Regression Methods for Epidemiologic Analysis

II.3

Sander Greenland

Introduction 3.1

Basic tabular and graphical methods are an essential component of epidemiologic analysis and are often sufficient, especially when one need consider only a few variables at a time. They are, however, limited in the number of variables that they can examine simultaneously. Even sparse-strata methods (such as Mantel–Haenszel) require that some strata have two or more subjects; yet, as more and more variables or categories are added to a stratification, the number of subjects in each stratum may eventually drop to 0 or 1.

Regression analysis encompasses a vast array of techniques designed to overcome the numerical limitations of simpler methods. This advantage is purchased at a cost of stronger assumptions, which are compactly represented by a *regression model*. Such models (and hence the assumptions they represent) have the advantage of being explicit; a disadvantage is that the models may not be well understood by the intended audience or even the user. Regression models can and should be tailored by the analyst to suit the topic at hand; the latter process is sometimes called *model specification*. This process is part of the broader task of *regression modeling*.

To ensure that the assumptions underlying the regression analysis are reasonable approximations, it is essential that the modeling process be actively guided by the scientists involved in the research, rather than be left solely to mechanical algorithms. Such active guidance requires familiarity with the variety and interpretation of models. Hence, the present chapter will focus primarily on forms of models and their interpretation, rather than on the more technical issues of model fitting and testing. Because this chapter provides only outlines of key topics, it should be supplemented by readings in more detailed treatments of regression analysis, as can be found in Breslow and Day (1980, 1987), McCullagh and Nelder (1989), Clayton and Hills (1993), and Hosmer and Lemeshow (2000). For an in-depth treatment of the difficulties and limitations of regression analysis in nonexperimental studies, see Leamer (1978) or Berk (2004).

Achieving working competence in regression analysis requires comfort with basic geometry and algebra. While the ensuing discussion attempts to be self-contained, readers who feel lacking or weak in mathematical skills would do well to review a textbook in high school mathematics or college algebra (focusing especially on functions, graphs, and natural logarithms) before studying regression methods.

Regression Functions 3.2

A regression *function* is distinct from a model for that function. A regression *model* is another, simpler function used to approximate or estimate the true regression function. This distinction is often obscured and even unrecognized in elementary treatments of regression, which in turn has generated much misunderstanding

of regression modeling. Therefore, this chapter provides separate discussions of *regression* functions and regression models.

There are two primary interpretations of regression functions, frequentist and Bayesian, which correspond to two different interpretations of probability (see Rothman and Greenland 1998, Chap. 12). The present chapter uses the frequentist interpretation, but briefly discusses the Bayesian interpretation at the end of this section. In both interpretations, the term *regression* is often used to refer to the regression function.

3.2.1 Frequentist Regression

In the frequentist view, the *regression* of a variable Y on another variable X is the function that describes how the average (mean) value of Y changes across population subgroups defined by levels of X. This function is often written as $E(Y|X = x)$, which should be read as "the average of Y when the variable X takes on the specific value x." The "E" part of the notation stands for "expectation", which here is just another word for "population mean".

As an example, suppose Y stands for "height" to the nearest centimeter at some time t, X stands for "weight" to the nearest kilogram at time t, and the population of interest is that of Denmark at time t. If we subclassify the Danish population at t into categories of weight X, compute the average height in each category, and tabulate or graph these average heights against the weight categories, the result displays the regression, $E(Y|X = x)$, of height Y on weight X in Denmark at time t. Several important points should be emphasized:

1. The *concept* of regression involves no modeling. Some would describe this fact by saying that the concept of regression is essentially "nonparametric". The regression of Y on X is just a graphical property of the physical world, like the orbital path of the earth around the sun.
2. There is nothing mathematically sophisticated about the regression function. Each point on a regression curve could be computed by taking the average of Y within a subpopulation defined as having a particular value of X. In the example, the value of the regression function at $X = 50$ kg, $E(Y|X = 50)$, is just average height at time t among Danes who weigh 50 kg at time t.
3. A regression function cannot be unambiguously computed until we carefully define X, Y, *and* the population over which the averages are to be taken. We will call the latter population the *target population* of the regression. This population is all too often left out of regression definitions, often resulting in confusion.

Some ambiguity is unavoidable in practice. In our example, is time t measured to the nearest year, day, minute, or millisecond? Is the Danish population all citizens, all residents, or all persons present in Denmark at t? We may decide that leaving these questions unanswered is tolerable, because varying the definitions over a modest range would not change the result to an important extent. But if we left time completely out of the definition, the regression would become hopelessly

ambiguous, for now we would not have a good idea of who to include or exclude from our average: Should we include people living in Denmark in prehistoric times, or in the time of King Canute (a thousand years ago), or in the distant future (a thousand years from now)? The choice could have a strong effect on our answer, because of the large changes in height-to-weight relations that have occurred over time.

Other Concepts of Population 3.2.2

It is important to distinguish between a "target population" and a "source population". The target population of regression is defined without regard to our observations; for example, the regression of diastolic blood pressure on cigarette usage in China is defined whether or not we conduct a study in China (the target for this regression). A source population is a source of subjects for a particular study and is defined by the selection methods of the study; for example, a random-sample survey of all residents of Beijing would have Beijing as its source population. The concepts of target and source populations connect only insofar as inferences about a regression function drawn from a study are most easily justified when the source population of the study is identical to the target population of the regression. Otherwise, issues of generalization from the source to the target have to be addressed (see Rothman and Greenland 1998, Chap. 8).

In some literature, regression functions (and many other concepts) are defined in terms of averages within a "superpopulation" or "hypothetical universe". A superpopulation is an abstraction of a target population, sometimes said to represent the distribution (with respect to all variables of interest) of all possible persons that ever were or ever could be targets of inference for the analysis at hand. Because the superpopulation approach focuses on purely hypothetical distributions, it has encouraged substitution of mathematical theory for the more prosaic task of connecting study results to populations of immediate public-health concern. Thus, the present chapter defines regression functions in terms of averages within real (target) populations.

Binary Regression 3.2.3

The concept of regression applies to variables measured on any scale: The regressand and the regressor may be continuous or discrete, or even binary. For example, Y could be an indicator of diabetes ($Y = 1$ for present, $Y = 0$ for absent), and X could be an indicator for sex ($X = 1$ for female, $X = 0$ for male). Then $E(Y|X = 1)$ would represent the average of the diabetes indicator Y among females, and $E(Y|X = 0)$ would represent the average of Y among males.

When the regressand Y is a binary indicator $(0, 1)$ variable, $E(Y|X = x)$ is called a *binary regression*, and this regression simplifies in a very useful manner. Specifically, when Y can be only 0 or 1, the average $E(Y|X = x)$ equals the proportion of population members who have $Y = 1$ among those who have $X = x$. For example,

if Y is the diabetes indicator, $E(Y|X = x)$ is the proportion with diabetes (i.e., with $Y = 1$) among those with $X = x$. To see this, let N_{yx} denote the number of population members who have $Y = y$ and $X = x$. Then the number of population members with $X = x$ is $N_{1x} + N_{0x} = N_{+x}$, and the average of Y among these members, $E(Y|X = x)$, is

$$\frac{N_{1x} \times 1 + N_{0x} \times 0}{N_{1x} + N_{0x}} = \frac{N_{1x}}{N_{+x}} ,$$

which is just the proportion with $Y = 1$ among those with $X = x$.

