Chapter 3
Basic Statistical Methods

Statistical analyses involving multiple predictors are generalizations of simpler techniques developed for investigating associations between outcomes and single predictors. Although many of these should be familiar from basic statistics courses, we review some of the key ideas and methods here as background for the methods covered in the rest of the book and to introduce some basic notation.

Sections 3.1–3.3 review basic methods for continuous outcomes, including the t-test and one-way ANOVA, the correlation coefficient and the linear regression model for a single predictor. Section 3.4 focuses on contingency table methods for investigating associations between binary outcomes and categorical predictors, including a discussion of basic measures of association. Section 3.5 introduces descriptive methods for survival time outcomes, including Kaplan–Meier survival curves and the logrank test. In Sect. 3.6, we introduce the use of the bootstrap as a method to obtain CIs for estimates in situations where traditional methods are inappropriate. Finally, Sect. 3.7 discusses the importance of properly interpreting negative findings from statistical analyses, focusing on the use of point estimates and CIs rather than P-values.

3.1 t-Test and Analysis of Variance

The t-test and one-way ANOVA are basic tools for assessing the statistical significance of differences between the average values of a continuous outcome across two or more samples. Both the t-test and one-way ANOVA can be seen as methods for assessing the association of a categorical predictor—binary in the case of the t-test, with more than two levels in the case of one-way ANOVA—with a continuous outcome. Both are based in statistical theory for normally distributed outcomes, but work well for many other types of data; and both turn out to be special cases of linear regression models.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_3,

Table 3.1 t-Test of difference in average glucose by exercise

```
. t-test glucose if diabetes == 0, by(exercise)

Two-sample t-test with equal variances

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      no |    1191    97.36104    .2868131    9.898169    96.79833    97.92376
     yes |     841    95.66825    .3258672    9.450148    95.02864    96.30786
---------+--------------------------------------------------------------------
combined |    2032    96.66043    .2162628     9.74863    96.23631    97.08455
---------+--------------------------------------------------------------------
    diff |            1.692789    .4375862                .8346243    2.550954
------------------------------------------------------------------------------
Degrees of freedom: 2030

                     Ho: mean(no) - mean(yes) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t =   3.8685               t =   3.8685              t =   3.8685
   P < t =   0.9999         P > |t| =   0.0001          P > t =   0.0001
```

3.1.1 t-Test

The basic t-test is used in comparing two independent samples. The t-statistic on which the test is based is the difference between the two sample averages, divided by the standard error of that difference. The t-test is designed to work in small samples, whereas Z-tests are not. Table 3.1 shows the result of a t-test comparing average fasting glucose levels among women without diabetes, according to exercise. This is the first of many examples in Chaps. 3 and 4 using data from the heart and estrogen/progestin study (HERS), a clinical trial of hormone therapy (HT) for prevention of recurrent heart attacks and death among 2,763 post-menopausal women with existing coronary heart disease (CHD) (Hulley et al. 1998). Average glucose is 97.4 mg/dL among the 1,191 women who do not exercise as compared to 95.7 mg/dL among the 841 women who do. The difference of 1.7 mg/dL is statistically significant ($P = 0.0001$) in the two-sided test shown in the column headed `Ha: diff != 0` (`!=` is Stata notation for "not equal to.") The P-value gives the probability—under the null hypothesis that mean glucose levels are the same in the two populations being compared—of observing a t-statistic more extreme, or larger in absolute value, than the observed value.

3.1.2 One- and Two-Sided Hypothesis Tests

In clinical research, unlike some other areas of science, two-sided hypothesis tests are almost always used. In the two-sided t-test, we are testing the null hypothesis (`Ho`) of equal population means against the alternative hypothesis (`Ha`) that the one mean is either smaller or larger than the other. The two-sided test is appropriate, for example, when a new treatment might turn out to be beneficial *or* to have adverse effects.

In contrast, only one of these alternatives is considered in a one-sided test. As a result, the smaller of the one-sided P-values is half the magnitude of the two-sided P-value. The resulting advantage of the one-sided test is that at a given significance level, less evidence in favor of the alternative hypothesis is required to reject the null. For example, using a one-sided test in a sample of 100 observations, we would declare statistical significance at the 5% level if the t-statistic exceeds 1.66; using a two-sided test it would need to exceed 1.98 (in absolute value). A direct benefit is that a somewhat smaller sample size is required when a study is designed to be analyzed using a one-sided test.

Use of a one-sided test is sometimes motivated by prior information that makes only one of the alternatives of interest. An example might be in testing an existing treatment known to be safe for evidence of benefit on a new endpoint. One-sided tests are also used in *noninferiority* trials comparing a new to a standard treatment; in this setting the alternative hypothesis is that the new treatment performs almost as well or better than the standard treatment, as against the null hypothesis of clearly performing worse.

However, in part because they make it possible to reject the null hypothesis on weaker evidence, one-sided tests are not commonly used in clinical research. Even in noninferiority trials where one-sided tests are clearly appropriate, a standard text on the conduct of clinical trials (Friedman et al.1998) recommends that the tests be carried out at a significance level of 2.5%. Thus to claim noninferiority, the same strength of evidence would be required as in a two-sided test. Furthermore, Fleiss (1988) argues that the other alternative *ought* generally to be of interest, and that in treatment trials adverse effects can rarely be ruled out with sufficient certainty to justify a one-sided test. We endorse this conservative view, and recommend using two-sided tests unless a one-sided test is strongly motivated by specific reasons.

The Stata `t-test` command gives P-values for both one-sided test as well as the two-sided test. In Table 3.1, the one-sided P-value on the right (`Ha: diff > 0`) gives the probability (again, under the null hypothesis) of observing a t-statistic larger than the observed value, while the one on the left (`Ha: diff < 0`) gives the probability of observing one that is smaller. In this example, there is strong evidence ($P = 0.0001$) that the mean glucose level is higher in the population of women who do not exercise, as compared to those who do, and essentially no evidence ($P = 1.0$) that it is smaller.

3.1.3 Paired t-Test

The paired t-test is for use in settings where individuals or observations are linked across the two samples. Examples include measurements taken at two time points on the same individuals, or on other naturally linked pairs, as in a clinical trial where one eye is treated and the other serves as a control. In this case, the two samples are not independent and failure to take account of the pairwise relationships wastes information and is potentially erroneous.

The paired t-test procedure first computes the pairwise differences for each individual or linked pair. In the first example, this is the change in the outcome from the first time point to the second, and in the second, the difference between the outcomes for the treated and control eyes. Then a t-test is used to assess whether the population mean of these paired differences differs from zero. An increase in power results because between-individual variability is eliminated in the first step. The paired t-test is also implemented using the `t-test` command in Stata. The more complicated case where we want to examine the influence of some other factor on within-individual changes is covered in Sect. 7.3.

3.1.4 One-Way Analysis of Variance

Suppose that we need to compare sample averages across the arms of a clinical trial with multiple treatments, or more generally across more than two independent samples. For this purpose, one-way ANOVA and the F-test take the place of the t-test. The F-test, presented in more detail in Sect. 4.3, assesses the null hypothesis that the mean value of the outcome is the same across all the populations sampled from, against the alternative that the means differ in at least two of the populations. For example, the one-way ANOVA shown in Table 3.2, the F-test for `Between groups` ($P = 0.0371$), suggests that mean SBP differs by ethnicity in the population represented in the HERS cohort.

3.1.5 Pairwise Comparisons in ANOVA

The statistically significant F-test in the one-way ANOVA indicates the overall importance of ethnicity for predicting SBP. In addition, Stata implements the Bonferroni, Scheffé, and Sidak procedures for assessing the statistical significance of all possible pairwise differences between groups, without inflation of the overall or family-wise type-I error rate (FER), which can arise from testing multiple null hypotheses. These and other methods for controlling the FER are discussed in Sects. 4.3.4 and 13.4.1. All three methods implemented in the `oneway` command show that the difference in average SBP between the African American and white groups is statistically significant after correction for multiple comparisons, but that the other pairwise differences are not; we show the Scheffé result.

Table 3.2 One-way ANOVA assessing differences in SBP by ethnicity

```
. oneway sbp ethnicity, tabulate scheffe

            | Summary of systolic blood pressure
  ethnicity |        Mean   Std. Dev.       Freq.
------------+------------------------------------
      White |   134.78376   18.831686        2451
   Afr Amer |   138.23394   19.992518         218
      Other |   135.18085   21.259767          94
------------+------------------------------------
      Total |   135.06949   19.027807        2763

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2384.26992      2   1192.13496      3.30     0.0371
 Within groups      997618.388   2760   361.455938
------------------------------------------------------------------------
    Total           1000002.66   2762   362.057443

                  Comparison of systolic blood pressure by ethnicity
                                (Scheffe)
Row Mean-|
Col Mean |      White   Afr-Amer
---------+----------------------
Afr-Amer |    3.45018
         |      0.037
   Other |    .397089   -3.05309
         |      0.980      0.429
```

3.1.6 *Multi-way ANOVA and ANCOVA*

Multi-way ANOVA is an extension of the one-way procedure to deal simultaneously with more than one categorical predictor, while analysis of covariance (ANCOVA) is commonly defined as an extension of ANOVA that includes continuous as well as categorical predictors. The t- and F-tests retain their central importance in these procedures. However, one-way ANOVA and the t-test implicitly estimate the different population means by the sample averages; in contrast, the population means in multi-way ANOVA and ANCOVA are usually *modeled*. Thus these procedures are most easily understood as multipredictor linear regression models, which are covered in Chap. 4.

3.1.7 *Robustness to Violations of Normality Assumption*

The t- and F-tests are fairly robust to violations of the normality assumption, especially in larger samples. By robust we mean that the type-I error rate, or probability of rejecting the null hypothesis when it holds, is not seriously affected. They are primarily sensitive to outliers, which tend to decrease efficiency and make it harder to detect real differences between groups. Thus the effect is conservative, in the sense of making it more likely that we will accept the null hypothesis when some real difference exists.

Large samples reduce sensitivity of the t-test to the assumption that the outcome is normally distributed because the distribution of the difference between the sample averages, which directly underlies the test, converges to a normal distribution even when the outcome itself has some other distribution. If violations of the normality assumption are mild to moderate, samples of 50–100 may be large enough, in particular with equal group sizes, but considerably larger samples might be needed with severe violations. Analogous large-sample behavior holds for the regression coefficients estimated in multipredictor linear models as well as the other regression models that are the primary topic of this book.

3.1.8 Nonparametric Alternatives

One commonly recommended solution for violations of the normality assumption is to use nonparametric Wilcoxon rank-sum or Kruskal–Wallis tests rather than the t-test or one-way ANOVA. Two other nonparametric methods are discussed below in Sect. 3.2 on the correlation coefficient.

While they avoid specific parametric distributional (i.e., normality) assumptions, these methods are not assumption-free. For example, the Wilcoxon and Kruskal–Wallis tests are based on the assumption that the outcome distributions being compared differ in *location* (mean and/or median) but not in *scale* (variance) or *shape*, as might be captured by a histogram, and can give misleading results if these assumptions are violated. Furthermore, these two tests do not provide an interpretable measure of the strength of the association. More generally, nonparametric methods sometimes result in loss of efficiency, and do not easily accommodate multiple predictors, unlike the regression methods which are the focus of this book.

