Chapter 5
Logistic Regression

Patients testing positive for a sexually transmitted disease at a clinic are compared to patients with negative tests to investigate the effectiveness of a new barrier contraceptive. One-month mortality following coronary artery bypass graft surgery is compared in groups of patients receiving different dosages of beta blockers. Many clinical and epidemiological studies generate outcomes which take on one of two possible values, reflecting presence/absence of a condition or characteristic at a particular time, or indicating whether a response occurred within a defined period of observation. In addition to evaluating a predictor of primary interest, it is important to investigate the importance of additional variables that may influence the observed association and therefore alter our inferences about the nature of the relationship. In evaluating the effect of contraceptive use in the first example, it would be clearly important to control for age in addition to behaviors potentially linked to infection risk. In the second example, a number of demographic and clinical variables may be related to both the mortality outcome and treatment regime. Both of these examples are characterized by binary outcomes and multiple predictors, some of which are continuous.

Methods for investigating associations involving binary outcomes using contingency table methods were briefly covered in Sect. 3.4. Although these techniques are useful for exploratory investigations, and in situations where the number of predictor variables of interest is limited, they can be cumbersome when multiple predictors are being considered. Further, they are not well suited to situations where predictor variables may take on a large number of possible values (e.g., continuous measurements). Similar to the way linear regression techniques expanded our arsenal of tools to investigate continuous outcomes, the logistic regression model generalizes contingency table methods for binary outcomes. In this chapter, we cover the use of the logistic model to analyze data arising in clinical and epidemiological studies. Because the basic structure of the logistic model mirrors that of the linear regression model, many of the techniques for model construction, interpretation, and assessment will be familiar from Chap. 4.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_5,

5.1 Single Predictor Models

Recall the example in Sect. 3.4 investigating the association between CHD and age for the WCGS. Table 5.1 summarizes the observed proportions (P) of CHD diagnoses for five categories of age, along with the estimated risk difference (RD), relative risk (RR), and odds ratio (OR). The last three measures are computed according to procedures described in Sect. 3.4, using the youngest age group as the baseline category. The estimates show a tendency for increased risk of CHD with increasing age. Although this information provides a useful summary of the relationship between CHD risk and age, the choice of five-year categories for age is arbitrary. A regression representation of the relationship would provide an attractive alternative and obviate the need to choose categories of age.

Recall that in standard linear regression, we modeled the average of a continuous outcome variable y as a function of a single continuous predictor x using a linear relationship of the form

$$\mathrm{E}\,[y|x] = \beta_0 + \beta_1 x.$$

We might be tempted to use the same model for a binary outcome variable. First, note that if we follow convention and code the values of a binary outcome as one for those experiencing the outcome and zero for everyone else, the observed proportion of outcomes among individuals characterized by a particular value of x is simply the mean (or "expected value") of the binary outcome in this group. In the notation introduced in Sect. 3.4, we symbolize this quantity by $P(x)$. The linear model for our binary outcome might then be expressed as

$$P(x) = \mathrm{E}\,[y|x] = \beta_0 + \beta_1 x. \tag{5.1}$$

This has exactly the same form as the linear regression model; the expected value of the outcome is modeled as a linear function of the predictor. Further, changes in the outcome associated with specified changes in the predictor x have a risk difference interpretation: For example, if x is a binary predictor taking on the values 0 or 1, the effect of increasing x one unit is to add an increment β_1 to the outcome. From (5.1),

$$P(1) - P(0) = \beta_1.$$

Referring back to Definition (3.14) in Sect. 3.4, we see that this is the risk difference associated with a unit increase in x. Models with this property are often referred to as *additive risk models* (Clayton and Hills 1993).

Table 5.1 CHD for five age categories in the WCGS sample

Age group	P	$1-P$	RD	RR	OR
35–40	0.057	0.943	0.000	1.000	1.000
41–45	0.050	0.950	−0.007	0.883	0.877
46–50	0.093	0.907	0.036	1.635	1.700
51–55	0.123	0.877	0.066	2.156	2.319
56–60	0.149	0.851	0.092	2.606	2.886

There are several limitations with the linear model (5.1) as a basis for regression analysis of binary outcomes. First, the statistical machinery which allowed us to use this linear model to make inferences about the strength of relationship in Chap. 4 required that the outcome variable follow an approximate normal distribution. For a binary outcome, this assumption is clearly incorrect. Second, the outcome in the above model represents a probability or risk. Thus, any estimates of the regression coefficients must constrain the estimated probability to lie between zero and one for the model to make sense. The first of these problems is statistical, and addressing it would require generalizing the linear model to accommodate a distribution appropriate for binary outcomes. The second problem is numerical. To ensure sensible estimates, our estimation procedure would have to satisfy the constraints mentioned.

Another issue is that in many settings, it seems implausible that outcome risk would change in a strictly linear fashion for the entire range of possible values of a continuous predictor x. Consider a study examining the likelihood of a toxicity response to varying levels of a treatment. We would not expect the relationship between likelihood of toxicity and dose to be strictly linear throughout the range of possible doses. In particular, the likelihood of toxicity should be zero in the absence of treatment and increase to a maximum level, possibly corresponding to the proportion of the sample susceptible to the toxic effect, with increasing dose.

Figure 5.1 presents four hypothetical models linking the probability $P(x)$ of a binary outcome to a continuous predictor x. In addition to the linear model (**a**), there is the exponential model (**b**) that constrains risk to increase exponentially with x, the "step function" model (**c**) that allows irregular (but piecewise-constant) change in risk with increasing values of x, and the smooth S-shaped curve in (**d**) known as the *logistic* model. The exponential model is also known as *log linear* because it specifies that the logarithm of the outcome risk is linear in x. It presents a problem similar to that noted for the linear model above: Namely, that risk is not obviously constrained to be less than one for large values of $\beta_0 + \beta_1 x$. The outcome probabilities for model (**c**) simply represent the estimated proportion of positive outcomes in each group specified by the categories of x, and has the desirable properties that risks are clearly constrained to fall in the interval [0, 1], and that the nature of the increase in the interval can be flexibly represented by different "step" heights. However, it lacks smoothness, a property that is biologically plausible in many instances. In addition, the choice of break points delineating the changes in risk is subjective. By contrast, the logistic model allows for a smooth change in risk throughout the range of x, and has the property that risk increases slowly up to a "threshold" range of x, followed by a more rapid increase and a subsequent leveling off of risk. This shape is consistent with many dose-response relationships (illustrated by the toxicity example from the previous paragraph). As we will see later in this chapter, all of these models represent valid alternatives for assessing how risk of a binary outcome changes with the value of a continuous predictor. However, most of our focus will be on the logistic model.

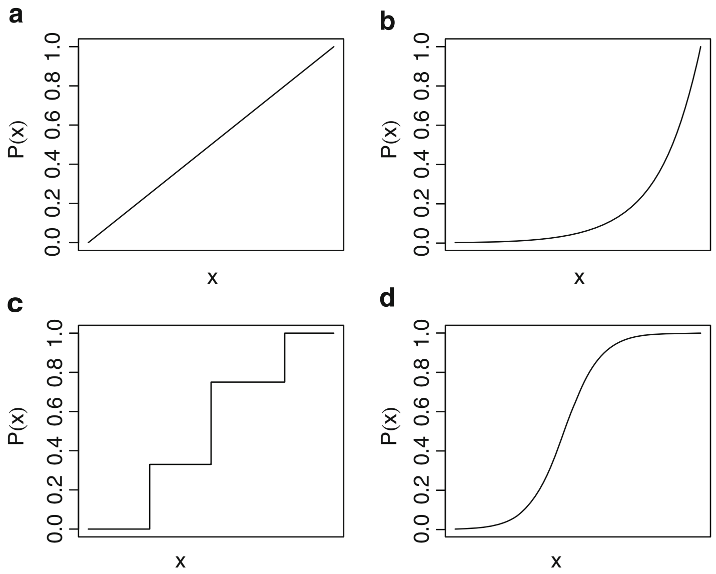

Fig. 5.1 Risk models for a binary outcome and continuous predictor (**a**) Linear (**b**) Exponential (**c**) Step function (**d**) Logistic

In addition to a certain degree of biological plausibility, the logistic model does not pose the numerical difficulties associated with the linear and log-linear models, and has a number of other appealing properties that will be described in more detail below. For these reasons, it is by far the most widely used model for binary outcomes in clinical and epidemiological applications, and forms the basis of logistic regression modeling. However, adoption of the logistic model still implies strong assumptions about the relationship between outcome risk and the predictor. In fact, expressed on a transformed scale, the model prescribes a linear relationship between the logarithm of the odds of the outcome and the predictor.

The logistic model plotted in Fig. 5.1d is defined by the equation

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{5.2}$$

In terms of the odds of the outcome associated with the predictor x, the model can also be expressed as

$$\frac{P(x)}{1 - P(x)} = \exp(\beta_0 + \beta_1 x). \tag{5.3}$$

Consider again the simple case where x takes on the values 0 or 1. From the last equation, the ratio of the odds for these two values of x are

$$\frac{P(1)/\left[1 - P(1)\right]}{P(0)/\left[1 - P(0)\right]} = \exp(\beta_1). \tag{5.4}$$

Expressed in this form, we see that the logistic model specifies that the ratio of the odds associated with these two values of x is given by the factor $\exp(\beta_1)$. Equivalently, the odds for $x = 1$ are obtained by multiplying the odds for $x = 0$ by this factor. Because of this property, the logistic model is an example of a *multiplicative risk model* (Clayton and Hills 1993). (Note that the log-linear model is also multiplicative in this sense, but is based on the outcome risks rather than the odds.)

Although not easily interpretable in the form given in (5.2) and (5.3), expressed as the logarithm of the outcome odds (as given in (5.3)), the model becomes linear in the predictor

$$\log\left[\frac{P(x)}{1 - P(x)}\right] = \beta_0 + \beta_1 x. \tag{5.5}$$

This model states that the log odds of the outcome is linearly related to x, with intercept coefficient β_0 and slope coefficient β_1 (i.e., the logistic model is an additive model when expressed on the log odds scale). The logarithm of the outcome odds is also frequently referred to as the *logit* transformation of the outcome probability.

In the language introduced in Chaps. 3 and 4, (5.2), (5.3), and (5.5) define the systematic part of the logistic regression model, linking the average $P(x)$ of the outcome variable y to the predictor x. The random part of the model specifies the distribution of the outcome variable y_i, conditional on the observed value x_i of the predictor (where the subscript i denotes the value for a particular subject). For binary outcomes, this distribution is called the *binomial* distribution and is completely specified by the mean of y_i conditional on the value x_i. To summarize, the logistic model makes the following assumptions about the outcome y_i:

(1) y_i follows a Binomial distribution.
(2) The mean $\mathrm{E}\left[y|x\right] = P(x)$ is given by the logistic function (5.2).
(3) Values of the outcome are statistically independent.

These assumptions closely parallel those associated with the linear regression (in Sect. 3.3), the primary difference being the use of the binomial distribution for the outcome y. Note that the assumption of constant variance of y across different values of x is not required for the logistic model. Another difference is that the random aspect of the logistic model is not included as an additive term in the regression equation. However, it is still an integral part of estimation and inference regarding model coefficients. (This is discussed further in Sect. 5.6.)

As we will see in the rest of this chapter, both of the alternative expressions (5.2) and (5.5) for the logistic model are useful: the linear logistic form (5.5) is the basis

Table 5.2 Logistic model for the relationship between CHD and age

```
. logistic chd69 age, coef

Logit estimates                                   Number of obs   =       3154
                                                  LR chi2(1)      =      42.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.17806                       Pseudo R2       =     0.0241

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0744226   .0113024     6.58   0.000      .0522703    .0965748
       _cons |  -5.939516    .549322   -10.81   0.000     -7.016167   -4.862865
------------------------------------------------------------------------------
```

for regression modeling, while the (nonlinear) logistic form (5.2) is useful when we want to express the outcome on its original scale (e.g., to estimate outcome risk associated with a particular value of x).

One of the most significant benefits of the linear logistic formulation (5.5) is that the regression coefficients are interpreted as log odds ratios. These can be expressed as odds ratios via simple exponentiation (as demonstrated above in (5.4)), providing a direct generalization of odds ratio methods for frequency tables to the regression setting. This property follows directly from the definition of the model, and is demonstrated in the next section. Finally, we note that there are a number of alternative regression models for binary outcomes that share similar properties to the logistic model. Although none of these comes close to the logistic model in terms of popularity, they offer useful alternatives in some situations. Some of these will be discussed in Sect. 5.5.

5.1.1 Interpretation of Regression Coefficients

Table 5.2 shows the fit of the logistic model (5.5) for the relationship between CHD risk and age in the WCGS study. The coefficient labeled _cons in the table is the intercept (β_0), and the coefficient labeled age is the slope (β_1) of the fitted logistic model. Since the outcome for the model is the log odds of CHD risk, and the relationship with age is linear, the slope coefficient β_1 gives the change in the log odds of chd69 associated with a one-year increase in age. We can verify this by using the formula for the model (5.5) and the estimated coefficients to calculate the difference in risk between a 56- and a 55-year-old individual:

$$\log\left[\frac{P(56)}{1-P(56)}\right] - \log\left[\frac{P(55)}{1-P(55)}\right]$$
$$= (-5.940 + 0.074 \times 56) - (-5.940 + 0.074 \times 55) = 0.074.$$

This is just the coefficient β_1 as expected; performing the same calculation on an arbitrary one-year age increase would produce the same result (as shown at the end of this section). The corresponding odds ratio for any one-year increase in `age` can then be computed by simple exponentiation:

$$\exp(0.074) = 1.077.$$

This odds ratio indicates a small (approximately 8%) but statistically significant increase in the odds of CHD for each one-year age increase. We can estimate the (clinically more relevant) odds ratio associated with a ten-year increase in age the same way, yielding:

$$\exp(0.074 \times 10) = 2.105.$$

Following the same approach we can use (5.5) to calculate the log odds ratio and odds ratio for an arbitrary $\varDelta$ unit increase in a predictor x as follows:

$$\log\left[\frac{\frac{P(x+\varDelta)}{1-P(x+\varDelta)}}{\frac{P(x)}{1-P(x)}}\right] = \beta_1\varDelta, \quad \frac{\frac{P(x+\varDelta)}{1-P(x+\varDelta)}}{\frac{P(x)}{1-P(x)}} = \exp(\beta_1\varDelta). \tag{5.6}$$

In addition to computing odds ratios, the estimated coefficients can be used in the logistic function representation of (5.2) to estimate the probability of having CHD during study follow-up for a individual with any specified age. For a 55-year-old individual:

$$P(55) = \frac{\exp(-5.940 + 0.074 \times 55)}{1 + \exp(-5.940 + 0.074 \times 55)}.$$

Of course, such an estimate only makes sense for ages near the values used in fitting the model.

The output in Table 5.2 also gives standard errors and 95% CIs for the model coefficients. The interpretation of these is the same as for the linear regression model. The fact that the interval for the coefficient of `age` excludes zero indicates statistically significant evidence that the true coefficient is different than zero. Similar to linear regression, the ratio of the coefficients to their standard errors forms the Wald (`z`) test statistic for the hypothesis that the true coefficients are different than zero. This statistic is assumed to approximately follow a normal distribution, and the associated P-value and 95% confidence intervals rely on this assumption. As introduced in Sect. 3.6, bootstrap confidence intervals are useful when the accuracy of this approximation is questionable. The logarithm of the likelihood for the fitted model along with a likelihood ratio (LR) statistic `LR chi2(1)` and associated P-value (`Prob > chi2`) are also provided. Maximum likelihood is the standard method of estimating parameters from logistic regression models, and is based on finding the estimates which maximize the joint probability (or *likelihood*—see Sect. 5.6) for the observed data under the chosen model.

Table 5.3 Effects of age differences of 1 and 10 years, by reference age

Age (x)	$P(x)$	$P(x+1)$	odds(x)	odds($x+1$)	OR	RR	ER
40	0.049	0.053	0.052	0.056	1.077	1.073	0.004
50	0.098	0.105	0.109	0.117	1.077	1.069	0.007
60	0.186	0.198	0.229	0.247	1.077	1.062	0.012
Age (x)	$P(x)$	$P(x+10)$	odds(x)	odds($x+10$)	OR	RR	ER
40	0.049	0.098	0.052	0.109	2.105	1.996	0.049
50	0.098	0.186	0.109	0.229	2.105	1.899	0.088
60	0.186	0.325	0.229	0.482	2.105	1.746	0.139

The LR statistic given in the table compares the likelihood from the fitted model with the corresponding model excluding age, and addresses the hypothesis that there is no (linear) relationship between age and the log odds of CHD occurrence. The associated P-value is obtained from the χ^2 distribution with one degree of freedom (corresponding to the single predictor used in the model). LR tests are covered in more detail in Sect. 5.2.1. Note that the `Pseudo R2` value in the table is intended to provide a measure paralleling that used in linear regression models, and is related to the LR statistic.

As an additional illustration of the properties of the logistic model, Table 5.3 presents a number of quantities calculated directly from the coefficients in Table 5.2 and (5.2) and (5.5). For the ages 40, 50, and 60, the table gives the estimated response probabilities and odds. These are also calculated for one- and ten-year age increases so that corresponding odds ratios can be computed. As prescribed by the model, the odds ratios associated with a fixed increment change in age remain constant across the age range. Estimates of RR and ER are also computed for one- and ten-year age increments to illustrate that the fitted logistic model can be used to estimate a wide variety of quantities in addition to odds ratios. Note that the estimated values of ER and RR are not constant with increasing age (because the model does not restrict them to be so). Note also that although measures such as ER and RR can be computed from the logistic model, the resulting estimates will not in general correspond to those obtained from a regression model defined on a scale on which ER or RR is assumed constant. We will return to this topic when we consider alternative binary regression approaches in Sect. 5.5, and again in Sect. 9.3, where we consider use of the logistic model to estimate response probabilities for binary predictors representing contrasting exposure scenarios in the context of causal inference.