The epidemiologic ramifications of the preceding relation are important. Let $\Pr(Y = y|X = x)$ stand for "the proportion (of population members) with $Y = y$ among those with $X = x$" (which is often interpreted as the probability of $Y = y$ in the subpopulation with $X = x$). If Y is a binary indicator, we have just seen that

$$E(Y|X = x) = \Pr(Y = 1|X = x) ,$$

that is, the average of Y when $X = x$ equals the proportion with $Y = 1$ when $X = x$. Thus, if Y is an indicator of *disease presence* at a given time, the regression of Y on X, $E(Y|X = x)$, provides the proportion *with* the disease at that time, or prevalence proportion, given $X = x$. For example, if $Y = 1$ indicates diabetes presence on January 1, 2010 and X is weight on that day, $E(Y|X = x)$ provides diabetes prevalence as a function of weight on that day. If Y is instead an indicator of *disease incidence* over a time interval (cf. Chap. I.2 of this handbook and Chap. 3 of Rothman and Greenland, 1998), the regression of Y on X provides the proportion getting disease over that interval, or incidence proportion, given $X = x$. For example, if $Y = 1$ indicates stroke occurrence in 2010 and X is weight at the start of the year, $E(Y|X = x)$ provides the stroke incidence (proportion) in 2010 as a function of initial weight.

3.2.4 Multiple Regression

The concept of multiple regression is a simple extension of the ideas discussed above to situations in which there are multiple (two or more) regressors. To illustrate, suppose Y is a diabetes indicator, X_1 stands for "sex" (coded 1 for females, 0 for males), and X_2 stands for "weight" (in kilograms). Then the regression of Y on X_1 and X_2, written $E(Y|X_1 = x_1, X_2 = x_2)$, provides the average of Y among population members of a given sex X_1 and weight X_2. For example, $E(Y|X_1 = 1, X_2 = 70)$ is the average diabetes indicator (and, hence, the diabetes prevalence) among women who weigh 70 kg.

We can use as many regressors as we want. For example, we could include age (in years) in the last regression. Let X_3 stand for "age". Then $E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$ would provide the diabetes prevalence among population members of a given sex, weight, and age. Continuing to include regressors produces a very clumsy notation, however, and so we adopt a simple convention: We will let $\boldsymbol{X}$ represent the ordered list of all the regressors we want to consider. Thus, in our diabetes

example, $\boldsymbol{X}$ will stand for the horizontal list (X_1, X_2, X_3) of "sex", "weight", and "age". Similarly, we will let $\boldsymbol{x}$ stand for the horizontal ordered list of values (x_1, x_2, x_3) for $\boldsymbol{X} = (X_1, X_2, X_3)$. Thus, if we write $E(Y|\boldsymbol{X} = \boldsymbol{x})$, it is merely a shorthand for

$$E\left(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3\right) ,$$

when there are three regressors under consideration.

More generally, if there are n regressors $X_1, \ldots, X_n$, we will write $\boldsymbol{X}$ for the ordered list $(X_1, \ldots, X_n)$ and $\boldsymbol{x}$ for the ordered list of values $(x_1, \ldots, x_n)$. The horizontal ordered list of variables $\boldsymbol{X}$ is called a *row vector* of regressors, and the horizontal ordered list of values $\boldsymbol{x}$ is called a *row vector* of values. Above, the vector $\boldsymbol{X}$ is composed of the $n = 3$ items "sex", "weight", and "age", and the list $\boldsymbol{x}$ is composed of specific values for sex (0 or 1), weight (kilograms), and age (years). The number of items n in $\boldsymbol{X}$ is called the length or dimension of $\boldsymbol{X}$.

The term *multivariate regression* is usually reserved for regressions in which there are multiple *regressands*. To illustrate, suppose Y_1 is an indicator of diabetes presence, Y_2 is diastolic blood pressure, and $\boldsymbol{Y}$ is the list (Y_1, Y_2) composed of these two variables. Also, let $\boldsymbol{X}$ be the list (X_1, X_2, X_3) composed of the sex indicator, weight, and age, as before. The multivariate regression of diabetes and blood pressure on sex, weight, and age provides the average diabetes indicator *and* average blood pressure for each combination of sex, weight, and age:

$$E\left(Y_1, Y_2|X_1 = x_1, X_2 = x_2, X_3 = x_3\right) = E(Y|X = x) .$$

In general, there may be any number of regressands in the list Y and regressors in the list $\boldsymbol{X}$ of a multivariate regression. Multivariate regression notation allows one to express the separate regressions for each regressand in one equation.

Regression and Causation

3.2.5

When considering a regression function $E(Y|X = x)$, the variable Y is termed the dependent variable, outcome variable, or *regressand*, and the variable X is termed the independent variable, predictor, covariate, or *regressor*. The "dependent/independent" terminology is common but also problematic because it invites confusion of distinct probabilistic and causal concepts of dependence and independence. For example, if Y is age and X is blood pressure, $E(Y|X = x)$ represents the average age of persons given blood pressure, X. But it is blood pressure X that causally depends on age Y, not the other way around.

More generally, for any pair of variables X and Y, we can consider either the regression of Y on X, $E(Y|X = x)$, or the regression of X on Y, $E(X|Y = y)$. Thus, the concept of regression does not necessarily imply any causal or even temporal relation between the regressor and the regressand. For example, Y could be blood pressure at the start of follow-up of a cohort, and X could be blood pressure after 1 year of follow-up; then $E(Y|X = x)$ would represent the average initial blood pressure among cohort members whose blood pressure after 1 year of follow-up is x. This is an example of a noncausal regression.

Because regression functions do not involve any assumptions of time order or causal relations, regression coefficients and quantities derived from them represent measures of association, not measures of effect. To interpret the coefficients as measures of causal effects, it is important that the regression function being modeled provide a representation of the effects of interest that is approximately unconfounded (for a general discussion of the concept of confounding see Chap. I.9 of this handbook and Chap. 4 of Rothman and Greenland, 1998).

To make this no-confounding assumption more precise, suppose $\boldsymbol{X}$ contains the exposures of interest and $\boldsymbol{Z}$ contains the other regressors. Following Pearl (1995), we may then write

$$E\left[Y|\,\mathrm{Set}(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{Z}=\boldsymbol{z}\right]$$

for the average value Y would have *if* everyone in the target population with $\boldsymbol{Z}=\boldsymbol{z}$ had their $\boldsymbol{X}$ value set to $\boldsymbol{x}$. This potentially counterfactual average can be very different from $E(Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z})$. The latter refers only to those population members with $\boldsymbol{X}=\boldsymbol{x}$ and $\boldsymbol{Z}=\boldsymbol{z}$, whereas the former refers to all population members with $\boldsymbol{Z}=\boldsymbol{z}$, including those who actually had $\boldsymbol{X}$ equal to values other than $\boldsymbol{x}$.