Nonparametric tests are most useful for unadjusted between-group comparisons where the P-value is of primary interest, in particular for variables with skewed distributions that cannot be normalized by transformation, or outliers that must be retained for substantive reasons.

3.1.9 Equal Variance Assumption

When sample sizes are unequal, the t-test is less robust to violations of the assumption of equal variance across samples than to violations of normality. Violations of this assumption can seriously affect the type-I error rate, not always in a conservative direction, and large samples do not make the test any more robust. In contrast, the overall F-test in ANOVA loses efficiency, but the type-I error rate is generally not increased. However, subsequent pairwise comparisons using t-tests remain vulnerable.

Table 3.3 *t*-Test allowing for unequal variances

```
. t-test glucose if diabetes == 0, by(exercise) unequal

Two-sample t-test with unequal variances

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      no |    1191    97.36104    .2868131    9.898169    96.79833    97.92376
     yes |     841    95.66825    .3258672    9.450148    95.02864    96.30786
---------+--------------------------------------------------------------------
combined |    2032    96.66043    .2162628     9.74863    96.23631    97.08455
---------+--------------------------------------------------------------------
    diff |            1.692789    .4341096                .8413954    2.544183
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  1858.33

                     Ho: mean(no) - mean(yes) = diff = 0

     Ha: diff < 0                Ha: diff != 0               Ha: diff > 0
       t =   3.8995                t =   3.8995                t =   3.8995
   P < t =   1.0000          P > |t| =   0.0001          P > t =   0.0000
```

In the two-sample case, this problem is easily addressed using a version of the t-test for unequal variances. This is based on a modified estimate of the standard error of the difference in sample averages. In the analysis shown in Table 3.1, the standard deviation of glucose is 9.9 mg/dL among women who do not exercise, as compared to 9.5 mg/dL among the women who do. In this case, the re-analysis allowing for unequal variances, shown in Table 3.3, gives qualitatively the same result ($P = 0.0001$). We recommend systematic use of this version of the t-test, since the increase in robustness comes at very little cost in efficiency. Analogous extensions of ANOVA in which the variance is allowed to vary by group are also possible, though not implemented in the Stata `one-way` or `anova` commands.

3.2 Correlation Coefficient

The Pearson correlation coefficient r is a scale-free measure of linear association between two variables x and y, and is defined as follows:

$$\begin{aligned} r(x, y) &= \frac{\mathrm{Cov}(x, y)}{\mathrm{SD}(x)\mathrm{SD}(y)} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}}. \end{aligned} \tag{3.1}$$

In (3.1), $\mathrm{Cov}(x, y)$ is the sample covariance of x and y, $\bar{x}$ and $\bar{y}$ are their sample means, $\mathrm{SD}(x)$ and $\mathrm{SD}(y)$ their standard deviations, and n is the sample size. The covariance reflects the degree to which observations on the two variables differ from their respective means in the same degree and direction. Dividing $\mathrm{Cov}(x, y)$ by the standard deviations of x and y in (3.1) gives the correlation $r(x, y)$, which is scale-free in the sense that it always takes on values between –1 and 1 and does not vary with the units of measurement used for either variable (Problem 3.2).

The correlation coefficient is a measure of *linear* association, in a sense that will become clearer in Sect. 3.3 on the simple linear model. Values of r near zero denote the absence of linear association, while values near 1 mean that x and y increase almost in lockstep, their paired values in a scatterplot falling close to a straight line with positive slope. Correlations between –1 and zero mean that y tends to *de*crease as x increases. Note that powerful *nonlinear* associations between x and y—for example, if y is proportional to x^2—are often consistent with correlations near zero; in the example, this can happen if $\bar{x} \approx 0$.

3.2.1 Spearman Rank Correlation Coefficient

Like the t-test (and the coefficients of the linear regression model described below), the correlation coefficient is sensitive to outliers. In this case, a robust alternative is the Spearman correlation coefficient, which is equivalent to the Pearson coefficient applied to the *ranks* of x and y. This measure of correlation also takes on values between –1 and 1. By rank, we mean position in the ordered sequence of the values of a variable; if x takes on values 1.2, 0.5, 18.3, and 2.7, then the ranks of these values are 2, 1, 4, and 3, respectively. Thus the rank of the outlier 18.3 is only 1 unit larger than the rank of the next largest value 2.7, the same distance that separates the ranks of any two sequential values of x, thus depriving the outlier of undue influence in estimating the correlation between x and y. Ties are handled by computing the average rank of the tied values. Ranks are used in a range of nonparametric methods, in no small part because of their robustness when the data include outliers. Their disadvantage is that any information contained in the measured values of the outcome beyond the ranks is lost.

3.2.2 Kendall's τ

Another rank-based alternative to Pearson's correlation coefficient is Kendall's τ, defined as the difference in the number of concordant and discordant pairs of data points, as a proportion of the number of evaluable pairs. In the absence of ties, the pair of data points (x_i, y_i) and (x_j, y_j) for observations i and j is concordant if $x_i > x_j$ and $y_i > y_j$, or if $x_i < x_j$ and $y_i < y_j$, and discordant otherwise. It is easy to see that we need only know the ranks of the x and y values, not their actual values, to evaluate the conditions for concordance. If the numbers of concordant and discordant pairs are about equal, then $\tau \approx 0$; essentially this means that the fact that $x_i > x_j$ gives little information about whether $y_i > y_j$. But as the proportion of concordant pairs grows, τ approaches 1, reflecting the fact that the ordering of the x pairs is highly associated with the ordering of the y pairs. Conversely, if most pairs are discordant, then τ approaches –1; again, the orderings of the x and y pairs are highly associated. Kendall's τ is sometimes used as a measure of correlation for time-to-event outcomes.

3.3 Simple Linear Regression Model

Here we present the simple linear regression model with a continuous outcome and a single continuous predictor variable.

3.3.1 Systematic Part of the Model

The main purpose of this model is to determine how the average value of the continuous outcome y varies with the value of a single predictor x. The average values of the outcome are assumed to lie on a "regression line" or "line of means." Figure 3.1 shows values of baseline SBP by age in the HERS trial of hormone therapy. To make the idea of a line of means more concrete, the square symbols in the plot show the average SBP within each decile of age. Naturally, there is some noise in these local means, although much less than in the raw data. Moreover, the continuous regression line, assumed to be linear, captures the increasing trend rather

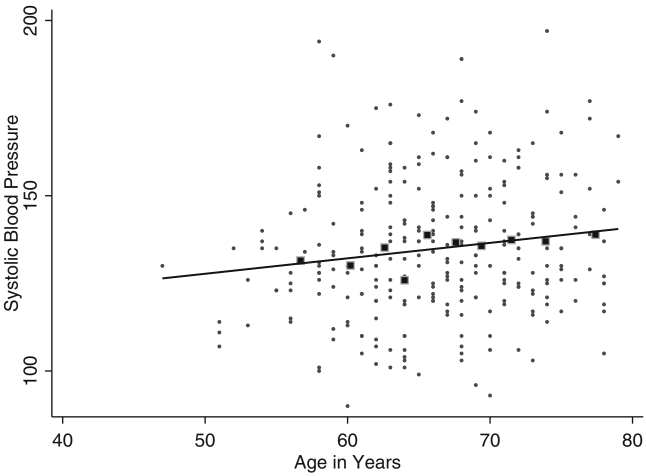

Fig. 3.1 Linear regression model for SBP and age

well. Its slope represents the systematic dependence of the outcome on the predictor, and is thus usually the focus of interest.

The formula for the regression line is simple and has interpretable parameters:

$$\begin{aligned} \mathrm{E}[y|x] &= \text{average value of SBP for a given age} \\ &= \beta_0 + \beta_1 \mathtt{age} \\ &= 105.7 + 0.44\mathtt{age}. \end{aligned} \tag{3.2}$$

In (3.2), $\mathrm{E}[y|x]$ is shorthand for the *E*xpected or average value of the outcome y at a given value of the predictor x. β_1 gives the slope of the regression line, and is interpretable as the change in average SBP for a one-year increase in age. The estimate of β_1 from the sample shown in the plot suggests that among women with heart disease, average SBP increases 0.44 mmHg for each one-year increase in age. This estimate is the best fitting value in a sense explained below in Sect. 3.3.4.

It is also easy to see that the estimate of the intercept parameter $\beta_0 = 105.7$ gives the average value of the outcome when age is zero. While not meaningless in this case, these data obviously provide no direct information about SBP at age zero. This illustrates the more general point that while regression models are often approximately true within the range of the observed data, extrapolation is usually risky. "Centering" the predictor by subtracting off a value within the range of the data can resolve this problem. One reasonable choice in this example would be the sample average age of 67; then the centered age variable would have value zero for women at age 67, and the new intercept, 135.2 mmHg, estimates average SBP among women this age. The slope estimate is unaffected by centering the age variable.

3.3.2 Random Part of the Model

It is also clear from Fig. 3.1 that at any given age, SBP varies considerably. Possible sources of this variability include measurement error, diurnal patterns, and a potentially broad range of unmeasured determinants of SBP, including the immediate circumstances when the measurement was made. These factors are combined in an error term ε, so that for observation i

$$\begin{aligned} \mathrm{SBP}_i &= \text{mean SBP for subjects of age}_i + \text{error}_i \\ &= \beta_0 + \beta_1 \mathtt{age}_i + \varepsilon_i. \end{aligned} \tag{3.3}$$

The statistical assumptions of the linear regression model concern the distribution of ε. Specifically, we assume that $\varepsilon_i \sim$ i.i.d $\mathcal{N}(0, \sigma_\varepsilon^2)$, meaning that ε is *i*ndependently and *i*dentically *d*istributed and has a

- Normal distribution
- Mean zero at every value of age
- Constant variance σ_ε^2 at every value of age
- Values that are statistically independent

In Sect. 4.7, we will see that the first assumption may sometimes be relaxed. The second assumption is important to checking whether the relationship between a numerical predictor and the outcome is linear, as assumed in (3.2), (3.3), and Fig. 3.1; violations can be examined and repaired using methods also introduced in Sect. 4.7. The third assumption, of constant variance, is sometimes called *homoscedasticity*; data which violate this assumption are called *heteroscedastic*, and can be dealt with using methods also discussed in Sect. 4.7 as well as Chap. 8. Chapters 7 and 12 introduce methods for data where the fourth assumption, of independence, does not hold. Some examples include samples with repeated measures on individuals, cluster samples where patients are selected from within a sample of physician practices, and complex survey samples such as the national health and nutrition examination survey (NHANES).

3.3.3 Assumptions About the Predictor

In contrast to the outcome, no distributional assumptions are made about the predictor in the linear regression model. In the case of the linear model with a single continuous predictor, we do not assume that the predictor has a normal distribution, although we will see in Sect. 4.7 that outlying values of the predictor can cause trouble in some circumstances. In addition, binary, categorical, and discrete numeric variables including counts are easily accommodated as predictors in these models.

Although we do not need to make assumptions about the distribution of the predictor, these models do perform better when it is relatively variable. For example, it would be more difficult to estimate the age trend in average SBP if the sample were limited to women aged 65–70. For binary and categorical predictors, the analogous limitation is that the subgroups defined by the predictor should not be too small. The impact of the variability of the predictor, or lack of it, is reflected in the standard error of the regression coefficient, as shown below in Sect. 3.3.7.