5.1.2 Categorical Predictors

Similar to the conventional linear regression model, the logistic model (5.5) is equally valid for categorical risk factors. For example, we can use it to look again at the relationship between CHD risk and the binary predictor arcus senilis as

Table 5.4 Logistic model for CHD and arcus senilis

```
. logistic chd69 i.arcus

Logistic regression                               Number of obs   =       3152
                                                  LR chi2(1)      =      12.98
                                                  Prob > chi2     =     0.0003
Log likelihood = -879.10783                       Pseudo R2       =     0.0073

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.arcus |    1.63528   .2195035    3.66   0.000        1.257    2.127399
------------------------------------------------------------------------------
```

shown in Table 5.4. The regression output in Table 5.4 summarizes the model fit in terms of the odds ratio for the included predictor, and does not include estimates of the regression coefficients. In particular, the model intercept is omitted. This is the default option in many statistical packages such as Stata. Specifying the `coef` option as illustrated in Table 5.2 provides coefficient estimates, including the intercept. Note also that the estimated odds ratio, P-value for the Wald test that the true value the odds ratio is one (or, equivalently that the coefficient is zero), and corresponding 95% CI are virtually the same as the results obtained in Table 3.5. Because `arcus` is a binary predictor (coded as one for individuals with the condition and zero otherwise), entering it directly into the model as if it were a continuous measurement produces the desired result: the coefficient represents the log odds ratio associated with a one-unit increase in the predictor. (In this case, only one, single unit increase is possible by definition.) For two-level categorical variables with levels coded other than zero or one, care must be taken so that they are appropriately treated as categories (and not continuous measurements) by the model-fitting software.

Categorical risk factors with multiple levels are treated similarly to the procedure introduced in Sect. 4.3 for linear regression. In this way, we can repeat the analysis in Table 5.1, dividing study participants into five age groups and taking the youngest group as the reference. In order to estimate odds ratios for each of the four older age groups compared to the youngest group, we need to construct four indicator variables corresponding to the levels of the categorical variable encoding the age groups. Stata does this automatically via the `i.` prefix for the categorical predictor `agec`, as shown in Table 5.5. This variable is constructed with categories corresponding to the age divisions shown in Table 5.1.

Note that the estimated odds ratios appear to be identical to those in the table. In fact, because we are estimating a parameter for each age category except the youngest (reference) group, we are not imposing any restrictions on the parameters (i.e., the logistic assumption does not come into play as it does for continuous predictors). Thus, we would expect the estimated odds ratios to be identical to those estimated using the contingency table approach.

The LR test for this model compares the likelihood for the model with four indicator variables for age with that from the corresponding model with no

Table 5.5 Logistic Model for CHD and age as a categorical factor

```
. logistic chd69 i.agec
Logistic regression                               Number of obs   =       3154
                                                  LR chi2(4)      =      44.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -868.14866                       Pseudo R2       =     0.0252
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |
          1  |   .8768215   .2025406   -0.57   0.569     .5575563    1.378903
          2  |    1.70019   .3800504    2.37   0.018     1.097046    2.634935
          3  |   2.318679   .5274963    3.70   0.000     1.484545    3.621494
          4  |   2.886314   .7462298    4.10   0.000     1.738895    4.790864
------------------------------------------------------------------------------
. testparm i.agec
           chi2(  4) =   44.08
         Prob > chi2 =    0.0000

. contrast agec, mcompare(sidak) eform effects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
        agec |          4       44.08     0.0000
------------------------------------------------
Note: Sidak-adjusted p-values are reported for
      tests on individual contrasts only.
---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
        agec |            4
---------------------------
------------------------------------------------------------------------------
             |                                Sidak               Sidak
             |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |
(1 vs base)  |   .8768215   .2025406   -0.57   0.966      .493201    1.558829
(2 vs base)  |    1.70019   .3800504    2.37   0.068     .9742722    2.966979
(3 vs base)  |   2.318679   .5274963    3.70   0.001     1.315633    4.086453
(4 vs base)  |   2.886314   .7462298    4.10   0.000     1.515851    5.495795
-----------------------------------------------------------------------------

. * Tests for linear trend
. test -1.agec + 3.agec + 2*4.agec = 0
 ( 1)  - [chd69]1.agec + [chd69]3.agec + 2*[chd69]4.agec = 0
           chi2(  1) =   31.45
         Prob > chi2 =    0.0000

. contrast {agec -2 -1 0 1 2}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
        agec |          1       31.45     0.0000
------------------------------------------------

. contrast q(1).agec, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
        agec |          1       31.45     0.0000
------------------------------------------------
```

predictors. In contrast to the individual Wald tests provided for each level of age, the LR test examines the overall effect of age represented as a five-level predictor. The results indicate that inclusion of age affords a statistically significant improvement in the fit of the model.

The table also includes output from the Stata `testparm` and `contrast` commands, used here to test the global hypothesis that the coefficients for the four older age categories are all equal to zero. This hypothesis is identical to the one addressed by the LR test in this case, and the resulting Wald `chi2` test statistic is quite similar to the LR statistic. The correspondence between these two tests is also discussed in Sects. 5.2.1 and 10.4.2.

We note that caution should be exercised in interpretation of significance results for individual Wald tests for categorical predictors with multiple levels, especially in cases where the overall hypothesis test is not statistically significant. As discussed in Sect. 4.3.4, the `mcompare` option allows for control of the familywise Type-1 error rate (FER) in making multiple pairwise comparisons, using Bonferroni, Sidak, and Scheffé procedures. In this case, we used the `contrast` command with option `mcompare(sidak)` to obtain more conservative P-values and CIs for the age effects (the odds-ratios are unchanged).

An additional test of interest in this example is evaluation of the presence of linear trend in the log odds of CHD with increasing age category. This test is implemented exactly as described for linear regression models in Sect. 4.3.5, using the contrast coefficients given in Table 4.8; the test is also obtained using both `contrast` commands introduced in Table 4.9. The result shown in Table 5.5 is quite significant, indicating evidence for a linear trend in the log odds of disease with increasing category of age, and confirming our impression of a regular increase in odds ratios with increasing age. The methods presented there for evaluating departure from linearity are also directly applicable to the logistic model.

Estimating regression coefficients for levels of a categorical predictor often involves specification of an appropriate reference category, especially for nominal categorical predictors. For the example in Table 5.5, this was chosen automatically by Stata as the age category with the smallest numerical label. (A similar procedure is followed by most major statistical packages.) Since age can be considered as ordinal, it makes sense in this case to preserve the ordering of the categories, especially if assessing trends in outcome odds with increasing age is of interest. However, in cases where a reference group different from the default is of interest, most statistics packages (including Stata and SAS) have methods for changing the default. For example, using `ib2.agec` rather than `i.agec` in the `logistic` command in Table 5.5 will result in the second age category being used as the reference. Alternatively, the model can be re-fit using a recoded version of the predictor. Note that it is also possible to compute odds ratios comparing arbitrary groups from the coefficients obtained using the default reference group. For example, the odds ratio comparing the fourth age group in Table 5.5 to the third can be shown to be $\frac{2.88}{2.32} = 1.24$. (This calculation is left as an exercise.)

Another important consideration in selecting a reference group for a categorical predictor are the sample sizes in each category. As a general rule, when individuals are unevenly distributed across categories it is desirable to avoid making the smallest group the reference category. This is because standard errors of coefficients for other categories will be inflated due to the small sample size in the reference group.

A final issue that arises in fitting models with ordinal categorical predictors formed based on an underlying continuous measurement is the choice of how many categories, and how these should be defined. In the example in Table 5.5, the choice of five-year age groups was somewhat arbitrary. In many cases, categories will correspond to pre-existing hypotheses or be suggested by convention (e.g., ten-year age categories in summaries of cancer rates). In the absence of such information, a good practice is to choose categories of equal size based on quantiles of the distribution of the underlying measure.

How many categories a given model will support depends on the overall sample size as well as the distribution of outcomes in the resulting groups. In the WCGS sample, a logistic model including a coefficient for each unique age (assigning the youngest age as the reference group) yields reasonable estimates and standard errors. There are 266 individuals in the smallest group. (A much simpler model that fits the data adequately can also be constructed using the methods discussed in Sect. 5.4.1.) Care must be taken in defining categories to ensure that there are adequate numbers in the subgroups (possibly by collapsing categories). In general, avoid categorizations that result in categories that are homogeneous with respect to the outcome or that contain fewer than ten observations. Problems that arise when this is not the case are discussed in Sect. 5.4.4.

5.2 Multipredictor Models

Clinical and epidemiological studies of binary outcomes typically focus on the potential effects of multiple predictors. When these are categorical and few in number, contingency table techniques suffice for data analyses. However, for larger numbers of potential predictors and/or when some are continuous measurements, regression methods have a number of advantages. For example, the WCGS study measured a number of potential predictors of CHD, including total serum cholesterol, diastolic and SBP, smoking, age, body size, and behavior pattern. The investigators recognized that these variables all may contribute to outcome risk in addition to being potentially associated with each other, and that in assessment of the influence of a selected predictor, it might be important to control for the potential confounding influence of others. Because there are a number of candidate predictors, some of which can be viewed as continuous measurements, multiple regression techniques are very appealing in analyzing such data.

The logistic regression model for multiple predictor variables is a direct generalization of the version for a single predictor introduced above (5.5). For a binary

Table 5.6 Multiple logistic model for CHD risk

```
. logistic chd69 age chol bmi sbp i.smoke if chol<645, coef

Logistic regression                             Number of obs   =       3141
                                                LR chi2(5)      =     159.80
                                                Prob > chi2     =     0.0000
Log likelihood = -807.19249                     Pseudo R2       =     0.0901

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0644476   .0119073     5.41   0.000     .0411097    .0877855
        chol |   .0107413   .0015172     7.08   0.000     .0077675     .013715
         bmi |   .0574361   .0263549     2.18   0.029     .0057814    .1090907
         sbp |   .0192938   .0040909     4.72   0.000     .0112759    .0273117
     1.smoke |   .6344778   .1401836     4.53   0.000     .3597231    .9092325
       _cons |  -12.31099    .977256   -12.60   0.000    -14.22638    -10.3956
------------------------------------------------------------------------------
```

outcome y, and p predictors $x_1, x_2, \cdots, x_p$, the systematic part of the model is defined as follows:

$$\log\left[\frac{P(x_1, x_2, \cdots, x_p)}{1 - P(x_1, x_2, \cdots, x_p)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{5.7}$$

This can be re-expressed in terms of the outcome probability as follows:

$$P(x_1, x_2, \cdots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}. \tag{5.8}$$

As with standard multiple linear regression, the predictors may include continuous and categorical variables. The multiple-predictor version of the logistic model is based on the same assumptions underlying the single predictor version. (These are presented in Sect. 5.1.) In addition, it assumes that multiple predictors are related to the outcome in an additive fashion on the log odds scale. The interpretation of the regression coefficients is a direct generalization of that for the simple logistic model:

- For a given predictor x_j, the coefficient β_j gives the change in log odds of the outcome associated with a unit increase in x_j, for arbitrary fixed values for the predictors $x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p$.
- The exponentiated regression coefficient $\exp(\beta_j)$ represents the odds ratio associated with a one unit change in x_j.

Table 5.6 presents the results of fitting a logistic regression model examining the impact on CHD risk of age, cholesterol (mg/dL), SBP (mmHg), BMI (computed as weight in kilograms divided by the square of height in meters), and a binary indicator of whether or not the participant smokes cigarettes, using data from the WCGS sample. This model is of interest because it addresses the question of whether a select group of established risk factors for CHD are independent predictors for the WCGS study.

Twelve observations were dropped from the analysis in Table 5.6 because of missing cholesterol values. An additional observation was dropped (via the `if` statement in the `regress` command) because of an unusually high cholesterol value (645 mg/dL) that is clearly an outlier. Note that all predictors are entered as continuous measurements in the model. The coefficient for any one of these (e.g., `chol`) gives the log odds ratio (change in the log odds) of CHD for a unit increase in the predictor, adjusted for the presence of the others. The small size of the coefficients for these measures reflects the fact that a unit increase on the measurement scale is a very small change, and does not translate to a substantial change in the log odds.

Log odds ratios associated with larger increases are easily computed as described in Sect. 5.1. The 95% CIs for coefficients of all included predictors exclude zero, indicating that each is a statistically significant independent predictor of outcome risk (as measured by the log odds). Of course, additional assessment of this model would be required before it is adopted as a "final" representation of outcome risk for this study. In particular, we would want to evaluate whether the linearity assumption is met for continuous predictors, evaluate whether additional confounding variables should be adjusted for, and check for possible interactions. These topics are discussed in more detail below.

As an example of an application of the fitted model in Table 5.6, consider calculating the log odds of developing CHD within ten years for a 60-year-old smoker, with 253 mg/dL of total cholesterol, SBP of 136 mmHg, and a BMI of 25. Applying (5.7) with the estimated coefficients from Table 5.6,

$$\begin{aligned}\log\left[\frac{P(60,253,136,25,1)}{1-P(60,253,136,25,1)}\right] &= -12.311 + .0644\times 60 + .0107\times 253\\ &\quad +.0193\times 136 + .0574\times 25 + .6345\times 1\\ &= -1.046.\end{aligned}$$

A similar calculation gives the corresponding log odds for a similar individual of age 50:

$$\begin{aligned}\log\left[\frac{P(50,253,136,25,1)}{1-P(50,253,136,25,1)}\right] &= -12.311 + .0644\times 50 + .0107\times 253\\ &\quad +.0193\times 136 + .0574\times 25 + .6345\times 1\\ &= -1.690.\end{aligned}$$

Finally, the difference between these gives the log odds ratio for CHD associated with a ten year increase in age for individuals with the specified values of all of the included predictors:

$$-1.046 - (-1.690) = 0.644.$$

Table 5.7 Multiple logistic model with rescaled predictors

```
. logistic chd69 age_10 chol_50 bmi_10 sbp_50 i.smoke if chol<645

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(5)      =     159.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -807.19249                       Pseudo R2       =     0.0901

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.904989   .2268333     5.41   0.000     1.508471    2.405735
     chol_50 |   1.710974   .1297977     7.08   0.000     1.474584    1.985259
      bmi_10 |   1.775995   .4680613     2.18   0.029     1.059518    2.976973
      sbp_50 |   2.623972   .5367142     4.72   0.000     1.757326    3.918016
     1.smoke |   1.886037   .2643914     4.53   0.000     1.432933    2.482417
------------------------------------------------------------------------------
```

Closer inspection reveals that this result is just ten times the coefficient for age in Table 5.6. In addition, we see that we could repeat the above calculations for any ten-year increase in age, and for any fixed values of the other predictors and obtain the same result. Thus, the formula (5.6) for computing log odds ratios for arbitrary increases in a single predictor applies here as well. The odds ratio for a ten-year increase in age (adjusted for the other included predictors) is given simply by

$$\exp(0.0644 \times 10) = \exp(.644) = 1.90.$$

Interpretation of regression coefficients for categorical predictors also follow that given for single predictor logistic models. For example, the coefficient (0.634) for the binary predictor variable smoke in Table 5.6 is the log odds ratio comparing smokers to nonsmokers for fixed values of age, chol, sbp, and bmi. The corresponding odds ratio

$$\exp(0.634) = 1.89$$

measures the proportionate increase in the odds of developing CHD for smokers compared to nonsmokers adjusted for age, cholesterol, SBP, and BMI.

The estimated coefficients for the first four predictors in Table 5.6 are all very close to zero, reflecting the continuous nature of these variables and the fact that a unit change in any one of them does not translate to a large increase in the estimated log odds of CHD. As shown above, we can easily calculate odds ratios associated with clinically more meaningful increases in these predictors. An easier approach is to decide on the degree of change that we would like the estimates to reflect and fit a model based on predictors rescaled to reflect these decisions. For example, if we would like the model to produce odds ratios for ten-year increases in age, we should represent age as the rescaled predictor age_10 = age/10. Table 5.7 shows the estimated odds ratios from the model including rescaled versions of the first four predictors in Table 5.6. (The numbers after the underscores in the variable names indicate the magnitude of the scaling.) We also "centered" these predictors before scaling them by subtracting of the mean value for each. (Centering predictors is discussed in Sects. 3.3.1 and 4.6.) Note that the log-likelihood and Wald test statistics for this model are identical to their counterparts in Table 5.6.

5.2.1 Likelihood Ratio Tests

In Sect. 5.1, we briefly introduced the concept of the likelihood, and the LR test for logistic models. The likelihood for a given model is interpreted as the joint probability of the observed outcomes expressed as a function of the chosen regression model. The model coefficients are unknown quantities and are estimated by maximizing this probability (hence the name maximum-likelihood estimation). For numerical reasons, maximum-likelihood estimation in statistical software is usually based on the logarithm of the likelihood. An important property of likelihoods from nested models (i.e., models in which predictors from one are a subset of those contained in the other) is that the maximized value of the likelihood from the larger model will always be at least as large as that for the smaller model.

Although the numerical value of the likelihood (or log-likelihood) for a single model does not have a particularly useful interpretation, the LR statistic assessing the difference in likelihoods from two nested models is a valuable tool in model assessment (analogous to the F tests introduced in Sect. 4.3.3). It is especially useful when investigating the contribution of more than one predictor, or for predictors with multiple levels.

For example, consider assessment of the contribution of self-reported behavior pattern to the model summarized in Table 5.7. In the WCGS study, investigators were interested in "type A" behavior as an independent risk factor for CHD. Behavior was classified as either type A or type B, with each type subdivided into two further levels A_1, A_2, B_3, and B_4 (coded as 1, 2, 3, and 4, respectively). The expanded model addresses the question of whether behavior pattern contributes to CHD risk when other established risk factors are accounted for.

Table 5.8 displays the results of including the four-level categorical variable `behpat` in the model from Table 5.7. The natural coding of the variable results in type A_1 behavior being taken as the reference level. Examination of the coefficients and associated 95% CIs for the remaining indicators reveals that although the second category of type A behavior appears not to differ from the reference level, both categories of type B behavior do display statistically significant differences, and are associated with lower outcome risk.

