Chapter 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 fat on average have higher low-density lipoprotein (LDL) cholesterol, a risk factor for 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 HbA_{1c}, a marker for poor control of glucose levels; however, older age and ethnic background also predict higher HbA_{1c}.

These are all examples of potentially complex relationships in observational data where a continuous outcome of interest, such as BMD, SBP, and HbA_{1c}, 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 CIs, hypothesis tests, and P-values are valid. For example, it is now standard practice to account for clinical center in the analysis of multisite clinical trials, often using the random effects methodology to be introduced in Chap. 7. 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.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_4,

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, hormone therapy (HT) has a smaller beneficial effect on LDL levels among postmenopausal women who are also taking statins, and its effect on BMD may be greater in younger postmenopausal women. These are examples of *interaction*, where the association of a factor of primary interest with an 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 Chap. 5, survival times in Chap. 6, and repeated measures in Chap. 7. 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 Chap. 3 is extended to accommodate multiple predictors. Section 4.3 shows how categorical predictors with multiple levels are coded and interpreted. Sections 4.4–4.6 describe how multipredictor regression models can be used to deal with confounding, mediation, and interaction, respectively. Section 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. Section 4.8 introduces sample size, power, and minimum detectable effect calculations for the multiple linear model. In Chap. 9, we use a *potential outcomes* view of *causal effects* to show how and under what conditions multipredictor regression models might be used to estimate them, and in Chap. 10 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

```
. regress 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 (HT) (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 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 Chap. 9, 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 nonusers. 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 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

Table 4.2 Adjusted regression of glucose on exercise

```
. regress 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
------------------------------------------------------------------------------
```

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 Chaps. 9 and 10 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 Sect. 9.1, 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.

4.2.1.1 Interpretation of Adjusted Regression Coefficients

In (4.2), the coefficient β_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.

4.2.2.1 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 (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).

4.2.2.2 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,

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

where, as before, $\sigma^2_{y|\mathbf{x}}$ is the residual variance of the outcome and $\sigma^2_{x_j}$ 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 $\sigma^2_{y|\mathbf{x}}$, 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.

4.2.2.3 t-Tests and Confidence Intervals

The t-tests of the null hypothesis H_0: $\beta_j = 0$ and CIs 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 CI 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 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 β_j^s for predictor x_j is defined in analogy to (3.12) as

$$\beta_j^s = \beta_j \sigma_{x_j} / \sigma_y, \tag{4.5}$$

where σ_{x_j} and σ_y are the 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 Chap. 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:

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

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 $\mathrm{E}[y|\mathbf{x}]$ 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 differs from zero. 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 variable `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 Chap. 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

Table 4.3 Coding of indicators for a multilevel categorical variable

	Indicator variables			
`physact`	`2.physact`	`3.physact`	`4.physact`	`5.physact`
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

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 `i.` variable prefix. By default, it uses the level with the lowest value as the reference group, although this is easily modified using a variable prefix of the form `ib`*k*, where *k* is the code of the alternative baseline category. 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,

$$\mathrm{E}[\texttt{glucose}|\mathbf{x}] = \beta_0 + \beta_2 \texttt{2.physact} + \cdots + \beta_5 \texttt{5.physact}. \tag{4.7}$$

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 E[`glucose`|**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 4.9). 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 β_5-β_4 (why?). All regression packages make it straightforward to estimate and test hypotheses about these *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.
- 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 (`5.physact`) 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 $\hat{\beta}_2$ (`2.physact`), or 97.6 mg/dL (95% CI 96.5–98.6 mg/dL). The `margins` command is then used to estimate the mean level in all five groups.

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 newer command `contrast{physact 0 0 −1 0 1}` is also used to compare groups 3 and 5. The contrast coefficients correspond in order to the five levels of `physact`. The two nonzero coefficients, −1 for group 3 and 1 for group 5, directly reflect the `lincom` command, and the three zeroes correspond to the omitted groups. The `effects` option is needed to obtain the estimated between-group difference and 95% confidence interval supplied by default by the `lincom` command. We explain *contrasts* in more detail in Sect. 4.3.5 below.

Table 4.4 Regression of physical activity on glucose

```
. regress glucose i.physact if diabetes == 0

      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 |
          2  |  -.8584489   1.084152    -0.79   0.429    -2.984617    1.267719
          3  |  -1.226199   1.011079    -1.21   0.225     -3.20906    .7566629
          4  |  -2.433855   1.010772    -2.41   0.016    -4.416114    -.451595
          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 + 2.physact
 ( 1)  2.physact + _cons = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   97.56211   .5414437   180.19   0.000     96.50027    98.62396
------------------------------------------------------------------------------

. margins physact
Adjusted predictions                                Number of obs   =       2032
Model VCE    : OLS
Expression   : Linear prediction, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     physact |
          1  |   98.42056   .9392676   104.78   0.000     96.57963    100.2615
          2  |   97.56211   .5414437   180.19   0.000      96.5009    98.62332
          3  |   97.19436   .3742409   259.71   0.000     96.46086    97.92786
          4  |   95.98671   .3734108   257.05   0.000     95.25483    96.71858
          5  |   95.14286   .6120416   155.45   0.000     93.94328    96.34244
------------------------------------------------------------------------------

. lincom 5.physact - 3.physact
 ( 1)  - 3.physact + 5.physact = 0
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.051505    .717392    -2.86   0.004    -3.458407   -.6446024
------------------------------------------------------------------------------
```

(continued)

Table 4.4 (continued)

```
. contrast {physact 0 0 -1 0 1}, effects
Contrasts of marginal linear predictions
Margins      : asbalanced
-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
     physact |          1        8.18     0.0043
-------------------------------------------------
------------------------------------------------------------------------------
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     physact |
        (1)  |  -2.051505    .717392    -2.86   0.004    -3.458407   -.6446024
------------------------------------------------------------------------------
```

Table 4.5 Overall physical activity effects on glucose

```
. quietly regress glucose i.physact if diabetes == 0

. testparm i.physact
       F(  4,  2027) =    4.43
            Prob > F =    0.0014

. contrast physact
Contrasts of marginal linear predictions
Margins      : asbalanced
-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
     physact |          4        4.43     0.0014
-------------------------------------------------
```

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 4.5 ($F(4, 2027) = 4.43$, $P = 0.0014$) shows that glucose levels clearly differ among the groups defined by `physact`. The same result is also obtained using the `contrast` command.

4.3.4 Multiple Pairwise Comparisons Between Categories

When the focus is on the difference between a single prespecified 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 prespecified 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 Chap. 7, when between-treatment differences are evaluated at multiple points in time. We also discuss the implications of multiple comparisons for model selection in Sect. 10.3.2, and more broadly in Sect. 13.4.1.

For this case, various methods are available for controlling the familywise error rate (FER) for the wider set of comparisons being made. 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 FER under the *complete null hypothesis* that all the group-specific population means are the same. However, it is subject to inflation of the FER under *partial null hypotheses*—that is, when there are some real population differences between subgroups.

More conservative procedures that protect the FER 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 FER and k is the number of preplanned comparisons to be made. The Sidak correction is slightly more liberal for small values of k, but otherwise equivalent. The Scheffé method is another, although 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.

As noted in Sect. 3.1.5, the Bonferroni, Sidak, and Scheffé procedures are available with the `oneway` ANOVA in Stata. In addition, beginning with Version 12, the `contrast` and `margins` postestimation commands implement analogous pairwise comparisons for all regression models discussed in this book, with control of FER using the Bonferroni, Sidak, and Scheffé procedures available via the `mcompare` option. These new commands have extensive capabilities for postestimation hypothesis testing, a few of which are illustrated below, and many others beyond the scope of this book. In Table 4.5, we obtained Bonferroni-corrected comparisons with the reference level of `physact` using the command `contrast`

Table 4.6 Bonferroni-corrected physical activity effects

```
. regress glucose i.physact if diabetes == 0

      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 |
          2  |  -.8584489   1.084152    -0.79   0.429    -2.984617    1.267719
          3  |  -1.226199   1.011079    -1.21   0.225     -3.20906    .7566629
          4  |  -2.433855   1.010772    -2.41   0.016    -4.416114    -.451595
          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
------------------------------------------------------------------------------

. contrast physact, mcompare(bonferroni) effects
Contrasts of marginal linear predictions
Margins      : asbalanced
-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
     physact |          4        4.43     0.0014
-------------------------------------------------
Note: Bonferroni-adjusted p-values are reported
      for tests on individual contrasts only.
---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
     physact |            4
---------------------------
------------------------------------------------------------------------------
             |                              Bonferroni            Bonferroni
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     physact |
(2 vs base)  |  -.8584489   1.084152    -0.79   1.000     -3.56876    1.851862
(3 vs base)  |  -1.226199   1.011079    -1.21   0.901    -3.753832    1.301434
(4 vs base)  |  -2.433855   1.010772    -2.41   0.065     -4.96072     .093011
(5 vs base)  |  -3.277704   1.121079    -2.92   0.014    -6.080331   -.4750759
------------------------------------------------------------------------------
```

`physact, compare(bonferroni)`. Note that while the estimates and overall F-test are unchanged, the P-values for the pairwise comparisons are larger and the CIs wider than in the regression output (Table 4.6).

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 FER under partial null hypotheses. It also illustrates the general principle that controlling the FER 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 FER for all possible 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 FER under partial null hypotheses.

4.3.5 Testing for Trend Across Categories

The coefficient estimates for the categories of `physact` shown in Table 4.4 decrease in order, suggesting that mean glucose levels are characterized by a linear trend across the levels of `physact`. Tests for linear trend are best performed using a *contrast* in the coefficients corresponding to the various levels of the categorical predictor.

> *Definition*: A *contrast* is a weighted sum of the regression coefficients of the form $a_1\beta_1 + a_2\beta_2 + \cdots + a_p\beta_p$ in which the weights, or *contrast coefficients*, sum to zero: that is, $a_1 + a_2 + \cdots + a_p = 0$.

The contrasts used to test for trend can be motivated as linear regressions of the adjusted means for each category on the categorical variable, treated as a continuous predictor, after centering and possibly rescaling the numeric codes used for each category. The resulting contrast coefficients used to test for linear trend have a simple pattern: they are

- Integer-valued
- Evenly spaced
- Symmetric about zero

Using integers is just a convenience. Underlying the even spacing is the assumption that the "distances" between adjacent categories are all the same; below, we briefly outline how this assumption can be relaxed. Symmetry about zero implies that they also sum to zero, as required.

To make this specific, the contrast coefficients that we would use to test for trend across the five levels of `physact` are -2, -1, 0, 1, and 2. More generally, when the number of levels is odd, the contrast coefficients are sequential integers (spacing of one), and by symmetry, the middle category has coefficient 0 and drops out. Thus for three categories, the coefficients are -1, 0, and 1, and for seven, follow in order from -3 to 3. When the number of levels is even, a spacing of two is the smallest that gives integer-valued contrast coefficients, and none of the categories

Table 4.7 Trend test in a model omitting the intercept

```
. regress glucose ibn.physact if diabetes == 0, noconstant

      Source |       SS       df       MS              Number of obs =    2032
-------------+------------------------------           F(  5,  2027) =40227.86
       Model |  18987135.4     5  3797427.08            Prob > F      =  0.0000
    Residual |  191344.609  2027  94.3979322            R-squared     =  0.9900
-------------+------------------------------           Adj R-squared =  0.9900
       Total |    19178480  2032  9438.22835           Root MSE      =  9.7159

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     physact |
           1 |   98.42056   .9392676   104.78   0.000     96.57853    100.2626
           2 |   97.56211   .5414437   180.19   0.000     96.50027    98.62396
           3 |   97.19436   .3742409   259.71   0.000     96.46043     97.9283
           4 |   95.98671   .3734108   257.05   0.000      95.2544    96.71902
           5 |   95.14286   .6120416   155.45   0.000     93.94256    96.34315
------------------------------------------------------------------------------