Finally, when we want to assess the relationship of the outcome with the true values of the predictor, we effectively assume that the predictors are measured without error. This is often not very realistic, and the effects of violations are the subject of ongoing statistical research. Random measurement errors unrelated to the outcome result in attenuation of estimated slope coefficients toward zero, sometimes called *regression dilution bias* (Frost and Thompson 2000). Despite some loss of efficiency, reasonable estimation is often possible in the presence of mild-to-moderate error in the measurement of the predictors. Moreover, for prediction of new outcomes, values of the predictor measured with error may suffice.

Table 3.4 OLS regression of SBP on age

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model |  2179.70702     1  2179.70702           Prob > F      =  0.0188
    Residual |  106991.347   274   390.47937           R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  109171.054   275  396.985652           Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286    .186455     2.36   0.019     .0734621    .8075952
       _cons |    105.713   12.40238     8.52   0.000      81.2969     130.129
------------------------------------------------------------------------------
```

3.3.4 Ordinary Least Squares Estimation

The model (3.3) refers to the population of women with heart disease from which the sample shown in Fig. 3.1 was drawn. The regression line in the figure is an estimate of the population regression line that was found using *ordinary least squares* (OLS). Of all the lines that could be drawn though the scatterplot of the data to represent the trend in SBP with increasing age, the OLS estimate minimizes the sum of the squared vertical differences between the data points and the line.

Since the regression line is uniquely determined by β_0 and β_1, the intercept and slope parameters, fitting the regression model amounts to finding estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ which meet the OLS criterion. In addition to being easy to compute, these OLS estimates have desirable statistical properties. If model assumptions hold, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of the population parameters.

> *Definition*: An estimate is *unbiased* if, over many repeated samples drawn from the population, the average value of the estimates based on the different samples would equal the population value of the parameter being estimated.

OLS estimates are also minimally variable and well behaved in large samples when the distributional assumptions concerning ε are not precisely met. However, a drawback of the OLS estimation criterion is sensitivity to outliers, which arises from squaring the vertical differences (Problem 3.1). Section 4.7 will show how to diagnose and deal with influential points.

Table 3.4 shows Stata results for an OLS regression of SBP on age. The estimate of β_1, the slope coefficient (`Coef.`) for `age`, is 0.44 mmHg per year, and the intercept estimate $\hat{\beta}_0$ is 105.7 mmHg (`_cons`).

3.3.5 Fitted Values and Residuals

The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in turn determine the *fitted value* $\hat{y}$ corresponding to every data point:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{3.4}$$

It should be plain that the fitted value $\hat{y}_i$ lies on the estimated regression line at the point where $x = x_i$. For a woman at the average age of 67, the fitted value is

$$105.713 + 0.4405286 \times 67 = 135.2 \text{ mmHg}. \tag{3.5}$$

The *residuals* are defined as the difference between observed and fitted values of the outcome:

$$r_i = y_i - \hat{y}_i. \tag{3.6}$$

The residuals are the sample analog of ε, the error term introduced earlier in Sect. 3.3, and as such are particularly important in fitting the model, in estimating the variability of the parameter estimates, and in checking model assumptions and fit (Sect. 4.7).

3.3.6 Sums of Squares

Various *sums of squares* are central to understanding OLS estimation and to reading the Stata regression model output in Table 3.4. First is the *total sum of squares* (TSS):

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{3.7}$$

where $\bar{y}$ is the sample average of the outcome y. TSS captures the total variability of the outcome about its mean. In Table 3.4, TSS = 109,171 and appears in the row and column labeled `Total` and `SS` (for Sum of Squares), respectively.

In an OLS model, TSS is split into two components. The first is the *model sum of squares* (MSS), or the part of the variability of the outcome about its mean that can be accounted for by the model:

$$\text{MSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{3.8}$$

The second component of outcome variability, the part that cannot be accounted for by the model, is the *residual sum of squares* (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.9}$$

By definition, RSS is minimized by the fitted regression line. In Table 3.4, MSS and RSS appear in the rows labeled `Model` and `Residual` of the `SS` column. The identity TSS = MSS + RSS is a central property of OLS, but more difficult to prove than it may seem.

3.3.7 Standard Errors of the Regression Coefficients

MSS and RSS also play an important role in estimating the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ and in testing the null hypothesis of central interest, H_0: $\beta_1 = 0$. These standard errors depend on the variance of ε—that is, the variance of the outcome about the regression line—which is estimated in our single predictor model by

$$\hat{\text{Var}}(\varepsilon) = \hat{\sigma}^2_{y|x} = \text{RSS}/(n-2). \tag{3.10}$$

In Table 3.4, $\hat{\sigma}^2_{y|x}$ equals 390.5, and appears in the column and row labeled `MS` (for Mean Square) and `Residual`, respectively.

The variance of $\hat{\beta}_1$ is estimated by

$$\hat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2_{y|x}}{(n-1)\hat{\sigma}^2_x}, \tag{3.11}$$

where $\hat{\sigma}^2_x$ is the sample variance of the predictor x. The square root of the variance of an estimate is referred to as its *standard error*, or SE($\hat{\beta}$). In Table 3.4, the standard error of the estimated slope coefficient for `age`, found in the column labeled `Std.Err.`, is approximately 0.186.

From the numerator and denominator of (3.11), it is clear that the variance of the slope estimate *increases* with the residual outcome variance not explained by the model, but *decreases* with larger sample size and with the variance of the predictor (as we pointed out earlier in Sect. 3.3.3). In our example of SBP and age, estimation of the trend in age is helped by the relatively large age range in the sample. It should also be intuitively clear that the precision of the slope estimate is increased in samples where the data are tightly clustered about the regression line—in other words, if the residual variance of the outcome is small. Figure 3.1 shows that this is not the case with our example; SBP varies widely about the regression line at every value of age.

3.3.8 Hypothesis Tests and Confidence Intervals

When the outcome is normally distributed, the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have a normal distribution, and the ratio of the slope estimate to its standard error has a t-distribution with $n-2$ degrees of freedom. This leads directly to a test of the null hypothesis of no slope: that is, H_0: $\beta_1 = 0$, or in substantive terms, no systematic relationship between predictor and outcome. In Table 3.4, the t-statistic and corresponding P-value for `age` are shown in the columns labeled `t` and `P>|t|`. In the example, we are able to reject the null hypothesis that SBP does not change with age at the usual 5% level of significance ($P = 0.019$).

The t-distribution also leads to 95% CIs for the population parameter β_1, shown in Table 3.4 in the columns labeled `[95% Conf. Interval]`. The confidence interval does not include 0, in accord with the result of the t-test of the null hypothesis. Under the assumptions of the model, a CI computed this way would, on average, include the population value of the parameter in 95 of 100 random samples. In a more intuitive interpretation, we could exclude with 95% confidence age trends in SBP smaller than 0.07 mmHg/year or larger than 0.81 mmHg/year.

3.3.8.1 Relationship Between Hypothesis Tests and Confidence Intervals

Hypothesis tests and CIs provide overlapping information about the parameter or association being assessed. Common ground is that when a two-sided test is statistically significant at $P < 0.05$, then the corresponding 95% CI will exclude the null parameter value. However, the P-value, especially if it is small, does give a more direct sense of the strength of the evidence against the null hypothesis. Likewise, only the confidence interval provides information about the range of parameter values that are consistent with the data. In Sect. 3.7 below, we argue that CIs are particularly important in the interpretation of negative findings—that is, cases where the null hypothesis is not rejected. Both the P-value and the CI are important for understanding statistical results in depth, and getting beyond the simple question of whether or not an association is statistically significant. This overlapping relationship between hypothesis tests and CIs holds in many settings in addition to linear regression.

3.3.8.2 Hypothesis Tests and Confidence Intervals in Large Samples

The hypothesis tests and CIs in this section follow from basic statistical theory for data with normally distributed outcomes. However, linear regression models are commonly used with outcomes that are at best approximately normal, even after transformation. Fortunately, in large samples the t-tests and CIs for $\hat{\beta}_0$ and $\hat{\beta}_1$ are valid even when the underlying outcome is not normal. How large a sample is required depends on how far and in what way the outcome departs from normality. If the outcome is uniformly distributed, meaning that every value in its range is equally likely, then the t-tests and CIs may be valid with as few as 30–50 observations. However, with long-tailed outcomes, samples of at least 100 and sometimes much larger may be required for hypothesis tests and CIs to be valid.

3.3.9 Slope, Correlation Coefficient, and R^2

The slope coefficient β_1 in a simple linear model is systematically related to the Pearson correlation coefficient r, reviewed in Sect. 3.2:

$$r = \beta_1 \sigma_x / \sigma_y, \tag{3.12}$$

where σ_x and σ_y are the standard deviations of the predictor and outcome, respectively. Thus we can get r from β_1 by factoring out the scales on which x and y are measured (Problem 3.3), scales which are reflected in the standard deviations. Furthermore, the t-test of H_0: $\beta_1 = 0$ is equivalent to a test of H_0: $r = 0$.

However, the correlation coefficient is not simply interchangeable with the slope coefficient in a simple linear model. In particular, the slope coefficient distinguishes the roles of the predictor x and outcome y, with differing assumptions applying to each, and would change if those roles were reversed, but $r(x, y) = r(y, x)$. Note that reversing the roles of predictor and outcome becomes even more problematic with multipredictor models. In addition, the slope coefficient β_1 depends on the units in which both predictor and outcome are measured, so that if either or both were measured in different units, β_1 would change. For example, our estimate of the age trend in SBP would be 4.4 mmHg per decade if age were measured in ten-year units. While both versions are interpretable, this dependence on the scale of both predictor and outcome can make it difficult to assess the strength of the association. In addition, the dependence on scale would make it hard to judge whether age is a stronger predictor of SBP than other variables. From this point of view, the scale-free correlation coefficient r is easier to interpret.

The correlation coefficient r and thus the slope coefficient β_1 are also systematically related to the *coefficient of determination* R^2

$$R^2 = r^2 = \frac{\text{MSS}}{\text{TSS}}. \tag{3.13}$$

R^2 is interpretable as the proportion of the total variability of the outcome (TSS) that is accounted for by the model (MSS). As such, it is useful for comparing models (Sect. 10.2). In Table 3.4, the value of `R-squared` is only 0.0200, which you can easily verify is equal to MSS/TSS = 2,179/109,171. This shows that age only explains a very small proportion of the variability of SBP, even though it is a statistically significant predictor in a sample of moderate size.

3.4 Contingency Table Methods for Binary Outcomes

In Chap. 2, we reviewed exploratory techniques for categorical outcome variables. We expand that review here to include contingency table methods for assessing associations between binary outcomes and categorical predictors.

Table 3.5 Two-by-two contingency table for CHD and arcus