As an example, suppose the target population is all persons born during 1901–1950 surviving to age 50, Y is an indicator of death by age 80, $\boldsymbol{X}$ contains only X_1 = pack-years of cigarettes smoked by age 50, and $\boldsymbol{Z}=(Z_1,Z_0)$ where $Z_1=1$ if female, 0 if male and Z_2 = year of birth. Then

$$E\left[Y|X_1=20,\boldsymbol{Z}=(1,1940)\right]$$

would be the average risk of dying by age 80 (mortality proportion) among women born in 1940 and surviving to age 50 who smoked 20 pack-years by age 50. In contrast,

$$E\left[Y|\,\mathrm{Set}(X_1=20),Z=(1,1940)\right]$$

would be the average risk of dying by age 80 among all women born in 1940 and surviving to age 50 *if* all such women had smoked 20 pack-years by age 50.

In regression analysis, we may define effect measures as contrasts of average outcomes (such as incidence) in the same population under different conditions. Consider the ratio effect measure contrasting the average of Y in the subpopulation with $\boldsymbol{Z}=\boldsymbol{z}$ when $\boldsymbol{X}$ is set to $\boldsymbol{x}^*$ versus that average when $\boldsymbol{X}$ is set to $\boldsymbol{x}$:

$$\frac{E\left[Y|\,\mathrm{Set}(\boldsymbol{X}=\boldsymbol{x}^*),\boldsymbol{Z}=\boldsymbol{z}\right]}{E\left[Y|\,\mathrm{Set}(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{Z}=\boldsymbol{z}\right]}.$$

In the example,

$$\frac{E\left[Y|\,\mathrm{Set}(X_1=20),\boldsymbol{Z}=(1,1940)\right]}{E\left[Y|\,\mathrm{Set}(X_1=0),\boldsymbol{Z}=(1,1940)\right]}$$

represents the *effect* of smoking 20 pack-years by age 50 versus no smoking on the risk of dying by age 80 among women born in 1940. On the other hand, the ratio measure

$$\frac{E\left[Y|\,\mathrm{Set}(X_1 = 20), \boldsymbol{Z} = (1, 1940)\right]}{E\left[Y|\,\mathrm{Set}(X_1 = 0), \boldsymbol{Z} = (1, 1940)\right]}$$

represents only the *association* of smoking 20 pack-years by age 50 versus no smoking with the risk among women born in 1940, because it contrasts two different subpopulations (one with $X_1 = 20$, the other with $X_1 = 0$).

To infer that all associational measures estimated from our analysis equal their corresponding effect measures, we would have to make the following assumption of no confounding given $\boldsymbol{Z}$ (which is sometimes expressed by stating that there is no residual confounding):

$$E\left(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}\right) = E\left[Y|\,\mathrm{Set}(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}\right] \ .$$

This assumption states that the average we observe or estimate in the subpopulation with both $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Z} = \boldsymbol{z}$ is equal to what the average in the larger subpopulation with $\boldsymbol{Z} = \boldsymbol{z}$ would have been if everyone had $\boldsymbol{X}$ set to $\boldsymbol{x}$. It is important to appreciate the strength of the assumption. In the above example, the no-confounding assumption would entail

$$E\left[Y|X_1 = 20, Z = (1, 1940)\right] = E\left[Y|\,\mathrm{Set}(X_1 = 20), \boldsymbol{Z} = (1, 1940)\right] \ ,$$

which states that the risk we will observe among women born in 1940 who smoked 20 pack-years by age 50 equals the risk we would have observed in *all* women born in 1940 if they all had smoked 20 pack-years by age 50. The social variables associated with both smoking and death should lead us to doubt that the two quantities are even approximately equal.

If only a single summary measure of effect is desired, the covariate-specific no-confounding assumption can be replaced by a less restrictive assumption tailored to that measure. To illustrate, suppose in the above example we are only interested in what the effect of smoking 20 versus zero pack-years would be on *everyone* in the target, regardless of sex or birth year, as measured by the causal risk ratio

$$E\left[Y|\,\mathrm{Set}(X_1 = 20)\right] / E\left[Y|\,\mathrm{Set}(X_1 = 0)\right] \ .$$

The corresponding measure of association is the risk ratio for 20 versus 0 pack-years, standardized to the total population:

$$\frac{\sum_z E\left(Y|X_1 = 20, \boldsymbol{Z} = \boldsymbol{z}\right) \Pr(\boldsymbol{Z} = \boldsymbol{z})}{\sum_z E\left(Y|X_1 = 0, \boldsymbol{Z} = \boldsymbol{z}\right) \Pr(\boldsymbol{Z} = \boldsymbol{z})} \ ,$$

where $\Pr(\boldsymbol{Z} = \boldsymbol{z})$ is the proportion with $\boldsymbol{Z} = \boldsymbol{z}$ in the target. The no-confounding assumption we need here is that the standardized ratio equals the causal ra-

tio. This summary assumption could hold even if there was confounding within levels of sex and birth year (although it would still be implausible in this example).

The dubiousness of no-confounding assumptions is often the chief limitation in using epidemiologic data for causal inference. This limitation applies to both tabular and regression methods. Randomization of persons to levels of $\boldsymbol{X}$ can largely overcome this limitation because it ensures that effect estimates follow a quantifiable probability distribution centered around the true effect. Randomization is not an option in most settings, however.

The default strategy is to ensure there are enough well-measured confounders in $\boldsymbol{Z}$ so that the no-confounding assumption is at least plausible. This strategy often leads to few subjects at each level $\boldsymbol{x}$ of $\boldsymbol{X}$ and $\boldsymbol{z}$ of $\boldsymbol{Z}$, which in turn lead to the sparse-data problems that regression modeling attempts to address (Robins and Greenland 1986; Greenland 2000a, b; Greenland et al. 2000). A major limitation of this strategy is that, often, key confounders are poorly measured or unmeasured, and so cannot be used in ordinary modeling; prior distributions for the missing confounders must be used instead (Greenland 2003a).

3.2.6 Frequentist versus Bayesian Regression

In frequentist theory, an expectation is interpreted as an average in a specific subgroup of a specific population. The regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$ thus represents an objective functional relation among theoretically measurable variables (the average of Y as a function of the variables listed in $\boldsymbol{X}$). It may be that this relation has not been observed, perhaps because it exists but we are unable to measure it, or because it does not yet exist. Examples of the former and latter are the regressions of blood pressure on weight in Spain 10 years ago and 10 years from now. In either situation, the regression is an external relation that one tries to estimate, perhaps by projecting (extrapolating) from current knowledge about presumably similar relations. For example, one might use whatever survey data one can find on blood pressure and weight to estimate what the regression of blood pressure on weight would look like in Spain 10 years ago or 10 years from now. In this approach, one tries to produce an *estimate* $\widehat{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ of the true regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$.

In subjective Bayesian theory, an expectation is what we would or should expect to see in a given target population. This notion of expectation corresponds roughly to a prediction of what we would see if we could observe the target in question. The regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$ does not represent an objective relation to be estimated, but instead represents a subjective (personal) expectation about how the average of Y varies across levels of $\boldsymbol{X}$ in the target population. Like the frequentist regression estimate, however, it is something one constructs from whatever data one may find that seems informative about this variation.