. * Tests for linear trend
. test -2*1.physact - 2.physact + 4.physact + 2*5.physact = 0
 ( 1)  - 2*1bn.physact - 2.physact + 4.physact + 2*5.physact = 0
       F(  1,  2027) =   12.11
            Prob > F =    0.0005

. contrast {physact -2 -1 0 1 2}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     physact |          1       12.11     0.0005
------------------------------------------------

. contrast q(1).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     physact |          1       12.11     0.0005
------------------------------------------------
```

are omitted. Thus with four categories, the contrast coefficients are −3, −1, 1, and 3, and with six, they are −5, −3, −1, 1, 3, and 5. So it is easy to figure out the contrast coefficients for any number of categories.

Table 4.7 shows a linear regression of glucose levels on physical activity, omitting the intercept, which we obtain by specifying `ibn.physact` in the `regress` command, in combination with the option `noconstant`. In this model, the group means for levels 1–5 of `physact` are given by β_1, β_2, β_3, β_4 and β_5, rather than by (4.8). The `test` command calculates the contrast using the contrast coefficients −2, −1, 0, 1, and 2, then compares it to the null value of zero; again, β_3, corresponding to the middle category, drops out. The result ($F(1, 2027) = 12.11$, $P = 0.0005$) leaves little doubt that there is a declining trend in mean glucose with increasing levels of physical activity.

Table 4.7 also shows two other methods for obtaining the test for linear trend. The first, using the command `contrast{physact −2 −1 0 1 2}`, incorporates the contrast coefficients for the five categories directly, in the same order as the levels of `physact`; this approach was also used to contrast levels 3 and 5 of `physact` in Table 4.4.

The second method uses the so-called *contrast operator* `q(1).` Including `(1)` as part of the operator specifies the test for linear trend; the default is to provide additional tests for quadratic, cubic, and quartic trends, plus a joint test for all four patterns. In both commands, the `noeffects` option prevents Stata from printing the numeric values of the contrasts, which are uninterpretable in this case.

The `q.` contrast operator treats the ordered categories as evenly spaced, regardless of the coding of the categorical variable. This assumption can be relaxed using the `p.` operator instead, in combination with a coding for the categorical variable that reflects the hypothesized spacing. For example, if we hypothesized spacings of 2, 1, 1, and 2 units between the categories of the physical activity variable, coding the levels as 1, 3, 4, 5, and 7, then testing for linear trend using the command `contrast p(1).physact, noeffects` would obtain the appropriate test.

Of course, the default in Stata and other statistical packages is to include the intercept in almost all regression models; in the Cox model, introduced in Chap. 6, the baseline hazard plays this role. When an intercept is included in the model, one level of the categorical variable must generally serve as the reference category and be omitted from the model. This default form of the model was laid out Table 4.3 and (4.8), and is obtained simply by specifying `i.physact` in the `regress` command.

Fortunately, we can easily adapt the integer-valued, evenly-spaced, symmetric, zero-sum contrast coefficients to the default form of the model with an intercept, simply by dropping the coefficient corresponding to the omitted reference category. To see why this works, and why the intercept does not figure in the contrast, we evaluate the contrast in the regression coefficients specifying the means for each level of `physact`, as shown in (4.8):

$$\begin{aligned} 0 &= -2\beta_0 - (\beta_0 + \beta_2) + (\beta_0 + \beta_4) + 2(\beta_0 + \beta_5) \\ &= -\beta_2 + \beta_4 + 2\beta_5 \end{aligned} \tag{4.9}$$

In (4.9), the mean for level three of `physact`, $\beta_0 + \beta_3$, is omitted because the contrast coefficient $a_3 = 0$, and β_0 disappears because the contrast coefficients sum to zero. Table 4.8 summarizes the resulting contrasts used to test for trend when the categorical variable has 3–6 levels and the lowest category is treated as the reference.

Table 4.9 shows the test for trend in glucose levels across the levels of `physact`, based on the default form of the model including an intercept. The trend test result is exactly the same as in Table 4.7, whether we use `test` or either version of the `contrast` command to obtain it.

Table 4.8 Trend contrasts for models with an intercept

Number of categories	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 4.9 Trend test in a model including the intercept

```
. regress glucose i.physact if diabetes == 0

      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 |
          2  |  -.8584489   1.084152    -0.79   0.429    -2.984617    1.267719
          3  |  -1.226199   1.011079    -1.21   0.225     -3.20906    .7566629
          4  |  -2.433855   1.010772    -2.41   0.016    -4.416114    -.451595
          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
------------------------------------------------------------------------------

. * Tests for linear trend
. test -2.physact + 4.physact + 2*5.physact = 0
 ( 1)  - 2.physact + 4.physact + 2*5.physact = 0
       F(  1,  2027) =   12.11
            Prob > F =    0.0005

. contrast {physact -2 -1 0 1 2}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     physact |          1       12.11     0.0005
------------------------------------------------

. contrast q(1).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     physact |          1       12.11     0.0005
------------------------------------------------
```

A few more details about these trend tests are worth noting:

- In (4.9), we showed why β_0 does not figure in the contrasts in Table 4.8. By extension, the effects of any adjustment variables held constant by the model would also drop out.

- If a different reference category is used, we simply drop that component of the contrast rather than the first. For example, suppose we specified level two as the reference category for `physact` using `ib2.physact` in the `regress` command. Then the appropriate contrast would be $-2\beta_1 + \beta_4 + 2\beta_5 = 0$. If we specified level three as the reference category, using `ib3.physact`, the contrast would be $-2\beta_1 - \beta_2 + \beta_4 + 2\beta_5 = 0$. The trend test results are unaffected by changing the reference category.
- As compared to a simpler approach in which the categorical variable is treated as a continuous predictor, using the categorical version of the model in conjunction with contrasts to test for trend can be more efficient when there is *both* trend and departure from it, a problem we examine next. This occurs because the model captures the departures from linear trend, reducing the residual variance, and thus making regression effects easier to detect.
- These contrasts are valid for the other models in this book, including logistic, survival, repeated measures, and GLMs. In GLMs and Cox models, treating a multilevel predictor as categorical rather than continuous achieves no efficiency gain of the kind sometimes seen in linear models. Nonetheless, in such cases, treating the predictor as categorical rather than continuous should achieve at least somewhat better fit.
- Similar contrasts are available for assessing quadratic, cubic, and quartic trends across categories, now easily accessible using the `contrast` command with the `q.` and `p.` contrast operators.

4.3.5.1 Departures from Linear Trend

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. After demonstrating a statistically significant trend as in Table 4.7 or 4.9, it is easy to test for such a departure. One method for doing this uses a model in which the categorical variable is treated both as continuous and categorical. In this set-up, the continuous version accounts for the trend, while the categorical version captures departure from it. Thus, in Table 4.10 the F-test for the overall 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.

Table 4.10 Testing for departure from linear trend

```
. quietly regress glucose physact i.physact if diabetes == 0
note: 5.physact omitted because of collinearity

. testparm i.physact
       F(  3,  2027) =    0.26
            Prob > F =    0.8511
```

Table 4.11 Testing for departure from linear trend

```
. quietly regress glucose i.physact if diabetes == 0

. contrast q(2/4).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
     physact |
(quadratic)  |          1        0.11     0.7411
    (cubic)  |          1        0.01     0.9415
  (quartic)  |          1        0.49     0.4859
       Joint |          3        0.26     0.8511
             |
    Residual |       2027
------------------------------------------------
```

Two additional comments about the model in Table 4.10:

- The omission of an additional category of `physcat` is expected, in fact necessary for the test for departure from trend to work. For this to occur, `physact` must precede `i.physact` in the regression command; with the reverse ordering, Stata would omit `physact` as continuous instead.
- *This model is only useful for testing from departure from trend*. Neither the coefficient nor the *t*-test for the effect of `physact` as continuous is interpretable. Estimation of the effects of the categorical variable as well as the test for trend must be carried out as in Table 4.7 or 4.9, using a model including the categorical version of the predictor only.

We can obtain exactly the same result from the original model including `physact` only as a categorical variable, using the contrast operator `q(2/4).physact`. This assesses evidence for quadratic, cubic, and quartic trends, as well as evidence for all three jointly. Because we omitted the test for linear trend, the 3 degree-of-freedom joint test is equivalent to the first approach using `physact` as both continuous and categorical. Note that the specific form of the contrast operator depends on the number of levels: for example, we would need to use `contrast q(2/3).physact` if `physact` had four levels, and `contrast q(2/5).physact` if it had six (Table 4.11).

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 comparison 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 coefficient may be useful for describing differences between 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. In Chap. 9, we use the concept of *potential outcomes* to define causal effects more precisely, and to show when multipredictor models can be used to estimate them in the presence of confounding, and when they cannot.

To summarize briefly, the overall point of Chap. 9 is that to assess confounding we first need a hypothesized causal framework. In particular, the potential confounder should be plausible as a cause of both the predictor of interest and the outcome, or as a proxy for such a cause. Within this hypothesized framework, the data provide support for confounding if we find that:

- The potential confounder is associated with the predictor of interest, and also independently associated with the outcome.
- The coefficient for the effect of the primary predictor on the outcome changes when we add the potential confounder to the model. Note, however, that analogous changes are also seen in logistic, Cox, and some other models, discussed in Chaps. 5, 6, and 8, when nonconfounders associated with the outcome but not the predictor of interest are added to the model.

4.4.1 Range of Confounding Patterns

Confounders often explain some of the association of a predictor of interest with the outcome, so that the adjusted effect, which may have a causal interpretation, is often weaker than the unadjusted effect. We 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 $\mathcal{E}$, the exposure of primary interest, is binary and coded 0 and 1, and the potential confounder, $\mathcal{C}$, 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, $\mathcal{E}$ 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 $\mathcal{C}$. The regression lines for $\mathcal{C}$ are the same for both groups defined by $\mathcal{E}$; in other words, there is no association with $\mathcal{E}$ after adjustment for $\mathcal{C}$.

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 $\mathcal{E}$ and y, but on the right the regression lines for $\mathcal{C}$

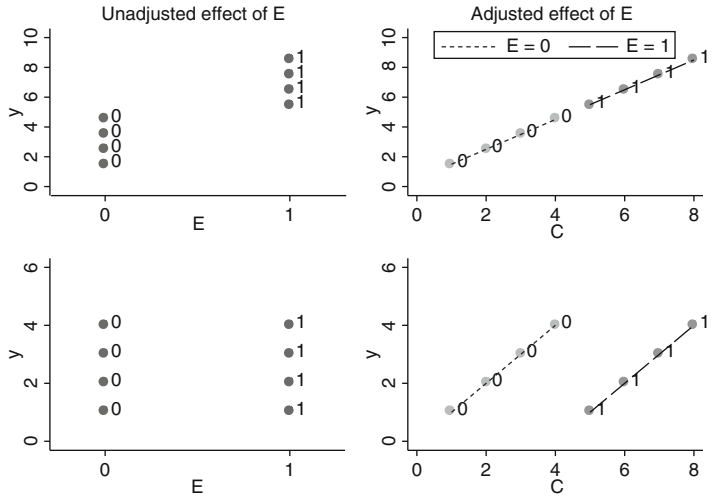

Fig. 4.1 Complete and negative confounding patterns

are very distinct for the groups defined by $\mathcal{E}$. In short, the association between $\mathcal{E}$ and y is unmasked by adjustment for $\mathcal{C}$. Negative confounding can occur under the following circumstances:

- 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.

4.4.2 Confounding Is Difficult to Rule Out

The problem of confounding can be more resistant to multipredictor regression modeling than the example in Table 4.12 might suggest. We assumed in that example that the model included all confounders of the effect of BMI on LDL. Of course, the multipredictor linear model (4.2) can (within limits imposed by sample size) include more than a few predictors, giving us considerable freedom to model the effects of other causal factors. 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
- Accurately represented in the systematic part of the model

Table 4.12 Unadjusted and adjusted regressions of LDL on BMI