```
. cs chd69 arcus, or

                 | arcus senilis          |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |       102         153  |        255
        Noncases |       839        2058  |       2897
-----------------+------------------------+----------
           Total |       941        2211  |       3152
                 |                        |
            Risk |  .1083953    .0691995  |   .080901
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .0391959       |  .0166915    .0617003
      Risk ratio |         1.566419       |  1.233865    1.988603
 Attr. frac. ex. |         .3616011       |  .1895387    .4971343
 Attr. frac. pop |         .1446404       |
      Odds ratio |          1.63528       |  1.257732    2.126197 (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =    13.64  Pr>chi2 = 0.0002
```

3.4.1 Measures of Risk and Association for Binary Outcomes

In the WCGS (Rosenman et al. 1964) of CHD introduced in Chap. 2, an association of interest to the original investigators was the relationship between CHD risk and the presence/absence of corneal arcus senilis among participants upon entry into the study. Because each participant could be unambiguously classified as having developed CHD or not during the ten-year course of the study, the indicator variable that takes on the value one or zero according to whether or not participants developed the disease is a legitimate binary outcome for the analysis. Corneal arcus is a whitish annular deposit around the iris that occurs in a small percentage of older adults, and is thought to be related to serum cholesterol level. Table 3.5 presents the results of a basic two-by-two table analysis for this example. The results were obtained using the `cs` command in Stata, which provides a number of useful quantities in addition to a simple crosstabulation of the binary CHD outcome `chd69` with the binary indicator of the presence of arcus.

The `Risk` estimates (0.108 and 0.069) summarize outcome risk for individuals with and without arcus and are simply the observed proportions of individuals with CHD in these groups at the baseline visit of the study. The output also includes several standard epidemiological measures of association between outcome risk and the predictor variable, along with corresponding 95% CIs. These are numerical comparisons of the risk estimates between the two groups defined by the predictor.

The `Risk difference` or *excess risk* is defined as the difference between the estimated risk in the groups defined by the predictor. For the table, we can verify that the risk difference is

$$0.1084 - 0.0692 = 0.039$$

The `Risk ratio` or *relative risk* is the ratio of these risks—for the example in the table,

$$0.1084/0.0692 = 1.57.$$

The `Odds ratio` is the ratio between the corresponding odds in the two groups. The odds of an outcome occurring are computed as the probability of occurrence divided by the complementary probability that the event does not occur. Since the denominators of these two probabilities are identical, the odds can be also be calculated as the ratio of the number of outcomes to nonoutcomes. Frequently used in games of chance, "even odds" obtains when these two probabilities are equal.

In Table 3.5, the odds of CHD occurrence in the two arcus groups are $0.1084/(1-0.1084) = 102/839$ and $0.0692/(1-0.0692) = 153/2058$, respectively. The ratio of these two numbers yields the estimated odds ratio (1.635) comparing the odds of CHD occurrence among participants with arcus to the odds of those without this condition. Although the odds ratio is somewhat less intuitive as a risk measure than the risk difference and relative risk, we will see that it has properties that make it useful in a wide range of study designs, and (in Chap. 5) that it is fundamental in the definition and interpretation of the *logistic regression* model.

Finally, note that Table 3.5 provides two auxiliary summary measures of *attributable risk* (i.e., `Attr. frac. ex.` and `Attr. frac. pop`), which estimate the fraction of outcomes which can be attributed to the predictor in the subgroup with the predictor (sometimes referred to as "exposed" individuals) and in the overall population, respectively. Although these measures can easily be estimated from the data in the table, their validity and interpretability depends on a number of factors, including study design and the causal connections between measured and unmeasured predictors and the outcome. See Rothman and Greenland (1998) for further discussion of these measures.

In the last example, we saw that the observed outcome proportions for groups defined by different values of a predictor are the fundamental components of the three summary measures of association: the excess risk, relative risk, and odds ratio. To discuss these further, it will be useful to have symbolic definitions. Following the notation introduced in Sect. 3.3 for a continuous outcome measure, we will denote the binary outcome variable CHD by y, and let the values 1 and 0 represent individuals with and without the outcome, respectively. We will symbolize the outcome probability for an individual associated with a particular value x of a single predictor as

$$P(x) = \Pr(y = 1|x)$$

and estimate this using the proportion of individuals with the outcome $y = 1$ among all those in the sample with the value x of the predictor. For example, $P(0)$ and $P(1)$ symbolize the outcome probability or risk associated with two levels of the binary predictor `arcus` in Table 3.5 (where we follow the usual convention that individuals possessing the characteristic have the values $x = 1$, and individuals without the characteristic have $x = 0$). The following equation defines all three summary risk measures introduced above using this notation:

$$\begin{aligned} ER &= P(1) - P(0) \\ RR &= P(1)/P(0) \\ OR &= \frac{P(1)/\left[1 - P(1)\right]}{P(0)/\left[1 - P(0)\right]}, \end{aligned} \tag{3.14}$$

where ER, RR, and OR denote the excess risk, relative risk, and odds ratio, respectively.

Like the correlation coefficient, these measures provide a convenient single number summary of the direction and magnitude of the association. The major distinction between them is that the ER is a measure of the difference in risk between the two groups (with no difference indicated by a value of zero), while both the RR and OR compare the risks in relative terms (with no difference indicate by a value of one). Note that because the component risks range between zero and one, the ER can take on values between -1 and 1. By contrast, the RR and OR range between 0 and ∞.

Relative measures are appealing because they are dimensionless, and convey a clear impression of how outcome risk is increased/decreased by exposure. The RR in particular is favored by epidemiologists because of its interpretability as a ratio of risks. However, relative measures are less desirable when the goal is to convey the "importance" of a particular risk in absolute terms: In the example, the estimated RR for the risk of CHD is approximately 1.6 times higher for men with arcus. The ER tells us that this corresponds to a 4% difference in absolute risk. Note that if the risk of the outcome were ten times lower in both groups, we would have the same estimated RR, but the corresponding ER would also be ten times smaller (or 0.4%).

A further feature of the RR worth remembering is that its maximum value is constrained by the level of risk in the comparison group. For example, if $\Pr(0) = 0.5$, $RR \le 2$ must hold. The OR has the advantages of a relative measure, and in addition is not constrained by the level of the risk in the reference group. However, being based on the odds of the outcome rather than the probability, the OR lacks the intuitive interpretation of RR. The only exception is when the outcome risk is quite small. For such rare outcomes, the OR closely approximates the RR and can be interpreted similarly. (This property can be seen from the above definition by noting that if outcome risk is close to zero, then $[1 - \Pr(0)]$ and $[1 - \Pr(1)]$ will both be approximately one.) Unfortunately, the odds ratio is often inappropriately reported as a relative risk even when this condition is not met (Holcomb et al. 2001). Because the value of the OR is always more extreme than the value of the RR (except when both equal one), this can be misleading. For these reasons, we recommend that the measure of association reported in research findings be that actually used in the analysis.

A final important property of all three measures of association introduced above is that their interpretation depends on the underlying study design. In the WCGS example, the outcome risks represent the *incidence proportion* of CHD over the entire duration of the study (approximately ten years). The measures of association in the table should be interpreted accordingly. By contrast, the sexually transmitted infection example mentioned at the beginning of this chapter referred to a cross-sectional sample. Outcome risk in this setting is measured by the *prevalence* of the outcome among the groups defined by the predictor. In this case, the terms "prevalence odds," "prevalence ratio," and "excess prevalence" provide unambiguous alternative labels for OR, RR, and ER, respectively.

The relative merits of the ER, RR, and OR are discussed at length in most epidemiology textbooks (e.g., Rothman and Greenland 1998). For our purposes, they are equally valid and the choice is dependent on the nature and goals of the research investigation. In fact, for prospective and cross-sectional study designs, we will see that we can freely convert between measures. (Case-control designs are a special case which will be covered in Sect. 5.3.) However, from the standpoint of regression modeling, we will see in Chap. 5 that the OR has clear advantages.

3.4.2 Tests of Association in Contingency Tables

Addressing the research question posed in the example presented in Table 3.5 involves more than simply summarizing the degree of the observed association between CHD and arcus. We would also like to account for uncertainty in our estimates before concluding that the association reflects more than just a chance finding in this particular sample of individuals. The 95% CIs associated with the measures of association in the table help in this regard. For example, the fact that the confidence interval for the odds ratio excludes the value 1.0 allows us to conclude that the true value for this measure is greater than one, and indicates a statistically significant positive association between the presence of arcus and CHD occurrence. This corresponds to testing the null hypothesis that the true odds ratio is equal to one, with the alternative hypothesis being that this odds ratio is different than one. The fact that the value of one is excluded from the CI corresponds to rejection of this hypothesis at the 5% significance level. Of course, establishing the possible causal connection between these two variables is a more complex issue.

The χ^2 *(chi-squared) test* of association is an alternative way to make inferences about an observed association. Note that the result of this test (presented in Table 3.5) agrees with the conclusions drawn for the 95% CIs for the various measures of association. The statistic addresses the null hypothesis of no association, and is computed using the squared differences between the observed proportions of individuals in each cell of the two-way table and the corresponding proportions that would be expected if the null hypothesis were true. Large values of the statistic indicate departure from this hypothesis, and the associated P-value is computed using the χ^2 distribution with degrees of freedom specified. The χ^2 statistic for a two-by-two table is less appealing as a measure of association than the alternative measures discussed above. However, in cases where predictors have more than two levels (as discussed below) and a single summary measure of association cannot be calculated, the χ^2 statistic is useful as a global indicator of whether or not an association may be present.

Table 3.6 Female partner's HIV status by AIDS diagnosis of male partner

```
. cs hivp aids, or exact

                 | AIDS diag. in male     |
                 | [1=yes/0=no]           |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |         3           4  |          7
        Noncases |         2          22  |         24
-----------------+------------------------+----------
           Total |         5          26  |         31
                 |                        |
            Risk |        .6    .1538462  |   .2258065
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .4461538       | -.0050928    .8974005
      Risk ratio |              3.9       |  1.233644    12.32933
 Attr. frac. ex. |         .7435897       |  .1893933    .9188926
 Attr. frac. pop |         .3186813       |
      Odds ratio |             8.25       |  1.200901     57.1864 (Cornfield)
                 +-----------------------------------------------
                               1-sided Fisher's exact P = 0.0619
                               2-sided Fisher's exact P = 0.0619
```

The validity of the χ^2 test is dependent on available sample size; like many commonly used statistical tests, the validity of the reference χ^2 distribution for the test statistic is approximate, with the approximation improving with increasing number of observations. Consequently, for small sample sizes, approximate P-values and associated inferences may be unreliable. An alternative in these cases is to base inferences on *exact* methods. Table 3.6 presents an example from a cross-sectional study of sexual transmission of human immunodeficiency virus (HIV) in monogamous female partners of males infected from contaminated blood products (O'Brien et al. 1994). The outcome of this study was HIV status of the female partner at recruitment. Males were known to have been infected first (via medical records) and exposure of females was limited to contact with male partners. The available sample size ($n = 31$) was limited by the availability of couples meeting the strict eligibility criteria.

Table 3.6 addresses the hypothesis that more rapid disease progression in the males (as indicated by an AIDS diagnosis occurring at or before the time of recruitment of the couple) is associated with sexual transmission of HIV to the female (represented by the binary indicator `hivp`). In addition to observed counts, the table includes proportions of the outcome by AIDS diagnosis in the male partners, and the measures of association described above. The table also presents the results of Fisher's exact test. Similar to the χ^2 test, the Fisher test addresses the hypothesis of independence of outcome and predictor. However, the P-value is computed exactly, conditioning on the observed marginal totals. The P-value for the χ^2 test applied to the data in Table 3.6 (not shown) is 0.029. Similarly, the lower 95% confidence limits for the RR and OR exclude the value one, also indicating

Table 3.7 CHD events by age in WCGS cohort