Both frequentist and Bayesian authors have noted that the two approaches often yield similar interval estimates (Cox and Hinkley 1974; Good 1983). It is increasingly recognized that divergences are usually due to differences in the criteria for a "good" point estimate: Frequentists traditionally prefer criteria of

unbiased prediction (e.g., having an average error of zero), whereas Bayesians more often prefer criteria of closeness (e.g., having the smallest average squared error possible). When analogous criteria are adopted in both approaches, Bayesian and frequentist methods can yield similar numeric results in standard epidemiologic applications.

Nonetheless, Bayesians and frequentists interpret their results differently. The Bayesian presents a prediction, denoted by $E(Y|\boldsymbol{X} = \boldsymbol{x})$, which he or she interprets as his or her "best bet" about the average of Y when $\boldsymbol{X} = \boldsymbol{x}$, according to some criteria for "best bet". The frequentist presents a prediction, denoted by $\widehat{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ (or, more commonly, $\widehat{Y}_{\boldsymbol{X}=\boldsymbol{x}}$), which he or she interprets as "the" best estimate of the average of Y when $\boldsymbol{X} = \boldsymbol{x}$, according to some criteria for "best estimate" (such as minimum variance among statistically unbiased estimators). Too often, the latter criteria are presumed to be universally shared, but are not really shared or even properly understood by epidemiologists; one could and would reach different conclusions using other defensible criteria (such as minimum mean squared error). For these reasons, when conducting regression analyses we find it valuable to consider both frequentist and Bayesian interpretations of methods and results.

Basic Regression Models 3.3

In any given instance, the true regression of Y on $\boldsymbol{X}$, $E(Y|\boldsymbol{X} = \boldsymbol{x})$, is an extremely complicated function of the regressors $\boldsymbol{X}$. Thus, even if we observe this function without error, we may wish to formulate simplified pictures of reality that yield *models* for this regression. These models, while inevitably incorrect, can be very useful. A classic example is the representation of the distance from the earth to the sun, Y, as a function of day of the year T. To the nearest kilometer, this distance is a complex function of T because of the gravitational effects of the moon and of the other planets in the solar system. If we represent the orbit of the earth around the sun as a circle with the sun at the center, our regression model will predict the distance $E(Y|T = t)$ by a single number (about 150 million kilometers) that does not change with t. This model is adequate if we need only predict the distances to 2% accuracy. If we represent the orbit of the earth as an ellipse, our regression model will predict the earth-sun distance as smoothly and cyclically varying over the course of a year (within a range of about 147 to 153 million kilometers). Although it is not perfectly accurate, this model is adequate if we need to predict the distances to within 0.2% accuracy.

Model Specification and Model Fitting 3.3.1

Our description of the above models must be refined by distinguishing between the *form* of a model and a *fitted* model. "Circle" and "ellipse" refer to forms, that is, general classes of shapes. The circular model form corresponds to assuming a constant earth-sun distance over time; the elliptical model form allows this

distance to vary over a temporal cycle. The process of deciding between these two forms is a simple example of *model specification*.

If we decide to use the circular form, we must also select a value for the radius (which is the earth-sun distance in the model). This radius specifies which circle (out of the many possible circles) to use as a representation of the earth's orbit and is an example of a model *parameter*. The process of selecting the "best" radius is an example of *model fitting*, and the circle that results is sometimes called the *fitted model* (although the latter term is sometimes used to refer to the model form instead). There are two important relations between a set of data and a model fit to those data. First, there is "distance" from the fitted model to the data; second, there is "resistance" or "stability" of the fitted model, which is the degree to which the parameter estimates change when the data themselves are changed.

Depending on our accuracy requirements, we may have on hand several simplified pictures of reality and hence several candidate models. At best, our choice might require a trade-off between simplicity and accuracy, as in the preceding example. There is an old dictum (often referred to as "Occam's razor") that one should not introduce needless complexity. According to this dictum, if we need only two percent accuracy in predicting the earth's distance from the sun, then we should not bother with the ellipse model and instead use the constant distance derived from the circle model.

There is a more subtle benefit from this advice than avoiding needless mental exertion. Suppose we are given two models, one (the more complex) containing the other (the more simple) as a special case, and some data with which to fit the two models. Then the more complex model will be able to fit the available data more closely than the simpler model, in the sense that the predictions from the more complex model will (on average) be closer to what was seen in the data than will the predictions from the simpler model. This is so in the above example because the ellipse contains the circle as a special case. Nonetheless, there is a penalty for this closeness to the data: The predictions obtained from the more complex model tend to be less stable than those obtained from the simpler model.

Consider now the use of the two different model forms to predict events outside of the data set to which the models were fit. An example would be forecasting the earth's distance from the sun; another would be predicting the incidence of AIDS five years in the future. Intuitively, we might expect that if one model is both closer to the data and more stable than the other, that model will give more accurate predictions. The problem is that the choice among models is rarely so clear-cut: Usually, one model will be closer to the data, while the other will be more stable, and it will be difficult to tell which will be more accurate. This is one dilemma we often face in a choice between a more complex and simpler model.

To summarize, model specification is the process of selecting a model form, while model fitting is the process of using data to estimate the parameters in a model form. There are many methods of model fitting, and the topic is so vast and technical that we will only superficially outline a few key elements. Nearly all commercial computer programs are based on one of just a few fitting methods, so that nearly all users (statisticians as well as epidemiologists) are forced to base their

analyses on the assumptions of these methods. We will briefly discuss specification and fitting methods below.

Background Example 3.3.2

The following epidemiologic example will be used at various points to illustrate specific models. At the time of this writing, there is a controversy over whether women with no history of breast cancer but thought to be of high risk (due to family history and perhaps other factors) should be given the drug tamoxifen as a prophylactic regimen. Current evidence suggests that tamoxifen might prevent breast cancer but also cause or promote endometrial and liver cancer.

One measure of the net impact of tamoxifen prophylaxis up to a given age is the change in risk of death by that age. Suppose the regressand Y is an indicator of death by age 70 ($Y = 1$ for dead, 0 for alive). The regressors $\boldsymbol{X}$ include

$$\begin{aligned} X_1 &= \text{years of tamoxifen therapy,} \\ X_2 &= \text{age (in years) at start of tamoxifen therapy,} \\ X_3 &= \text{age at menarche,} \\ X_4 &= \text{age at menopause,} \\ X_5 &= \text{parity.} \end{aligned}$$

The target population is American women born during 1945–1950 who survive to age 50 and do not use tamoxifen before that age. If tamoxifen is not taken during follow-up, we set age at tamoxifen start (X_2) to 70 because women who start at 70 or later and women who never take tamoxifen have the same exposure history during the age interval under study.

In this example, the regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$ is just the average risk, or incidence proportion, of death by age 70 among women in the target population who have $\boldsymbol{X} = \boldsymbol{x}$. Therefore, we will write $R(\boldsymbol{x})$ as a shorthand for $E(Y|\boldsymbol{X} = \boldsymbol{x})$. We will also write R for the crude (overall) average risk $E(Y)$, $R(x_1)$ for the average risk $E(Y|X_1 = x_1)$ in the subpopulation defined by having $X_1 = x_1$ (without regard to the other variables), and so on.