```
. regress 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
------------------------------------------------------------------------------

. regress 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
------------------------------------------------------------------------------
```

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 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.

4.4.3 Adjusted Versus Unadjusted $\hat{\beta}$s

Uncontrolled confounding induces bias in unadjusted (or inadequately adjusted) estimates of the causal effects that are commonly the focus of our attention. This suggests 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, although it almost surely does not have a causal interpretation. Thus, it should not be expected to have the same value as the causal parameter.

4.4.4 Example: BMI and LDL

We turn to a relatively simple example, again using data from the HERS cohort. BMI and LDL cholesterol are both established heart disease risk factors. It is reasonable to hypothesize that higher BMI leads to higher LDL in some causal sense, to be made more specific in Chap. 9. An unadjusted model for BMI and LDL is shown in Table 4.12. 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 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 versus 0.32 mg/dL) in the fully adjusted model shown in Table 4.12 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 4.12 can all be shown to meet sample diagnostic criteria for potential confounding of the effect of 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 significant. 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 Chap. 10.

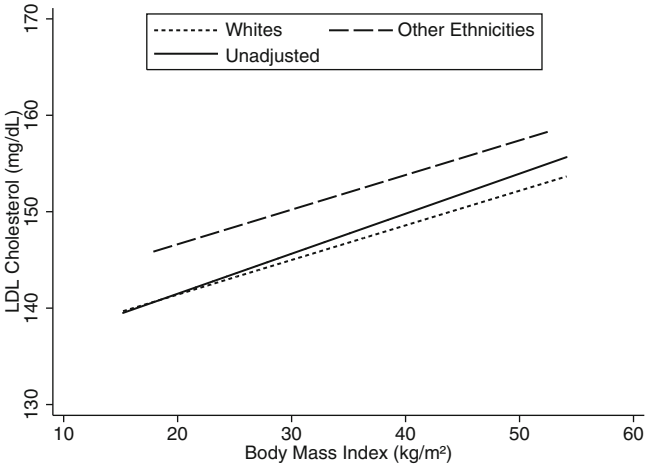

Fig. 4.2 Unadjusted and adjusted regression lines

Figure 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 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 the adjusted model for LDL shown in Table 4.12, we assumed that age, race/ethnicity, smoking, and alcohol use might confound the effect of BMI, because they affect both BMI and LDL levels, or are proxies for factors that do. However,

if the primary predictor is a cause of one of the covariates, which in turn affects the outcome, this would be an instance of *mediation*. 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.

Thus a potential mediator, like a potential confounder, must make sense in terms of a hypothetical causal framework. In particular, it should be plausible as an *effect* of the predictor of interest and as a *cause* of the outcome, or as a proxy for the true intermediary factor. Within this framework, the data support mediation if we find that:

- The potential mediator is associated with the predictor of interest *and* with the outcome, controlling for the predictor of interest.
- The coefficient for the effect of the primary predictor on the outcome changes when we add the potential mediator to the model. However, as with confounders, analogous changes are also seen in logistic, Cox, and some other models when nonmediators associated with the outcome but not the predictor of interest are added to the model.

Thus mediators behave like confounders in regression models, and can only be distinguished by the hypothesized causal framework—the data have little to tell us about the direction of the causal effects.

4.5.1 Indirect Effects via the Mediator

The effect of the primary predictor on the mediator, and of the mediator on the outcome, together comprise the hypothesized *indirect causal pathway* via the mediator. If the models used to estimate these effects adequately control confounding of both relationships, then the two effects may together have a causal interpretation as the *indirect effect* of the primary predictor; additional assumptions underlying this interpretation are discussed in Sect. 9.6. Accordingly, primary evidence for the indirect effect via the mediator is given by a test of the effect of the primary predictor on the mediator, in combination with a second test of the effect of the mediator on the outcome. The overall null hypothesis of no indirect effect is rejected only if *both* underlying null hypotheses are rejected at the nominal α level, preventing inflation of the type-I error rate.

4.5.2 Overall and Direct Effects

If the indirect pathway exists, and confounding has been controlled, then the coefficient for the primary predictor before adjustment for the mediator has a causal interpretation as the *overall effect* of the primary predictor on the outcome. The coefficient adjusted for the mediator is interpretable as the so-called *direct effect*

of the primary predictor via other pathways that do not involve the mediator. Finally, the *difference* between overall and direct effects of the primary predictor is interpretable as the indirect effect.

Tests for the difference between the overall and direct effects can also be used to assess mediation. However, these tests are complicated by the need to compare coefficient estimates for the primary predictor from two different models, but estimated using the same data. As a result, the two estimates are correlated, which must be taken into account. Surprisingly, these tests are less powerful in some cases than the joint test of the indirect pathway just discussed.

It is important to note that these interpretations may hold only under additional conditions in the generalized linear and Cox models discussed in Chaps. 5, 6, and 8. In particular, tests for the difference between the overall and direct effects can give false-positive results, because the *collapsibility* issue first introduced in Sect. 3.4.5. As we have already pointed out, in these models the coefficient for the primary predictor will generally change if a powerful predictor is added to the model. This holds even if the new covariate is not associated with the primary predictor, implying that it plays no mediating role.

4.5.3 Percent Explained

The *relative* difference between the overall and direct effects is sometimes referred to as the *percent explained* (PE) and used as an additional summary measure of the indirect effect. Direct estimation of PE rests on the assumption that the primary predictor and mediator do not interact (Robins and Greenland 1992; Freedman et al. 1992). This assumption can be checked using methods explained in Sect. 4.6, and possibly relaxed (Li et al. 2001; Vansteelandt 2009; VanderWeele 2009) as discussed briefly in Sect. 9.6. Testing and CI estimation for PE are even more complicated and problematic than for the difference between the overall and direct effects of the primary predictor.

4.5.4 Example: BMI, Exercise, and Glucose

We examined the extent to which the effects of BMI on glucose levels might be mediated through its effects on likelihood of exercise. Although exercise may in some cases affect BMI, in HERS exercise was weakly associated ($P = 0.06$) with a small *increase* in BMI over the first year of the study. As a result, we would argue that in this population of older women with established heart disease, BMI mainly affects likelihood of exercise, with very little feedback. Thus, mediation of the effects of BMI by exercise makes sense in terms of a hypothesized causal framework. We recognize that our simple models might not completely control confounding of the relationships among BMI, exercise, and glucose, and could be improved with expert input.

To assess mediation of the effects of BMI by exercise, we assessed both links in the hypothesized indirect pathway. Specifically, we first used a logistic model (Chap. 5) to assess the independent effects of BMI on likelihood of exercise, adjusting for age, race/ethnicity, smoking, alcohol use, and poor or fair self-reported health. Results in Table 4.13 show that each kg/m^2 increase in BMI is associated with an 8% decrease in the odds of exercise (95% CI 4–10%, $P < 0.0005$). In addition, the linear model for glucose levels establishes the second link in the indirect pathway, showing that exercise is independently associated with a decrease in average glucose of about 1 mg/dL (95% CI 0.1–1.9, $P = 0.027$). So the proposed mediator is associated with both the primary predictor and independently with the outcome. Since both null hypotheses are rejected at the nominal 2-sided 5% level, there is evidence for the indirect causal pathway via exercise.

On the other hand, 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. We manipulated regression results stored as so-called `scalars` to calculate PE as $(0.5025557 - 0.4859684)/0.5025557 \times 100 = 3.3\%$. Thus, while our joint test of the indirect pathway shows that we can rule out chance at the nominal 5% level, only a very small part of the effect of BMI on glucose levels appears to be mediated by its effects on likelihood of exercising.

4.5.5 Pitfalls in Evaluating Mediation

Evaluating mediation, in particular estimating direct effects and PE, has many potential difficulties. In particular, bias can arise from uncontrolled confounding of the association between the mediator and the outcome (Robins and Greenland 1992; Cole and Hernán 2002)—even in clinical trials where the primary predictor is randomized treatment assignment. In observational data, we obviously need to control confounding of the effects of the primary predictor as well. Additional difficulties arise if a confounder of the mediator/outcome relationship is affected by treatment, and thus a causal intermediate (Petersen et al. 2006). We briefly cover these issues in Sect. 9.6.

4.5.5.1 Temporality

In addition, it is often difficult to infer causal direction in cross-sectional data. Longitudinal data may provide stronger support for the hypothesized indirect pathway by 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, if these changes all occur more or less simultaneously, and between sequential longitudinal observations, the temporal ordering can easily be obscured. Furthermore, as discussed in Sect. 6.3.1, longitudinal analyses set up to

Table 4.13 Indirect pathway from BMI to glucose levels via exercise

```
. * Overall effect of BMI on glucose, adjusting for age and alcohol use
. regress glucose BMI age10 nonwhite smoking drinkany poorfair if diabetes == 0

      Source |       SS       df       MS              Number of obs =    2025
-------------+------------------------------           F(  6,  2018) =   25.48
       Model |   13529.786     6  2254.96434           Prob > F      =  0.0000
    Residual |  178590.143  2018  88.4985842           R-squared     =  0.0704
-------------+------------------------------           Adj R-squared =  0.0677
       Total |  192119.929  2024  94.9209135           Root MSE      =  9.4074
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .5025557   .0414832    12.11   0.000     .4212013    .5839102
       age10 |   .7093964   .3259568     2.18   0.030     .0701494    1.348643
    nonwhite |   .8801519   .7610825     1.16   0.248    -.6124377    2.372741
     smoking |   .1812593   .6135155     0.30   0.768    -1.021931    1.384449
    drinkany |   .7137293   .4305044     1.66   0.097    -.1305502    1.558009
    poorfair |  -.2052528   .5394217    -0.38   0.704    -1.263134    .8526288
       _cons |   77.63278   2.687214    28.89   0.000     72.36278    82.90279
------------------------------------------------------------------------------
. * Store coefficient for BMI as estimate of overall effect
. scalar overall = _b[BMI]

. * First link: logistic model for BMI effect on exercise
. logistic exercise BMI age10 nonwhite smoking drinkany poorfair if diabetes == 0

Logistic regression                             Number of obs   =       2025
                                                LR chi2(6)      =     158.56
                                                Prob > chi2     =     0.0000
Log likelihood = -1294.4669                     Pseudo R2       =     0.0577
------------------------------------------------------------------------------
    exercise | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .9235428   .0093154    -7.89   0.000     .9054643    .9419822
       age10 |   .8171735   .0600467    -2.75   0.006     .7075662    .9437597
    nonwhite |   .8012592   .1416865    -1.25   0.210     .5665721    1.133159
     smoking |   .3012331   .0470011    -7.69   0.000     .2218658    .4089921
    drinkany |   .9159856   .0883199    -0.91   0.363      .758255    1.106527
    poorfair |    .523097   .0671846    -5.05   0.000      .406684    .6728331
------------------------------------------------------------------------------

. * Second link: fully adjusted model for effect of exercise on glucose levels
. regress glucose BMI age10 nonwhite smoking drinkany poorfair exercise ///
     if diabetes == 0

      Source |       SS       df       MS              Number of obs =    2025
-------------+------------------------------           F(  7,  2017) =   22.59
       Model |  13964.2063     7  1994.88661           Prob > F      =  0.0000
    Residual |  178155.723  2017  88.3270811           R-squared     =  0.0727
-------------+------------------------------           Adj R-squared =  0.0695
       Total |  192119.929  2024  94.9209135           Root MSE      =  9.3982
```

(continued)

Table 4.13 (continued)