The LR statistic is computed as twice the difference between log likelihoods from the two models, and can be referred to the χ^2 distribution for significance testing. Because the likelihood for the larger model must be larger than the likelihood for the smaller (nested) model, the difference will always be positive. Twice the difference between the log likelihood for the model including `behpat` (Table 5.8) and that for the model excluding this variable (Table 5.6) is

$$2 \times [-794.81 - (-807.19)] = 24.76.$$

Table 5.8 Logistic model for WCGS behavior pattern

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 i.smoke i.behpat if chol<645

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(8)      =     184.57
                                                  Prob > chi2     =     0.0000
Log likelihood =    -794.81                       Pseudo R2       =     0.1040

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |    1.83375   .2198681     5.06   0.000     1.449707    2.319529
     chol_50 |   1.704097   .1301391     6.98   0.000     1.467201    1.979243
      sbp_50 |   2.463504   .5086518     4.37   0.000     1.643621    3.692369
      bmi_10 |   1.739415   .4620341     2.08   0.037     1.033479    2.927551
     1.smoke |   1.830672   .2583097     4.29   0.000      1.38837    2.413882
             |
      behpat |
          2  |   1.068257   .2363271     0.30   0.765     .6924157    1.648103
          3  |   .5141593   .1245593    -2.75   0.006     .3198064    .8266243
          4  |    .572071   .1826117    -1.75   0.080     .3060107    1.069457
------------------------------------------------------------------------------
. estimates store mod1
```

Table 5.9 Likelihood ratio test for four-level WCGS behavior pattern

```
. lrtest mod1

likelihood-ratio test                                  LR chi2(3)  =     24.76
(Assumption: . nested in mod1)                         Prob > chi2 =    0.0000
```

This value follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables present in the larger model (three in this case). Statistical packages like Stata can often be used to compute the LR test directly by first fitting the larger model (in Table 5.8), and saving the likelihood in the user-defined variable (in this case, in the variable `mod1` created in the last line of the table). Next, the reduced model eliminating `behpat` is fit, followed by a command to evaluate the LR test as displayed in the Table 5.9. (See Table 5.6 for the full regression output for this model.) The result agrees with the calculation above, and the associated P-value indicates that collectively, the four-level categorical representation of behavior pattern makes a statistically significant independent contribution to the model.

The similarity between the two odds ratios for type A (the reference level and the second indicator for type A_2 behavior) and type B (the indicators representing types B_3 and B_4 behavior) in Table 5.8 suggests that a single binary indicator distinguishing the A and B patterns might suffice. Note that the logistic model that represents behavior pattern as a two-level indicator (with type B behavior as the reference category) is actually nested within the model in Table 5.8. (The model including the two-level representation is a special case of the four-level version when the coefficients for the two levels of type B and type A behavior, respectively, are identical.) Table 5.10 displays the fitted model and LR test results for this reduced model including the two-level binary indicator `dibpat`. The fact that the difference between the likelihoods for the two models is not statistically significant

Table 5.10 Likelihood ratio test for two-level WCGS behavior pattern

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 i.smoke i.dibpat if chol<645

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(6)      =     184.34
                                                  Prob > chi2     =     0.0000
Log likelihood = -794.92603                       Pseudo R2       =     0.1039

----------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
      age_10 |   1.830252   .2190623    5.05   0.000      1.44754    2.314147
     chol_50 |   1.702406   .1299562    6.97   0.000     1.465835    1.977157
      sbp_50 |   2.467919   .5084377    4.38   0.000     1.648039    3.695681
      bmi_10 |   1.732349   .4596114    2.07   0.038     1.029917    2.913859
     1.smoke |   1.829163   .2580698    4.28   0.000     1.387265    2.411822
    1.dibpat |   2.006855   .2897341    4.82   0.000     1.512259    2.663212
----------------------------------------------------------------------------
. lrtest mod1

likelihood-ratio test                                 LR chi2(2)  =      0.23
(Assumption: . nested in mod1)                        Prob > chi2 =    0.8904
```

confirms our suspicion that modeling the effect of behavior pattern as a two-level predictor is sufficient to capture the contribution of this variable.

As demonstrated above, the LR test is a very useful tool in comparing nested logistic regression models. Note that alternate tests based on Wald statistics can also be used, as illustrated in Tables 4.4 and 5.5. In moderate to large samples, the results from the LR and Wald tests for the effects of single predictors will agree quite closely. However, in smaller samples the results of these two tests may differ substantially. In general, the LR test is more reliable than the Wald test, and is preferred when both are available. Finally, note that because the likelihood is computed based on the observations used to fit the model, it is important to ensure that the same observations are included in each candidate model considered in LR testing. This was accomplished in the examples by insuring that the fitted models excluded 12 observations with missing values for cholesterol, and another with an outlying value of 645. Likelihoods from models fit on differing sets of observations are not comparable. A more complete discussion of the concepts of likelihood and maximum-likelihood estimation is given in Sect. 5.6.

5.2.2 Confounding

A common goal of multiple logistic regression modeling is to investigate the association between a primary predictor and the outcome, accounting for the possible mediating or confounding influence of additional measured predictors. For example, in evaluating the observed association between behavior pattern (considered in the previous section) and CHD risk, it is important to consider the effects of additional variables that might be related to both behavior and

Table 5.11 Logistic model for type A behavior pattern and selected predictors

```
. logistic dibpat age_10 chol_50 sbp_50 bmi_10 i.smoke

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(5)      =      53.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -2150.1739                       Pseudo R2       =     0.0124

------------------------------------------------------------------------------
      dibpat | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   1.324032   .0881552     4.22   0.000      1.16205    1.508594
     chol_50 |   1.084241   .0464136     1.89   0.059     .9969839    1.179135
      sbp_50 |   1.461247   .1876433     2.95   0.003     1.136104    1.879442
      bmi_10 |   1.123846   .1672474     0.78   0.433     .8395252    1.504459
     1.smoke |    1.26933   .0930786     3.25   0.001     1.099403    1.465522
------------------------------------------------------------------------------
```

CHD occurrence. Recall from Chap. 4 that regression models can account for potential confounding or mediation influences of such variables by considering the adjusted and unadjusted associations between the outcome and predictor of primary interest. In this section, we briefly review these issues in the logistic regression context.

Consider again the assessment of behavior pattern as a predictor of CHD in the WCGS example considered in the previous section. In the analysis summarized in Table 5.10, we concluded that a two-level indicator (dibpat) distinguishing type A and B behaviors adequately captures the effects of this variable on CHD (in place of a more complex, four-level summary of behavior). The discussion in Chap. 9 will suggest that we should consider the possible causal relationships of the additional variables in the model with both the outcome and behavior pattern before making any conclusions about the possible causal connection between behavior type and the outcome.

Recall the discussion of confounding and mediation presented in Sects. 4.4 and 4.5. To be a confounder of an association of primary interest, a variable must be associated with both the outcome and the primary predictor. From Table 5.10, all of the predictors in addition to dibpat are independently associated with the CHD outcome. Since dibpat is a binary indicator, we can examine its association with these predictors via logistic regression as well. Table 5.11 presents the resulting model. With the exception of BMI (bmi_10), all appear to be associated with behavior pattern. In deciding which variables to adjust for in summarizing the CHD-behavior pattern association, it is worth considering the possible causal relationships to help identify or distinguish variables with confounding influence from those that could be potential mediators or effect modifiers.

Causal connections are likely to be very complex. For example, age can be considered as a possible confounder of the relationship between behavior type and CHD. However, BMI, cholesterol, SBP (hypertension), and smoking could either exert a confounding influence or be viewed as mediating variables in the pathway between behavior and CHD. The unadjusted odds ratio (95% CI) for the association between type A behavior and CHD is 2.36 (1.79, 3.10). By contrast, the adjusted odds ratio in Table 5.10 is 2.01 (95% CI 1.51, 2.66). Note that dropping any of the additional predictors from the model singly results in little change to the estimated OR for type A behavior (less than 5%). Thus if any of these variables acts as a mediator, the influence appears to be weak. This suggests that the influence of type A behavior on CHD may act partially through another unmeasured pathway. (Or that this characterization of behavior is itself mediated through other unmeasured behavioral characteristics.) In this case, adjustment for the other variables is appropriate if they are considered as confounders. However, if they (with the possible exception of age) are regarded as mediators, then the effect assessed on the adjusted model can be viewed as an estimate of the direct effect of behavior not mediated through the pathways mediated by these variables. See Sects. 9.6 and 10.2 for further discussion of these issues. Of course, before concluding that we have adequately modeled the relationship between behavior pattern and CHD we need to account for possible interactions between included predictors (Sect. 5.2.4), and conduct diagnostic assessments of the model fit including nonlinearity in relationships with continuous predictors (Sect. 5.4).

5.2.3 Mediation

As an example of assessment of mediation in the context of a binary outcome, we consider an example from the FIT study, a randomized trial investigating the effect of a treatment for reducing spinal fracture risk in postmenopausal women with prior history of fracture due to osteoporosis (Black et al. 1996b). We are interested in evaluating possible mediation of treatment effects through changes in bone mineral density (BMD). A finding that much of the beneficial effect of treatment operated through this pathway would be of practical interest in development of future treatments.

Table 5.12 presents two logistic regression models for the effect of randomized treatment assignment on a binary indicator of spinal fracture occurrence. The first model gives the marginal effect of assignment to treatment in the entire sample of 5,470 women. Assuming that randomization was effective, the unadjusted odds ratio for treatment in this model represents an *intention to treat* estimate of the effectiveness of treatment assignment. The second model in the table includes predictors for change in BMD (in standard deviation units) between follow-up and baseline, baseline level of BMD (also in standard deviation units), baseline smoking status (former and current smokers compared to nonsmokers as the reference category), and a binary indicator of a history of previous spinal fracture (`frac_base`). Age (in years) is also included as a restricted cubic spline with three knots. Note that since the follow-up level of BMD reflects changes that occurred postrandomization, these baseline measures represent potential confounders of the association between change in BMD and new fracture occurrence. As discussed in Sect. 4.5, interpretation of the apparent attenuation of the effect of treatment in this model

Table 5.12 Logistic regression estimation of marginal and direct effect of treatment assignment on new fracture risk in the FIT study example

```
*** Marginal treatment effect ***
. logistic frac_new i.treat

Logistic regression                               Number of obs   =       5470
                                                  LR chi2(1)      =      32.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -1163.5889                       Pseudo R2       =     0.0136

------------------------------------------------------------------------------
    frac_new | Odds Ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.treat |   .5052736   .0624452   -5.52   0.000    .3965785      .64376
------------------------------------------------------------------------------

*** Direct treatment effect not mediated by change in BMD ***
. logistic frac_new i.treat bmd_diff bmd_base i.frac_base i.smoking age_spl*

Logistic regression                               Number of obs   =       5339
                                                  LR chi2(8)      =     311.04
                                                  Prob > chi2     =     0.0000
Log likelihood =  -982.6019                       Pseudo R2       =     0.1366

------------------------------------------------------------------------------
    frac_new | Odds Ratio   Std. Err.     z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.treat |   .5966412   .0829632  -3.71   0.000     .4543112    .7835616
    bmd_diff |   .7062953   .0505978  -4.85   0.000     .6137729    .8127648
    bmd_base |   .6885569   .0412505  -6.23   0.000     .6122735    .7743444
  1.frac_base |   3.428229   .4569538   9.24   0.000     2.640048    4.451719
             |
     smoking |
          1  |   1.141699   .1555701   0.97   0.331     .8741083    1.491207
          2  |   1.379136   .2722494   1.63   0.103     .9366451    2.030669
             |
    age_spl1 |   1.123983   .0413332   3.18   0.001     1.045822    1.207986
    age_spl2 |   .9476609   .0329655  -1.55   0.122     .8852031    1.014526
------------------------------------------------------------------------------
```

relative to the first (unadjusted) model requires assumptions about the causal nature of the relationships represented. In this example, a plausible interpretation is that treatment effects are mediated through treatment-induced changes in BMD.

Following the approach introduced in Sect. 4.5, we can assess whether the conditions for mediation are met by fitting two models: the first, a linear regression for the dependence of change in BMD on treatment assignment; the second, a logistic regression of the dependence of the outcome on change in BMD. In both cases, we adjust for the possible confounders displayed in Table 5.12. Both models yield highly significant results for the Wald tests of the coefficients representing the key components of the mediating relationships. Further, there is no evidence for interaction between treatment assignment and change in BMD. This, and the observed attenuation in the estimated effect of treatment in the second model in Table 5.12 provides evidence for the possible mediating role of change in BMD.

As also discussed in Sect. 4.5, it may also be of interest to make separate estimates of the direct and indirect components of the overall effect of treatment assignment on fracture risk, and to estimate the proportion of the treatment effect explained (PTE) by the mediating influence of changes in BMD. Similar to the examples presented in that section, the odds ratio of 0.597 for treatment assignment in the second model shown in Table 5.12 can be interpreted as an estimate of the direct effect of treatment not mediated through effects on BMD.

By contrast to the results presented in Sect. 4.5, decomposing the relationships between outcome, treatment, and a mediator into overall, indirect, and direct effect components poses additional difficulties in the context of logistic regression models. This results from the use of the odds ratio as a measure of association, as discussed in Sect. 3.4.4. (A similar phenomenon occurs for the Cox regression model introduced in the next chapter.) Performing analyses using an alternative binary regression model based on relative risks rather than odds ratios (see Sect. 5.5.3) avoids this difficulty. Chapter 9 presents further discussion of this topic, including an introduction to more general techniques for assessment of mediation based on causal inference methods. In particular, these methods allow estimation of the causal direct effect of treatment, not mediated through the mediating variable. This estimate will generally differ from the regression estimate described here and has a clearer causal interpretation, especially when additional confounding variables play a role.

5.2.4 Interaction

Recall from Chap. 4 that an interaction between two predictors in a regression model means that the degree of association between each predictor and the outcome varies according to levels of the other predictor. The mechanics of fitting logistic regression models including interaction terms is quite similar to standard linear regression (see Sect. 4.6). For example, to fit an interaction between two continuous predictors x_1 and x_2, we include the product x_1x_2 as an additional predictor in a model containing x_1 and x_2 as shown in (5.9):

$$\log\left[\frac{P(x_1, x_2, x_1 \times x_2)}{1 - P(x_1, x_2, x_1x_2)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2. \tag{5.9}$$

Fitting interactions between categorical predictors and between continuous and categorical predictors also follows the procedures outlined in Chap. 4. However, because of the log odds ratio interpretation of regression coefficients in the logistic model, interpreting results of interactions is somewhat different. We review several examples below.

For an illustrative example of a two-way interaction between two binary indicator variables from the WCGS study, consider the regression model presented in Table 5.13. The fitted model includes the indicator `arcus` for arcus senilis (defined in Sect. 3.4), a binary indicator `bage_50` for participants over the age of 50, and the product between them, `bage_50#arcus`, made automatically by the `##` operator in the `logistic` command. The research question addressed is whether the

Table 5.13 Logistic model for interaction between arcus and age as a categorical predictor

```
. logistic chd69 i.bage_50##i.arcus, coef

Logistic regression                               Number of obs   =       3152
                                                  LR chi2(3)      =      40.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -865.43251                       Pseudo R2       =     0.0228
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.bage_50 |   .8932677   .1721239     5.19   0.000     .5559111    1.230624
     1.arcus |   .6479628   .1788637     3.62   0.000     .2973964    .9985293
             |
bage_50#arcus |
         1 1 |  -.5920552   .2722269    -2.17   0.030     -1.12561   -.0585002
             |
       _cons |  -2.882853   .1089261   -26.47   0.000    -3.096344   -2.669362
------------------------------------------------------------------------------
```

association between arcus and CHD is age dependent. The statistically significant result of the Wald test for the coefficient associated with the product of the indicators for age and arcus indicates that an interaction is present. This means that we cannot interpret the coefficient for `arcus` as a log odds ratio without specifying whether or not the participant is older than 50. (A similar result holds for the interpretation of `bage_50`.)

The procedure for obtaining the component odds ratios is similar to the methods for obtaining main and interaction effects for linear regression models, and is straightforward using the regression model. If we represent `1.arcus` and `1.bage_50` as x_1 and x_2 in (5.9), we can compute the log odds for any combination of values of these predictors using coefficients from Table 5.13. For example, the log odds of CHD occurrence for an individual over 50 years old without arcus is given by

$$\begin{aligned}\log\left[\frac{P(0,1,0)}{1-P(0,1,0)}\right] &= \beta_0+\beta_2\\ &= -2.883+0.893=-1.990.\end{aligned}$$

Similarly, the log odds for an individual between 39 and 49 years old without arcus is

$$\log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right]=\beta_0.$$

With these results, we see that the five expressions below define the component log odds ratios in the example:

$$\begin{aligned}\log\left[\frac{P(1,0,0)}{1-P(1,0,0)}\right]-\log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right]&=\beta_1=0.648\\ \log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right]-\log\left[\frac{P(0,1,0)}{1-P(0,1,0)}\right]&=\beta_1+\beta_3=0.056\end{aligned}$$

Table 5.14 Component odds ratios for arcus-age interaction model

Odds ratio	Groups compared
$\exp(\beta_1) = 1.91$	Arcus vs. no arcus, age 39–49
$\exp(\beta_1 + \beta_3) = 1.06$	Arcus vs. no arcus, age 50–59
$\exp(\beta_2) = 2.44$	Age 50–59 vs. age 39–49, no arcus
$\exp(\beta_2 + \beta_3) = 1.35$	Age 50–59 vs. age 39–49, arcus
$\exp(\beta_1 + \beta_2 + \beta_3) = 2.58$	Arcus and age 50–59 vs. no arcus and ages 39–49

Table 5.15 Example odds ratio for arcus-age interaction model

```
. lincom 1.bage_50 + 1.bage_50#1.arcus

 ( 1)  [chd69]1.bage_50 + [chd69]1.bage_50#1.arcus = 0
------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.351497   .2850372     1.43   0.153     .8939071    2.043325
------------------------------------------------------------------------------
```

$$\log\left[\frac{P(0,1,0)}{1-P(0,1,0)}\right] - \log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right] = \beta_2 = 0.893$$
$$\log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right] - \log\left[\frac{P(1,0,0)}{1-P(1,0,0)}\right] = \beta_2 + \beta_3 = 0.301$$
$$\log\left[\frac{P(1,1,1)}{1-P(1,1,1)}\right] - \log\left[\frac{P(0,0,0)}{1-P(0,0,0)}\right] = \beta_1 + \beta_2 + \beta_3 = 0.949. \quad (5.10)$$

The corresponding odds ratios are then easily calculated by exponentiation, as shown in Table 5.14.