Vacuous Models 3.3.3

A model so general that implies nothing at all, but simply re-expresses the overall average risk R in a different notation, is

$$E(Y) = R = \boldsymbol{\alpha}\,, \quad 0 < \boldsymbol{\alpha} < 1\,. \tag{3.1}$$

(this model does exclude $R = 0$ or 1, but it allows R to be arbitrarily close to 0 or 1, so this exclusion is of no practical consequence). There is only one regression parameter (or coefficient) $\boldsymbol{\alpha}$ in this model, and it corresponds to the average risk

in the target population. A model such as model (3.1) that has no implication (i.e., that imposes no restriction or constraint) is said to be *vacuous*.

Two models are said to be equivalent if they have identical implications for the regression. A model equivalent to model (3.1) is

$$E(Y) = R = \exp(\alpha) , \quad \alpha < 0 . \tag{3.2}$$

This model has no implication. In this model, α is the natural logarithm of the overall average risk:

$$\alpha = \ln(R) .$$

Another model equivalent to models (3.1) and (3.2) is

$$E(Y) = R = \text{expit}(\alpha) , \tag{3.3}$$

where $\text{expit}(\alpha)$ is the *logistic* transform of α, defined as

$$\text{expit}(\alpha) = \frac{\exp(\alpha)}{1 + \exp(\alpha)} .$$

Again, model (3.3) has no implication. Now, however, the parameter α in model (3.3) is the logit (log odds) of the overall average risk:

$$\alpha = \ln\left(\frac{R}{1 - R}\right) = \text{logit}(R) .$$

For an introduction of risk measures in general see Chap. I.2 of this handbook and Chap. 3 of Rothman and Greenland (1998).

3.3.4 Constant Models

In comparing the complexity and implications of two models A and B, we say that model A is more general, more flexible, or more complex than model B, or that A contains B, if all the implications of model A are also implications of model B, but not vice-versa (that is, if B imposes some restrictions beyond those imposed by A). Other ways of stating this relation are that B is simpler, stronger, or stricter than A, B is contained or nested within A, or B is a special case of A. The following model is superficially similar to model (3.1), but is in fact much more strict:

$$E\left(Y|X_1 = x_1\right) = R(x_1) = \alpha \tag{3.4}$$

for all x_1. This model implies that the average risks of the subpopulations defined by years of tamoxifen use are identical. The parameter α represents the common value of these risks. This model is called a *constant* regression because it allows no variation in average risks across levels of the regressor. To see that it is a special case of model (3.1), note that $E(Y)$, the overall average, is just an average of all the X_1-specific averages $E(Y|X_1 = x_1)$. Hence, if all the X_1-specific averages equal α, as in model (3.4), then the overall average must equal α as well, as in model (3.1).

The following two models are equivalent to model (3.4):

$$R(x_1) = \exp(\alpha) , \tag{3.5}$$

which can be rewritten

$$\ln [R(x_1)] = \alpha ,$$

and

$$R(x_1) = \text{expit}(\alpha) = e^{\alpha}/(1 + e^{\alpha}) , \tag{3.6}$$

which can be rewritten

$$\text{logit} [R(x_1)] = \alpha .$$

In model (3.5), α is the common value of the log risks $\ln[R(x_1)]$, while in model (3.6), a is the common value of the logits, $\text{logit}[R(x_1)]$. Each of the equivalent models (3.4)–(3.6) is a special case of the more general models (3.1)–(3.3).

A constant regression is of course implausible in most situations. For example, age is related to most health outcomes. In the above example, we should expect the average death risk to vary across the subgroups defined by age at start (X_2). There is an infinitude of ways to model these variations. The problem of selecting a useful model from among the many choices is discussed below. For now, we only describe some of the more common choices, focusing on models for average risks (incidence proportions), incidence odds, and person-time incidence rates. The models for risks and odds can also be used to model prevalence proportions and prevalence odds.

Linear Risk Models 3.3.5

Consider the model

$$R(x_1) = \alpha + \beta_1 x_1 . \tag{3.7}$$

This model allows the average risk to vary across subpopulations with different values for X_1, but only in a linear fashion. The model implies that subtracting the average risk in the subpopulation with $X_1 = x_1$ from that in the subpopulation with $X_1 = x_1 + 1$ will always yield β_1, *regardless* of what x_1 is. Under model (3.7),

$$R(x_1 + 1) = a + \beta_1 (x_1 + 1)$$

and

$$R(x_1) = \alpha + \beta_1 x_1 ,$$

so

$$R(x_1 + 1) - R(x_1) = \beta_1 .$$

Thus, in our example, β_1 represents the difference in risk between the subpopulation defined by having $X_1 = x_1 + 1$ and that defined by having $X_1 = x_1$. The model implies that this difference does not depend on the reference level x_1 for X_1, used for the comparison.

Model (3.7) is an example of a *linear* risk model. It is a special case of model (3.1); it also contains model (3.4) as a special case (model (3.4) is the special case of model (3.7) in which $\beta_1 = 0$ and so average risks do not vary across levels of X_1). Linear risk models (such as model (3.7)) are easy to understand, but have a severe technical problem that makes them difficult to fit in practice: There are combinations of α and β_1 that would produce impossible values (less than 0 or greater than 1) for one or more of the risks $R(x_1)$. Several models partially or wholly address this problem by transforming the linear term $\alpha + \beta_1 x_1$ before equating it to the risk. We will study two of these models below.

3.3.6 Recentering

Under model (3.7),

$$R(0) = \alpha + \beta \times 0 = \alpha \ ,$$

so α represents the average risk for the subpopulation with $X_1 = 0$. In the present example, 0 is a possible value for X_1 (tamoxifen) and so this interpretation of α presents no problem. Suppose, however, we modeled X_3 (age at menarche) instead of X_1:

$$R(x_3) = \alpha + \beta_3 x_3 \ .$$

Because age at menarche cannot equal zero, α would have no meaningful interpretation in this model. In order to avoid such interpretational problems, it is a useful practice to recenter a variable for which zero is impossible (such as X_3) by subtracting some frequently observed value from it before putting it in the model. For example, age 13 is a frequently observed value for age at menarche. We can redefine X_3 to be "age at menarche minus 13 years". With this redefinition, $R(x_3) = \alpha + \beta_3 x_3$ refers to a different model, one in which $R(0) = \alpha$ represents the average risk for women who were age 13 at menarche. We will later see that such recentering is advisable when using any model, and is especially important when product terms ("interactions") are used in a model.

3.3.7 Rescaling

A simple way of describing β_1 in model (3.7) is that it is the difference in risk per unit increase in X_1. Often the units used to measure X_1 are small relative to exposure increases of substantive interest. Suppose, for example, that X_1 was diastolic blood pressure (DBP) measured in mm Hg; β_1 would then be the risk difference per mm increase in DBP. A 1 mm Hg increase would, however, be of no clinical interest; instead, we would want to consider increases of at least 5 and

possibly 10 or 20 mm Hg. Under model (3.7), the difference in risk per 10 mm Hg increase would be $10\beta_1$. If we wanted to have β_1 represent the difference in risk per 10 mm Hg, we need only redefine X_1 as DBP divided by 10; X_1 would then be DBP in cm Hg.