```
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .4859684   .0421125    11.54   0.000     .4033798     .568557
       age10 |   .6655835   .3262395     2.04   0.041     .0257819    1.305385
    nonwhite |   .8315359   .7606607     1.09   0.274    -.6602267    2.323299
     smoking |  -.0612536   .6225991    -0.10   0.922    -1.282258    1.159751
    drinkany |   .6954023   .4301665     1.62   0.106    -.1482147    1.539019
    poorfair |  -.3387946   .5422525    -0.62   0.532    -1.402228     .724639
    exercise |  -.9762492   .4402026    -2.22   0.027    -1.839548   -.1129499
       _cons |   78.86342    2.74136    28.77   0.000     73.48723    84.23961
------------------------------------------------------------------------------
. * Store coefficient for BMI as estimate of direct effect, and calculate PE
. scalar direct = _b[BMI]
. scalar PE = round((overall-direct)/overall*100, 0.1)
. scalar list PE
       PE =        3.3
```

examine such temporal patterns can be misleading if the mediator also potentially confounds the association between the primary predictor and outcome (Hernán et al. 2001).

4.5.5.2 Problems with PE

Finally, while PE is a popular and relatively interpretable measure of mediation, CIs for this measure can be wide and unreliable if the overall effect of the primary predictor is weak or noisily estimated. In addition, while PE is nominally a percentage, values outside the interval from 0% to 100% are possible. In particular, this occurs if the direct and indirect effects of the primary predictor are in opposite directions—for instance, if a treatment has both beneficial and adverse effects on the outcome, via different pathways. Even when PE is between 0% and 100%, confidence bounds are commonly outside this range. In addition, Molenberghs et al. (2002) show that estimates of PE are also influenced by the precision of measurements of both the mediator and outcome, potentially leading to highly misleading results.

4.6 Interaction

In Sect. 4.4, we gave examples in which a multipredictor linear model might be used to reduce or eliminate confounding of the effects of a primary predictor. So far, we have made the assumption that causal effect of the primary predictor was the same within strata defined by the covariates. However, this may not hold. In this section, we show how to use regression to model the resulting *interaction*, so that we can estimate causal effects that differ according to the level of a covariate. Interaction is also referred to as *effect modification* or *moderation*, and must be distinguished from both confounding and mediation (Baron and Kenny 1986).

Table 4.14 Model for interaction of HT and statins

Group	HT	statins	HT#statins	E[LDL|**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	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

4.6.1 Example: Hormone Therapy and Statin Use

As an example of interaction, we examine whether the effect of HT on LDL cholesterol differs according to baseline statin use, using data from HERS. To do this, a constructed interaction variable is useful. Suppose both assignment to HT and use of statins at baseline are coded using indicator variables. Then, the product of these two variables is also an indicator, equal to one only for the subgroup of women who reported using statins at baseline and were randomly assigned to HT, and zero for everyone else. Now, consider the regression model

$$\mathrm{E}[\texttt{LDL}|\mathbf{x}] = \beta_0 + \beta_1 \texttt{HT} + \beta_2 \texttt{statins} + \beta_3 \texttt{HT\#statins}, \tag{4.10}$$

where `HT` is the indicator of assignment to HT, `statins` the indicator of baseline statin use, and `HT#statins` the interaction term, which Stata calculates automatically.

Table 4.14 shows the values of (4.10) 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 nonusers. Note that both overall and within the strata defined by baseline statin use, we can assume that the groups randomly assigned to HT and placebo are comparable.

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 nonusers among women assigned to placebo, while $\beta_2 + \beta_3$ gives the analogous difference among women assigned to HT. However, women were not randomized to statin use, so 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 4.15 shows that there is some evidence for a smaller effect of HT on LDL among women reporting statin use at study baseline. The command `i.HT##i.statins` instructs Stata to include both so-called main effects, shown as `1.HT` and `1.statins` in the output, as well as the interaction term `HT#statins`, which it calculates only for the purposes of running the regression and does not retain in the data.

Table 4.15 Interaction of hormone therapy and statin use

```
. reg LDL1 i.HT##i.statins

      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 |   3969848.8  2607  1522.76517            Root MSE      =  37.912

------------------------------------------------------------------------------
        LDL1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.HT |  -17.72836   1.870629    -9.48   0.000    -21.39643   -14.06029
   1.statins |  -13.80912    2.15213    -6.42   0.000    -18.02918   -9.589065
             |
  HT#statins |
       1 1   |   6.244416   3.076489     2.03   0.042     .2118042    12.27703
             |
       _cons |   145.1567   1.325549   109.51   0.000     142.5575     147.756
------------------------------------------------------------------------------

. lincom 1.HT + 1.HT#1.statins
 ( 1)  1.HT + 1.HT#1.statins = 0
------------------------------------------------------------------------------
        LDL1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -11.48394   2.442444    -4.70   0.000    -16.27327   -6.694615
------------------------------------------------------------------------------
```

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 `HT#statins` (that is, $\hat{\beta}_1 + \hat{\beta}_3$) using the `lincom` command. Note that in contrast to the `regress` command itself, where we used `##` to obtain both main effects and interaction term, in the `lincom` command we used a single `#` to specify the interaction term only. The result 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 ($t = 2.03$, $P = 0.042$). 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.

Finally, we note that in the `lincom` command shown in Table 4.15, we have to specify the values of each variable—in this case, 1 and 1—to which the interaction term applies. If either of the two main effects is a multicategory predictor, then the interaction would also have more than one level. For example, if we wanted to assess interaction between HT and level of physical activity, we would use the commands shown in Table 4.16. The `testparm` command is used to obtain a global test of the

Table 4.16 Interaction of hormone therapy and physical activity

```
. regress LDL1 i.HT##i.physact

      Source |       SS       df       MS              Number of obs =    2608
-------------+------------------------------           F(  9,  2598) =   12.19
       Model |  160857.353     9  17873.0393            Prob > F      =  0.0000
    Residual |  3808991.44  2598   1466.1245            R-squared     =  0.0405
-------------+------------------------------           Adj R-squared =  0.0372
       Total |   3969848.8  2607  1522.76517            Root MSE      =   38.29

------------------------------------------------------------------------------
        LDL1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.HT |  -4.973552   5.810288    -0.86   0.392    -16.36681    6.419711
             |
     physact |
          2  |   4.386916   4.612377     0.95   0.342     -4.65739    13.43122
          3  |    6.96232   4.338071     1.60   0.109    -1.544106    15.46875
          4  |   8.797315   4.378699     2.01   0.045     .2112231    17.38341
          5  |   6.793914   5.040489     1.35   0.178    -3.089867    16.67769
             |
  HT#physact |
       1 2   |  -6.714054   6.799605    -0.99   0.324    -20.04725    6.619138
       1 3   |  -10.71075   6.367042    -1.68   0.093    -23.19573    1.774244
       1 4   |  -13.15391   6.411071    -2.05   0.040    -25.72523   -.5825811
       1 5   |  -12.96408   7.314865    -1.77   0.076    -27.30763    1.379473
             |
       _cons |   133.4211   3.928472    33.96   0.000     125.7178    141.1243
------------------------------------------------------------------------------

. testparm i.HT#i.physact
       F(  4,  2598) =    1.42
            Prob > F =    0.2258

. contrast HT#physact
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
  HT#physact |          4        1.42     0.2258
------------------------------------------------
```

interaction, which is not statistically significant, despite nearly significant P-values for the interaction terms for HT and levels 3, 4, and 5 of physical activity. The `contrast` command gives an equivalent result.

4.6.2 Example: BMI and Statin Use

Similar approaches can be used to assess modification of the effects of continuous predictors. For example, the association between BMI and baseline LDL cholesterol levels was shown in Sect. 4.4.4 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 nonusers.

In examining this interaction, centering `BMI` about its mean value of 28.6 kg/m^2 makes the parameter estimate for statin use more interpretable, as shown below. Then, to implement the analysis, we would first compute `BMIc`, the new centered BMI variable. Note that because `statins` is an indicator variable coded 1 for users and 0 for nonusers, the interaction variable `statins#c.BMIc` automatically made by Stata is by definition equal to `BMIc` in statin users, but equal to zero for nonusers. 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

$$\begin{aligned} \mathrm{E}[\texttt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_1\texttt{statins} + \beta_2\texttt{BMIc} + \beta_3\texttt{statins\#c.BMIc} \\ & + \beta_4\texttt{age} + \beta_5\texttt{nonwhite} + \beta_6\texttt{smoking} + \beta_7\texttt{drinkany}. \end{aligned} \tag{4.11}$$

Thus, among women who do not use statins,

$$\begin{aligned} \mathrm{E}[\texttt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_2\texttt{BMIc} \\ & + \beta_4\texttt{age} + \beta_5\texttt{nonwhite} + \beta_6\texttt{smoking} + \beta_7\texttt{drinkany}, \end{aligned} \tag{4.12}$$

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

$$\begin{aligned} \mathrm{E}[\texttt{LDL}|\mathbf{x}] = {} & \beta_0 + \beta_1\texttt{statins} + \beta_2\texttt{BMIc} + \beta_3\texttt{statins\#c.BMIc} \\ & + \beta_4\texttt{age} + \beta_5\texttt{nonwhite} + \beta_6\texttt{smoking} + \beta_7\texttt{drinkany} \\ = {} & \beta_0 + \beta_1 + (\beta_2 + \beta_3)\texttt{BMIc} \\ & + \beta_4\texttt{age} + \beta_5\texttt{nonwhite} + \beta_6\texttt{smoking} + \beta_7\texttt{drinkany}. \end{aligned} \tag{4.13}$$

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 nonusers depends on BMI. Subtracting (4.12) from (4.13), the difference in average LDL in statin users as compared to nonusers is $\beta_1 + \beta_3\texttt{BMIc}$.

Table 4.17 shows the results of the interaction model for statin use and BMI. The estimated coefficients have the following interpretations:

- `statins`: Among women with `BMIc = 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.

Table 4.17 Interaction model for BMI and statin use

```
. regress LDL i.statins##c.BMIc 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]
-------------+----------------------------------------------------------------
   1.statins |  -16.25301   1.468788   -11.07   0.000    -19.13305   -13.37296
       BMIc |   .5821275    .160095     3.64   0.000     .2682082    .8960468
            |
     statins#|
      c.BMIc |
          1  |   -.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    -.3883704    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    .8004355
       _cons |   162.4052   7.583312    21.42   0.000     147.5356    177.2748
------------------------------------------------------------------------------

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

- `BMIc`: 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 significant (t=3.64, $P < 0.0005$).
- `statins#c.BMIc`: 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 nonusers plus the estimated difference in slopes. The estimate of −0.12 mg/dL per unit increase in BMI is not statistically significant ($t = -0.54$, $P = 0.59$), but the 95% CI (−0.55 to 0.31 mg/dL per unit increase in BMI) is fairly wide.

Figure 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 the adverse effect of higher BMI on LDL may be blocked by statin use.

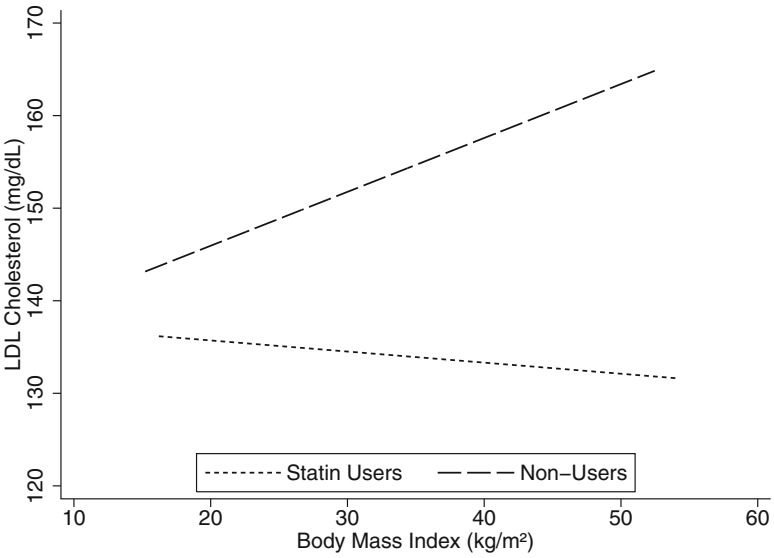

Fig. 4.3 Stratum-specific regression lines

4.6.3 Interaction and Scale

Interaction models are often distinguished from simpler *additive* models which do not include interaction terms. Moreover, the simpler additive model is generally treated as the default in predictor selection, with an interaction 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. 5) the *logit* transformation of $\mathrm{E}[Y|\mathbf{x}]$ is modeled, while in some generalized linear models (GLMs; Chap. 8), 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.