Referring back to Table 5.13, we see that all of the component odds ratios aren't immediately obvious from standard regression output. However, the log odds ratio and associated 95% CIs for `arcus` among individuals in the younger age group and for older individuals among those without arcus can be read directly. This is because when we set either variable to zero (the reference level), the interaction term evaluates to zero and is eliminated. Estimated log odds ratios corresponding to the nonreference levels of these variables involve the interaction term, and differ from their counterparts by the value of its coefficient (–0.592). Standard errors and 95% CIs for these estimates require additional calculations that cannot be completed without further information about the fitted model. Fortunately, many statistical packages have facilities that greatly simplify these calculations. Table 5.15 illustrates the use of the `lincom` command in Stata to compute the odds ratio comparing the odds of CHD in individuals of age 50 and over with the odds among those under 50, among individuals with arcus.

By specifying the correct combination of coefficients (corresponding to those in Table 5.14), the output in the Table 5.15 provides the desired odds ratio estimate along with the 95% CI. Results of the accompanying hypothesis test that the underlying log odds ratio is zero are also provided.

Table 5.16 Logistic model for interaction between arcus and age as continuous

```
. logistic chd69 i.arcus##c.age, coef

Logistic regression                               Number of obs     =      3152
                                                  LR chi2(3)        =     53.33
                                                  Prob > chi2       =    0.0000
Log likelihood = -858.93362                       Pseudo R2         =    0.0301

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.arcus |   2.754185   1.140118     2.42   0.016     .5195952    4.988774
         age |    .089647   .0148904     6.02   0.000     .0604623    .1188317
             |
 arcus#c.age |
          1  |  -.0498298   .0233431    -2.13   0.033    -.0955814   -.0040782
             |
       _cons |  -6.788086   .7179977    -9.45   0.000    -8.195335   -5.380836
------------------------------------------------------------------------------
```

Interactions between a continuous and categorical variable are handled in a similar fashion to those involving binary predictors. In the previous example, the categorization of age was somewhat arbitrary. In fact, because age was represented by two categories, essentially the same results could have been obtained using frequency table techniques (as illustrated in Table 3.9). A more complete assessment of the interaction can be obtained by considering age as a continuous variable (previously considered in Table 5.2). For example, this would allow us to investigate whether increase in CHD risk with increasing age differs in individuals with and without arcus. The logistic model addressing this question is displayed in Table 5.16.

Note the use of the ## operator in Stata, introduced in Sect. 4.6, which instructs the program to include an interaction term between the two variables. This is accomplished by inclusion of the product of arcus and age (`arcus#c.age`) as well as the individual predictors `age` and `1.arcus`. For a fixed age (e.g., 55), the log odds ratio associated with having arcus is calculated as follows, using the estimated coefficients from Table 5.16:

$$\log\left[\frac{P(1,55,55)}{1-P(1,55,55)}\right] - \log\left[\frac{P(0,55,0)}{1-P(0,55,0)}\right]$$
$$= (-6.788 + 2.754 + (0.090 - 0.050) \times 55) - (-6.788 + 0.090 \times 55)$$
$$= (2.754 - 0.050 \times 55) = 0.014.$$

We see that this corresponds to an odds ratio of exp(0.014) = 1.01, which is similar to that calculated for the corresponding age group in Table 5.14. We can obtain this estimate and its 95% CI directly as shown in Table 5.17.

Note that because age is represented as a continuous variable, its value must be specified in interpreting the effect of arcus on the log odds of CHD risk. Similarly, among individuals with arcus, log odds ratios can be computed for any specified increase in age. Figure 5.2 displays the estimated log odds as a function of age,

Table 5.17 Logistic model for interaction between arcus and age as a continuous predictor

```
. lincom 1.arcus + 55*1.arcus#c.age

 ( 1)  [chd69]1.arcus + 55*[chd69]1.arcus#c.age = 0

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.013637   .2062336     0.07   0.947     .6802954    1.510313
------------------------------------------------------------------------------
```

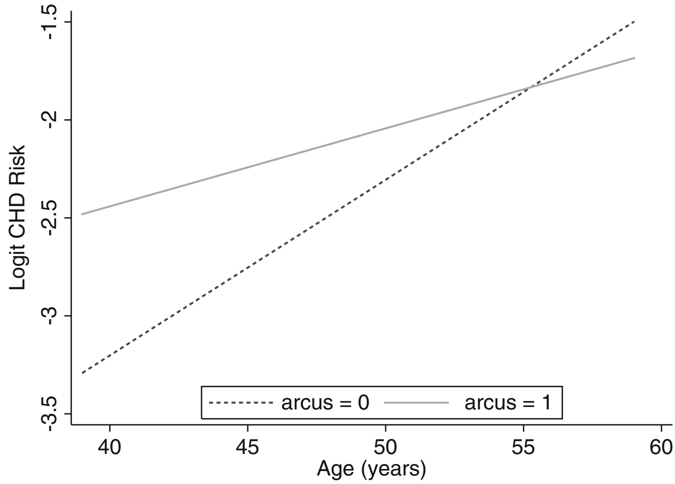

Fig. 5.2 Log odds of CHD and age for individuals with and without arcus senilis

separately for individuals with and without arcus. The equations for these two lines can be obtained directly from the coefficients in Table 5.16 and are printed below for individuals with and without arcus, respectively:

$$\log\left[\frac{P(\mathsf{age})}{1-P(\mathsf{age})}\right] = (-6.788 + 2.754) + (0.090 - 0.050) \times \mathsf{age}$$
$$= -4.034 + 0.040 \times \mathsf{age}.$$

and

$$\log\left[\frac{P(\mathsf{age})}{1-P(\mathsf{age})}\right] = -6.788 + 0.0896 \times \mathsf{age}.$$

Figure 5.2 displays the results obtained above, indicating that CHD risk is higher for younger participants with arcus. However, older participants with arcus seem to be at somewhat lower risk than those without arcus. Of course, further interpretation of these equations should be preceded by thorough checking of the linearity of the relationship between age and the log odds of the outcome, including whether more complicated, higher-order interaction terms are needed.

Recall the discussion in Sect. 5.1 where we motivated the logistic model as an example of a multiplicative risk model (see (5.4)). By contrast, the risk difference model (introduced in (5.1) and discussed further in Sect. 5.5.3) is an example of an additive risk model. In addition to defining two distinct ways in which a predictor can act to modify outcome risk, this distinction turns out to be very important in the context of interaction: For a specified outcome and predictor pair, it is possible to have interaction under the multiplicative model and not under the additive model, and vice versa.

For example, if we fit the additive risk model to the data from the age/arcus example in Table 5.16, the Wald test P-value for inclusion of the product term (`age_50arcus`) is 0.15. (The corresponding value from the logistic model was 0.03.) The implications of this are that we should not necessarily regard interaction as mirroring a biological mechanism, but rather as a property of the data and model being fit. In the example, we would want to account for the interaction if we were using the logistic model but not necessarily if we were analyzing the WCGS data using the additive model. The additive regression model is described further in Sect. 5.5.3. Also, see Clayton and Hills (1993) and Jewell (2004) for more detailed discussions of the distinction between multiplicative and additive interaction.

5.2.5 Prediction

Frequently, the goal of fitting a logistic model is to predict risk of the binary outcome given a set of risk factors. Recall that in Sect. 5.2.1, we fit a logistic model for the CHD outcome in the WCGS sample, using age, cholesterol level, systolic blood pressure, BMI, a binary indicator of current cigarette smoking (with nonsmokers composing the reference group), and an indicator of type A behavior as predictors. Table 5.10 summarizes the results. Table 5.18 presents an expanded version of this model that includes two additional predictors `bmichol` and `bmisbp` for the interactions between BMI and serum cholesterol level and BMI and SBP (both centered and scaled as described in Sect. 5.2). These were both found to make statistically significant contributions to the model in further analyses investigating two way interactions between the original predictors in Table 5.10.

As shown in Sect. 5.2, the estimated coefficients from the model in Table 5.18 can be used directly in the logistic formula (5.8) to compute the log odds (or the corresponding probability) of CHD for an arbitrary individual by specifying the desired values for the predictors. Table 5.19 displays a few such predictions (labeled `prchd`) for five individuals in the WCGS sample (obtained using the `predict` command in Stata).

Table 5.18 Expanded logistic model for CHD events

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp,
coef

Logistic regression                               Number of obs   =       3141
                                                  LR chi2(8)      =     198.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -788.01957                       Pseudo R2       =     0.1117

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_10 |   .5949713   .1201092     4.95   0.000     .3595615     .830381
     chol_50 |   .5757131     .07779     7.40   0.000     .4232474    .7281787
      sbp_50 |   1.019647   .2066014     4.94   0.000     .6147159    1.424579
      bmi_10 |   1.048839   .2998176     3.50   0.000     .4612074    1.636471
       smoke |   .6061929   .1410533     4.30   0.000     .3297335    .8826523
      dibpat |   .7234267   .1448996     4.99   0.000     .4394288    1.007425
     bmichol |  -.8896932   .2746471    -3.24   0.001    -1.427992   -.3513948
      bmisbp |  -1.503455    .631815    -2.38   0.017     -2.74179   -.2651208
       _cons |  -3.416061   .1504717   -22.70   0.000     -3.71098   -3.121142
------------------------------------------------------------------------------
```

Table 5.19 Sample predictions from the logistic model in Table 5.18

```
     +-------------------------------------------------------------------+
     | chd69   age   chol   sbp        bmi       smoke   dibpat      prchd |
     |-------------------------------------------------------------------|
  1. |    no    49    225   110   19.78795      smoker    A1,A2   .0433952 |
  2. |    no    42    177   154    22.9551      smoker    A1,A2   .0708145 |
  3. |    no    42    181   110   23.62529   nonsmoker    B3,B4   .0082533 |
  4. |    no    41    132   124     23.109      smoker    B3,B4   .0089318 |
  5. |   yes    59    255   144   21.52041      smoker    B3,B4   .1926046 |
     |-------------------------------------------------------------------|
```

5.2.6 *Prediction Accuracy*

In some applications, we may be interested in using a logistic regression model as a tool to classify outcomes of newly observed individuals based on values of measured predictors. For the WCGS example just considered, this may involve deciding on treatment strategy based on prognosis as measured by the predicted probability from the logistic model in Table 5.18. Similar to the goals of developing diagnostic tests for detecting diseases, this approach requires us to choose a cut-off or threshold value of the predicted outcome probability above which treatment would be initiated. A fundamental consideration in choosing this threshold is in evaluating the degree of misclassification of outcomes incurred by the choice. For a binary outcome, misclassification can be quantified by calculating the proportion of individuals incorrectly classified as either having the outcome or not. These are known as the *false-positive* and *false-negative* rates, respectively, and are standard measures of prediction error in the logistic regression context. Rather than state prediction performance in terms of misclassification, the following complementary measures are frequently used in assessment of prediction rules for binary outcomes:

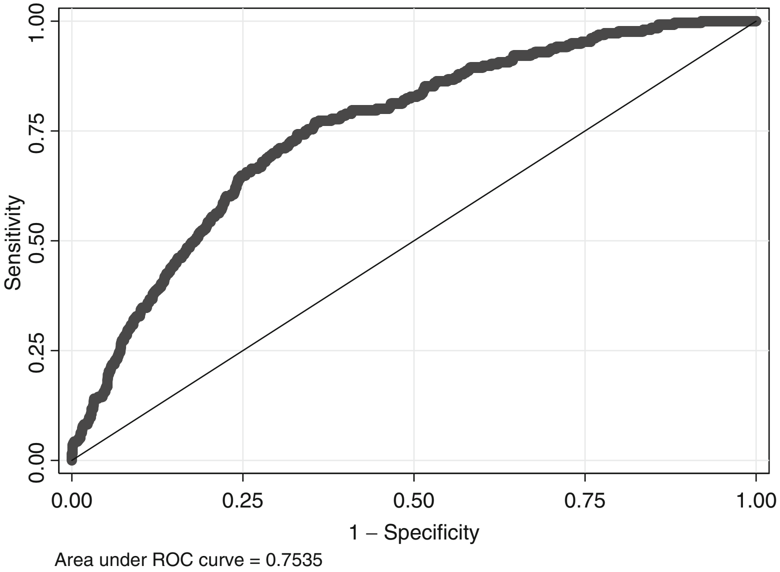

Fig. 5.3 ROC curve for logistic prediction of CHD events

Sensitivity The proportion of individuals with the outcome that are correctly classified, calculated as the complement of the false-negative rate.

Specificity The proportion of individuals without the outcome that are correctly classified, calculated as the complement of the false-positive rate.

As the threshold value of a prediction rule varies between zero and one, these quantities can be calculated and compared to evaluate overall performance. A *receiver operating characteristic* (ROC) curve plots the sensitivity against the false-positive rate (i.e., one minus the specificity) for a range of thresholds to help visualize test performance. Figure 5.3 shows the ROC curve for the current example (obtained using the `lroc` command in Stata), along with a diagonal reference line, usually interpreted as representing the ROC curve for a test that is no better than the flip of a coin.

ROC curves for tests with overall good performance (i.e., low misclassification rates for both positive and negative outcomes) will lie close to the left and topmost margins of the plot. In Fig. 5.3, a test with a sensitivity of around 75% is close to optimal in this sense. (The threshold value corresponding to a sensitivity of 0.75 and a specificity of 0.64 in Fig. 5.3 is about 0.07.) Note that in most practical situations, assessment of test performance has a subjective component: The cost of misclassifying an individual as positive may be deemed more serious than the alternative situation, or vice versa. These considerations weigh into evaluation of test results. The area under an ROC curve (also known as the *C-statistic*) provides an overall measure of classification accuracy, with the value of one representing perfect accuracy. In the present case, the value of 0.754 does not indicate very impressive performance.

A clear limitation with the example above is that the individuals used to evaluate the performance are the same as those used to fit the model on which the classification rule is based. Alternative techniques that do not share this limitation include cross-validation and learning set/test set validation (both described in Sect. 10.1). Finally, note that although logistic regression is a valid approach for development of prediction tools, alternative techniques are available. Classification trees are an example of a larger class of tree-based methods, and involve fewer modeling assumptions than the logistic approach. See Goldman et al. (1996) for an example of their application in a clinical context. Prediction is discussed in greater detail in Sect. 10.1.

5.3 Case-Control Studies

In situations where binary outcomes are rare or difficult to observe, it is not always feasible to collect a large enough sample to investigate the relationship between the outcome and predictors of interest. Consider the problem of evaluating dietary risk factors for stomach cancer. Because this disease is relatively rare (accounting for approximately 2% of annual cancer deaths in the United States), only a very large cross-sectional or prospective sample would include sufficient numbers of cases to evaluate associations with predictors of interest. Case-control studies address this problem by recruiting a fixed number of individuals with the outcome of interest (the cases) and a number of comparable control individuals free of the outcome. Retrospective histories of predictor variables of interest are then collected via questionnaire after recruitment.

A well-known example of a case-control study is the Ille-et-Vilaine study of cancer conducted in France between 1972 and 1974. It includes 200 cases and 775 comparable controls, and was designed to investigate alcohol, diet, and tobacco consumption as risk factors for esophageal cancer in men. This is known as an *unmatched* study since cases and controls were sampled separately in predetermined numbers. An alternative type of case-control study is based on *matching* a fixed number of controls to each sampled case based on selected characteristics. Methods for matched studies are different and will be covered briefly below in Sect. 5.3.1.

Because the overall proportion of individuals is fixed by design in a case-control study (e.g., $200/995$, or approximately five controls per case for Ille-et-Vilaine), it is not meaningful to make direct comparisons of outcome risk (estimated as the proportion of individuals with the outcome) between groups defined by predictor variables, as is conventional in studies where participants are not sampled based on their outcome status. Rather, analyses are based on the distribution of predictors variables compared across case/control status. At first glance, this approach does not seem to address the fundamental question of whether or not the predictor is associated with increased risk of developing the outcome. For example, observing that

Table 5.20 Odds ratio for smoking and esophageal cancer

```
. tabodds case ditob, or

---------------------------------------------------------------------------
       ditob |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
   0-9 g/day |    1.000000          .            .            .           .
   10+ g/day |   10.407051      64.89       0.0000      5.119049   21.157585
---------------------------------------------------------------------------

. tabodds ditob case, or

---------------------------------------------------------------------------
        case |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
           0 |    1.000000          .            .            .           .
           1 |   10.407051      64.89       0.0000      5.119049   21.157585
---------------------------------------------------------------------------
```

self-reported alcohol consumption differed between cases and controls in Ille-et-Vilaine does not seemingly translate into a clear statement about esophageal cancer risk associated with alcohol use. Further, application of conventional measures of association to settings where the role of the outcome and predictor are reversed seemingly leads to unintuitive results. For example, observing that individuals with esophageal cancer risk are twice as likely (in terms of the relative risk) as cancer-free individuals to report a specified degree of alcohol consumption does not state the association in a way that makes the possible causal connection clear.

Recall that our definitions of the relative risk, risk difference, and odds ratios in Chap. 3 were stated in terms of the outcome probabilities. This limits their usefulness in retrospective settings such as case-control studies. However, it is a unique property of the odds ratio that it retains its validity as a measure of outcome risk, even for case-control sampling. To demonstrate this for a simple example, Table 5.20 presents odds ratios for the Ille-et-Vilaine study estimated using the `tabodds` procedure in Stata. The first part of the table gives the odds of the binary case-control status indicator `case` compared in two groups defined by the binary indicator `ditob` of moderate to heavy level of smoking (10+ grams/day of tobacco smoked), and the second part gives the corresponding odds ratio comparing moderate-to-heavy level of smoking between cases and controls. The estimated odds ratios are identical. This property does not hold for the risk difference and relative risk.

We can also demonstrate this property directly using the definition of the odds ratio. Table 5.21 presents a hypothetical 2 × 2 table for a binary outcome and predictor in terms of the frequencies of n individuals in the four possible cross-categorizations (labeled a, b, c, and d). We estimate the outcome probability among individuals with and without the predictor with the proportions $a/(a+c)$ and $b/(b+d)$, respectively, and the corresponding odds of the outcome as

$$\frac{a/(a+c)}{c/(a+c)} \text{ and } \frac{b/(b+d)}{d/(b+d)}. \tag{5.11}$$

The resulting odds ratio is then ad/bc.

Table 5.21 Outcome by predictor status for a case-control study

	Predictor		
Outcome	Yes	No	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
Total	$a + c$	$b + d$	n

Similarly, we can estimate the exposure probability among individuals with and without the outcome as $a/(a + b)$ and $c/(c + d)$, and the corresponding odds as above. It is easy to verify that the odds ratio based on these is also ad/bc. This property of the odds ratios is central to the wide use of case-control studies, and suggests that logistic regression may be applicable as well. The additional fact that the odds ratio approximates the relative risk for rare outcomes (e.g., many forms of cancer) increases its appeal.

Recall that in the logistic regression model, the intercept coefficient β_0 is interpreted as the "baseline" log odds of outcome risk obtained when no predictors are included in the model (or, equivalently, when all predictors take on the value zero). As we have stated above, this quantity cannot be meaningfully estimated from case-control studies. As a result, the intercept coefficient in logistic regression models for case-control data can not be interpreted as providing an estimate of baseline risk in the population from which the sample was drawn. It is a remarkable fact that the logistic model is nonetheless directly applicable to data from case-control studies, and that estimated regression coefficients for included predictors provide valid estimates of log odds ratios, sharing the interpretation from other study types. Note that the logistic is the only binary regression model with this property.

A primary hypothesis underlying the Ille-et-Vilaine study was that alcohol consumption was related to esophageal cancer. Alcohol consumption was measured in average total daily consumption in grams, estimated directly from questionnaire responses on a number of different types of alcoholic beverages. The investigators recognized that age and smoking were potential confounding influences, and should be accounted for in assessing the association between alcohol consumption and cancer risk. (Dietary factors were also considered, but are not discussed here.)

Table 5.22 presents the results of a logistic regression model fit to these data, including a four-level categorization `alcgp` of average daily alcohol consumption and controlling for the dichotomous indicator `ditob` of moderate-to-heavy smoking (introduced above) and `age` (in years) as a continuous predictor. The lowest level of alcohol consumption (0–39 g/day) is taken as the reference category, and the three included indicators represent 40–79, 80–119, and 120+ g/day, respectively. The results indicate a clear increase in cancer risk with increasing alcohol consumption, and that this effect is evident when age and smoking are accounted for.

Estimated odds ratios in Table 5.22 are larger than 1.0, and the associated 95% CIs exclude 1.0, indicating that each of the predictors is associated with statistically

Table 5.22 Logistic model for alcohol consumption and esophageal cancer