Division of a variable by a constant, as just described, is sometimes called *rescaling* of the variable. Such rescaling is advisable whenever it changes the measurement unit to a more meaningful value. Unfortunately, rescaling is often done in a way that makes the measurement unit *less* meaningful, by dividing the variable by its sample standard deviation (SD). The sample SD is an irregular unit unique to the study data, and depends heavily on how subjects were selected into the analysis. For example, the SD of DBP might be 12.7 mm Hg in one study and 15.3 mm Hg in another study. Suppose each study divided DBP by its SD entering it in model (3.7). In the first study β_1 would refer to the change in risk per 12.7 mm Hg increase in DBP, whereas in the second study β_1 would refer to the change in risk per 15.3 mm Hg. The rescaling would thus have rendered the coefficients interpretable only in peculiar and different units, so that they could not be compared directly to one another or to coefficients from other studies.

We will later see that rescaling is even more important when product terms are used in a model. We thus recommend that rescaling be done using simple and easily interpreted constants for the divisions. Methods that involve division by sample SDs (such as transformations of variables to Z-scores), however, should be avoided.

Exponential Risk Models

3.3.8

Consider the following model:

$$R(x_1) = \exp\left(\alpha + \beta_1 x_1\right) \ . \tag{3.8}$$

Since the exponential function (exp) is always positive, model (3.8) will produce positive $R(x_1)$ for any combination of $\alpha + \beta_1$. Model (3.8) is sometimes called an *exponential* risk model. It is a special case of the vacuous model (3.2); it also contains the constant model (3.5) as the special case in which $\beta_1 = 0$.

To understand the implications of the exponential risk model, we can recast it in an equivalent form by taking the natural logarithm of both sides:

$$\ln\left[R\left(x_1\right)\right] = \ln\left[\exp\left(\alpha + \beta_1 x_1\right)\right] = \alpha + \beta_1 x_1 \ . \tag{3.9}$$

Model (3.9) is often called a *log-linear* risk model. The exponential/log-linear model allows risk to vary across subpopulations defined by X_1, but only in an exponential fashion. To interpret the coefficients, we may compare the log risks under model (3.9) for the two subpopulations defined by $X_1 = x_1 + 1$ and $X_1 = x_1$:

$$\ln\left[R\left(x_1 + 1\right)\right] = \alpha + \beta_1\left(x_1 + 1\right)$$

and

$$\ln\left[R\left(x_1\right)\right] = \alpha + \beta_1 x_1 \ ,$$

so

$$\ln\left[R\left(x_1+1\right)\right]-\ln\left[R\left(x_1\right)\right]=\ln\left[R\left(x_1+1\right)/R\left(x_1\right)\right]=\beta_1\ .$$

Thus, under models (3.8) and (3.9), β_1 represents the log risk ratio comparing the subpopulation defined by having $X_1 = x_1+1$ and that defined by $X_1 = x_1$, regardless of the chosen reference level x_1. Also, $\ln[R(0)] = \alpha + \beta \times 0 = \alpha$ if $X_1 = 0$; thus, α represents the log risk for the subpopulation with $X_1 = 0$ (and so is meaningful only if X_1 can be zero).

We can derive another (equivalent) interpretation of the parameters in the exponential risk model by noting that

$$R\left(x_1+1\right)=\exp\left[\alpha+\beta_1\left(x_1+1\right)\right]$$

and

$$R\left(x_1\right)=\exp\left(\alpha+\beta_1 x_1\right)$$

so

$$R\left(x_1+1\right)/R\left(x_1\right)=\exp\left[\alpha+\beta_1\left(x_1+1\right)-\left(\alpha+\beta_1 x_1\right)\right]=\exp\left(\beta_1\right)\ .$$

Thus, under models (3.8) and (3.9), β_1 represents the ratio of risks between the sub-populations defined by $X_1 = x_1+1$ and $X_1 = x_1$, and this ratio does not depend on the reference level x_1 (because x_1 does not appear in the final expression for the risk ratio). Also, $R(0) = \exp(\alpha + \beta \times 0) = e^{\alpha}$, so e^{α} represents the average risk for the subpopulation with $X_1 = 0$.

As with linear risk models, exponential risk models have the technical problem that some combinations of α and β_1 will yield risk values greater than 1, which are impossible. This will not be a practical concern, however, if all the fitted risks and their confidence limits fall well below 1.

3.3.9 Logistic Models

Neither linear nor exponential risk models can be used to analyze case-control data if no external information is available to allow estimation of risks in the source population, whereas the following model can be used without such information:

$$R\left(x_1\right)=\text{expit}\left(\alpha+\beta_1 x_1\right)=\frac{\exp\left(\alpha+\beta_1 x_1\right)}{1+\exp\left(\alpha+\beta_1 x_1\right)}\ . \tag{3.10}$$

This model is called a *logistic* risk model, after the logistic function (expit) in the core of its definition. Because the range of the logistic function is between 0 and 1, the model will only produce risks between 0 and 1, regardless of the values for α, β_1, and x_1. The logistic model is perhaps the most commonly used model in epidemiology, so we examine it in some detail. Model (3.10) is a special case of model (3.3), but unlike model (3.3) it is not vacuous because it constrains the

X_1-specific risks to follow a particular (logistic) pattern. The constant model (3.6) is the special case of the logistic model in which $\beta_1 = 0$.

To understand the implications of the logistic model, it is helpful to recast it as a model for the odds. First, note that, under the logistic model (3.10),

$$1 - R\left(x_1\right) = 1 - \frac{\exp\left(\alpha + \beta_1 x_1\right)}{1 + \exp\left(\alpha + \beta_1 x_1\right)} = \frac{1}{1 + \exp\left(\alpha + \beta_1 x_1\right)} .$$

Since $R(x_1)/[1 - R(x_1)]$ is the odds, we divide each side of (3.10) by the last term and find that, under the logistic model, the odds of disease $O(x_1)$ when $X_1 = x_1$ is

$$O\left(x_1\right) = \frac{R\left(x_1\right)}{1 - R\left(x_1\right)} = \frac{\dfrac{\exp\left(\alpha + \beta_1 x_1\right)}{1 + \exp\left(\alpha + \beta_1 x_1\right)}}{\dfrac{1}{1 + \exp\left(\alpha + \beta_1 x_1\right)}} = \exp\left(\alpha + \beta_1 x_1\right) . \qquad (3.11)$$

This equation shows that the logistic risk model is equivalent to an exponential *odds* model.

Taking logarithms of both sides of (3.11), we see that the logistic model is also equivalent to the log-linear odds model

$$\ln\left[O\left(x_1\right)\right] = \alpha + \beta_1 x_1 . \qquad (3.12)$$

Recall that the logit of risk is defined as the log odds:

$$\text{logit}\left[R\left(x_1\right)\right] = \ln\left[R\left(x_1\right)/\left(1 - R(x_1)\right)\right] = \ln\left[O\left(x_1\right)\right] .$$

Hence, from (3.12), the logistic model can be rewritten in one more equivalent form,

$$\text{logit}\left[R\left(x_1\right)\right] = \alpha + \beta_1 x_1 . \qquad (3.13)$$

This equivalent of the logistic model is often called the logit-linear risk model, or *logit model*.