Table 4.18 Interaction model for HT effects on absolute change in LDL

```
. regress LDLch HT##c.cLDL0

      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]
-------------+----------------------------------------------------------------
        1.HT |  -15.47703   1.196246   -12.94   0.000    -17.82273   -13.13134
       cLDL0 |  -.3477064   .0225169   -15.44   0.000    -.3918593   -.3035534
             |
  HT#c.cLDL0 |
          1  |  -.0786871   .0316365    -2.49   0.013    -.1407226   -.0166517
             |
       _cons |  -4.888737   .8408392    -5.81   0.000    -6.537522   -3.239953
------------------------------------------------------------------------------
```

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 $\mathcal{C}$ confounds $\mathcal{E}$, 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.

4.6.4 Example: Hormone Therapy and Baseline LDL

The effect of HT 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 4.18. 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 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.
- `HT#c.cLDL0`: 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 4.19 Interaction model for HT effects on percent change in LDL

```
. regress LDLpctch HT##c.cLDL0

      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]
-------------+----------------------------------------------------------------
       1.HT |  -10.79035   .8514335   -12.67   0.000    -12.45991   -9.120789
      cLDL0 |  -.2162436   .0160265   -13.49   0.000    -.2476697   -.1848176
            |
  HT#c.cLDL0 |
          1 |   .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 4.19. In particular, the interaction term `HT#c.cLDL0` 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.

4.6.5 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 interaction terms involving a categorical predictor will almost always give mistaken results. In contrast, the `i.` and `##` operators in Stata will handle this situation. However, suppose one of the predictors has four levels and the other three levels. Then the interaction is modeled using an extra $(4-1)(3-1) = 6$ indicator variables. Many different patterns are subsumed by the alternative hypothesis of interaction, only a few of which may be of interest or biologically plausible; moreover, the F-test for interaction may have low power.

- 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 4.6) 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 CIs 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. Many of these procedures rely heavily on the transformations of both predictor and outcome that were introduced in Chap. 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.

4.7.1.1 Component-Plus-Residual Plots

In unadjusted analysis, checks for departures from linearity could be carried out using LOWESS, the nonparametric scatterplot smoother introduced in Chap. 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.

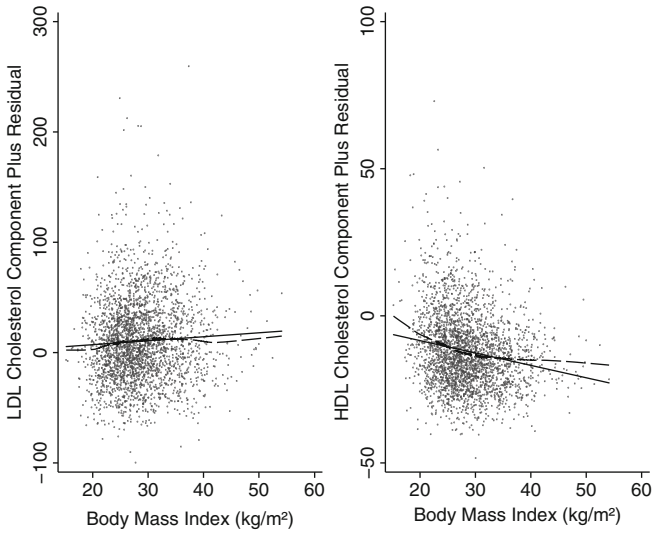

Fig. 4.4 CPR plots for multiple regressions of LDL and HDL on BMI

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`. Rather than the residuals of the RVP plot, the residuals plus the component of the fitted values due to BMI are plotted and smoothed against BMI.

Figure 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, with solid lines representing the linear fits for BMI, and the dashed lines the LOWESS smooths of the plotted component-plus-residuals (CPRs) against BMI. If the linear fits for BMI were satisfactory, then there would be no nonlinear pattern across values of BMI in the CPRs. 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 fitted model is shown in Table 4.20.

For interpretability, we centered the linear term `BMIc` on the sample mean of 28.6 kg/m^2 before calculating the quadratic term, `BMIc2`, and also centered age. The linear and quadratic terms in centered BMI are both clearly needed ($P < 0.0005$). In this model, the intercept 47.6 estimates expected HDL for a 67-year old, white nonsmoking abstainer with BMI $=$ 28.6 kg/m^2. The `BMIc` coefficient

Table 4.20 Linear plus quadratic model for effect of BMI on HDL

```
. regress HDL BMIc BMIc2 agec nonwhite smoking drinkany

      Source |       SS       df       MS              Number of obs =    2745
-------------+------------------------------           F(  6,  2738) =   39.99
       Model |  38474.0925     6  6412.34874            Prob > F      =  0.0000
    Residual |   439006.42  2738  160.338356            R-squared     =  0.0806
-------------+------------------------------           Adj R-squared =  0.0786
       Total |  477480.512  2744  174.008933            Root MSE      =  12.662
------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        BMIc |  -.5272063   .0507626   -10.39   0.000    -.6267432   -.4276693
       BMIc2 |   .0242527   .0053231     4.56   0.000      .013815    .0346904
        agec |   .1893209   .0380347     4.98   0.000     .1147414    .2639005
    nonwhite |   2.494766   .7815733     3.19   0.001     .9622325    4.027299
     smoking |  -2.070298   .7449086    -2.78   0.005    -3.530938   -.6096584
    drinkany |   4.345096   .5041409     8.62   0.000     3.356561    5.333631
       _cons |   47.86615   .3794279   126.15   0.000     47.12215    48.61014
------------------------------------------------------------------------------
```

estimate of −0.53 estimates the decrease in average HDL per unit increase in BMI, *at the point where BMI* = *28.6* kg/m^2, while the coefficient for `BMIc2` captures the (upward) curvature of the regression line.

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.

4.7.1.2 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.

However, other transformations of the predictor may sometimes be more successful and should be considered. Figure 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 transformations are useful for S-shapes. A drawback is that these lines often fit badly in the tails of the predictor distribution, especially if the data there are sparse. As in the HDL example in Table 4.20, lower order terms are generally retained in polynomial models: specifically, we would include the linear term along with the quadratic term in the upper left panel, as well as with the quadratic plus cubic terms on the upper right.

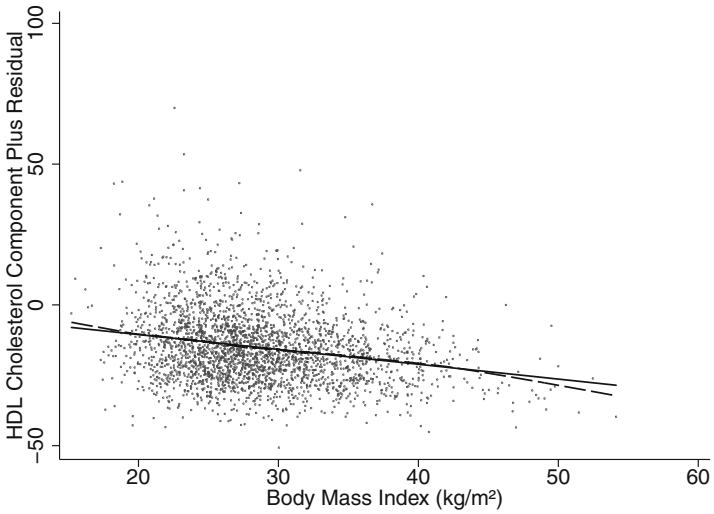

Fig. 4.5 CPR plot for HDL model with linear and quadratic terms in BMI

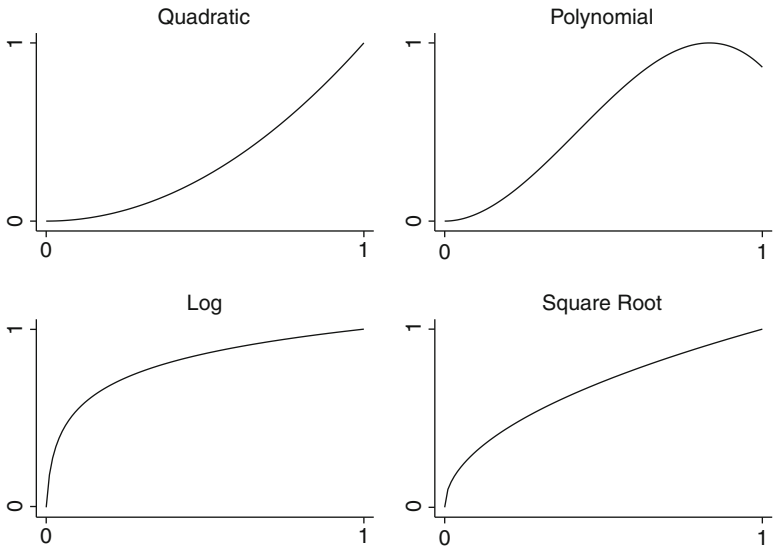

Fig. 4.6 Linearizing predictor transformations

The lower panels of Fig. 4.6 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. Each of

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

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.

4.7.1.3 Restricted Cubic Splines

Improving on the flexibility of polynomial transformations but with better behavior in the tails, *restricted cubic splines* are now implemented in Stata and other packages. This transformation requires selecting a small number of *knots*, or cutpoints, usually placed at symmetric percentiles of the predictor distribution. If there are k knots, the predictor is represented in the model by $k - 1$ spline variables. The effect of the predictor on the mean of the outcome is then modeled as cubic polynomials in the intervals between knots (achieving flexibility), is smooth at each knot (avoiding unrealistic sharp bends), but is constrained to be linear beyond the extreme knots (improving behavior in the tails). Suppose that in the model for the effect of BMI on HDL, we represent BMI by a restricted cubic spline with the default five knots. The results are shown in Table 4.21.

A primary advantage of restricted cubic splines is that the first of the $k-1$ spline variables is just the untransformed predictor, so that all nonlinearity is captured by the other $k-2$ variables. This affords a straightforward statistical test for departure from linearity, analogous to the tests for the contribution of quadratic and cubic terms in a polynomial model. The first F-test in Table 4.21 for the joint effect of the nonlinear components `BMIsp2, BMIsp3`, and `BMIsp4` confirms that the departure from linearity is important, despite the large t-test P-values. The second F-test confirms the overall importance of BMI for predicting HDL.

Another big advantage of restricted cubic splines is that graphical diagnostics for nonlinearity are considerably more difficult with the logistic, Cox, repeated measures, and GLMs presented in later chapters. However, departures from linearity can be conveniently assessed and modeled using restricted cubic splines in all of these settings.

Table 4.21 Restricted cubic spline model for effect of BMI on HDL

```
. mkspline BMIsp = BMI, cubic
. regress HDL BMIsp1 BMIsp2 BMIsp3 BMIsp4 age10 nonwhite smoking drinkany