```
. logistic case i.alcgp i.ditob age

Logistic regression                               Number of obs   =        975
                                                  LR chi2(5)      =     280.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -354.34556                       Pseudo R2       =     0.2838

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       alcgp |
          2  |   4.063502   1.024363     5.56   0.000      2.47926     6.66007
          3  |   7.526931   2.138602     7.10   0.000     4.312895    13.13612
          4  |   32.07349   11.58611     9.60   0.000     15.80015    65.10752
             |
     1.ditob |   7.375744   2.732364     5.39   0.000      3.56842    15.24529
         age |   1.068417   .0087666     8.07   0.000     1.051372    1.085738
------------------------------------------------------------------------------
```

significant increases in risk of esophageal cancer. Further, since esophageal cancer is relatively rare in the general population on which this study was conducted, interpreting the odds ratios as estimated relative risks is approximately correct.

A single summary of the contribution of alcohol consumption to a model including age and smoking can be obtained by fitting the same model excluding the indicators for alcohol, and performing a likelihood ratio test, as shown in Table 5.23. This procedure assumes that the full model including alcohol in Table 5.22 is fit first, and the model log likelihood is stored for future reference as `mod1` (in the second line of the output in Table 5.23). The results indicate a substantial contribution of the categorical summary `alcgp` of alcohol consumption to the overall fit of the model as summarized by the large log LR statistic (128.7). Further analyses might investigate the relationship between alcohol, smoking, and the log odds of cancer risk in more detail, possibly including these variables as continuous measures. We would naturally want to evaluate the linearity assumption implicit in including the variables (and age) in this form as well.

5.3.1 Matched Case-Control Studies

Consider the issues that would arise in designing a case-control study investigating esophageal cancer in a different population than Ille-et-Vilaine, possibly focusing on exposures other than alcohol as potential risk factors: We certainly would like to take into account known confounding factors such as those considered above as part of our design. If there are many such variables, we may be concerned that they will not be well represented in our chosen sample, and/or that analyses accounting for their influence may be overly complex. If we could recruit study subjects accounting for their profiles for these suspected confounders, we might be able to avoid some of these difficulties. This is the rationale for *matching*. We can

Table 5.23 Likelihood ratio test for contribution of `alcgrp`

```
. quietly logistic case i.alcgp i.ditob age
. est store mod1
. logistic case i.ditob age

Logistic regression                               Number of obs   =        975
                                                  LR chi2(2)      =     152.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -418.68894                       Pseudo R2       =     0.1537

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.ditob |   9.463852   3.362354     6.33   0.000     4.716825     18.9883
         age |   1.055568   .0073642     7.75   0.000     1.041232      1.0701
------------------------------------------------------------------------------
. lrtest mod1

likelihood-ratio test                                  LR chi2(3)  =    128.69
(Assumption: . nested in mod1)                         Prob > chi2 =    0.0000
```

build in control for confounding by incorporating knowledge of known confounders into the design of the study. By matching cases with controls that have the same values of these variables, we ensure control for confounding by comparing cases and controls within strata defined by the matching factors. In one of the simplest matched designs, disease cases are paired with controls into *matched sets* having similar values of the matching variables.

Because cases and controls within matched sets are sampled together based on shared values of the matching variables, the structure of the overall sample differs from that of an unmatched study. If we were to try to account for the sampling design via a standard logistic model that accounted for the matched sets with indicator variables, the number of parameters would frequently be too large for reliable estimation. For example, in a matched pair study with 200 matched pairs, as many as 199 parameters would be needed to account for the matching criteria. Clearly another regression approach is called for.

Regression modeling for matched data is based on a modification of the maximum-likelihood estimation approach used for the conventional logistic model (and described in more detail in Sect. 5.6). The *conditional logistic regression model* avoids estimating parameters accounting for the matching via *conditioning*. The parameters for predictors in this model have the log odds ratio interpretation familiar from the standard logistic model. The result is that we can conduct regression analyses exactly as before. However, the variables used in matching are controlled for automatically and not used directly in modeling. The `clogit` command in Stata provides a very convenient way to fit conditional logistic regression models. Most major statistical packages have similar facilities.

Matching is not always a good idea and should never be undertaken lightly. Effective matching (in cases where matching variables are strong confounders) can yield more precise estimates of the disease/exposure relationship. However, in cases where the matching variables do not actually confound the relationship between the exposure of interest and the outcome, the matching can lead to estimates with decreased precision relative to those obtained from an unmatched study. Further, satisfying matching criteria can be difficult and may result in a loss of cases. Good basic references for statistical analysis of data from matched case-control studies include Breslow and Day (1984) and Jewell (2004).

5.4 Checking Model Assumptions and Fit

Section 4.7 presented a number of techniques for assessing model fit and assumptions for linear regression models. Here, we cover many of the same topics for logistic models. Fortunately, many of the issues and techniques are similar and the methods from linear models apply more or less directly. One simplification of model assessment for binary outcomes is that no checks of distributional assumptions analogous to normally distributed residuals and constant variance are required. This is because the probability distribution for binary outcomes has a simple form that does not include a separate variance parameter. All required parameters are included in the model for the relationship between the log odds of the outcome and the predictors as described in Sects. 5.1 and 5.2. By contrast, construction and interpretation of graphical methods of assessment are more complex because of the nature of residuals from logistic models. We focus here on issues that differ from the approaches discussed in Sect. 4.7. We also note that additional issues arise in assessment of models for repeated or longitudinal binary outcomes such as those introduced in Chap. 7, due to the nature of the assumed dependence between outcomes.

5.4.1 Linearity

In Table 5.2, we fit a simple logistic regression model relating CHD risk and age for the WCGS data. In addition to providing a simple description of the relationship, the model makes it easy to compute the log odds associated with an arbitrary value of age. However, as in simple linear regression (Sect. 4.7), the uncritical adoption of the assumption that variables are linearly related to the outcome can lead to biased estimates and incorrect inferences. LOWESS scatterplot smoothing methods (introduced in Chap. 2) offer an exploratory approach to assessing the form of relationship between the log odds of the outcome and age that obviates the need to impose a particular parametric form. In the case of binary outcomes, these average the outcome proportions (or the corresponding log odds) over groups whose size is specified the bandwidth of the selected smoothing method. Figure 5.4 displays the log odds estimated by LOWESS (obtained using the `lowess` command in Stata with the `logit` option) along with the linear logistic fit. The latter is represented by the dashed line, obtained by simply plotting the log odds estimated by the model

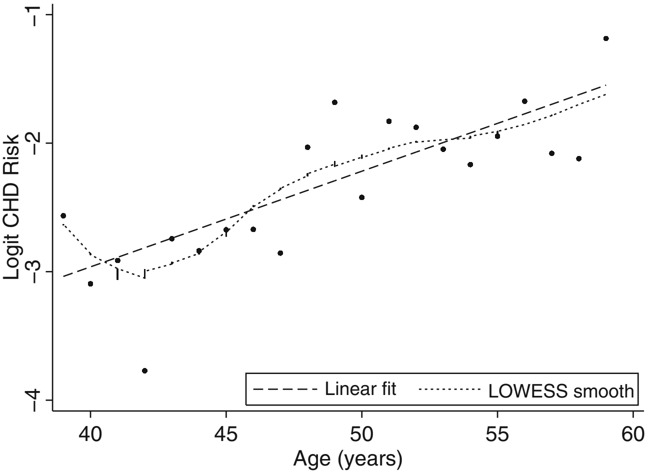

Fig. 5.4 Assessing linearity in the relationship between CHD risk and age

for all the (3,154) individuals in the sample. The smoothed estimated is given by the dotted line. The plotted points are the empirical log odds of the outcome for each of the unique values of age observed in the sample.

Although not conclusive, the results indicate that the linear logistic model fits the data reasonably well. However, the smoothed estimate suggests an initial decrease in the log odds of CHD risk for ages less than 42, followed by a fairly regular increase. The decrease might be due to elevated CHD risk among younger participants. In fact, 7% of the 39-year-olds ($n = 266$) in the study had CHD compared to 4% of the 40-year-old participants. The initial decline in the smoothed estimate is clearly influenced by the observed 2% rate of CHD among the 42-year-olds as well. A reasonable approach to evaluating this further would be to test for particular departures from linearity by adding polynomial terms in age or using restricted cubic splines (similar to the approach described in Sect. 4.10). Table 5.24 displays results from a model including a quadratic term in age (centered to reduce possible collinearity with the linear term). The Wald test statistic clearly indicates that the addition of this term does not afford a statistically significant improvement in the fit over the linear model. We can conclude that the linear model is adequate.

If the role of age in modeling is primarily as an adjustment factor, we would also want to examine whether the assumption of linearity impacts inferences about other predictors. Adoption of the linear form is acceptable if no impacts are seen, but predictions of outcome risk based on the linear model may yield biased results for ages not well represented in the data. Diagnostics for checking linearity in the context of multiple predictor models are somewhat less well developed for logistic models than for linear models. For example, tools like the component plus residual (CPR) plots presented in Sect. 4.7 are not generally available. However, the techniques presented here in combination with LR comparisons of models are

Table 5.24 Logistic model incorporating a quadratic effect of age

```
. logistic chd69 age agesq, coef

Logistic regression                               Number of obs   =       3154
                                                  LR chi2(2)      =      42.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -869.14333                       Pseudo R2       =     0.0241

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0769963   .0150015     5.13   0.000     .0475938    .1063987
       agesq |  -.0005543   .0021066    -0.26   0.792    -.0046831    .0035745
       _cons |   -6.04301    .678737    -8.90   0.000    -7.37331    -4.71271
------------------------------------------------------------------------------
```

usually sufficient to diagnose and correct nonlinearity problems. The increased availability of nonparametric regression approaches for binary regression (discussed briefly in Sect. 5.5) is rapidly expanding the arsenal of available tools in this area.

5.4.2 Outlying and Influential Points

Similar to the definition of residuals for linear regression (in Sect. 4.7), *standardized Pearson residuals* for logistic regression models are based on comparing observed values of the outcome variable with predictions from a fitted model. However, because outcomes in logistic models are binary, the values of these residuals cluster in two groups corresponding to the two values of the outcome. This makes graphical displays of residuals more difficult to interpret than in the linear regression case. An exception occurs when there are relatively few unique covariate patterns in the data (e.g., when predictors are categorical) and residuals and predictions can be grouped.

Figure 5.5 shows standardized Pearson residuals for the model in Table 5.18, plotted against the ordered observation number for the individual subjects. This *index plot* allows observations with unusually large residuals relative to other observations to be identified and investigated as potential outliers. The grouping of residuals based on outcome status is evident from the plot. In this case, although a number of observations have fairly large residuals (i.e., greater than two), none appear to be indicative of outlying observations. A number of other plots based on residuals are possible. In our experience, these are less useful in general than the investigation of influential points discussed in the next paragraph.

Diagnostic techniques for identifying influential observations in logistic regression models are also quite similar in definition and interpretation to their counterparts for linear regression. Most statistical packages that feature logistic regression allow computation of influence statistics that measure how much the estimated coefficients for a fitted model would change if the observation were deleted. Figure 5.6 shows influence statistics (often called DFBETA values) for the model in Table 5.18, plotted against the estimated outcome probabilities.

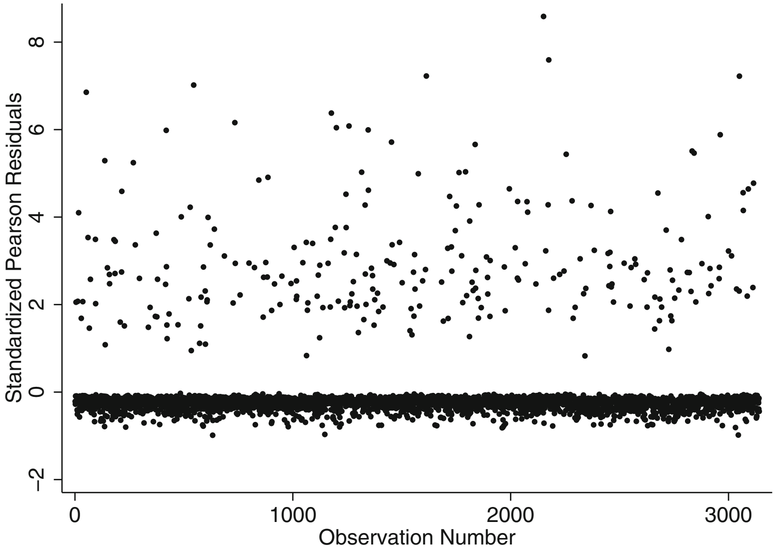

Fig. 5.5 Standardized pearson residuals for logistic model in Table 5.18

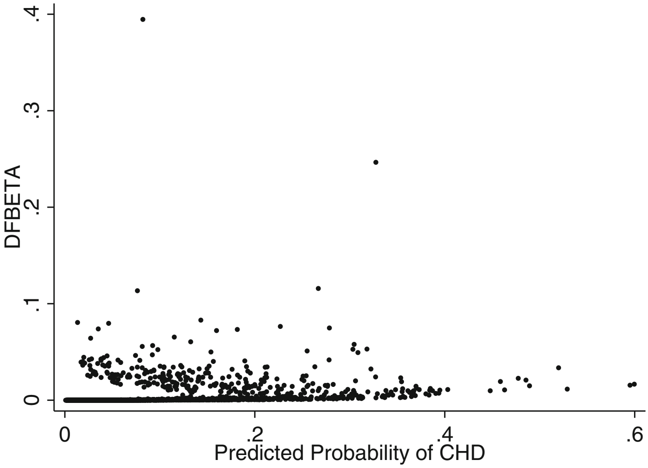

Fig. 5.6 Influence statistics for logistic model in Table 5.18

Two observations appear to have more influence than the rest. The most extreme observation is for an individual who is a nonsmoker with CHD, characterized by below average cholesterol (188) and a very high BMI value (39). Deletion of either observation (or both) resulted in no noticeable changes to model coefficients. Since there is no reason to suspect that any of the data are incorrect, both observations were retained.

Table 5.25 Link test for logistic model in Table 5.18

```
. linktest

Logit estimates                                   Number of obs   =       3141
                                                  LR chi2(2)      =     200.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -786.89258                       Pseudo R2       =     0.1130