As a general caution regarding terms, note that "log-linear model" can refer to any of several different models, depending on the context: In addition to the log-linear risk model (3.9) and the log-linear *odds* model (3.12) given above, there are also log-linear *rate* models and log-linear *incidence-time* models, which will be described below.

We can derive two equivalent interpretations of the logistic model parameters. First,

$$\ln\left[O\left(x_1 + 1\right)\right] = \alpha + \beta\left(x_1 + 1\right) ,$$

$$\ln\left[O\left(x_1\right)\right] = \alpha + \beta_1 x_1 ,$$

so

$$\ln\left[O\left(x_1 + 1\right)\right] - \ln\left[O\left(x_1\right)\right] = \ln\left[O\left(x_1 + 1\right)/O\left(x_1\right)\right] = \beta_1 .$$

Thus, under the logistic model (3.10), β_1 represents the log odds ratio comparing the subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$. Also, $\ln[O(0)] = \alpha + \beta_1 \times 0 = \alpha$; thus, α is the log odds (logit) for the subpopulation with $X_1 = 0$ (and so is meaningful only if X_1 can be zero). Equivalently, we have

$$O\left(x_1 + 1\right) / O\left(x_1\right) = \exp\left(\beta_1\right)$$

and

$$O(0) = \exp(\alpha) ,$$

so that $\exp(\beta_1)$ is the odds ratio comparing the subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$, and $\exp(\alpha)$ is the odds for the subpopulation with $X_1 = 0$.

Logistic models may be applied to case-control studies by re-interpreting the odds $O(\boldsymbol{x})$ as the case-control ratio in the study; see Breslow and Day (1980, Chap. 6) or Rothman and Greenland (1998, pp 416–422) for details. For an introduction to case-control studies we refer to Chap. I.6 of this handbook and Chap. 7 of Rothman and Greenland (1998).

3.3.10 A Graphical Example

Suppose a particular cohort has a 1-year risk of a cardiovascular event that is 0.02 at age 50 rising to 0.32 at age 80, an absolute risk increase of 0.30, a ratio risk increase of 0.32/0.02 = 16-fold, and a ratio odds increase of (0.32/0.68)/(0.02/0.98) = 23.06. The average annual absolute risk increase is 0.30/30 = 0.01, but the way this increase is distributed over ages could be quite different under different models.

If the risk increase is linear in age and x is age, the linear model for the risk from age 51 to 80 would be $R(x) = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$ we get $\alpha_1 = 0.02$ and $\beta_1 = 0.30/30$ year $= 0.01/$ year, a constant absolute increase in risk of 0.01 for each of age.

Now suppose the increase is exponential rather than linear. The loglinear form of the exponential model would be $\ln[R(x)] = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$ we now get $\alpha_1 = \ln(0.02) = -3.912$ and $\beta_1 = \ln(16)/30$ year $= 0.09242/$ year, corresponding to a constant proportionate risk increase of $e^{0.09242} = 1.097$ or about 9.7% for each year of age. This corresponds to an absolute risk increase of only about 0.002 going from age 50 to 51, but of about 0.03 (15 times more) going from age 79 to 80.

Finally, suppose the increase is logistic. The logit version of the logistic model would be $\text{logit}[R(x)] = \alpha_1 + \beta_1(x - 50)$. Solving $R(50) = 0.02$ and $R(80) = 0.32$ we now get $\alpha_1 = \text{logit}(0.02) = -3.892$ and $\beta_1 = \ln(23.06)/30$ years $= 0.1046/$ years, corresponding to a constant proportionate odds increase of $e^{0.1046} = 1.11$ or about 11% for each year of age. This corresponds to an absolute risk increase of only about 0.002 going from age 50 to 51, but of about 0.022 (11 times more) going from age 79 to 80.

Figure 3.1a gives plots of the risks from the above three models from age 50 to 80. The linear model produces a straight line, whereas the exponential model produces

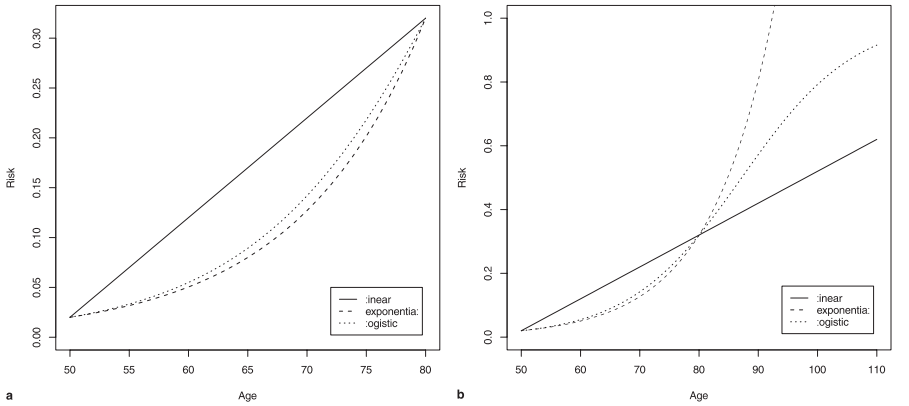

Figure 3.1. (a) Risks from linear, exponential and logistic model from age 50 to age 80 with a 1-year risk of 0.02 at age 50 and 0.32 at age 80; (b) risks from linear, exponential and logistic model extrapolated to age 110 with a 1-year risk of 0.02 at age 50 and 0.32 at age 80

an exponential curve; these shapes will always hold when x is not transformed. The logistic curve is between the two, but is much closer in shape to the exponential for risks below 0.25, and almost the same as the exponential for risks below 10%. As shown in Fig. 3.1b as a projection of the above example, the logistic curve gradually straightens out and is close to linear for risks between 40% and 60%; above that point it begins to level off, becoming nearly flat (horizontal) as it approaches 1. In contrast, the linear and exponential curves will eventually continue on above 1, and so produce impossible values for risks (which is a problem if the actual risks could get large). For negative β_1 the curves would instead go downward from left to right.

Other Risk and Odds Models

3.3.11

In addition to those given above, several other risk models are occasionally mentioned but rarely used in epidemiology. The linear odds model is obtained by replacing the average risk by the odds in the linear risk model:

$$O(x_1) = \alpha + \beta_1 x_1 \,. \tag{3.14}$$

Here, β_1 is the *odds* difference between subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$, and α is the odds for the subpopulation with $X_1 = 0$. Like risk, the odds cannot be negative; unfortunately, some combinations of α and β_1 in model (3.14) will produce negative odds. As a result, this model (like the linear risk model) is difficult to fit and gives unsatisfactory results in many settings.

Another model replaces the logistic transform (expit) in the logistic model (3.10) by the inverse of the standard normal distribution, which also has a range between 0 and 1. The resulting model, called a *probit* model, has seen much use in bioassay. Its absence from epidemiologic use may stem from the fact that (unlike

the logistic) its parameters have no simple epidemiologic interpretation, and the model appears to have no general advantage over the logistic in epidemiologic applications.