```
. tabulate chd69 agec, col chi2

           |                         agec
 CHD event |     35-40      41-45      46-50      51-55      56-60 |     Total
-----------+-------------------------------------------------------+----------
        no |       512      1,036        680        463        206 |     2,897
           |     94.29      94.96      90.67      87.69      85.12 |     91.85
-----------+-------------------------------------------------------+----------
       yes |        31         55         70         65         36 |       257
           |      5.71       5.04       9.33      12.31      14.88 |      8.15
-----------+-------------------------------------------------------+----------
     Total |       543      1,091        750        528        242 |     3,154
           |    100.00     100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(4) =  46.6534   Pr = 0.000
```

a statistically significant association. By contrast, the (two-sided) P-value for the Fisher's exact test for Table 3.6 is 0.062, indicating failure to reject the hypothesis of independence at the 5% level.

A commonly cited rule-of-thumb is that the Fisher's exact test should be used whenever any of the expected cell counts are less than 5. Note that Fisher's exact test applies to tables formed by variables with more than two categories. Although it can almost always be used in place of the χ^2 test, the associated computations can be lengthy for large sample sizes, especially for tables with dimensions larger than 2×2. Given the increased speed of modern desktop computers and the availability of more computationally efficient algorithms, we recommend using the exact P-value whenever it can easily be computed (i.e., in a matter of minutes) or is provided, and especially in cases where either actual or expected minimum cell counts are less than 5.

3.4.3 Predictors with Multiple Categories

In the WCGS study discussed above, one potentially important predictor of CHD risk is age at entry into the study. Despite the fact that this can be considered as a continuous variable for the purpose of analyses, we might begin investigating the relationship by grouping age into multiple categories and summarizing CHD risk in the resulting groups. Table 3.7 shows the results obtained by dividing subjects into five-year age intervals using a constructed five-level categorical variable `AGEC`. With the exception of the first two columns, the estimated percentages of individuals with CHD in the second row of the table clearly increase with increasing age. In addition, the accompanying χ^2 test indicates that age and CHD risk are associated.

As mentioned above, the conclusion of association based on the χ^2 test does not reveal anything about the nature of the relationship between these variables. More insight could be gained by computing measures of association between age and CHD risk. However, unlike the two-by-two table case, the fact that age is represented

Table 3.8 Odds ratios for CHD events by age group

```
. tabodds chd69 agec, or

---------------------------------------------------------------------------
        agec |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       35-40 |    1.000000          .            .            .           .
       41-45 |    0.876822       0.32       0.5692     0.557454    1.379156
       46-50 |    1.700190       5.74       0.0166     1.095789    2.637958
       51-55 |    2.318679      14.28       0.0002     1.479779    3.633160
       56-60 |    2.886314      18.00       0.0000     1.728069    4.820876
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(4)  =    46.64
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    40.76
                                  Pr>chi2  =   0.0000
```

with five levels means that a single measure will not suffice here. In fact, odds ratios can be computed to compare any two age groups. For example, the ER, RR, and OR comparing CHD risk in 56 to 60-year-olds with that in 35 to 40-year-olds are calculated by applying the formulas in (3.14) as follows:

$$\begin{aligned} ER &= (36/242) - (31/543) = 0.092 \\ RR &= \frac{36/242}{31/543} = 2.606 \\ OR &= \frac{\dfrac{36/242}{206/242}}{\dfrac{31/543}{512/543}} = 2.886. \end{aligned} \tag{3.15}$$

The results in Table 3.8 further reinforce our observation that CHD risk is increasing with increasing age. The odds ratios in the table are all computed using the youngest age group as the reference category. The pattern of increase in estimated odds ratios mirrors that seen in Table 3.7. Note that each odds ratio in the table is accompanied by a 95% confidence interval and associated hypothesis test. In addition, two global tests providing additional information are provided: The `Test of homogeneity` addresses the null hypothesis that odds ratios do not differ across age categories. In this case, the P-value indicates rejection, confirming the observed difference in the odds ratios mentioned above. Since age can be viewed as a continuous variable, and the categorical version considered here is ordinal, more specific alternatives to nonhomogeneity of odds are of greater scientific interest. The `Score test for trend` in Table 3.8 addresses the alternative hypothesis that there is a linear trend in the odds of CHD with increasing age categories. The statistically significant results indicate support for this hypothesis, and represent a stronger conclusion than nonhomogeneity. Note that this test is not applicable to nominal categorical variables.

Despite the useful information gained from the analysis in Tables 3.7 and 3.8, we may be concerned that our conclusions depend on the arbitrary choice of grouping age into five categories. Increasing the number of age categories may provide more information on how risk varies with age, but will also reduce the number of individuals in each category and lead to more variable estimates of risk in each group. This dilemma is one of the primary motivations for introducing a regression model for the dependence of outcome risk on a continuous predictor variable. Another motivation (which will be explored briefly below and more fully in Chap. 5) arises when we consider the joint effects on risk of multiple (categorical and/or continuous) predictor variables.

3.4.4 Analyses Involving Multiple Categorical Predictors

A common feature of observational clinical and epidemiological studies is that investigators do not experimentally control the distributions of characteristics of interest among participants in the sample. Unlike randomized trials in which random allocation serves to balance the distributions of characteristics across treatment arms, observational data are usually characterized by differing distributions across subgroups defined by predictors of primary interest. For example, observational studies of the relationship between dietary factors and cancer typically adjust for age since it is frequently related to both diet and cancer risk. A fundamental part of drawing inferences regarding the relationship between the outcome and key predictors in observational studies is to consider the potential influence of these other characteristics. This topic will be covered in detail for regression models in Chaps. 4–6, 9, and 10. Here we give a brief introduction for binary outcomes and categorical predictors.

Consider the cross-tabulation of a binary indicator 20-year mortality and self-reported smoking presented in Table 3.9. These data represent women participating in a health survey in Whickham, England, in 1972–1974 (Vanderpump et al. 1996). Deaths were ascertained via follow-up of participants over a 20-year period. The results indicate a statistically significant negative association between smoking and mortality (where `Cases` denote deceased women).

Before concluding that this somewhat unintuitive inverse relationship between smoking and mortality may reflect a real association in the population being studied, we need to consider the possibility that it may be due to the influence of other characteristics of women in the sample. The standard approach for controlling for the influence of additional categorical predictors in contingency tables is via a *stratified* analysis, where a relationship of interest is examined in subgroups defined by a additional variable (or variables).

Table 3.10 presents the same analysis stratified by a three-level categorical variable `agegrp` representing three categories of participant age (as ascertained in the original survey). The age-specific odds ratios and associated 95% CIs indicate a positive (but not statistically significant) association between smoking and vital status in two of the three age groups. The crude odds ratio reproduces the result obtained in Table 3.9, while the age-adjusted (`M-H combined`,

Table 3.9 Twenty-year vital status by smoking behavior

```
. cs vstatus smoker [freq = nn], or

                 | smoker                |
                 |   Exposed   Unexposed |    Total
-----------------+-----------------------+---------
           Cases |       139         230 |      369
        Noncases |       443         502 |      945
-----------------+-----------------------+---------
           Total |       582         732 |     1314
                 |                       |
            Risk |  .2388316    .3142077 | .2808219
                 |                       |
                 |      Point estimate   | [95% Conf. Interval]
                 |-----------------------+--------------------
 Risk difference |         -.075376      | -.1236536  -.0270985
      Risk ratio |         .7601076      |  .6347365   .9102415
 Prev. frac. ex. |         .2398924      |  .0897585   .3652635
 Prev. frac. pop |         .1062537      |
      Odds ratio |         .6848366      |  .5354784   .8758683  (Cornfield)
                 +---------------------------------------------
                            chi2(1) =    9.12  Pr>chi2 = 0.0025
```

Table 3.10 Twenty-year vital status by smoking behavior, stratified by age

```
. cs vstatus smoker [freq = nn], or by(agegrp)

          agegrp |      OR     [95% Conf. Interval]   M-H Weight
-----------------+----------------------------------------------
           18-44 |  1.776666    .8727834   3.615113     5.568471 (Cornfield)
           45-64 |  1.320359    .8728567   1.997089     19.55856 (Cornfield)
             64+ |  1.018182    .4240727    2.43359     4.772727 (Cornfield)
-----------------+----------------------------------------------
           Crude |  .6848366    .5354784   .8758683
    M-H combined |  1.357106    .9710409   1.896662
----------------------------------------------------------------
Test of homogeneity (M-H)    chi2(2) =   0.945  Pr>chi2 = 0.6234