------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .5646788    .306056     1.85    0.065    -.0351799    1.164538
      _hatsq |  -.1002356   .0688901    -1.46    0.146    -.2352576    .0347865
       _cons |  -.3983753   .3230497    -1.23    0.218    -1.031541    .2347904
------------------------------------------------------------------------------
```

5.4.3 *Model Adequacy*

The techniques discussed above address potential nonlinearity in the relationship between the log odds of the outcome and the predictor, but implicitly assume that the logistic model is correct. Recall from Sect. 4.7 that transformations of the outcome variable can be used to ensure that the distribution of the errors in a regression model are normally distributed. In a similar way, we can investigate the adequacy of the logistic model.

5.4.3.1 Specification Tests

A simple (and rather crude) approach to evaluating whether a given logistic model provides an adequate description of the data is through the use of a *specification test*. The `linktest` procedure in Stata is an example. Table 5.25 presents the results of applying `linktest` immediately after fitting the model in Table 5.18. This test involves fitting a second model, using the estimated right-hand side (i.e., the linear predictor) from the previously fitted model as a predictor. We would expect that the Wald test result for this predictor (labeled `_hat`) to be statistically significant if the original model provided a reasonable fit. The model fit by `linktest` also includes the square of this predictor (labeled `_hatsq`). The Wald test for inclusion of the latter variable is used to evaluate the hypothesis that the model is adequate; that is, the inclusion of the squared linear predictor should not improve prediction if the original model was adequate. Rejection indicates that the model is inadequate, and that an alternative binary regression model should be considered. Inadequacy may reflect the fact that even though important predictors are included and modeled correctly, the logistic model is not an appropriate representation of the relationship between outcome and predictors. It may also indicate that important predictors have been omitted, or are represented incorrectly in the model. The test can not distinguish between these two alternative explanations. It also does not suggest what alternate model form might be preferable.

In the example, the P-value for the Wald test for the predictor `_hatsq` does not provide strong evidence of inadequacy of the logistic model. However, the fact that the P-value for the predictor `_hat` in Table 5.25 is also not very small provides some indication that the overall fit may not be very good. (This is consistent with the large residuals noted in Sect. 5.4.2.)

Possible alternatives to the logistic model were discussed in Sect. 5.1, and will be covered in more detail in Sect. 5.5. Because these typically involve the use of specialized methods of estimation and result in coefficients with different interpretations, they are rarely used in practice. Fortunately, differences between results from alternative models are often small, and the logistic model applies in a very wide range of problems involving binary outcomes. Problems with fit can frequently be addressed using judicious selection and appropriate transformations of predictors.

5.4.3.2 Goodness of Fit Tests

Another approach to assessing model adequacy is provided by *goodness of fit* tests. The *Hosmer–Lemeshow* test is an example of this approach applicable to binary regression models such as the logistic. The test works by forming groups of the ordered, estimated outcome probabilities (e.g., ten equal-size groups based on deciles of the distribution of the outcome probabilities) and evaluating the concordance of the expected outcome frequencies in these groups with their empirical counterparts. The underlying hypothesis is that the estimated and observed frequencies agree. Thus, a statistically significant finding (i.e., rejection) indicates lack of fit. A nonsignificant finding rules out gross lack of fit.

Table 5.26 displays results of the Hosmer–Lemeshow test for the regression model fitted in Table 5.18. The `table` option requests that the observed and expected frequencies of the binary outcome (ones and zeros) for the requested groups be printed as well. The nonsignificant results do not indicate evidence for gross lack of fit. Increasing the number of groups to 20 yields a larger P-value (0.35), illustrating the sensitivity of the test to the number of groups chosen, and raising the possibility that judicious choice of group size may allow an investigator to choose the number of groups resulting in the most favorable P-value. To avoid this subjectivity, ten groups are generally recommended.

The Hosmer–Lemeshow test has a number of serious limitations. First, it is not sensitive to a number of sources of lack of fit such as misspecification of the model, and lacks power in these situations as a consequence. Further, the results of the test depend on the number of groups specified as well as the distribution of predictor values within these groups. Finally, the test can be very sensitive to fairly small fit discrepancies in large samples. Thus, a significant result may not signal a serious fit problem in such cases. Similarly, failure to find a statistically significant result does not necessarily mean that the model fits the data well. This test is most useful as a very crude way to screen for fit problems, and should not be taken as a definitive diagnostic of a "good" fit. Use in conjunction with a specification test (such as

Table 5.26 Hosmer–Lemeshow goodness of fit test

```
. lfit, group(10) table

Logistic model for chd69, goodness of fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0160 |     1 |   3.3 |   314 | 311.7 |   315 |
  |     2 | 0.0251 |     6 |   6.5 |   308 | 307.5 |   314 |
  |     3 | 0.0344 |    11 |   9.3 |   303 | 304.7 |   314 |
  |     4 | 0.0450 |    12 |  12.5 |   302 | 301.5 |   314 |
  |     5 | 0.0575 |    18 |  16.0 |   296 | 298.0 |   314 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.0728 |    10 |  20.4 |   304 | 293.6 |   314 |
  |     7 | 0.0963 |    28 |  26.5 |   286 | 287.5 |   314 |
  |     8 | 0.1268 |    44 |  34.7 |   270 | 279.3 |   314 |
  |     9 | 0.1791 |    50 |  46.7 |   264 | 267.3 |   314 |
  |    10 | 0.5996 |    76 |  80.3 |   238 | 233.7 |   314 |
  +--------------------------------------------------------+

       number of observations =      3141
             number of groups =        10
      Hosmer--Lemeshow chi2(8) =        11.36
                  Prob > chi2 =        0.1824
```

the one described above) may provide a bit broader screen to detect problems. However, results of either approach should not be relied on to guarantee model fit in the absence of supplementary investigations, including diagnostic assessment of residuals and influential observations.

5.4.4 Technical Issues in Logistic Model Fitting

In some cases, measures of association for binary outcomes such as odds ratios and relative risks take on the value zero, or are infinite. This happens when subgroups formed by the predictors are homogeneous with respect to outcome status. This translates to estimation problems in regression models, where parameters are typically represented as the logarithm of the underlying association measures.

Table 5.27 presents an example from the WCGS study using a four-level categorization of cholesterol level (0–150, 151–200, 201–250, and 251+) as a predictor of CHD outcome. Note the missing odds ratio estimates and the `note` explaining that "0.cholc dropped and 89 obs not used." Examination of the data reveals that there are no observed CHD cases among the 89 individuals with cholesterol in the default reference category (0–150 mg/dL). Because the odds of CHD are zero for this group, it is not possible to estimate valid odds ratios for the other categories. Choosing an alternate reference group allows valid estimates to be made. However, the odds ratio of zero for the lowest category still causes a fitting issue: the log odds ratio is infinite, and the parameter can not be estimated.

The problem raised in this example can be addressed by choosing a different categorization of cholesterol. However, this approach changes the interpretation of

Table 5.27 Logistic model for CHD and categorized cholesterol level

```
. logistic chd69 i.cholc
note: 0.cholc != 0 predicts failure perfectly
      0.cholc dropped and 89 obs not used

note: 3.cholc omitted because of collinearity

Logistic regression                               Number of obs   =       3053
                                                  LR chi2(2)      =      52.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -855.50635                       Pseudo R2       =     0.0299

------------------------------------------------------------------------------
       chd69 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       cholc |
         0   |  (empty)
         1   |   .2884527   .0574556   -6.24   0.000     .1952214    .4262082
         2   |   .4514053   .0642673   -5.59   0.000     .3414914    .5966966
         3   |  (omitted)
------------------------------------------------------------------------------
```

the categorized variable, and will not work in all cases. In small samples, frequently no amount of regrouping or recategorizing will eliminate these issues. In these situations, exact logistic regression methods (discussed in Sect. 5.5.4) should be considered. When exact methods are not computationally feasible, the penalized maximum likelihood approach proposed by Firth (1993) provides another possible alternative. This is available for Stata in a downloadable, user-defined module entitled *firthlogit*. We recommend that a statistician be consulted to diagnose the exact nature of the problem and suggest appropriate solutions.

Another issue to consider in fitting logistic regression models with multiple predictors is deciding how many predictors is appropriate. Fitting too many predictors can result in biased estimates and incorrect inferences. The severity of these problems is directly related to the sample size, the number of observed outcomes, and the distribution of outcomes over the included predictors. Chapter 10 discusses model building strategies in detail, and Sect. 10.4.2 provides guidelines for selection of appropriate number of predictors.

5.5 Alternative Strategies for Binary Outcomes

A review of current clinical and epidemiological research studies involving binary outcomes will reveal that the overwhelming majority of regression analyses are based on the logistic model. In some instances, specific knowledge about a disease-exposure relationship may suggest a different model. Alternatively, it may be desirable to summarize observed associations using measures such as the relative risk or risk difference in preference to the odds ratio. Because the logistic model yields only the latter, there are situations where alternative regression approaches may be preferred. Finally, diagnostic evaluations may lead to the conclusion that the logistic model is simply not right for a particular data set. In this section, we review some examples of alternative approaches to binary regression. We also briefly discuss models for categorical outcomes with more than two levels.

5.5.1 Infectious Disease Transmission Models

Recall the CDC transmission study data discussed in Sect. 3.4 (O'Brien et al. 1994). The goal of this study was to investigate risk factors for sexual transmission of HIV in susceptible female partners of previously infected males. Although the outcomes were restricted to prevalent HIV serostatus measured at enrollment, the infection dates of the male partners were approximately known from transfusion records. In addition, self-reported information on number of unprotected sexual contacts was also collected. These data pertain to contacts that occurred between the time of infection of the male partner and the time of enrollment. (Note that monogamy was an eligibility criterion, to reduce the possibility of infection from other sources.)

Unlike many chronic diseases, the mechanism of acquisition of many infectious diseases is well understood. In these cases, simple probabilistic *transmission models* linking outcomes with exposures are frequently used to quantify infection risk. One of the most basic such models links the cumulative probability of escaping infection following a series of exposed contacts. The model assumes that each contact carries an identical risk λ of infection, and that outcomes of successive contacts are independent. Under these assumptions, the chance of escaping infection following k contacts is

$$(1-\lambda)^k,$$

with the complementary probability of being infected following k contacts given by

$$P(k) = 1-(1-\lambda)^k.$$

This model corresponds well to the observed data from the CDC study: each female partner can be characterized by the binary infection status and the reported number of exposed contacts k (the predictor), with the outcome probability given above. This suggests that a binary regression approach linking these two variables would be ideal for estimating the per-contact transmission probability λ. Unfortunately, the logistic model does not provide a direct estimate. By contrast, an alternative transformation of $P(k)$, known as the complementary log–log, provides a model with a more appealing structure:

$$\log\{-\log[1-P(k)]\} = \log[-\log(1-\lambda)] + \log(k). \tag{5.12}$$

This model is similar to the familiar linear model

$$\log\{-\log[1-P(x)]\} = \beta_0 + \beta_1 x, \tag{5.13}$$

Table 5.28 Complementary log–log regression model for per-contact risk

```
. glm hivp, family(binomial) link(cloglog) offset(logcontacts)

Generalized linear models                          No. of obs      =        31
Optimization     : ML: Newton-Raphson              Residual df     =        30
                                                   Scale parameter =         1
Deviance         =   40.8340195                    (1/df) Deviance =  1.361134
Pearson          =  84.90572493                    (1/df) Pearson  =  2.830191

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(-ln(1-u))             [Complementary log-log]
Standard errors  : OIM

Log likelihood   = -20.41700975                    AIC             =  1.381743
BIC              = -62.18559663

----------------------------------------------------------------------------
        hivp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
       _cons |  -7.033126   .3803284   -18.49   0.000    -7.778556   -6.287696
 logcontacts |   (offset)
----------------------------------------------------------------------------


. bootstrap "glm hivp, family(binomial) link(cloglog) offset(logcontacts)"
  _b _se, reps(1000)

command:      glm hivp, family(binomial) link(cloglog) offset(logcontacts)
statistics:   b_cons     = [hivp]_b[_cons]

Bootstrap statistics                            Number of obs    =        31
                                                Replications     =      1000
----------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err.  [95% Conf. Interval]
-------------+--------------------------------------------------------------
      b_cons |  1000 -7.033126 -.0629388  1.163788   -8.216878  -6.296359
----------------------------------------------------------------------------
```

where the intercept coefficient $\beta_0 = \log[-\log(1-\lambda)]$, but includes the predictor $x = \log(k)$ as a fixed *offset*, with corresponding coefficient $\beta_1 = 1$ as specified by model (5.12). Predictors with fixed coefficients are referred to as *offsets*, and can be easily accommodated by standard statistical software packages. (Part of the model evaluation procedure in this case may include checking whether this is reasonable in terms of fit.) Similar to the logistic model, an inverse transformation allows us to represent this model on the probability scale as follows:

$$P(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)], \tag{5.14}$$

Table 5.28 shows the results of fitting model (5.12) using the generalized linear model estimation program `glm` in Stata, which we explain in greater detail in Chap. 8. Note that the logarithm of the number of contacts `logcontacts` appears as an offset, and no coefficient for this predictor was estimated.

An additional calculation inverting the complementary log–log transform of the intercept `_cons` provides the estimate of λ:

$$\lambda = 1 - \exp[-\exp(-7.033)] = 0.0009.$$

The approximate 95% CI (0.0004, 0.0019) can be obtained via a similar calculation applied to confidence limits given in the regression output. Because of the small sample size ($n = 31$), the approximate CIs may not be reliable. For comparison, Table 5.28 also gives bias-corrected 95% bootstrap CIs (calculated using 1,000 bootstrap samples) for the same model. The bias-corrected CI (0.0003, 0.0018) for the parameter λ can be obtained from the interval for the intercept coefficient β_0 (represented by `b_cons` in the table) via the calculation used for the approximate interval. The lower bound of this interval is only slightly more conservative than the approximate interval, but otherwise they are remarkably similar. The bootstrap interval should still be considered a better summary of uncertainty about λ.

Clearly, model (5.12) is very simple, and a number of the underlying assumptions are questionable (e.g., that the per-contact risk λ is constant). However, it is a useful "null" model to which more complex alternatives may be compared. Further, the parameter λ is an important ingredient in more complex mathematical epidemic models. This model is also interesting because it is an example of a *proportional hazards model*. These arise frequently in studies where controlling for duration of follow-up is an important consideration in data analyses, and are the subject of the next chapter. Finally, model (5.13) and the conventional logistic model are examples of the family of GLMs that includes most of the regression models considered in this book.

5.5.2 Pooled Logistic Regression

The MIRA study was a randomized trial designed to investigate the effectiveness of diaphragms as a means of prevention of sexual transmission of HIV in women in sub-Saharan Africa (Padian et al. 2007). Here, we consider data on 1,000 randomly selected individuals participating in a substudy investigating risk factors for infection with herpes simplex virus type 2 (HSV-2), conducted among women testing negative for infection at enrollment (de Bruyn et al. 2011). The study design is characterized by visits at three month intervals following enrollment, with infection outcome and predictor information collected at each. HSV-2 infection can occur at most once, so individual outcomes can be summarized by a binary indicator of whether or not infection has occurred during follow-up. Also, the interval of infection occurrence is informative about the possible time of infection. For example, individuals at higher risk for infection at any time during follow-up may also tend to be infected earlier. Direct application of the logistic model described in previous sections would not account for this, or the fact that multiple observations of both predictors and outcomes are available. Methods for regression analysis of survival outcomes (as discussed in Chap. 6) based on precisely measured times of outcome occurrence also don't apply unless we make an assumption about the actual occurrence times of infections (e.g., the midpoint of the interval between visits). *Pooled logistic regression* provides a hybrid approach that avoids such assumptions, and also allows the information on outcome occurrence collected in multiple study visits to be used appropriately.

Table 5.29 Example data from MIRA study

```
. list id mos hsv2 age stihx newparts if id==2 | id==54

      +-----------------------------------------+
      | id   mos   hsv2   agecat   stihx   newparts |
      |-----------------------------------------|
   4. |  2     3      0        1       0          0 |
   5. |  2     6      0        1       0          0 |
   6. |  2     9      0        1       0          0 |
   7. |  2    12      0        1       0          1 |
   8. |  2    15      0        1       0          0 |
      |-----------------------------------------|
   9. |  2    18      0        1       0          0 |
  10. |  2    21      0        1       0          0 |
  11. |  2    24      0        1       0          0 |
 409. | 54     3      0        2       0          1 |
 410. | 54     6      0        2       0          1 |
 411. | 54     9      1        2       0          0 |
      +-----------------------------------------+
```

Table 5.29 illustrates observations of key study variables for two selected participants from the MIRA study. In addition to indicators of outcome occurrence `hsv2`, each individual contributes observations of predictors that are fixed (`agecat`, a categorical representation of age at enrollment; `stihx`, a binary indicator of self-reported history of prior sexually transmitted infections) or time varying (`newparts`, a binary indicator of self report of recent new sexual partners). In addition to outcome and predictor values, the follow-up duration (`mos`) is recorded as the number of months elapsed since enrollment. The first individual remained uninfected for all study visits, and provides measures of the HSV-2 outcome and fixed and time-varying predictors for each interval. The actual time of infection is said to be *right censored* at the time of the last visit. The second individual was first observed to be infected at the fifth visit (12 months), and was removed from observation for treatment thereafter. The time of infection in this case is censored into the interval between the fourth and fifth visits. This data structure is typical for application of the pooled logistic model, and also shares features with survival data that are the subject of the next chapter. We would like the analysis to assess the association between predictors and outcome occurrence, and also account for the duration of follow-up.

Table 5.30 displays the results of fitting a logistic regression model to the data partially shown in Table 5.29. Because we want to make as few assumptions as possible about how infection risk varies with duration of follow-up, the model uses a restricted cubic spline (discussed in Chap. 4) with three knots to account for the effects of time. The estimated odds ratios for the spline predictors (`spl1` and `spl2`), together with the intercept odds (not shown) can be regarded as the "baseline" infection odds that applies to individuals with the additional predictors of interest set to zero. The significant result for the `testparm` command, which evaluates the Wald test for the hypothesis that both spline coefficients equal zero, indicates that accounting for time variation via a spline results in a significantly improved fit to the data relative to a model including only an intercept term. The

Table 5.30 Pooled logistic regression model for MIRA study example

```
. logistic hsv2 spl* i.agecat i.stihx newparts

Logistic regression                               Number of obs   =       6069
                                                  LR chi2(6)      =      33.69
                                                  Prob > chi2     =     0.0000
Log likelihood = -509.18109                       Pseudo R2       =     0.0320

------------------------------------------------------------------------------
        hsv2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        spl1 |   .9256021   .0379368   -1.89   0.059     .8541555    1.003025
        spl2 |   1.035493   .0556198    0.65   0.516     .9320218     1.15045
             |
      agecat |
          2  |   .6090384   .1369677   -2.20   0.027     .3919372     .946396
          3  |   .5522162   .2213153   -1.48   0.138     .2517489    1.211297
             |
     1.stihx |    1.92962   .4938946    2.57   0.010     1.168431    3.186695
    newparts |   1.826313   .4372661    2.52   0.012     1.142288    2.919945
------------------------------------------------------------------------------