      Source |       SS       df       MS              Number of obs =    2745
-------------+------------------------------           F(  8,  2736) =   30.35
       Model |  38913.5934     8  4864.19917            Prob > F      =  0.0000
    Residual |  438566.919  2736  160.294926            R-squared     =  0.0815
-------------+------------------------------           Adj R-squared =  0.0788
       Total |  477480.512  2744  174.008933            Root MSE      =  12.661
------------------------------------------------------------------------------
         HDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      BMIsp1 |  -1.008258   .2823244    -3.57   0.000    -1.561849   -.4546676
      BMIsp2 |   1.139488   2.424866     0.47   0.638    -3.615266    5.894242
      BMIsp3 |  -.4761041   9.557886    -0.05   0.960    -19.21751     18.2653
      BMIsp4 |  -1.757718   11.21143    -0.16   0.875    -23.74145    20.22601
       age10 |   1.882574   .3807256     4.94   0.000     1.136035    2.629113
    nonwhite |   2.469817   .7823079     3.16   0.002     .9358431    4.003791
     smoking |  -2.097091   .7452066    -2.81   0.005    -3.558315   -.6358663
    drinkany |   4.376239   .5041816     8.68   0.000     3.387624    5.364854
       _cons |    62.2474   6.939817     8.97   0.000     48.63959    75.85521
------------------------------------------------------------------------------

. * test for departure from linearity
. test BMIsp2 BMIsp3 BMIsp4
       F(  3,  2736) =    7.84
            Prob > F =    0.0000

. * test for overall effect of BMI
. test BMIsp1 BMIsp2 BMIsp3 BMIsp4
       F(  4,  2736) =   27.67
            Prob > F =    0.0000
```

The primary disadvantage of restricted cubic splines is that the numeric results for `BMIsp1, BMIsp2, BMIsp3`, and `BMIsp4` in Table 4.21 are uninterpretable. The resulting fit can only be adequately represented graphically, as in Fig. 4.7. The `adjustrcspline` command, part of the downloadable `postrcspline` package, can also be used to plot restricted cubic spline fits with CIs, for logistic and GLMs as well as standard linear models.

In addition, spline fits can be sensitive to the number of knots (Stone 1986). The flexibility of the fit increases with the number and placement of the knots, just as LOWESS smooths become more flexible with smaller bandwidths. In Stata, the default number is 5, but with datasets with fewer than 100 observations, 4 or 3 knots may work better. More than 5 knots are seldom necessary in large datasets unless the response to the predictor is unusually complicated. Plotting the fitted regression line is useful for judging the plausibility of the fit.

4.7.1.4 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

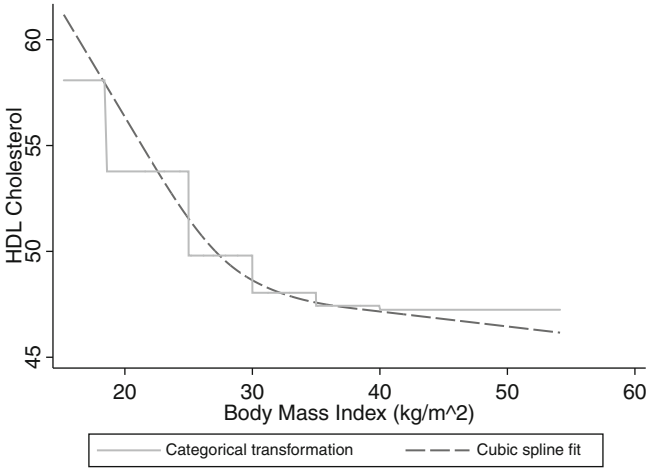

Fig. 4.7 HDL model with restricted cubic spline and categorical transformations of BMI

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. This method models the association between the ordinal categories and the outcome as a *step function*, also shown in 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, including polynomials and restricted cubic splines, are harder to motivate, present, and interpret.

4.7.1.5 Nonlinearity, Interaction, and Covariate Overlap

Apparent nonlinearity can sometimes mask interactions. For example, suppose that both the average value of a continuous predictor and its effect on the outcome differ across subgroups defined by a binary covariate. If we fail to model the interaction, the effect of the continuous predictor will appear nonlinear, even if its effects are completely linear *within* each subgroup. Furthermore, we show in Sect. 9.2.3 that unless there is considerable overlap in the values of the continuous predictor in the two subgroups—Fig. 9.1 is an extreme example—it can be difficult to distinguish non-linearity from effect modification by the covariate. This illustrates the difficulty of identifying a reasonably accurate model, especially if the sample size is small-to-moderate.

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.

4.7.2.1 Residual Plots

Various graphical methods introduced in Chap. 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.

Figure 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 dashed 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 Chap. 2, the upward curvature of the normal 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, 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.

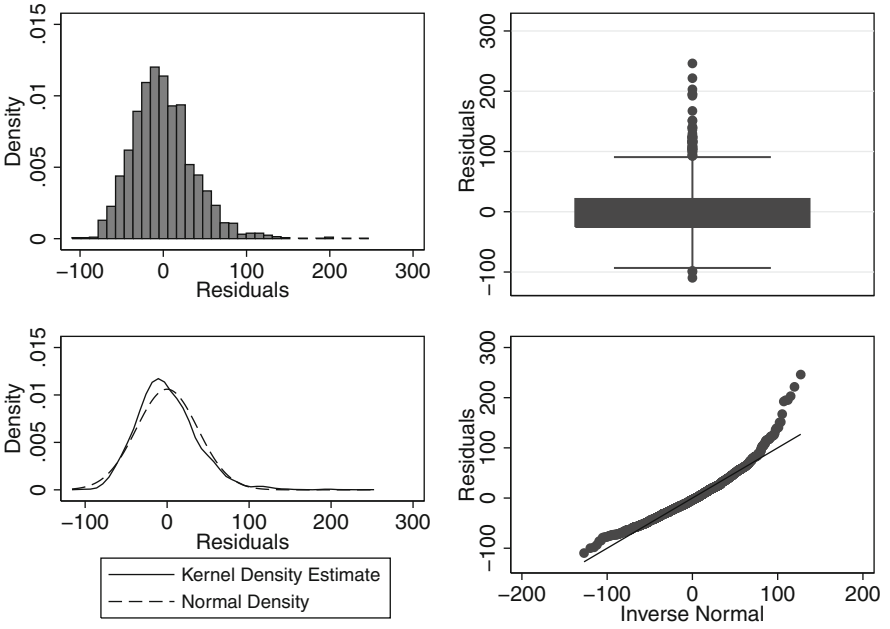

Fig. 4.8 Residuals with untransformed LDL

4.7.2.2 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.

4.7.2.3 Normalizing 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

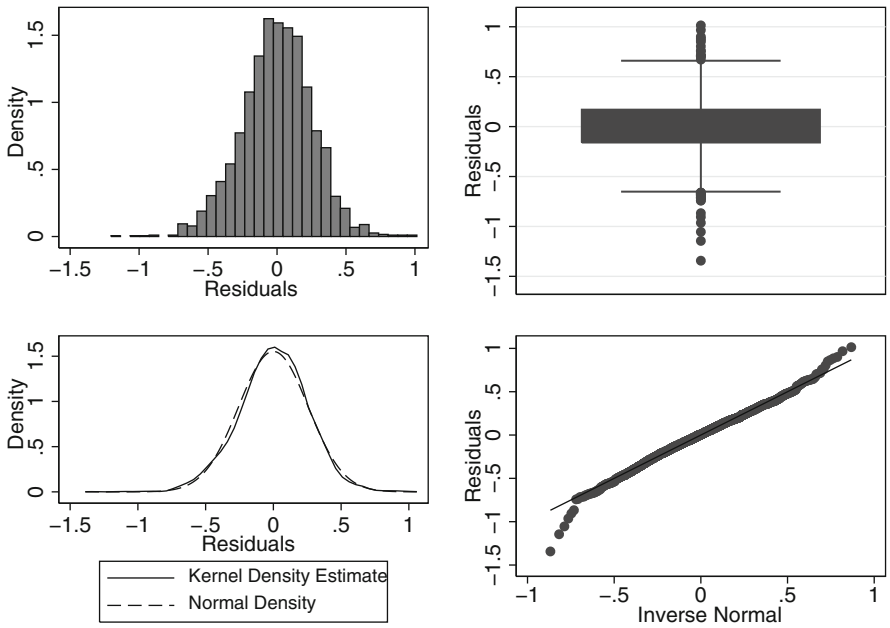

Fig. 4.9 Residuals with log-transformed LDL

positive, although this may complicate interpretation. The log transformation 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. Figure 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 nonnegative 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, providing Q–Q plots for each candidate.

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.

4.7.2.4 Alternatives to Transformation: Bootstrap and GLMs

Some outcome variables cannot be satisfactorily normalized by transformation, or there may be compelling reasons to analyze them on the original scale. Bootstrap CIs, as introduced in Sects. 3.6 and 4.5.4, are a useful alternative, implemented for most Stata procedures. We recommend use of percentile-based intervals, obtained using the `estat bootstrap` postestimation command, preferably based on 500 or more bootstrap samples, rather than the default of 50. These should be more reliable than the default intervals provided by the `vce(bootstrap)` option, which are based on the assumption that the coefficient estimate is normally distributed and use only the bootstrap estimate of the standard error.

Another good alternative is provided by the GLMs discussed in Chap. 8, in particular the gamma model, suitable for some badly skewed variables. Second-line options include dichotomizing the outcome, with analysis using logistic models, or categorizing the outcome into at least 3 ordered categories, then using proportional-odds or continuation-ratio models (Ananth and Kleinbaum 1997; Greenland 1994), as briefly described in Chap. 5.

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 CIs 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, and 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 estimates biased, some precision can be lost.

4.7.3.1 Residual Plots

Diagnostics for violations of the constant variance assumption also use the 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

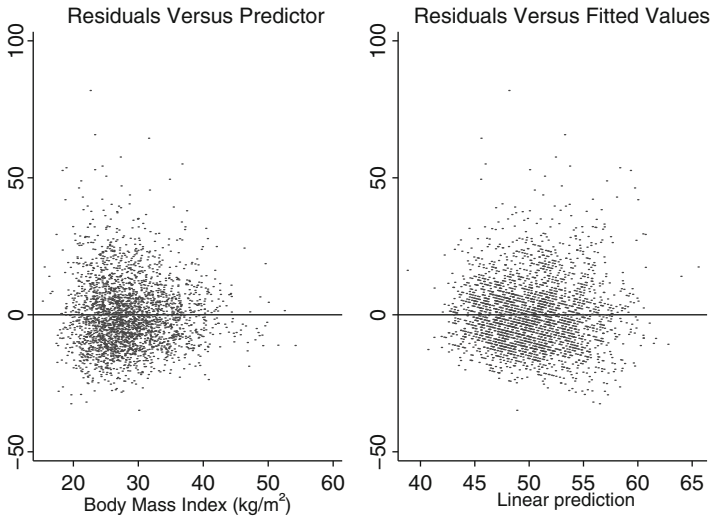

Fig. 4.10 Checking for constant residual variance

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.

Figure 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 nonconstant variance is mirrored in the slightly wider spread on the right in the facing plot of the residuals against the fitted values.

4.7.3.2 Subsample 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 versus 90). Even after log transformation of glucose, the variance would still have differed by a factor of

10 (0.097 versus 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.

4.7.3.3 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.

4.7.3.4 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 of 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.

4.7.3.5 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, nonconstant 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 nonconstant variance we found in Fig. 4.10. However, in this case our qualitative conclusions would be unchanged by log transformation of HDL.

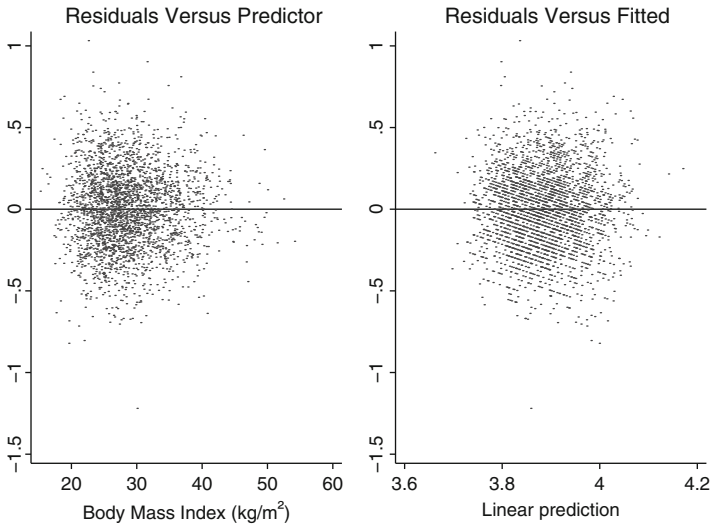

Fig. 4.11 Rechecking constant variance after log-transforming HDL

4.7.3.6 Robust Standard Errors

So-called robust or "sandwich" standard errors (Huber 1967), available with many Stata regression procedures using the option `vce(robust)`, are another convenient means of dealing with nonconstant residual variance. This method will provide more reliable inferences when the constant-variance assumption is violated, provided the model for $E[y|\mathbf{x}]$ is approximately correct. However, some caution is warranted in using these standard errors in smaller samples. In extensive simulations, Long and Ervin (2000) show that robust standard errors can be too small in samples as large as 250 observations. They find that a more conservative alternative developed by MacKinnon and White (1985) has the best properties; this can be specified using the option `vce(hc3)` with the `regress` command. Table 4.22 shows linear models for glucose levels, successively estimated using model-based, robust, and HC3 standard errors. While the very large difference in glucose levels according to diabetes status is unambiguous, even in this small sample, the robust standard errors are considerably larger. Moreover, evidence for the adverse effect of BMI appears considerably weaker with the more conservative robust SEs.

4.7.3.7 GLMs

GLMs are another important alternative when transformation of the outcome fails to rectify substantial violations of the assumption of constant variance. For example,

Table 4.22 Models with conventional, robust, and HC3 standard errors