                   Test that combined OR = 1:
                                Mantel--Haenszel chi2(1) =      3.24
                                                Pr>chi2 =    0.0719
```

or *Mantel–Haenszel*) estimate is computed via a weighted average of the age-specific estimates, where the stratum-specific weights are given in the right table margin (`M-H Weight`). Because this estimate is based on separate estimates made in each age stratum, the weighted average adjusts for the influence of age.

Comparison of the crude estimate with the adjusted estimate reveals that adjusting for age reverses the direction (and alters the significance) of the unadjusted result. Considering that none of the stratum-specific estimates indicate reduced risk associated with smoking, the crude estimate is surprising. This seemingly paradoxical result is often referred to as *Simpson's paradox*. To aid in further interpretation, Table 3.10 also includes results from two hypothesis tests of properties of the stratum-specific and combined odds ratios. The *test of homogeneity* addresses the null hypothesis that the three age-specific odds ratios are identical. Rejection of this hypothesis would provide evidence that the stratum-specific odds ratios differ, and may indicate a differential effect of smoking on mortality across different age groups. This phenomenon is also known as *interaction* or *effect modification*. In this case, the results indicate that the data do not support rejecting the null hypothesis in favor of the alternative hypothesis of differing age-specific odds ratios. We conclude that there is no strong evidence of interaction and that the age-specific odds ratios are similar. However, note that if we base the analysis in Table 3.10 on the relative risk rather than the odds ratio, the P-value for the test of homogeneity equals 0.045, indicating the presence of interaction. This illustrates that the presence or absence of statistical interaction may reflect our choice to work with a particular measure of association rather than some underlying causal phenomenon.

The second test result presented in Table 3.10 addresses the null hypothesis that the true age-adjusted ("combined") odds ratio for the association between vital status and smoking is different than one. This hypothesis is meaningful if we have already failed to reject the hypothesis of homogeneity. In this case, we have already concluded that we do not have strong evidence that the age-specific odds ratios differ, and the results of the test for an age-adjusted association indicate failure to reject the null hypothesis at the 5% significance level. We conclude that the observed unadjusted negative association between vital status and smoking is at least partially explained by age adjustment. In fact, adjusting for age results in a positive association between smoking and vital status, that is more in accordance with our expectations that smokers may experience more health problems.

The results of the Whickham example are an instance of a more general phenomenon in observational studies known as *confounding*. In the example, the seemingly paradoxical finding of a positive association (albeit not statistically significant) after adjustment for age can be explained by differences between age groups in the proportion of women who were smokers (women in the intermediate age group were more likely to smoke than women in the other groups), and the fact that mortality was much higher in the older women. Of course, other measured or unmeasured factors may also influence the relationship between smoking and vital status. A complete analysis would consider these. Also, it would be a good idea to consider alternate measures of age and smoking if available (e.g., treating them as continuous variables in a regression model). The phenomena of confounding and interaction will be discussed extensively in the regression context in the remaining chapters of the book.

3.4.5 Collapsibility of Standard Measures of Association

Following the discussion in the previous section, it is tempting to conclude that in situations where interaction can be ruled out, the presence of confounding can be assessed via observed differences between the crude and adjusted measures of association obtained from the Mantel–Haenszel approach for stratified contingency tables. Conversely, agreement between the stratum-specific estimates and the crude (unadjusted) estimate would seem to imply a lack of confounding.

There are two primary issues to consider when assessing absence/presence of confounding based on comparing unadjusted and adjusted association measures: the first is that because confounding is fundamentally tied to the causal interpretation given the associations involved, its presence can never be confirmed solely on statistical grounds. In the Whickam example from Table 3.10, interpreting age as a confounder of the smoking–mortality association as measured by odds ratios seems plausible. However, in many situations, the direction of the causal link between a risk factor and a suspected confounder is less clear. In these settings, observed differences between crude and adjusted association measures may reflect causal relationships other than confounding. Section 4.5 provides examples of *mediation* of the causal effects of an exposure variable on an outcome by an intermediate variable, and points out that this cannot be distinguished from confounding solely by observing differences between crude and adjusted measures of association.

The second issue is that different measures of association may exhibit different properties with respect to adjustment and pooling across strata, and these properties complicate simple interpretation of observed differences between pooled and adjusted measures. Intuitively, we might expect that in the absence of confounding and interaction, the association between a binary outcome and a single binary predictor at levels defined by a third categorical predictor would be homogeneous, and that the observed association in the strata would equal the crude association from the pooled table ignoring the third variable. A measure of association with this property is called *strictly collapsible*. Both the risk difference and the relative risk are collapsible in this sense. However, the odds ratio is not strictly collapsible. In some situations, the crude odds ratio may differ from the corresponding stratum specific and adjusted measures even when confounding is demonstrably absent.

Noncollapsibility of the odds ratio is illustrated in Table 3.11, in which the odds ratios measuring the association between a binary outcome variable Y and a binary predictor X are equal in strata defined by a third binary variable Z, and also equal to the adjusted measure. Yet, the crude odds ratio ignoring Z is different from the stratum specific measures, even though there is no marginal association between X and Z (i.e., confounding cannot be present). Note that both the crude and adjusted odds ratios are valid measures in this example. The crude measure is interpreted as the *marginal* odds ratio for the association between Y and X, while the adjusted measure is interpreted as the *conditional* odds ratio for a fixed value of Z.

We will see in Chap. 5 that noncollapsibility is also manifested in logistic regression models for binary outcomes, where regression coefficients have a log odds ratio interpretation, and in proportional hazards regression models for survival outcomes (Chap. 6), with coefficients interpretable as log hazard ratios. Note that in the case of rare outcomes, the close correspondence between odds ratios and relative risks noted above minimizes this distinction, and these cases analyses based on either measure will agree closely. Chapter 9 is entirely devoted to the topic of making valid causal inferences using data from observational studies, and provides a framework for understanding confounding that further clarifies the issues raised here.

Table 3.11 Example illustrating inequality of the odds ratio for the association between a binary outcome Y and a binary predictor X when stratified by a binary variable Z versus pooled across values of Z

```
. tabulate Y X if Z==0

           |           X
         Y |         0          1 |     Total
-----------+----------------------+----------
         0 |        20         10 |        30
         1 |        25         25 |        50
-----------+----------------------+----------
     Total |        45         35 |        80


. tabulate Y X if Z==1

           |           X
         Y |         0          1 |     Total
-----------+----------------------+----------
         0 |        25         25 |        50
         1 |        10         20 |        30
-----------+----------------------+----------
     Total |        35         45 |        80


. cs Y X, or by(Z)

               Z |      OR       [95% Conf. Interval]   M-H Weight
-----------------+----------------------------------------------
               0 |          2    .7897239   5.05171        3.125 (Cornfield)
               1 |          2    .7897239   5.05171        3.125 (Cornfield)
-----------------+----------------------------------------------
           Crude |   1.653061    .8873163  3.079631
    M-H combined |          2    1.028901  3.887644
----------------------------------------------------------------
Test of homogeneity (M-H)    chi2(1) =   0.000  Pr>chi2 = 1.0000