. testparm spl*

 ( 1)  [hsv2]spl1 = 0
 ( 2)  [hsv2]spl2 = 0

           chi2(  2) =   10.59
         Prob > chi2 =    0.0050
```

odds ratios for the additional predictors are interpreted similarly to the conventional logistic model studied in previous sections. For example, increased age is associated with decreased odds of infection relative to the youngest age group. Also, report of a recent new partner is associated with an approximate doubling of infection odds. As discussed previously, the form of the model specifies that the predictors act to increase or decrease baseline infection odds by fixed increments. This is an example of a *proportional odds model*. For outcomes that are rare, the odds ratios closely approximate *relative hazards* estimated from the closely related *proportional hazards model* that is discussed in the next chapter.

Although the pooled logistic regression model is widely applicable and easy to fit, it suffers from several disadvantages relative to the regression methods for survival data that are the subject of Chap. 6: First, it requires explicit modeling of the effects of time in the analysis, a feature not shared with the Cox proportional hazards regression model. Second, because event times are not known precisely, the causal links between time-varying predictors and event occurrence are less clear than in settings where such information is completely observed. Finally, in situations with missed and/or irregularly spaced intervals the estimates are susceptible to bias because of the need to make assumptions about the behavior of predictors and outcomes in unobserved periods of follow-up. Despite these issues, this approach is becoming increasingly popular for longitudinal data of the type described here, and, as discussed in Chap. 9, is commonly used in applications of regression to causal inference. More information about the approach, including a comparison to the Cox proportional hazards model is included in D'Agostino et al. (1990).

5.5.3 Regression Models Based on Risk Differences and Relative Risks

A recent study of prevalent human T-cell leukemia/lymphoma virus (HTLV) infection in infants born to mothers in the United Kingdom identified a number of factors associated with infection, including the parent's country of birth and ethnicity of the mother (Ades et al. 2000). The authors found that a regression model based on risk differences provided a better fit to the data than the logistic model, and reported their results accordingly.

Recall the linear regression model defined in (5.1) that relates risk for a binary outcome to a single predictor x:

$$P(x) = \beta_0 + \beta_1 x.$$

As noted in Sect. 5.1, the coefficient β_1 measures the risk difference associated with a unit increase in x. This model is often referred to as the "additive risk model" because the effect of any unit increase in the predictor x is to add an increment β_1 to the outcome risk. This was the model employed in the HTLV example. Although it provides a valid alternative to logistic regression, it is important to keep in mind the potential problems with fitting and interpretation (raised in Sect. 5.1).

As discussed in Sect. 3.4, the odds ratio is known to approximate the relative risk in the rare outcome setting. Consequently, odds ratios are frequently reported as relative risks in research findings. Unfortunately, this practice is not limited to rare outcomes, and has been the subject of considerable debate in the research literature (Holcomb et al. 2001). This has led many investigators to advocate that regression models based on the relative risk be used in preference to the logistic model (other than in case-control designs where standard regression approaches other than the logistic model do not directly apply). This is possible using the following regression model:

$$\log\left[P(x)\right] = \beta_0 + \beta_1 x. \tag{5.15}$$

This is the *log linear* model discussed in Sect. 5.1. The regression coefficient β_1 has the interpretation of the logarithm of the relative risk associated with a unit increase in x. Analogous to the procedure for obtaining odds ratios from logistic models, exponentiated coefficients yield relative risk estimates in this case. Although this model can be fitted with many standard software packages, numerical difficulties may arise because of the constraint that the sum of terms on the right-hand side must be no greater than zero for the results to make sense (due to the constraint that the outcome probability $P(x)$ must lie in the interval $[0, 1]$). In such cases, treating the observed binary responses as if they were distributed according to the Poisson distribution, and using estimation methods for GLMs (Chap. 8) generally yields very similar log relative risk estimates. If robust variances are used to estimate variability, the resulting inferences have been shown to yield results very similar to the conventional binomial estimation procedure and to avoid the associated

Table 5.31 Generalized linear models for CHD risk (P) and age (x)

Model	β_1 (95% CI)	Log-likelihood	$P(55)$
$P(x)$	$0.005(0.004, 0.007)$	-869.96	0.129
$\log[P(x)] - \textit{Binomial}$	$0.067(0.047, 0.087)$	-869.24	0.136
$\log[P(x)] - \textit{Poisson}$	$0.067(0.048, 0.087)$	-881.86	0.136
$\log\{-\log[1 - P(x)]\}$	$0.071(0.050, 0.092)$	-869.21	0.136
$\log\{P(x)/[1 - P(x)]\}$	$0.074(0.052, 0.097)$	-869.18	0.136

convergence problems with the latter (Zou 2004; Yelland et al. 2011). The Poisson approach is generally recommended in cases where relative risk estimates are desired and the log binomial model fails to converge.

Alternative approaches for obtaining adjusted relative risks from odds ratios estimated using logistic regression have been proposed in the literature (Zhang and Yu 1998). These are based on simple transformations of the estimated coefficients similar to the illustrative calculations demonstrated in Sect. 5.1.1. Unfortunately, such calculations can produce incorrect estimates for models including multiple predictors and should be avoided in favor of fitting appropriately defined regression models as described above (McNutt et al. 2003).

Table 5.31 presents the results of fitting five alternative GLMs for the relationship between CHD and age using the WCGS data. (Results were obtained with the Stata GLMs procedure `glm`, also applied in Table 5.28.) These correspond to the binomial regression models considered in this section (i.e., (5.1), (5.2), (5.13), and (5.15)) and the alternative Poisson regression approach. Results for the intercept parameter β_0 are similar. Note that the estimated regression coefficients cannot be directly compared because the models are based on different representations of the outcome. However, since all of them are based on the same number of parameters, comparison of the likelihoods provides a cursory look at how well they describe the data in relative terms. Although the likelihood for the logistic model is slightly larger, there is very little overall difference between the models. Similarly, the estimated coefficients for the log, complementary log–log, and logit models are remarkably similar. (The coefficients for the risk difference model differ because the outcome is modeled without transformation.) Finally, the estimated probabilities for a 55-year-old individual ($P(55)$) are also quite similar. Based on these results, there would be no particular reason to prefer any alternatives over the logistic model.

The results in Table 5.31 illustrate that a variety of models other than the logistic may be appropriate for a given problem. We note that additional binary regression models also exist that are useful in other contexts. For example, the *probit model* is used in the context of instrumental variable methods for binary outcomes in Sect. 9.7. However, given the ease of interpretation, wide use, and software availability of the logistic model, it is by far the most common choice in practice. In general, we advocate fitting the logistic model unless another model is preferable on scientific grounds. Lack of fit can often be dealt with via the techniques discussed in

Sect. 4.7, obviating the need to investigate alternative model formulations. Finally, note that the approaches discussed here are not directly applicable to data from case-control studies (Scott and Wild 1997).

5.5.4 Exact Logistic Regression

Recall the HIV transmission example considered in Tables 3.6 and 5.28. The dataset contains binary outcomes for 31 monogamous female sexual partners of males previously infected with HIV. With so few observations, the reliability of statistical inference relying on conventional statistical procedures such as the Wald and χ^2 test is questionable. The Fisher's exact test, discussed in Sect. 3.4 provides a useful alternative for outcome–predictor comparisons addressable using a two-by-two table. However, this approach limits inference to problems involving a single categorical predictor. The *exact logistic regression* model allows exact inferences to be applied in the regression setting, including models with continuous predictors.

In the example presented in Table 3.6, interest focuses on the possible association between presence of an AIDS diagnosis in the male partner and transmission to the female partner. In Table 5.28, we considered a specialized model linking the degree of sexual contact measured by the logarithm number of contacts reported by each partnership to transmission risk. In Table 5.32, we show the results of fitting both standard and exact logistic regression models to these data using Stata, including both the logarithm of the number of contacts and the indicator of AIDS diagnosis as predictors. Although no exact procedure is available to fit the GLM considered in Table 5.32, there is still interest in examining whether the observed effect of AIDS diagnosis from Table 3.6 is influenced by controlling for degree of exposure to infection. The estimated odds ratios from the two models are comparable. However, the degree of precision for the estimated effect of AIDS appears to be overstated in the standard logistic model. Note that in place of the Wald test results, the exact logistic reports columns labeled `Suff.` and `2*Pr(Suff.)`. These are based on *sufficient statistics* for each predictor in the model conditional on the values of the other predictor(s). Exact inference is based directly on these conditional distributions. If interest focuses on a particular predictor in a model (e.g., AIDS) it is possible to restrict inference to that variable, resulting in some computational savings.

Computational procedures for exact logistic regression are intensive, and frequently it will be unfeasible to fit models with multiple continuous covariates, even for datasets as small as considered in the above example. For this reason, the exact approach is recommended for small samples (typically less than 100), especially when *P*-values from standard asymptotic approaches such as the Wald test are in the range of plausible significance. Exact logistic regression can also be useful in situations where standard models fail to yield valid estimates, such as those discussed in Sect. 5.4.4. Finally, note that most of the procedures discussed for model assessment and postestimation inference that are applicable for the standard logistic model are not available for exact logistic regression.

Table 5.32 Conventional and exact logistic regression models for transmission risk in female partner for the CDC example

```
. logistic hivp logcontacts i.aids

Logistic regression                               Number of obs   =         31
                                                  LR chi2(2)      =       4.38
                                                  Prob > chi2     =     0.1119
Log likelihood = -14.368496                       Pseudo R2       =     0.1323

----------------------------------------------------------------------------
        hivp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
 logcontacts |   1.183255   .3653485     0.54   0.586     .6460355    2.167206
      1.aids |   8.091292    8.64643     1.96   0.050     .9963591    65.70824
----------------------------------------------------------------------------

. exlogistic hivp logcontacts aids

Exact logistic regression                          Number of obs =        31
                                                   Model score   =  4.855402
                                                   Pr >= score   =    0.0813
---------------------------------------------------------------------------
        hivp | Odds Ratio       Suff.  2*Pr(Suff.)     [95% Conf. Interval]
-------------+-------------------------------------------------------------
 logcontacts |   1.169055       35.86      0.6319      .6776411    2.229598
        aids |   8.085458           3      0.1547      .5835864    492.6288
---------------------------------------------------------------------------
```

5.5.5 Nonparametric Binary Regression

The examples of alternative techniques for binary regression considered above represent only a small subset of the available possibilities for estimating the relationship between a binary outcome and a predictor variable. The goal of *nonparametric regression* methods is to provide estimates of this relationship based on minimal assumptions about its form.

Recall the assessment of linearity for the logistic model for the relationship between CHD and age in the WCGS data in Sect. 5.4.1. The smoothed LOWESS estimate displayed in Fig. 5.4 is an example of a nonparametric logistic regression model for this relationship. Although the assumption that the predictor is related to the disease outcome in an additive fashion via the log odds is retained, this technique allowed us to relax the assumption that the relationship is linear by assuming only that the change in CHD risk with age has a certain degree of smoothness. This can prove very useful in exploring the form of the relationship between outcome and predictor, but does not yield readily interpretable parameter estimates or generalize easily to models including more than one predictor. The class of *generalized additive models* provide an extension to the LOWESS technique, allowing multiple predictors to be fit simultaneously, each of which can be represented as a smooth function (Hastie and Tibshirani 1999). Although very useful in evaluating outcome–predictor relationships, these models are frequently difficult to fit and interpret.

Methods for significance testing, CIs, and model evaluation are less well developed for nonparametric alternatives than for conventional logistic regression. In addition, decisions about degree of smoothness and interpretation of resulting estimates is often very complex. Finally, practical implementations of nonparametric binary regression that handle multiple predictors are not widely available in standard statistical packages. For these reasons, we recommend that flexible parametric approaches be used in accounting for nonlinearities in the relationship between predictor and outcome, and that nonparametric alternatives be used primarily for exploratory purposes.

Classification trees (Breiman et al. 1984) are another popular approach to nonparametric binary regression. As discussed in Sect. 5.2.5, these lack the linear and additive structure shared by other approaches, and have been used to develop prediction tools for using measured characteristics to correctly distinguish binary outcomes. However, classification trees can also be used to explore complex relationships between multiple predictors and a binary response. Because they do not yield estimates of association parameters, interpretation of the contribution of individual predictors to the outcome risk is complex. However, like the nonparametric regression approaches discussed above, they are very useful tools in exploratory analyses and can be very helpful in discovering and interpreting interaction.

5.5.6 More Than Two Outcome Levels

Research studies frequently yield outcomes that have multiple categories. (See Chap. 2 for definitions of categorical variable types.) Consider the back pain example introduced in Sect. 1.1, where pain intensity was measured on an ordered, ten-point scale. In addition to the *ordinal* categorical outcome just considered, *nominal* categorical outcome measures are also commonplace in clinical research. For example, the outcome in a study of cancer outcomes by cell type is a nominal categorical variable. Both type of outcomes can be investigated using contingency table methods. The limitations of these when multiple predictors are involved are clear. For certain questions, considering a binary representation might also be reasonable. For example, to investigate factors that distinguish patients suffering from severe pain from all others in the pain example. In this case, logistic regression is an appropriate tool to consider. However, there is clearly information lost in reducing ten levels down to two. In the remainder of this section we briefly review regression methods for nominal and ordinal categorical outcomes.

5.5.6.1 Ordinal Categorical Outcomes

The *proportional odds model* is a commonly used generalization of the logistic model that accommodates a multilevel categorical response with ordered categories. Rather that modeling the probability of response in a particular category, this model is based on the cumulative probability that the response is not greater than a chosen category. The dependence of this response on predictors is identical to the form of the logistic model. For the back pain example, (assuming a ten-level response and a single predictor x), the form of this model for a response probability of severity no greater than 5 is given by

$$\log\left[\frac{\Pr(y \leq 5)}{\Pr(y > 5)}\right] = \alpha_5 - \beta x.$$

A similar expression applies to all ten levels of the response. (We assume that the levels of the response are coded $1, 2, \ldots, 10$.)

Note that the intercept parameter α_5 is unique to this response level, and represents the probability of a response of no more than 5 among individuals with $x = 0$. Because the response is expressed as a cumulative probability, the intercept coefficients are constrained as $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{10}$. The coefficient β is interpreted as the log odds ratio associated with a unit increase in x, assumed to be constant across response levels. (i.e., response levels are parallel, each with slope β.) This assumption amounts to a strong restriction on the effect of the predictor on the response, and needs to be validated.

Note that there are many alternatives to the proportional odds model, including the *continuation ratio model*. We refer the reader to the references provided below for additional information on these.

5.5.6.2 Nominal Categorical Outcomes

When there is no natural ordering implicit in a categorical response, or when the assumptions implicit in the models above do not apply to an ordinal outcome, the *multinomial logistic* model (also known as the polytomous logistic model) can be used for regression analyses. For a single predictor x, the model specifies that each response level follows a logistic regression model for x, with a selected level specified as the reference. The regression coefficients for each level are unique; so for the pain example the model would include nine intercept and slope coefficients. For level 5, and specifying the first level as the reference category, the model would take the form

$$\log\left[\frac{\Pr(y = 5)}{\Pr(y = 1)}\right] = \alpha_5 + \beta_5 x.$$

Notice that when there are more than two outcome levels, the two levels specified in the model are not binary alternatives. The outcome then represents a log relative risk rather than a log odds. Thus, the coefficient β_5 represents the change in the log relative risk for level 5 (relative to the reference level 1) associated with a unit increase in x. The exponentiated value of this coefficient is interpreted as a *relative risk ratio* rather than an odds ratio. Because this model does not involve the restrictions implicit in the proportional odds model, it is an attractive alternative when the proportional odds assumption is not satisfied. However, because of the potentially large number of parameters and the flexibility of choice for the reference group, the multinomial logistic model can be challenging to interpret.

The models outlined here represent a few of those available for analyzing categorical responses. For further information on these and other models, including examples and a description of available software resources, see Ananth and Kleinbaum (1997) and Greenland (1994).

5.6 Likelihood

One of the common themes uniting methods presented in this book is the principle of using observed data to estimate unknown quantities of interest. The majority of the methods presented are regression models relating outcome and predictor variables measured on a sample of individuals. The principal unknown quantities in the models are the regression parameters. Once these are estimated, inferences can be made about the true values of these parameters and related quantities of interest such as predicted outcomes. All available information about the parameters is contained in the observed data. A standard approach to estimating parameters in models like the ones covered here is known as *maximum likelihood estimation*. Although not required for applications, a basic understanding of this topic helps in unifying the concepts underlying estimation and inference in most of the regression models covered in this book. Here, we provide a brief discussion of some of the key ideas in the binary regression context.

The *likelihood* associated with a set of independent observations of an outcome is just the product of their respective probabilities of occurrence under the assumed model relating outcomes to predictors. Because this represents the joint probability of observing all of the outcomes in the sample, the likelihood can then be interpreted as a measure of support provided for the model by the data. The maximum-likelihood estimate of the parameter(s) is the value for the parameter(s) that yields the maximum value of the likelihood for the observed data.

To take a very simple example from the binary outcome context, consider the problem of estimating the prevalence of HIV for the sample of 31 female partners of previously infected males from the CDC transmission study considered in the examples presented above and in Sect. 3.4. The assumed model is that the actual prevalence in the target population is represented by a constant that we can symbolize by P (similar to the definition introduced earlier in this chapter). We can think of P as the probability that a randomly sampled individual will test positive. The corresponding probability of observing a negative is $1 - P$. However, P is unknown. The observed data consist of the 31 indicators of HIV status, and the likelihood, as defined above, is just the product of the individual outcome probabilities:

$$P^7 \times (1 - P)^{23}.$$

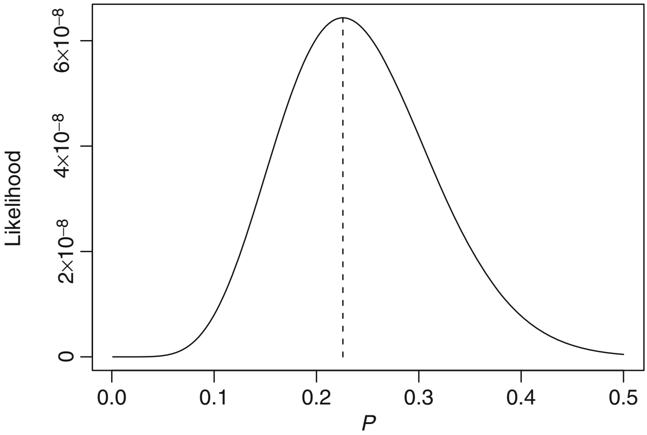

Fig. 5.7 Likelihood function for HIV prevalence

The likelihood is formed as the product of the individual outcome probabilities because these are independent events. It is a function of the unknown constant P, with the observed infection indicators providing the number of positive and negative individuals. Figure 5.7 presents a plot of this function for a range of values for P. The maximum-likelihood estimator of P is just the value of P that maximizes the likelihood function. This value is indicated in the figure. The maximum can be found easily in this example using calculus. Not surprisingly, it corresponds exactly to the intuitive estimate of the actual prevalence of HIV-positive individuals in the sample of 31: Because there are seven such individuals in the sample, the estimated prevalence is 0.226. For more complicated models (e.g., regression models with multiple predictors) computing the maximum typically involves iterative calculations on a computer.

Likelihood functions for binary regression models are defined following the procedure used above, but the outcome probability P for each individual is replaced with the form defined by the logistic model (5.2). To take another example from the CDC study, consider a regression model relating HIV status of the female partners to a binary indicator of presence of an AIDS diagnosis in the male. (This example was already considered in Sect. 3.4.) Following our conventional notation, we will represent the outcome as Y and the predictor as x. The observed data now include both Y and the binary predictor x for each individual in the sample. The likelihood takes exactly the same form as in the last example, except the constant P is replaced with the expression for the logistic model, substituting in each individual's value of the predictor (i.e., x_i for the i_{th} individual):

$$\prod_{i=1}^{31}\left[\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right]^{Y_i}\times\left[1-\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right]^{1-Y_i}.$$

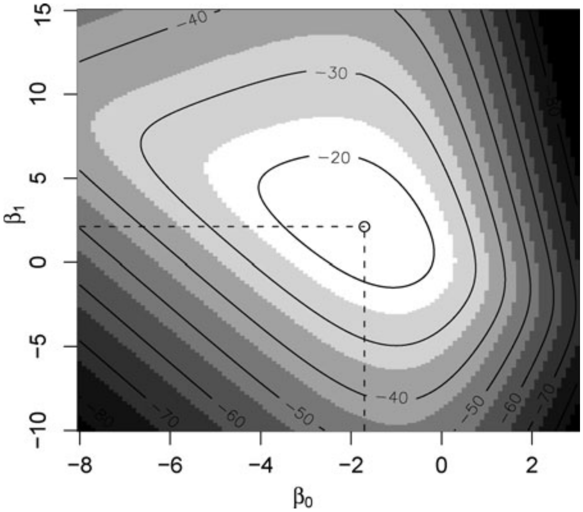

Fig. 5.8 Likelihood function for a two-parameter logistic model

Since both Y and x (the indicator of AIDS status) are observed, the only unknown quantities are the regression parameters β_0 and β_1. These are generally estimated using an iterative maximization algorithm. Figure 5.8 presents a plot of the logarithm of this function for a range of values for P. Because the likelihood function depends on two unknown parameters, it has the form of a "surface" when plotted in three dimensions. The two-dimensional figure represents the contours of this surface as seen from above. The maximum value is indicated, and the corresponding maximum-likelihood estimates for β_0 and β_1 are –1.705 and 2.110, respectively.

Because likelihoods are formed from the product of outcome probabilities for all individuals in a sample, the numerical value of a given likelihood depends on the sample size and is not particularly interpretable by itself. However, comparing likelihoods from nested models is a direct way to evaluate improvements in fit. This is the basis of the LR test.

Finally, we note that although the discussion here is limited to the binary outcome context, estimation methods for most of the regression models presented in this book are likelihood based. For example, least squares estimation and F-testing for comparing nested models in linear regression and analysis of variance models are examples of likelihood methods. Further, likelihood methods are fundamental to the family of GLMs discussed in Chap. 9.

5.7 Sample Size, Power, and Detectable Effects

Section 4.8 provides formulas for calculating sample size, power, and minimum detectable effects for the linear model. Analogous results hold for the logistic model. To compute the sample size that will provide power of γ in two-sided tests with type-1 error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j, accounting for the loss of precision due to multiple predictors, we can use

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2}{(\beta_j^a \sigma_{x_j})^2 p(1-p)\left(1-\rho_j^2\right)}, \tag{5.16}$$

where β_j^a is the hypothesized value of β_j under the alternative, $z_{1-\alpha/2}$ and z_γ are the quantiles of the standard normal distribution corresponding to the specified type-1 error and power, σ_{x_j} is the standard deviation of X_j and ρ_j is its multiple correlation with the other covariates, and p is the marginal prevalence of the outcome (Hsieh et al. 1998; Hsieh and Lavori 2000). For problems with predetermined n, power is given by

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a|\sigma_{x_j}\sqrt{np(1-p)\left(1-\rho_j^2\right)}\right] \tag{5.17}$$

and the minimum detectable effect (on the log-odds scale) by

$$\pm\,\beta_j^a = \frac{z_{1-\alpha/2} + z_\gamma}{\sigma_{x_j}\sqrt{np(1-p)\left(1-\rho_j^2\right)}}. \tag{5.18}$$

Some additional points:

- When X_j is binary with prevalence f_j, $\sigma_{x_j} = \sqrt{f_j(1-f_j)}$ in (5.16)–(5.18).
- 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 (6.26). 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-mg increase in the predictor is 1,000 times smaller than the effect of a 1-g 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 (5.18).
- Sample size (5.16) and minimum detectable effect (5.18) 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,

$$n = \frac{7.849}{(\beta_j^a)^2 p(1-p)}. \tag{5.19}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2.802}{\sqrt{np(1-p)}}. \tag{5.20}$$

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 log odds-ratio for treatment and $s_{x_j}^2 = 0.25$, we can calculate

$$n = 4 \times \frac{7.849}{(\beta_j^a)^2 p(1-p)}. \tag{5.21}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2 \times \frac{2.802}{\sqrt{np(1-p)}}. \tag{5.22}$$

- Power calculations using (5.17) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- As in calculations for the linear model, we need to use $|\beta_j^a|$ in (5.17) if $\beta_j^a < 0$. It follows that the negative of the value given by (5.18) is also a valid solution for the minimum detectable effect.
- These computations are valid for unmatched case-control studies, in which p, the sample prevalence of the outcome, is controlled by design. However, special methods are required for matched case-control studies.
- Sample size and power (but not minimum detectable effects) can be calculated using the `sampsi` command in Stata as well as many other statistical packages. Alternatively, (5.16)–(5.18) can easily be programmed in Stata, R, or Excel, or evaluated by hand if values of $z_{1-\alpha/2}$, z_γ, and $\Phi(\cdot)$ are available.
- The use of the factor $1S-\rho_j^2$ to account for covariate adjustment carries over from linear to logistic models. However, there is no analog to the reduction in residual variance that can result from including covariates in linear models, so that the adjustment to these calculations using $1-\rho_j^2$ is less likely to be conservative.
- The `sampsi` command does not incorporate the factor $1-\rho_j^2$ for covariate adjustment. An unadjusted estimate of n should be inflated by $1/(1-\rho_j^2)$; similarly the unadjusted 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 logistic models with a continuous predictor, `sampsi` can be made to work by reversing the role of predictor and outcome, as we show in an example below.
- These calculations were derived in Chap. 4 from the Wald test of $\beta_j = 0$. Calculations based on the more reliable LR test (Self and Mauritsen 1992) have been implemented in the Egret statistical package.

- In Sect. 4.8, we showed how the standard error $\mathrm{SE}(\hat{\beta}_j)$ plays a central role in sample size, power, and minimum detectable effect calculations for regression problems. $\mathrm{SE}(\hat{\beta}_j)$ is a large-sample approximation for the logistic model, and more exact small-sample computations using the noncentral t-distribution do not carry over from the linear model. Simulations of power may be a more reliable guide when the calculated or available sample size is small.
- Equations (5.16)–(5.18) are based on the assumption that the conditional mean of the outcome does not vary strongly across observations, which would hold if X_j is a relatively weak predictor, or equivalently if $|\beta_j^a|$ is small. Methods based on simulation avoid this simplification and perform slightly better in some circumstances (Vittinghoff et al. 2009). However, errors from these sources are usually small compared to errors arising from uncertainty about the required inputs.
- The alternative calculations (4.22)–(4.24) presented in Sect. 4.8, which use an estimate $\tilde{\mathrm{SE}}(\hat{\beta}_j)$ based on published results for an appropriately adjusted model using $\tilde{n}$ observations, carry over directly. There we showed that

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} \left[\tilde{\mathrm{SE}}(\hat{\beta}_j)\right]^2}{(\beta_j^a)^2}. \tag{5.23}$$

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

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a|/[\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j)]\right]. \tag{5.24}$$

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

$$\pm\, \beta_j^a = (z_{1-\alpha/2} + z_\gamma)\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j). \tag{5.25}$$

  In implementing these calculations, care must be taken to obtain the SE of the regression coefficient β_j, not the SE of the odds ratio e^{β_j}. Since results are usually available only for the odds ratio, this can computed as $\tilde{\mathrm{SE}}(\hat{\beta}_j) = \log(UL/LL)/3.92$, where UL and LL are the upper and lower 95% confidence bounds for the odds ratio. We must also ensure that X_j is measured on the same scale as in the published results.

To illustrate these methods, we first use the `sampsi` command to estimate the sample size providing 80% power in two-sided tests with α of 5% for a clinical trial of a new technique hypothesized to reduce the incidence of an adverse postsurgical outcome from 15% to 5%. We specify the proportion with the outcome in each group, which are equivalent to the means of a continuous outcome. By omitting the `sd1` option, we signal that the outcome is binary, with SD determined under the statistical model as $\sqrt{p(1-p)}$. With equal allocation to treatment and control, the `r()` option, which specifies the ratio of the sizes of the groups being compared,

Table 5.33 Sample size calculation for randomized trial