```
. regress glucose diabetes BMI age drinkany
      Source |       SS       df       MS              Number of obs =     137
-------------+------------------------------           F(  4,   132) =   37.43
       Model |  84874.7167     4  21218.6792            Prob > F      =  0.0000
    Residual |  74823.7504   132  566.846594            R-squared     =  0.5315
-------------+------------------------------           Adj R-squared =  0.5173
       Total |  159698.467   136  1174.25343            Root MSE      =  23.809

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   50.64445   4.585857    11.04   0.000     41.57318    59.71573
         BMI |   1.033281   .3662364     2.82   0.006     .3088297    1.757733
             ......
------------------------------------------------------------------------------

. regress glucose diabetes BMI age drinkany, vce(robust)
Linear regression                                       Number of obs =     137
                                                        F(  4,   132) =   19.32
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.5315
                                                        Root MSE      =  23.809

------------------------------------------------------------------------------
             |               Robust
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   50.64445   6.527487     7.76   0.000     37.73244    63.55647
         BMI |   1.033281   .4967385     2.08   0.039     .0506837    2.015879
             ......
------------------------------------------------------------------------------

. regress glucose diabetes BMI age drinkany, vce(hc3)
Linear regression                                       Number of obs =     137
                                                        F(  4,   132) =   17.96
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.5315
                                                        Root MSE      =  23.809

------------------------------------------------------------------------------
             |             Robust HC3
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   50.64445   6.715182     7.54   0.000     37.36116    63.92775
         BMI |   1.033281   .5244014     1.97   0.051    -.0040363    2.070599
             ......
------------------------------------------------------------------------------
```

Poisson and negative binomial models have now mostly taken the place of linear models for count outcomes using the variance-stabilizing square root transformation. In GLMs, including the logistic model (Chap. 5), the variance of the outcome is modeled as a function of its mean (Table 8.8); in the Poisson model, for example, the variance is assumed *equal* to the mean. Furthermore, the mean-variance assumption can be relaxed using variants of these models allowing for so-called *overdispersion*, or using robust standard errors, as just described.

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.

> *Definition*: *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 nonsignificant, 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. 10.1), 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 its statistical significance are almost unchanged 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

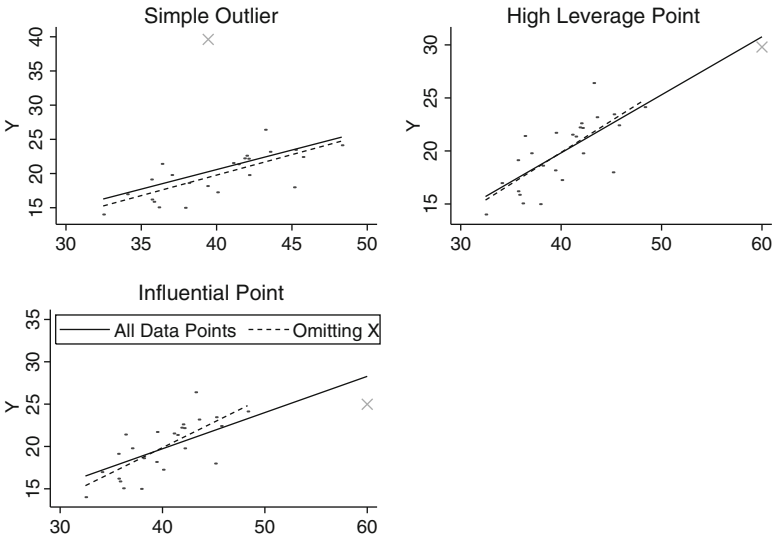

Fig. 4.12 Outlying, high-leverage, and influential points

make a nonsignificant 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.

4.7.4.1 DFBETAs

To check for sensitivity of the conclusions of an analysis 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.

Boxplots are convenient for identifying a small set of extreme outliers among the DFBETA values for each predictor. DFBETAs often have a very small interquartile 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. Figure 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.

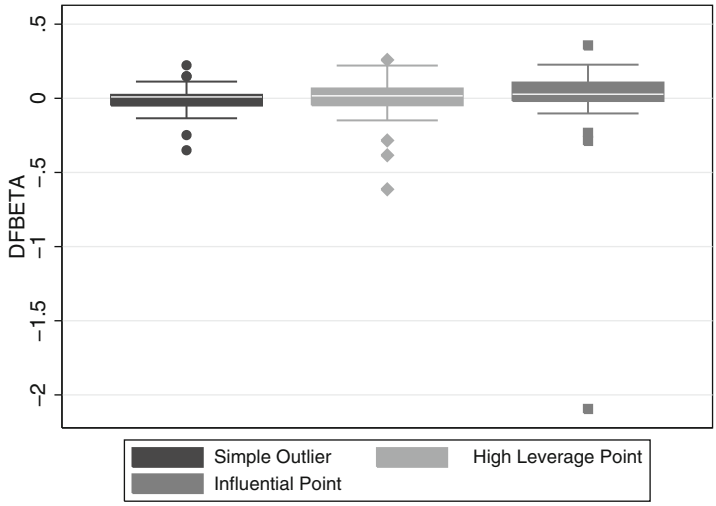

Fig. 4.13 DFBETAs for data sets shown in Fig. 4.12

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 estimated prediction error (Sect. 10.1). 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.

Figure 4.14 above 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.

The sensitivity of model results to the omission of these four points is summarized in Table 4.23. 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.

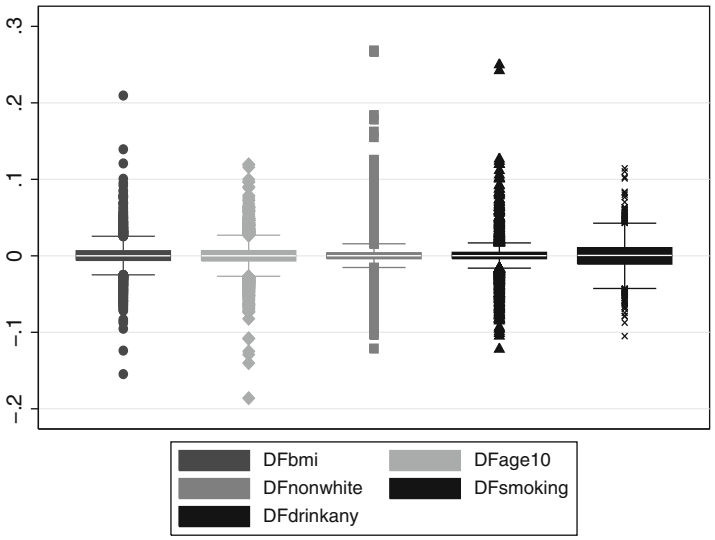

Fig. 4.14 DFBETAs for LDL model

Table 4.23 Sensitivity of LDL model to omission of four most influential points

	All observations			Omitting four observations		
Predictor 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	0.081
Alcohol use	−2.72	−5.66, 0.22	0.069	−2.64	−5.51, 0.23	0.072

A 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.

4.7.4.2 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 western collaborative group study (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.

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.

4.7.5.1 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 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.

4.7.5.2 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 interpretation. 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.

4.7.5.3 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 Sample Size, Power, and Detectable Effects

Section 4.2.2 presented the t-test of the null hypothesis $\beta_j = 0$, in which we compare $\hat{\beta}_j / \mathrm{SE}(\hat{\beta}_j)$ to the t-distribution with $n - (p+1)$ degrees of freedom. This test leads directly to methods for estimating sample size and power for analyses using the linear model. Suppose we would like to calculate the sample size that would provide power of γ to reject $\beta_j = 0$ in a two-sided test with type-I error rate α, under the alternative hypothesis $\beta_j = \beta_j^a$, assuming for now that $\beta_j^a > 0$. We begin with an expression for power, relying on the large-sample equivalence of the t and standard normal Z-distributions:

$$\begin{aligned}
\gamma &= \mathrm{P}\left[|\hat{\beta}_j|/\mathrm{SE}(\hat{\beta}_j) > z_{1-\alpha/2}\right] \\
&\approx \mathrm{P}\left[\hat{\beta}_j/\mathrm{SE}(\hat{\beta}_j) > z_{1-\alpha/2}\right] \\
&= \mathrm{P}\left[(\hat{\beta}_j - \beta_j^a)/\mathrm{SE}(\hat{\beta}_j) > z_{1-\alpha/2} - \beta_j^a/\mathrm{SE}(\hat{\beta}_j)\right] \\
&= 1 - \Phi\left[z_{1-\alpha/2} - \beta_j^a/\mathrm{SE}(\hat{\beta}_j)\right] \\
&= \Phi\left[\beta_j^a/\mathrm{SE}(\hat{\beta}_j) - z_{1-\alpha/2}\right]. \qquad (4.14)
\end{aligned}$$

In (4.14), $|\cdot|$ denotes absolute value; $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution (1.96 for a two-sided test with type-I error rate of 5%); and $\Phi(\cdot)$ is the cumulative distribution function for a standard normal variate Z, so that $\Phi(z_{1-\alpha/2}) = \mathrm{P}(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$. The first approximation in (4.14) holds because if β_j is positive, $\mathrm{P}(\hat{\beta}_j/\mathrm{SE}(\hat{\beta}_j) < z_{\alpha/2}) \approx 0$. The second step is simple algebra. The third follows because $(\hat{\beta}_j - \beta_j^a)/\mathrm{SE}(\hat{\beta}_j)$ has an approximate Z-distribution in large samples, and the fourth because of the symmetry of the Z-distribution about zero. Using (4.4) (with n in place of $n-1$) to evaluate $\mathrm{SE}(\hat{\beta}_j)$, then applying the inverse transformation Φ^{-1} to both sides of (4.14), and solving for n gives

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \sigma^2_{y|\mathbf{x}}}{(\beta_j^a \sigma_{x_j})^2 (1 - \rho_j^2)}. \qquad (4.15)$$

In (4.15), z_γ is the quantile of the standard normal distribution for power (0.84 for 80% power, 1.28 for 90%), $\sigma^2_{y|\mathbf{x}}$ is the residual variance of the outcome, σ_{x_j} is the standard deviation of X_j, and ρ_j is its multiple correlation with the other covariates. The variance inflation factor $1/(1-\rho_j^2)$ in (4.15) accounts for the potential loss of precision due to the inclusion of other predictors in the model (Hsieh et al. 1998).

In some problems, including secondary analyses of existing data, n is fixed. In that case, (4.15) can be solved to calculate power, if we specify β_j^a:

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a|\sigma_{x_j}\sqrt{n(1-\rho_j^2)}/\sigma_{y|\mathbf{x}}\right]. \tag{4.16}$$

Similarly, we can calculate the *minimum detectable effect*—that is, the smallest value of β_j^a for which a sample of size n would provide power of γ to reject the null hypothesis $\beta_j = 0$ in a two-sided test with type-I error of α. The minimum detectable effect is

$$\pm\,\beta_j^a = \frac{(z_{1-\alpha/2} + z_\gamma)\sigma_{y|\mathbf{x}}}{\sigma_{x_j}\sqrt{n(1-\rho_j^2)}}. \tag{4.17}$$

Some additional points:

- When X_j is binary with prevalence f_j, $\sigma_{x_j} = \sqrt{f_j(1-f_j)}$ in (4.15)–(4.17).
- When X_j is continuous with standard deviation σ_{x_j}, it is important to recognize that sample size, power, and minimum detectable effects do not depend in any real way on the units in which X_j is measured. This is most clearly seen in (4.17). Suppose X_j is usually measured in grams. Changing the unit to milligrams increases σ_{x_j} by a factor of 1,000, and shrinks β_j^a by the same factor. But of course the effect on the outcome of a 1-milligram increase in the predictor is 1,000 times smaller than the effect of a 1-gram increase. One way to avoid confusion is to consider the minimum detectable effect size for a one standard deviation change in X_j, which is often a reasonable-sized change to consider. That effect size is obtained by setting $\sigma_{x_j} = 1$ in (4.17).
- If $\beta_j^a < 0$ under the alternative, we have to use $|\beta_j^a|$ in (4.16) to get the correct result. It follows that the negative of the value given by (4.17) is also a valid solution for the minimum detectable effect.
- Because they are based on the standard normal distribution, (4.15)–(4.17) are only approximate. Exact solutions involve the noncentral t-distribution and iterative calculations. Numerous packages supply these estimates for small as well as large sample sizes; the `sampsi` and `sampsi_reg` commands in Stata work for binary and continuous predictors respectively. An approximate correction is to add 2 to the estimate of n provided by (4.15) for tests with α of 5%, and add 4 with α of 1% (Snedecor and Cochran 1989, page 104). The correction can be important when $n < 50$ and especially when $n < 25$.
- Sample size (4.15) and minimum detectable effect (4.17) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j, and we do not need to penalize for covariate adjustment. In that standard case,

$$n = 7.849 \times \sigma_{y|x}^2/(\beta_j^a)^2. \tag{4.18}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2.802 \times \sigma_{y|x}/\sqrt{n}. \tag{4.19}$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a 2-arm clinical trial with equal allocation to arms, so that β_j^a is the between-group difference in means and $s_{x_j}^2 = 0.25$, we can calculate

$$n = 4 \times 7.849 \times \sigma_{y|x}^2/(\beta_j^a)^2. \tag{4.20}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2 \times 2.802 \times \sigma_{y|x}/\sqrt{n}. \tag{4.21}$$

- Power calculations using (4.16) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- The Stata commands `sampsi` and `sampsi_reg` can also be used to compute power, but not minimum detectable effects.
- In using sample size calculators that do not allow for covariate adjustment, including the `sampsi` and `sampsi_reg` commands, the unadjusted sample size estimate should be inflated by $1/(1-\rho_j^2)$; similarly, the minimum detectable effect estimate should be inflated by $\sqrt{1/(1-\rho_j^2)}$. To calculate power, use $n(1-\rho_j^2)$ in place of n as an input.
- For the linear model, the proposed adjustment may be conservative, since adjustment for covariates will also reduce the residual variance $\sigma_{y|\mathbf{x}}^2$, to some extent offsetting the loss of precision due to the correlation ρ_j between X_j and the other covariates. This is particularly relevant in calculations for stratified randomized trials with continuous outcomes, since the stratification factor may account for a large proportion of the variance of the outcome, but is in expectation uncorrelated with treatment assignment.

To illustrate these calculations, suppose we are planning a randomized trial with equal allocation to active treatment and control ($f = 0.5$) to assess the effect of a new lipid-lowering agent on LDL levels. From pilot data, the residual standard deviation $\sigma_{y|x}$ for LDL is expected to be ≈ 38 mg/dL, and we hypothesize that the agent will lower average LDL levels about 40 mg/dL. Because this is a clinical trial, it is unlikely that we will need to adjust for covariates, so we can assume $\rho_j = 0$. The sample size must provide 80% power in a two-sided test with α of 5%.

We first calculate the sample size using the `sampsi` command in Stata, then using its capacity as a calculator to evaluate (4.15). Table 4.24 shows the results. In using `sampsi`, any values of the means for populations 1 and 2 that differ by 40 mg/dL would give the same answer, so for convenience we used 0 and 40. With the Snedecor and Cochran correction, using Stata to evaluate (4.15) gives about the same result as `sampsi`.

Table 4.24 Sample size calculations for a small clinical trial