                Test that combined OR = 1:
                             Mantel-Haenszel chi2(1) =      4.18
                                             Pr>chi2 =    0.0409
```

3.5 Basic Methods for Survival Analysis

In the previous section, we considered binary outcomes—that is, whether or not an event has occurred. Survival data represent an extension in which we take into account the time until the event occurs—or until the end of follow-up, if the event has not yet occurred at that point. These more complex outcomes are studied using techniques collectively known as *survival analysis*. The term reflects the origin of these methods in demographic studies of life expectancy.

3.5.1 Right Censoring

To illustrate the special characteristics of survival data, we consider a study of 6-mercaptopurine (6-MP) as maintenance therapy for children in remission from

Table 3.12 Weeks in remission among leukemia patients

Placebo:	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12, 12,15,17 22,23
6-MP:	6,6,6,6*,7,9*,10,10*,11*,13,16,17*, 19*,20*,22,23,25*,32*,32*,34*,35*

acute lymphoblastic leukemia (ALL) (Freireich et al. 1963). Forty-two patients achieved remission from induction therapy and were then randomized in equal numbers to 6-MP or placebo. The survival time studied was from randomization until relapse. At the time of the analysis, all 21 patients in the placebo group had relapsed, whereas only 9 of 21 patients in the 6-MP group had.

One crucial characteristic of these survival times is that for the 12 patients in the 6-MP group who remained in remission at the time of the analysis, the exact time to relapse was unobserved; it was only known to exceed the follow-up time. For example, one patient had only been under observation for six weeks, so we only know that the relapse time is longer than that. Such a survival time is said to be *right-censored*—"right" because on a graph the relapse time would lie somewhere to the right of the censoring time of six weeks.

> *Definition*: A survival time is said to be *right-censored* at time t if it is only known to be greater than t.

Table 3.12 displays follow-up times in the leukemia study. Asterisks mark the right-censored remission times.

Because of the censoring, we could not validly estimate the effects of 6-MP on time to relapse simply by comparing average follow-up times in the two groups (say, with a t-test). This simple approach would not work because the right-censored follow-up times in the 6-MP group are shorter, possibly much shorter, than the actual unobserved times to relapse for these patients. Furthermore, five of the right-censored values in the 6-MP group exceed the largest follow-up time in the placebo group; to ignore this would be throwing away valuable evidence for the effectiveness of the treatment. Survival analysis makes it possible to analyze right-censored data like these without bias or losing information contained in the length of the follow-up times.

3.5.2 Kaplan–Meier Estimator of the Survival Function

Suppose we would like to describe the probability of remaining in remission during each of the first ten weeks of the leukemia study. This probability is called the *survival function*.

> *Definition*: The *survival function* at time t, denoted $S(t)$, is the probability of being event-free at t; equivalently, the probability that the survival time is greater than t.

Table 3.13 Follow-up table for placebo patients in the leukemia study

Week of follow-up	No. followed	No. relapsed	No. censored	Conditional prob. of remission	Survival function
1	21	2	0	$19/21 = 0.91$	0.91
2	19	2	0	$17/19 = 0.90$	$0.90 \times 0.91 = 0.81$
3	17	1	0	$16/17 = 0.94$	$0.94 \times 0.81 = 0.76$
4	16	2	0	$14/16 = 0.88$	$0.88 \times 0.76 = 0.67$
5	14	2	0	$12/14 = 0.86$	$0.86 \times 0.67 = 0.57$
6	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
7	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
8	12	4	0	$8/12 = 0.67$	$0.67 \times 0.57 = 0.38$
9	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$
10	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$

We will first show how the survival function can be estimated for the 21 placebo patients. Because there is no right-censoring in the placebo group, we could simply estimate the survival function by the sample proportion in remission for each week. However, we will use a more complicated method because it accommodates right-censored data. This method depends on writing the survival function in any given week as a chain of conditional probabilities.

In Table 3.13 the placebo data are summarized by consecutive one-week intervals. The number of subjects who remain both in remission and in follow-up at the start of the week is given in the second column. The third and fourth columns list the numbers who relapse and who are censored during the week, respectively. Since none are censored, the number in follow-up is reduced only during weeks when a patient relapses. From the table, we see that in the first week, 19 of 21 patients remained in remission, so a natural estimate of the probability of being in remission in the first week is $19/21 = 0.91$. In the second week, 2 of the 19 placebo patients still in remission in the first week relapsed, and the remaining 17 remained in remission. Thus the probability of not relapsing in the second week, conditional on not having relapsed in the first, is estimated by $17/19 = 0.90$. It follows that the overall probability of remaining in remission in the second week is estimated by $19/21 \times 17/19 = 17/21 = 0.81$. Likewise, the probability of remaining in remission in the third week is estimated by $19/21 \times 17/19 \times 16/17 = 16/21 = 0.76$. In this case where there is no censoring, our chain of conditional probabilities reduces to the overall sample proportion in remission at the end of every week. You can easily verify that after ten weeks, the survival function estimate given by the chain of conditional probabilities is equal to the sample proportion still in remission.

Now we show how the survival function estimate based on the chain of conditional probabilities accommodates the censoring in the 6-MP group, as shown in Table 3.14. The problem we have to address is that two 6-MP subjects are censored prior to week 10. Since it is unknown whether they would have relapsed before the end of that week, we can no longer estimate the survival function at week 10 by the sample proportion still in remission at that point.

Table 3.14 Follow-up table for 6-MP patients in the leukemia study

Week of follow-up	No. followed	No. relapsed	No. censored	Condition. prob. of remission	Survival function
1	21	0	0	21/21 = 1.00	1.00
2	21	0	0	21/21 = 1.00	1.00 × 1.00 = 1.00
3	21	0	0	21/21 = 1.00	1.00 × 1.00 = 1.00
4	21	0	0	21/21 = 1.00	1.00 × 1.00 = 1.00
5	21	0	0	21/21 = 1.00	1.00 × 1.00 = 1.00
6	21	3	1	18/21 = 0.86	0.86 × 1.00 = 0.86
7	17	1	0	16/17 = 0.94	0.94 × 0.86 = 0.81
8	16	0	0	16/16 = 1.00	1.00 × 0.81 = 0.81
9	16	0	0	16/16 = 1.00	1.00 × 0.81 = 0.81
10	16	0	1	16/16 = 1.00	1.00 × 0.81 = 0.81

The rows of Table 3.14 for weeks 6 and 7 show how the method works with right-censored data. In week 6, three patients are observed to relapse, and one is censored (by assumption at the end of the week). Thus the probability of remaining in remission in week 6, conditional on having remained in remission in week 5, is $18/21 = 0.86$. Then we estimate the probability of remaining in remission in week 7, conditional on having remained in remission in week 6, as $16/17$: in short, the patient censored during week 6 has disappeared from the denominator, and does not contribute to the calculations for any subsequent week. Using this method for dealing with the censored observations, the conditional probabilities can still be estimated. As a result, we obtain a valid estimate of the probability of remaining in remission at the end of week 10, even though it is unknown whether the two censored patients remained in remission at that time. This approach allows us to extrapolate the survival experience of censored observation by those followed longer. This method requires modification in the case of *competing risks data* (Sect. 6.5) where *cumulative incidence functions* define the probability of failure in the presence of other causes of failure.

In essence, we have estimated the survival functions in the placebo and 6-MP groups using the well-known Kaplan–Meier estimator to deal with right censoring. In this example, the follow-up times have been grouped into weeks, but the method also applies to cases where they are observed more exactly. In Sect. 6.6.4, we examine the important assumption of *independent censoring* which underlies these procedures.

3.5.3 Interpretation of Kaplan–Meier Curves

Plots of the Kaplan–Meier estimates of $S(t)$ for the 6-MP and placebo groups in the leukemia study are shown in Fig. 3.2. Note that the curves drop at observed relapse times and are flat in the intervening periods. As a result, we can infer periods of

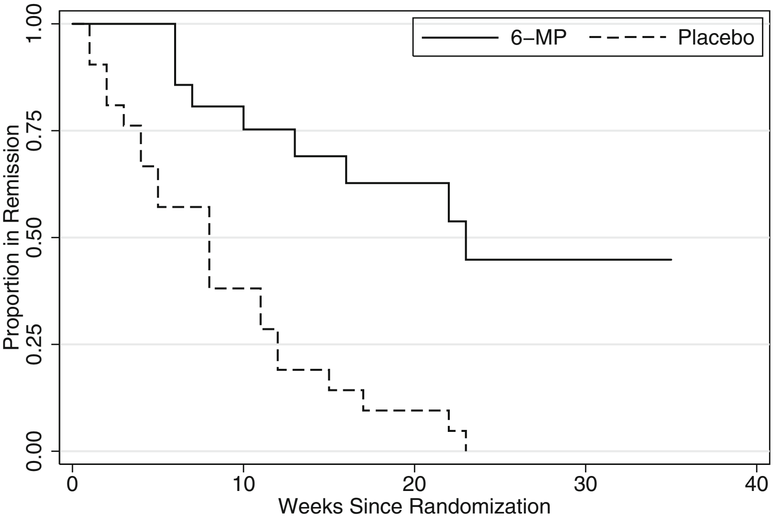

Fig. 3.2 Survival curves by treatment for leukemia patients

high risk, when the survival curve descends rapidly, as well as periods of lower risk, when it remains relatively flat. In particular, placebo patients appear to be at high risk of relapse in the first five weeks.

In addition, the estimated survival function for the 6-MP group is above the placebo curve over the entire follow-up period, giving evidence for higher probability of remaining in remission, or equivalently longer times in remission and lower risk of relapse in patients treated with 6-MP. In Sect. 3.5.6 below, we show how to test the null hypothesis that the survival functions are the same in the two groups.

3.5.4 Median Survival

The Kaplan–Meier results may also be used to obtain estimates of the median survival time, defined as the time at which half the relevant population has experienced the outcome event. In the absence of censoring, with every survival time observed exactly, the median survival time could be simply estimated by the sample median of survival times: that is, the earliest time at which half the study participants have experienced the event. From Table 3.13, we can see that median time to relapse is eight weeks in the placebo group—the first week in which at least half the sample $(12/21)$ have relapsed.

In the presence of censoring, however, we need to use the Kaplan–Meier estimate $\hat{S}(t)$ to estimate the median. In this case, the median survival time is estimated by the earliest time at which the Kaplan–Meier curve dips below 0.50. In the leukemia example, Fig. 3.2 shows that estimated median time to relapse is 23 weeks for 6-MP group, as compared to eight weeks for placebo—more evidence for the effectiveness of 6-MP as maintenance therapy for ALL.

By extension, other quantiles of the distribution of survival times can be obtained from the Kaplan–Meier estimate $\hat{S}(t)$. The pth quantile is estimated as the earliest time at which the Kaplan–Meier curve drops below $1 - p$. For instance, the lower quartile (i.e., the 0.25 quantile) is the earliest time at which the curve drops below $1 - 0.25 = 0.75$. The lower quartiles for the 6-MP and placebo groups are 13 and 4 weeks, respectively. However, a limitation of the Kaplan–Meier estimate is that when the curve does not reach $1 - p$, the pth percentile cannot be estimated. For example, Fig. 3.2 makes it clear that for the 6-MP group, quantiles of the distribution of remission times larger than the 0.6th cannot be estimated using the Kaplan–Meier method.

Note that while we can estimate the median and other quantiles of the distribution of survival times using the Kaplan–Meier results, we are unable to estimate the mean of the distribution in the typical case, as in the 6-MP group, where the longest follow-up time is censored (Problem 3.7).

A final note: graphs are useful for giving overall impressions of the survival function, but it is difficult to read quantities from them (e.g., median survival time or $\hat{S}(t)$ for some particular t). To obtain precise values, the results in Tables 3.13 and 3.14 can be printed in Stata using the `sts list` and `stsci` commands.

3.5.5 Cumulative Event Function

Another useful summary of survival data is the probability of having experienced the outcome event by time t. In terms of our leukemia example, this would mean estimating the probability of having relapsed by the end of each week of the study.

> *Definition*: The *cumulative event function* at time t, denoted $F(t)$, is the probability that the event has occurred by time t, or equivalently, the probability that the survival time is less than or equal to t. Note that $F(t) = 1 - S(t)$.

The cumulative event function is estimated by the complement of the Kaplan–Meier estimate of the survival function: that is, $\hat{F}(t) = 1 - \hat{S}(t)$. If t has the same value τ for all study participants, then $F(\tau)$ is interpretable as the outcome risk discussed in Sect. 3.4 on contingency table methods for binary outcomes. The cumulative event plots shown in Fig. 3.3 are also easily obtained in Stata by specifying the `failure` option.

Note that parametric methods can also be used to estimate survival distributions, as well as quantities that are not immediately available from the Kaplan–Meier approach (e.g., the mean and specified quantiles). However, because they rest on explicit assumptions about the form of these distributions, they are somewhat less robust than the methods presented here. For example, the mean can be poorly estimated in situations where a large proportion of the data are censored, with the result that the right tail of the survival function is only “known” by extrapolation.

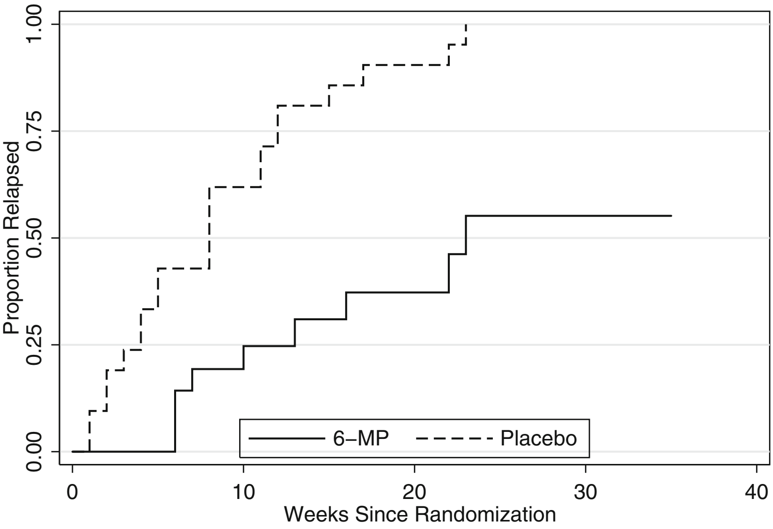

Fig. 3.3 Cumulative event curves by treatment for leukemia patients

3.5.6 Comparing Groups Using the Logrank Test

The Kaplan–Meier estimator provides an interpretable description of the survival experience of two treatment groups in the study of 6-MP as maintenance therapy for ALL. With those descriptions in hand, how do we go on to formally test for differences in relapse between the treatments?

The primary tool for the comparison of the survival experience of two or more groups is the *logrank test*. The null hypothesis for this test is that the survival distributions being compared are equal at all follow-up times. In the leukemia example, this implies that the population survival curves for 6-MP and placebo coincide. The alternative hypothesis is that the two survival curves differ at one or more points in time. Like the Kaplan–Meier estimator, the logrank test accommodates right-censoring. It works by comparing observed numbers of events in each group to the number expected if the survival functions were the same. The comparison accounts for differences in length of follow-up in calculating the expected numbers of events. Results are shown in Table 3.15.

There are a total of 30 events in the sample, 21 in the placebo group and 9 in the 6-MP group. The column labeled `Events expected` gives the expected number of events in the two groups under the null hypothesis of equal survival functions. In the leukemia data, average follow-up was considerably shorter in the placebo group and hence fewer events would be expected in that group. Clearly there were many more events than expected among placebo participants, and many fewer than expected in the 6-MP group. The resulting χ^2 statistic of 16.8 is statistically significant ($P < 0.00005$), in accord with our earlier impression that 6-MP is effective maintenance therapy for patients with ALL.

Table 3.15 Logrank test for leukemia example

```
Logrank test for equality of survival functions
-----------------------------------------------
         |   Events         Events
group    |  observed       expected
---------+-------------------------
6 MP     |         9          19.25
Placebo  |        21          10.75
---------+-------------------------
Total    |        30          30.00