```
. sampsi 0.05 0.15, power(0.8)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                    and p2 is the proportion in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            p1 =   0.0500
            p2 =   0.1500
         n2/n1 =   1.00

Estimated required sample sizes:

            n1 =      160
            n2 =      160

. display log((0.05/0.95)/(0.15/0.85))
  -1.2098379

. display (invnormal(0.975)+invnormal(.8))^2/((-1.2098379)^2*0.25*0.075*
  (1-0.075)) 309.17921
```

can be left at the default value of 1. In addition, we can safely assume that $\rho_j = 0$, so no adjustment for covariates is likely to be necessary in a randomized trial.

Table 5.33 shows the results. The `sampsi` command estimates that we need 160 participants per group. We also used (5.16) to estimate sample size. For that calculation, β_j^a, the hypothesized log-odds ratio for the effect of the new technique, is $\log(0.05/0.95)/(0.15/0.85) \approx -1.2098$. With equal allocation ($f = 0.5$) to treatment and control, $\sigma_x^2 = f(1-f) = 0.25$, and the marginal prevalence p of the outcome $\approx 7.5\%$. This gives an overall sample size estimate of 309.

Now, suppose we would like to estimate the sample size that will provide 80% power in two-sided tests with α of 5% to detect an independent association of SBP with CHD, adjusting for age, smoking, BMI, cholesterol levels, and behavioral patterns, as suggested by the results in Table 5.10. From pilot data, we estimate that the prevalence of CHD in the new sample of high risk men will be 30%, that SBP will be approximately 5 mmHg higher among the men with CHD, that the within-group SD of SBP will be 15 mmHg, and finally that $\rho_j \approx 0.33$. To do this computation using the `sampsi` command, we reverse the role of the outcome and predictor, so f is now the prevalence of CHD. Pre-calculation of the local variable `ratio` is required because the `r()` option will not allow the fractional input 3/7.

Table 5.34 first shows the calculation using the `sampsi` command, with the adjustment using the variance inflation factor then carried out based on the unadjusted results. In addition, we computed the sample size using (5.16), relying on the fact that $\beta_j \sigma_x$, the log-odds per SD increase in SBP, is approximately equal to the standardized (in SDs) difference in mean SBP between the subgroups with and without CHD. The two sample size estimates are close.

Table 5.34 Sample size calculation for the effect of SBP on risk of CHD

```
. display 3/7

. 42857143

. sampsi 0 5, sd1(15) r(.42857143) 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 =        5
           sd1 =       15
           sd2 =       15
         n2/n1 =     0.43

Estimated required sample sizes:

            n1 =      236
            n2 =      102

. display (236+102)/(1-0.33^2) 379.30648

. display (invnormal(.975)+invnormal(0.8))^2/((5/15)^2*.3*.7*(1-.33^2))
  377.48913
```

5.8 Summary

The logistic regression model extends frequency table techniques for investigating the association between a binary outcome and categorical predictor to include continuous predictors and allow simultaneous consideration of multiple (continuous and categorical) predictors.

Modeling techniques for logistic regression mirror those for linear regression, allowing many of the concepts and methods learned in Chap. 4 to be applied directly to studies involving binary outcomes. However, interpretation of logistic regression models is slightly more complex due to the model's nonlinear relationship between outcome risk and predictors. In particular, regression coefficients need to be transformed to be interpretable as odds ratios.

Although a powerful and useful tool, there are a number of situations where logistic regression is not the best method for analyzing binary outcome data. As we have seen in several examples, when attention is restricted to one or a few categorical predictors, regression techniques are not needed. In other situations, an alternative binary regression model linked to alternate measures of association such as relative risks or risk differences may be preferred. We refer readers to Chap. 6 for methods appropriate for regression analysis for event time outcomes. Although we have provided a brief illustration in Sect. 5.5.2 of how logistic regression can be used to investigate the effects of predictors on binary outcomes that are duration dependent, we refer readers to Chap. 9 for a more complete coverage of regression methods for event time outcomes. Finally, we note that when analysis focuses on causal inference about the effect of a particular binary predictor representing a treatment or exposure, the methods covered in Chap. 9 are generally preferred.

5.9 Further Notes and References

There are a number of excellent textbooks on logistic regression, including Breslow and Day (1984), Hosmer and Lemeshow (2000), Kleinbaum (2002), and Collett (2003). All of these provide more details and cover a broader range of topics than provided here. Although we have focused on Stata in our example analyses, most modern statistical software packages provide extensive facilities for fitting and interpretation of logistic models, including R, SAS, S-PLUS, and SPSS. More extensive facilities for exact logistic regression and contingency table methods are available in the programs StatXact and LogXact.

Throughout this chapter, we have concentrated on analysis of data where the outcomes and predictors were measured without substantial error and missing observations were not considered a major problem. In many studies, we cannot assume that this is the case. There is an extensive literature on the impacts of misclassified outcomes and measurement error in predictors in the context of logistic regression (Carroll et al. 1995; Magder and Hughes, 1997).

Missing data are an issue in most studies involving binary outcomes, and arise through a variety of mechanisms. When relatively few observations are involved, the problem can be handled via the default procedure in most available software programs (i.e., to eliminate any observations with one or more missing values among the predictors). The validity of this approach rests on the assumption that the individuals dropped from the analysis are "missing completely at random." However, when a substantial fraction of observations involve missing values, more care is required. In addition to the obvious problem of the reduction in power incurred by dropping observations there are substantial concerns that the results based on the remaining complete data may be biased. There are a number of approaches to handling missing observations, including sensitivity analyses, imputation, and modified maximum likelihood estimation methods. (See Jewell 2004 for a more complete discussion.) These tend to be complex to apply and are not generally well represented in standard software.

5.10 Problems

Problem 5.1. Verify that the numerical average (mean) of the following sample of 25 binary outcomes equals the proportion of positive outcomes (ones) in the sample:

$$(1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0)$$

Table 5.35 Logistic model for CHD and age

```
. logistic chd69 i.agec, coef

Logistic regression                               Number of obs   =       3154
                                                  LR chi2(4)      =      44.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -868.14866                       Pseudo R2       =     0.0252
------------------------------------------------------------------------------
       chd69 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |
          1  |  -.1314518   .2309941    -0.57   0.569    -.5841919    .3212882
          2  |   .5307399   .2235341     2.37   0.018     .0926211    .9688586
          3  |   .8409976   .2274986     3.70   0.000     .3951085    1.286887
          4  |    1.05998   .2585408     4.10   0.000     .5532496    1.566711
             |
       _cons |  -2.804337   .1849627   -15.16   0.000    -3.166858   -2.441817
------------------------------------------------------------------------------
```

Problem 5.2. Use the regression coefficients from the logistic model presented in Table 5.2 in the logistic formula (5.2) to estimate the quantities in Table 5.3 for a 65-year-old individual. Use additional calculations to add a new section to Table 5.3 for an age increment of five years.

Problem 5.3. Perform the basic algebra necessary to verify the properties of the logistic regression coefficient β_1 stated in (5.6).

Problem 5.4. The output in the Table 5.35 provides the regression coefficients corresponding to the model fitted in Table 5.5. Use the coefficients and calculations similar to those illustrated in Sect. 5.1.1 to compute the log odds ratio comparing CHD risk in the fourth age category (`4.agec`) with the third (`3.agec`). Also, compute the odds ratio for this comparison. Comment on how we might obtain an estimated standard error and 95% CI for this quantity.

Problem 5.5. For the fitted logistic regression model in Table 5.6, calculate the log odds for a 60-year-old smoker with cholesterol, SBP, and BMI values of 250 mg/dL, 150 mmHg, and 20, respectively. Now calculate the log odds for an individual with a cholesterol level of 200 mg/dL, holding the values of the other predictors fixed. Use these two calculations to estimate an odds ratio associated with a 50 mg/dL increase in cholesterol. Repeat the above calculations for a 70-year-old individual with identical values of the other predictors. Comment on any differences between the two estimated odds ratios.

Problem 5.6. Use the regression output in Table 5.16 and a calculation similar to that presented in (5.11) to compute the odds ratio comparing the odds of CHD in a 55-year-old individual with arcus to the corresponding odds for a 40-year-old who also has arcus.

Problem 5.7. Use the WCGS data set to fit the regression model presented in Table 5.18. Perform the Hosmer–Lemeshow goodness of fit test for the following number of groups: 10, 15, 20, and 25. Comment on the differences. The data set is available at http://www.biostat.ucsf.edu/vgsm.

Problem 5.8. Verify that the odds ratio formed from the two odds presented in (5.11) is given by ad/bc. Verify that the same odds ratio is obtained if the two component odds are computed based on the probability of exposure conditional on outcome status.

Problem 5.9. Compute the approximate 95% CI for the following per-contact infection risk based on the intercept coefficient and associated standard errors given in Table 5.28:

$$1 - \exp\left[-\exp(-7.033)\right].$$

5.11 Learning Objectives

(1) Describe situations in which logistic regression analysis is needed.
(2) Translate research questions appropriate for a logistic regression model into specific questions about model parameters.
(3) Use logistic regression models to test hypotheses about relationships between a binary outcome variable and a continuous or categorical predictor.
(4) Describe the logistic regression model, its key assumptions, and their implications.
(5) State the relationships between:

- Odds ratios and logistic regression coefficients.
- A two × two table analysis of the association between a binary outcome and single categorical predictor and a logistic regression model for the same variables.

(6) Know how a statistical package is used to fit a logistic regression model to continuous and categorical predictors.
(7) Interpret logistic regression model output, including:

- Regression parameter estimates, hypothesis tests, CIs.
- Statistics which quantify the fit of the model.