```
. sampsi 0 40, sd1(38) alpha(0.05) power(0.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =        0
            m2 =       40
           sd1 =       38
           sd2 =       38
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =       15
            n2 =       15

. * solution using Snedecor and Cochran correction
. display (invnormal(.975)+invnormal(.8))^2*38^2/(40^2*0.5*(1-0.5))+2
30.334456
```

When the predictor of interest is continuous, we can use the downloadable `sampsi_reg` command in Stata. Suppose, for example, that we would like to estimate the power of a study with 485 participants to detect an effect of higher BMI on SBP, controlling for age, race/ethnicity, smoking, alcohol use, and physical activity levels. From pilot data, we estimate that $\sigma_{y|\mathbf{x}} \approx 18.5$ mmHg, $\sigma_x \approx 5.5$ kg/m^2, and $\rho_j \approx 0.33$. We hypothesize that average SBP increases 0.5 mmHg for every kg/m^2 increase in BMI—that is, $\beta_j^a = 0.5$. What is the power of the study to detect this effect of BMI on SBP in a two-sided test with α of 5%?

Table 4.25 shows results of the computation using `sampsi_reg` in Stata, as well as a direct implementation of (4.16). Since `sampsi_reg` does not allow for the adjustment based on the variance inflation factor, we first deflate the available sample size by $1 - \rho_j^2$. The two estimates of power are in close agreement.

4.8.1 Calculations Using Standard Errors Based on Published Data

Equations (4.15)–(4.17) depend on $\sigma_{y|\mathbf{x}}$, σ_{x_j}, and ρ_j, for which it may be hard to obtain estimates. However, the derivation using (4.4) suggests a solution. Suppose an estimate $\tilde{\text{SE}}(\hat{\beta}_j)$ for the standard error of $\hat{\beta}_j$ is available, based on a multiple linear regression model with appropriate covariates and estimated using $\tilde{n}$ observations. For example, we could compute $\tilde{\text{SE}}(\hat{\beta}_j)$ from a published article as

Table 4.25 Power calculation for independent effect of BMI on SBP

```
. display 485*(1-.33^2)
432.1835
. sampsi_reg, alt(0.5) n1(432.1835) s(power) sx(5.5) sd1(18.5)

Estimate power for linear regression
Test Ho: Alt. Slope = Null Slope, usually Null Slope is 0

Assumptions:
          Alpha =    0.0500  (two-sided)
              N =  432.1835
     Null Slope =    0.0000
      Alt Slope =    0.5000
    Residual sd =   18.5000
      SD of X's =    5.5000

Estimated power:
       Power = .86934271

. display 1-normal(invnormal(0.975)-0.5*5.5*sqrt(485*(1-.33^2))/18.5)
.8708243
```

the width of the 95% CI for $\hat{\beta}_j$, divided by $2z_{.975} \approx 3.92$. Care must taken to ensure that the hypothesized value of β_j^a corresponds to the same measurement scale for X_j as in the source article. Then, (4.15) can be simplified as

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} \left[\tilde{\mathrm{SE}}(\hat{\beta}_j)\right]^2}{(\beta_j^a)^2}. \tag{4.22}$$

Similarly, power in a new sample of size n is given by

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a|/[\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j)]\right]. \tag{4.23}$$

Finally, the minimum detectable effect in a new sample of size n can be obtained as

$$\pm\, \beta_j^a = (z_{1-\alpha/2} + z_\gamma)\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j). \tag{4.24}$$

As an example, we could use the multiple linear model in Table 4.2 to obtain sample size, power, and minimum detectable effect estimates for a new study of the effect of BMI on glucose levels in nondiabetic women. Based on the HERS data with $\tilde{n} = 2028$, $\tilde{\mathrm{SE}}(\hat{\beta}_j) = (0.5707328 - 0.4077512)/3.92 \approx 0.0415528$. Suppose we hypothesize that glucose levels increase 0.5 mg/dL for each kg/m^2 increase in BMI, so $\beta_j^a = 0.5$.

In Table 4.26, we first use (4.22) to estimate that a new sample of 147 participants would provide 90% power in a 2-sided test with α of 5% to detect the hypothesized increase in glucose of 0.5 mg/dL for each kg/m^2 increase in BMI. Then, using (4.23), we find that a sample of 200 participants would provide almost 97% power to detect

Table 4.26 Calculations based on regression output

```
. * sample size for a new study providing 90% power
. display (invnormal(.975)+invnormal(.9))^2*2028*0.0415528^2/0.5^2
147.17185

. * power in a new study with 200 participants
. display 1-normal(invnormal(0.975)-0.5/(sqrt(2028/200)*0.0415528))
.96552967

. * minimum effect detectable with 80% power in a new study with 100 participants
. display (invnormal(.975)+invnormal(.8))*sqrt(2028/100)*0.0415528
.5242496
```

the hypothesized effect. Finally, using (4.24) suggests that a smaller sample of 100 participants would provide 80% power to detect a minimum effect of 0.52 mg/dL for each kg/m^2 increase in BMI.

4.9 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. To avoid misleading conclusions, it is important to check assumptions, including normality of the residuals, especially in small samples; transformations of the outcome, bootstrapping, and GLMs can be used to address violations. Nonconstant variance of the residuals is a potentially serious concern even in large samples, but can be resolved using robust standard errors. As with the models discussed in later chapters, nonlinear effects of continuous predictors can be accommodated using predictor transformations, including restricted cubic splines, and interactions modeled using product terms. Finally, it is important to recognize outcomes for which linear regression is not appropriate; these include binary, time-to-event, count, and repeated measures or clustered outcomes, and are addressed in subsequent chapters.

4.10 Further Notes and References

For more detailed information on the linear regression model, first-rate books include Weisberg (1985) and Draper and Smith (1981). 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.

4.10.1 Generalized Additive Models

Methods have also been developed for fitting linear as well as logistic (Chap. 5) and other GLMs (Chap. 8) 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 Chap. 6.

4.11 Problems

Problem 4.1. Using the 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`:

```
. regress 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. 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 4.5. 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 (4.11)–(4.13). Explain why the statin use variable needs to be included in this model.

Problem 4.6. 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. 4.6.2 allowing for interaction in predicting LDL?

Problem 4.7. 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 4.8. 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 4.9. 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 4.10. 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 substantive interest?

Problem 4.11. 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 4.12. Suppose you have used a restricted cubic spline to model a nonlinear response to your predictor of primary interest, similar to one of the models for HDL in Fig. 4.7. Figure out how to use the spline basis variables, which in Stata would be made by the `mkspline` command, and corresponding regression coefficients to plot the shape of the response estimated by the regression model.

Problem 4.13. 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 4.14. 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?

4.12 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 the coefficients of the model.
(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 including regression coefficient estimates, hypothesis tests, CIs, and statistics which quantify the fit of the model.
(10) Interpret regression coefficients when the predictor, outcome, or both are log transformed.