              chi2(1) =       16.79
              Pr>chi2 =      0.0000
```

The logrank test is easily generalized to the comparison of more than two groups. The logrank test statistic for $K > 2$ groups follows an approximate χ^2 distribution with $K - 1$ degrees of freedom. In this more general case, the null hypothesis is

$$H_0 : S_1(t) = \ldots = S_K(t) \quad \text{for all } t \tag{3.16}$$

where $S_k(t)$ is the survival function for the kth group at time t. In analogy to the F-test discussed in Sect. 4.3.3, the alternative hypothesis is that some or all of the survival curves differ at one or more points in time.

When the null hypothesis is rejected, visual inspection of the Kaplan–Meier plots can help to determine where the important differences arise. Another common procedure for understanding group differences is to conduct pairwise logrank tests. This requires cautious interpretation; see Sect. 4.3.4 for approaches to handling potential difficulties with multiple comparisons.

If there are more than two groups which are defined by ordered categories (e.g., disease stage) or categories based on a numerical variable (e.g., number of positive nodes), then a trend test based on the logrank is available. In Stata, this is obtained by using the `trend` option for the command `sts test`.

Like some other nonparametric methods reviewed earlier in this chapter, and as its name implies, the logrank test only uses information about the *ranks* of the survival times rather than their actual values. The semi-parametric Cox proportional hazards model covered in Chap. 6 also works this way. In every instance, the nonparametric approach reduces the need for making restrictive and sometimes hard-to-verify assumptions, with a view toward making estimates more robust.

There is an extensive literature on testing differences in survival between groups. These tests have varying levels of similarity to the logrank test. The most popular are extensions of the Wilcoxon test for censored data; these tests can be viewed as a weighted versions of the logrank test. Such weighting can make sense, for example, if early events are judged to be particularly important. However, in the absence of compelling and prespecified reasons, we recommend the logrank test as a default test.

Chapter 6 covers censoring and other types of missing data in greater depth, and also presents more comprehensive methods of analysis for survival data, including the multipredictor Cox proportional hazards regression model.

3.6 Bootstrap Confidence Intervals

Bootstrapping is a widely applicable method for obtaining standard errors and CIs in cases where approximate methods for computing valid CIs have been developed but not conveniently implemented in statistical packages; other situations where development of such methods has turned out to be intractable; and datasets where the assumptions underlying the established methods are badly enough violated that the resulting CIs would be unreliable.

In general, standard errors and CIs reflect the sampling distribution of statistics of interest, such as regression coefficient estimates: that is, their relative frequency if we repeatedly drew independent samples of the same size from the source population, and recalculated the statistics in each new sample. In standard problems such as linear regression, the sampling distribution of the regression coefficient estimates is well known on theoretical grounds, provided the data meet underlying assumptions.

Bootstrap procedures approximate the sampling distribution of statistics of interest by a *resampling* procedure. Specifically, the actual sample is treated as if it were the source population, and bootstrap samples are repeatedly drawn from it. Bootstrap samples of the same size as the actual sample—a key determinant of precision—are obtained by resampling *with replacement*, so that in a given bootstrap sample some observations appear more than once, some once, and some not at all. We use the sample to represent the population and hence resampling from the actual data mimics drawing repeated samples from the source population. Then, from each of a large number of bootstrap samples, the statistics of interest are computed. For example, if our focus was on the difference between the coefficient estimates for a predictor of interest before and after adjustment for a covariate, the two models would be estimated in each bootstrap sample, and the difference between the two coefficient estimates tabulated across samples. The result would be the bootstrap distribution of the difference, which can in turn be regarded as an estimate of its actual sampling distribution. CIs for the statistic of interest would then be computed from the bootstrap distribution. Stata calculates bootstrap CIs using three procedures:

- *Normal approximation*: If the bootstrap distribution of the statistic of interest is reasonably normal, it may be enough to compute its standard deviation, then compute a conventional CI centered on the observed statistic, simply substituting the bootstrap SD for the usual model-based standard error of the statistic. The bootstrap SD is a relatively stable estimate of the standard error, since it is based on the complete set of bootstrap samples, so a relatively small number of bootstrap samples may suffice. However, we often resort to the bootstrap

Table 3.16 Bootstrap confidence interval for association of age with SBP

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model | 2179.70702     1  2179.70702            Prob > F      =  0.0188
    Residual | 106991.347   274   390.47937            R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total | 109171.054   275  396.985652            Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286     .186455     2.36   0.019     .0734621    .8075952
       _cons |    105.713    12.40238     8.52   0.000      81.2969     130.129
------------------------------------------------------------------------------

.  bootstrap '"reg SBP age"' _b, reps(1000)

command:      reg SBP age
statistics:   b_age      = _b[age]

Bootstrap statistics                               Number of obs    =       276
                                                   Replications     =      1000

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       b_age |  1000  .4405287 -.0078003  .1744795  .0981403    .782917   (N)
             |                                      .0655767   .7631486   (P)
             |                                      .0840077   .7690148  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

precisely because the sampling distribution of the statistic of interest is unlikely to be normal, particularly in the tails. Thus this method is less reliable for constructing CIs than for estimating the standard error of the statistic.

- *Percentile Method*: The CI for the statistic of interest is constructed from the relevant quantiles of the bootstrap distribution. Because the extreme percentiles of a sample are very noisy estimates of the corresponding percentiles of a population distribution, a much larger number of bootstrap samples is required. If 1,000 samples were used, then a 95% CI for the statistic of interest would span the 25th to 975th largest bootstrap estimates.
- *Bias-Corrected Percentile Method*: The percentile-based confidence interval is shifted to account for bias, as evidenced by a difference between the observed statistic and the median of the bootstrap estimates. Again, a relatively large number of bootstrap samples is required.

Table 3.16 shows Stata output for the simple linear regression model for SBP shown earlier in Table 3.4, now with a bootstrap CI. In this instance, all three bootstrap results are fairly consistent with the parametric 95% CI (0.73–0.81 mmHg). See Sects. 4.5.4, 5.5.1, 6.6.1, and 7.9.1 for other examples where bootstrap CIs are computed.

3.7 Interpretation of Negative Findings

Confidence intervals obtained either by standard parametric methods or by the bootstrap play a particularly important role when the data do not enable us to reject a null hypothesis of interest. It is easy to overstate such negative findings. Recall that $P > 0.05$ does not prove the null hypothesis; it only indicates that the observed result could have arisen by chance, not that it necessarily did. A negative result worth discussing is best interpreted in terms of the point estimate and CI. In the following example, we can distinguish four possible cases, in increasing order of the strength of the negative finding. Suppose that a 20% reduction risk of recurrent heart attacks would justify the risks and costs of a possible new treatment, but that a risk reduction of only 5% would not meet this standard. The four cases are:

- The estimated risk reduction was large enough to be substantively important, but the CI spanned the null value and was thus too wide to provide strong evidence for effectiveness. Example: treatment reduced recurrence risk an estimated 20% (95% CI –1% to 37%). In this case, we might conclude that the study gives inconclusive evidence *for* the potential importance of the treatment; but it would be also important to note that the CI includes effects too small to be worthwhile.
- The estimated risk reduction was too small to be important, but the CI extended to values that could be important. Example: treatment reduced recurrence risk an estimated 5% (95% CI –15% to 22%). In this case the point estimate provides little support for the importance of the treatment, but the CI does not clearly rule out a potentially important effect.
- The estimated risk reduction was too small to be important, and while the CI did not include the null (i.e., $P < 0.05$), it did exclude values that could be important. Example: treatment reduced recurrence risk an estimated 3% (95% CI: 1% to 5%). In this case, we can definitively say that the treatment does not have a clinically important benefit, even though we can also rule out no effect.
- The estimated risk reduction was too small to be important, and the CI both included the null and excluded values that could be important. Example: treatment reduced recurrence risk an estimated 1% (95% CI –2% to 4%). Again, we can definitively say that the treatment does not have a clinically important benefit.

This approach using the point estimate and CI is preferable to interpretations based on *ex post facto* power calculations, which are driven by assumptions about the true effect size, and often inappropriately based on treating the observed effect size as if it were the true population value (Hoenig and Heisey 2001). A variant of this approach is to suggest that with a larger sample, the observed effect would have been statistically significant. But of course the CI for most negative findings tells us that the true effect size may well be nil or worse, which a larger sample might also firmly establish. In contrast to these problematic interpretations, the point estimate and CI can together be used to summarize what the data at hand have to tell us about the strength of the association and the precision of our information about it.

3.8 Further Notes and References

Among the best introductory statistics books are Freedman et al. (1991), Devore and Peck (1986), and Pagano and Gavreau (1993). Consult these for more complete coverage of basic statistical inference, ANOVA, and linear regression. Good references on methods for the analysis of contingency tables include Fleiss et al. (2003) and Jewell (2004). Two applied survival analysis texts with a biomedical orientation are Miller et al. (1981) and Marubini and Valsecchi (1995). Finally, for a review of bootstrap methods, see Efron and Tibshirani (1986, 1993).

3.9 Problems

Problem 3.1. An alternative to OLS is least absolute deviation (LAD) regression, in which the regression line is selected to minimize the sum of the absolute vertical differences (rather than squared differences) between the line and the data. Explain how this might reduce sensitivity to outliers.

Problem 3.2. To create a new age variable `age10` in units of ten years, we would divide the original variable age (in years) by ten, so that a woman of age 67 would have `age10` $= 6.7$. Similarly, the standard deviation of `age10` is changed by the same factor: that is, the SD of age is 6.38, so the SD of `age10` is 0.638. Suppose we want to estimate the effect of age in SD units, as is commonly done. How do we compute the new variable and what is *its* SD?

Problem 3.3. Using (3.12) and a statistical analysis program, demonstrate with your own data that the slope coefficient in a univariate linear model with continuous predictor and outcome is a rescaled transformation of the sample correlation between predictor and outcome.

Problem 3.4. The correlation coefficient is a measure of *linear* association. Suppose x takes on values evenly over the range from –10 to 10, and that $E[y|x] = x^2$. In this case, the correlation of x and y is zero, even though there is clearly a systematic relationship. What does this suggest about the need to test model assumptions? Using a statistical package, generate a random sample of 100 values of x uniformly distributed on [–10, 10], compute $E[y|x]$ for each value of x, add randomly generated standard normal errors to get the 100 values of y, and check the sample correlation of x and y.

Problem 3.5. Verify the estimates for the excess risk, relative risk, and odds ratio for the HIV example presented in Table 3.6.

Problem 3.6. The data presented below are from a case-control study of esophageal cancer. (The study and data are described in more detail in Sect. 5.3.)

```
. tabulate case ditob

       Case |
     status |
   (1=case, |          tobacco
  0=control) | 0-9 g/day  10+ g/day |     Total
-----------+----------------------+----------
          0 |       255        520 |       775
          1 |         9        191 |       200
-----------+----------------------+----------
      Total |       264        711 |       975
```

The rows (labeled according to `Case status`) represent 200 cancer cases and 775 cancer-free controls selected from the same population as the cases. The columns represent a binary indicator of reported consumption of more than ten grams of tobacco per day.

Compute the odds ratio comparing the risk of cancer in individuals who report consuming more than ten grams of tobacco per day with the corresponding risk in the group reporting less or no consumption. Next, compute the odds ratio comparing the proportion of individuals reporting higher levels of consumption among cases with that among the controls. Comment.

Problem 3.7. Suppose we could estimate the value of the survival function $S(t)$ for every possible survival time from $t = 0$ onward. Clearly $S(t) \rightarrow 0$ as t becomes large. It can be shown that the mean survival time is equal to the area under this "complete" survival curve. Why are we unable to estimate mean survival from the Kaplan–Meier result when the largest follow-up time is censored? To gain insight, contrast the survival curves for the 6-MP and placebo groups in Fig. 3.2.

Problem 3.8. In the leukemia study, the probability of being relapse-free at 20 weeks, conditional on being relapse-free at 10 weeks, can be estimated by the Kaplan–Meier estimate for 20 weeks, divided by the corresponding estimate for 10 weeks. In the placebo group, those estimates are 0.38 and 0.10, respectively. Verify that the estimated conditional probability of remission at week 20, conditional on being in remission at week 10, is 0.25. In the 6-MP group, estimated probabilities of remaining in remission are 0.81, 0.63, and 0.45 at 10, 20, and 30 weeks, respectively. Use these values to estimate the probabilities of remaining in remission at 20 and 30 weeks, conditional on being in remission at 10 weeks.

3.10 Learning Objectives

(1) Be familiar with the t-test (including versions for paired and unequal-variance data), one-way ANOVA, the correlation coefficient r, and some nonparametric alternatives.

(2) Describe the assumptions and mechanics of the simple linear model for continuous outcomes, and interpret the results.

(3) Define the basic measures of association (i.e., excess risk, relative risk, and odds ratio) for binary outcomes.

(4) Be familiar with standard contingency table approaches to evaluating associations between binary outcomes and categorical predictors, including the χ^2 test and the Mantel–Haenszel approach to estimating odds ratios adjusted for the confounding influence of additional predictors.

(5) Define right-censoring.

(6) Interpret Kaplan–Meier survival and cumulative event curves.

(7) Calculate median survival from an estimated survival curve.

(8) Interpret the results of a logrank test.