Models

The Stata package implements a convenient and often more biologically plausible alternative to the categorical transformations presented in Sect. 4.7.1 for addressing departures from linearity, called the *linear spline*. We again specify cutpoints, usually called *knots* in this context. The resulting fitted regression line is continuous at each of the knots and linear in the intervals between them. The `mkspline` command in Stata can be used to set up the transformed predictor variables, one for each interval defined by the cutpoints. As with the categorical transformation, selection of the knots is a non-trivial problem.

Methods have also been developed for fitting linear as well as logistic (Chap. 6) and other generalized linear models (Chap. 9) in which the adjusted response to each predictor can be flexibly modeled as a smooth (piecewise cubic rather than piecewise linear) spline, or alternatively using a LOWESS curve. In both cases the degree of smoothness is under the control of the analyst. Known as *generalized additive models* (Hastie and Tibshirani, 1986, 1999), implementations in the R statistical package make it easy to model and test the statistical significance of departures from linearity. Implementations in R of smooth spline transformations of predictors are also available for the Cox model, discussed in Chapter 7.

4.10 Problems

Problem 4.1. Using the Western Collaborative Group Study (WCGS) data for middle-aged men at risk for heart disease, fit a multipredictor model for total cholesterol (`chol`) that includes the binary predictor `arcus`, which is coded 1 for the group with *arcus senilis*, a milky ring in the iris associated with high cholesterol levels, and 0 for the reference group. Save the fitted values. Now refit the model with the code for the reference group changed to 2. Compare the coefficients, standard errors, P-values, and fitted values from the two models. The WCGS data are available at `http://www.biostat.ucsf.edu/vgsm`.

Problem 4.2. Using (4.2), show that β_j gives the difference in $\mathrm{E}[y|\mathbf{x}]$ for a one-unit increase in x_j, no matter what the values of x_j or the other predictors. *Hint:* Write the value of (4.2) for $x_j = x$ and then for $x_j = x+1$, for arbitrary (unspecified) values of the other predictors, all of which are held fixed, and subtract the first value from the second.

Problem 4.3. Using the WCGS data referenced in Problem 4.1, extract the fitted values from the multipredictor linear regression model for cholesterol and show that the square of the sample correlation between the fitted values and the outcome variable is equal to R^2. In Stata the following code saves the predicted values from the regression model in Table 4.2 to a new variable `yhat`:

```
. reg glucose exercise BMI smoking drinkany
. predict yhat
```

Then use the `pwcorr` and `display` commands to get the correlation between `yhat` and the predictor and square it.

Problem 4.4. Give an alternative coding for the unadjusted model predicting glucose from the five-level physical activity variable in which no intercept parameter is included in the model. In this case, there is no reference group, and all five group-specific indicators are included in the model. What is the interpretation of the βs in this model? How could the Stata `lincom` command be used to compare groups?

Problem 4.5. Use the `test` command in Stata or an equivalent command in another statistical package to show that $F = t^2$ for a pairwise contrast between any other level of a categorical predictor and the reference group used in the model.

Problem 4.6. In the model including an interaction between BMI and statin use, define a second new BMI variable so that estimates for BMI specific to women who do and do not use statins can be obtained directly from the regression coefficients, rather than having to compute sums of the coefficients for one of these groups. Define the values of the new BMI variable in the two groups, and then write down the regression equations analogous to (4.28), (4.29), and (4.30). Explain why the statin use variable needs to be included in this model.

Problem 4.7. If we "center" `age` – that is, replace it with a new variable defined as the deviation in age from the sample mean, what would be the interpretation of the intercept in the model for SBP (3.2)? If `BMI` had *not* been centered, how would the interpretation of the statin use variable change in the model in Sect. 4.6.5 allowing for interaction in predicting LDL?

Problem 4.8. Consider the associations between exercise and glucose levels among women without diabetes. What are the interpretations of the coefficient for exercise–

- in a simple linear model for glucose levels
- in a multipredictor linear regression model for glucose adjusting for all known confounders of the exercise association

Suppose factor X had been identified as a mediator of the exercise/glucose association. What would be the interpretation of the exercise coefficient in a multipredictor regression model that also adjusted for factor X, supposing that the exercise coefficient remained statistically significantly different from zero?

Problem 4.9. Suppose that in a clinical trial of the effects of a new treatment on glucose levels, the randomization is stratified on diabetes, an important predictor of this outcome. By virtue of randomization, the treatment is uncorrelated with diabetes. Using (4.4), explain why including diabetes in the analysis should provide a more efficient estimate of the treatment effect. Would it be a good idea to check for interaction between treatment and diabetes in this analysis? Why?

Problem 4.10. Using Stata (or another statistical package) and the WCGS data set referenced above in Problem 4.1 (or your own data set), verify that you get equivalent results from

- a t-test and a simple linear model with one binary predictor
- one-way ANOVA and a linear model with one multilevel categorical predictor.

Problem 4.11. What is the difference between showing that an interaction is statistically significant and showing that an association is statistically significant in one group but not in the other? Describe a pattern where the second condition holds but there would clearly be no interaction. Is that pattern of clinical interest?

Problem 4.12. Consider a predictor of interest for an important outcome in your field of expertise. Are there other predictors that might be hypothesized *a priori* to interact with the predictor of interest? Why?

Problem 4.13. Suppose a quadratic term in BMI is added to the model for HDL to rectify the departure from linearity and improve fit. How would you summarize this more complex association in presentations or a paper?

Problem 4.14. Consider a right-skewed outcome variable that could be adequately normalized using an unfamiliar fractional power transformation (say, the cube root). A simpler alternative is just to dichotomize the variable. Why would you expect this to be a costly choice in terms of efficiency? Now consider birth weights. Why might analysis of an indicator of low birth weight be worth the loss of efficiency in this case?

Problem 4.15. Suppose you fit a model with an influential point. With the point, the association of interest is just statistically significant, and without it, it is clearly not. What would you do?

4.11 Learning Objectives

1. Describe situations in which multipredictor analysis is needed. Given an analysis situation, decide if linear regression is appropriate.
2. Translate research questions appropriate for a regression model into specific questions about model parameters.
3. Use linear regression models to test hypotheses about relationships between variables, including confounding, mediation, and interaction.
4. Describe the linear regression model, its key assumptions, and their implications.
5. Explain why the estimates are called least squares estimates.
6. Define regression line, fitted value, residual, and influence.
7. State the relationships between
    - correlation and regression coefficients
    - the two-sample t-test and a regression model with one binary predictor
    - ANOVA and a regression model with categorical predictors.
8. Know how a statistical package is used to estimate the parameters in a regression model and make diagnostic plots to assess how well model assumptions are met.
9. Interpret regression model output output including regression parameter estimates, hypothesis tests, confidence intervals, and statistics which quantify the fit of the model.
10. Interpret regression coefficients when the predictor, outcome, or both are log-transformed.