Finally, several attempts have been made to use models that are mixtures of different basic models, especially for multiple regressions (discussed below). These mixtures have various drawbacks, including difficulties in fitting the models and interpreting the parameters (Moolgavkar and Venzon 1987). We thus do not discuss them here.

3.3.12 Rate Models

Instead of modeling average risks, we could model person-time incidence rates. If we let Y denote the *rate* observed in a study subpopulation (so that Y is the observed number of cases per unit of observed person-time), the regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$ represents the average number of cases per unit of person-time in the target subpopulation defined by $\boldsymbol{X} = \boldsymbol{x}$. We will denote this expected rate or "average rate" by $I(\boldsymbol{x})$.

Most rate models are analogues of risk and odds models. For example, the model

$$I(x_1) = E(Y|X_1 = x_1) = \alpha + \beta_1 x_1 \tag{3.15}$$

is a linear rate model, analogous to (but different from) the linear risk and odds models (3.7), (3.14). This rate model implies that the difference in average rates between subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$ is β_1, regardless of x_1. Also, α is the average rate for the subpopulation with $X_1 = 0$. This model can be problematic, because some combinations of α and β_1 in model (3.15) would produce negative rate values, which are impossible.

To prevent the latter problem, most rate modeling begins with an exponential *rate* model such as

$$I(x_1) = \exp(\alpha + \beta_1 x_1)\,. \tag{3.16}$$

Because the exponential (exp) can never be negative, this model will not produce negative rates, regardless of α, β_1, or x_1. The model is equivalent to the log-linear *rate* model

$$\ln[I(x_1)] = \alpha + \beta_1 x_1\,. \tag{3.17}$$

The parameter β_1 in models (3.16) and (3.17) is the log of the rate ratio comparing the subpopulation with $X_1 = x_1 + 1$ to the subpopulation with $X_1 = x_1$, regardless of x_1; hence, $\exp(\beta_1)$ is the corresponding rate ratio $I(x_1 + 1)/I(x_1)$. Also, α is the log of the rate for the subpopulation with $X_1 = 0$; hence, $\exp(\alpha)$ is the average rate $I(0)$ when $X_1 = 0$. The exponential rate model (3.16) is analogous to, but different from, the exponential risk model (3.8) and the exponential odds model (3.11).

Incidence-Time and Hazard Models

We can also model the average time to occurrence of an event, starting from some designated zero time such as birth (in which case "time" is age), start of treatment, or some calendar date. These are called incidence-time, waiting-time, failure-time, or survival-time models (cf. Chap II.4 of this handbook). Let T stand for time of the event measured from zero. One approach to incidence time regression is to use a linear model for log incidence time, such as

$$E[\ln(T)|X_1 = x_1] = \alpha - \beta_1 x_1 . \tag{3.18}$$

Because T is always positive, $\ln(T)$ is always defined. In this model, α is the average log incidence time in the subpopulation with $X_1 = 0$, and $-\beta_1$ is the difference in average log incidence times when comparing the subpopulation with $X_1 = x_1 + 1$ to the subpopulation with $X_1 = x_1$ (regardless of the value x_1). Model (3.18) is a generalization of the basic *accelerated-life* model (Cox and Oakes 1984).

Note that the sign of β_1 in the model is reversed from its sign in earlier models. This reversal is done so that, if the outcome event at T is undesirable, then as in earlier models positive values of β_1 will correspond to harmful effects from increasing X_1, and negative values will correspond to beneficial effects. For example, under the model, if T is death time and β_1 is positive, an increase in X_1 will be associated with earlier death.

Another generalization of the basic accelerated-life model, similar but not identical to model (3.18), is the log-linear model for expected incidence time

$$\ln[E(T|X_1 = x_1)] = \alpha - \beta_1 x_1 . \tag{3.19}$$

Model (3.19) differs from model (3.18) because the log of an average is greater than the average of the logs (unless T does not vary). Model (3.19) can be rewritten

$$E(T|X_1 = x_1) = \exp(\alpha - \beta_1 x_1) = \exp(-\beta_1 x_1)\, e^{\alpha} ,$$

$$= \exp(-\beta_1 x_1) T_0 ,$$

where $T_0 = E(T|X_1 = 0) = e^{\alpha}$. Under model (3.19) e^{α} is the average incidence time in the subpopulation with $X_1 = 0$, and $e^{-\beta_1}$ is the ratio of average incidence times in the subpopulation with $X_1 = x_1 + 1$ and the subpopulation with $X_1 = x_1$. As with model (3.18) the sign of β_1 is negative so that positive values of β_1 will correspond to harmful effects.

More common approaches to modeling incidence times impose a model for the risk of the event up to each point in time, or for the rate of the event at each point in time. The most famous such model is the *Cox model*, also known as the *proportional hazards model*. We can give an approximate description of this model as follows: Suppose we specify a time span Δt that is small enough so that the risk of having the event in any interval t to $t + \Delta t$ among those who survive to t without the event is very small. The Cox model then implies that the rates in any such short

interval will follow an exponential model like (3.16) with α but not β_1 allowed to vary with time t.

If we write $I(t; x_1)$ for the average rate in the interval t to $t + \Delta t$ among persons who survive to t and have $X_1 = x_1$, the Cox model implies that

$$I(t; x_1) \approx \exp(\alpha_t + \beta_1 x_1) . \tag{3.20}$$

Under the model, the approximation ($\approx$) improves as Δt gets smaller. Note that the intercept a_t may vary with time, but in this simple Cox model the X_1-coefficient β_1 is assumed to remain constant. This means that, at any time t, the rate ratio comparing subpopulations with $X_1 = x_1 + 1$ and $X_1 = x_1$ will be

$$I(t; x_1 + 1)/I(t; x_1) \approx \exp[\alpha_t + \beta_1(x_1 + 1)]/\exp(\alpha_t + \beta_1 x_1) = \exp(\beta_1) ,$$

so that β_1 is the log rate ratio per unit of X_1, regardless of either the reference level x_1 *or* the time t at which it is computed.

Under the Cox model (3.20) the rate at time t for the subpopulation with $X_1 = 0$ is given by $I(t; 0) = \exp(\alpha_t)$. If we denote this "baseline" rate by $\lambda_0(t)$ instead of $\exp(\alpha_t)$, we have

$$I(t; x_1) \approx \exp(\alpha_t + \beta_1 x_1) = \exp(\alpha_t + \beta_1 x_1) = \lambda_0(t) \exp(\beta_1 x_1) = \exp(\beta_1 x_1)\lambda_0(t) .$$

The last expression is the standard form of the model given in most textbooks. The term "Cox model" has become fairly standard, although a special case of the model was proposed by Sheehe (1962) some 10 years before Cox (1972).

The approximate form of the Cox model (3.20) may be seen as an extension of the exponential rate model (3.16) in which the rates may vary over time. In statistical theory, the assumption is made that, at each time t, the rate $I(t; x_1)$ approaches a limit $\lambda(t; x_1)$ as Δt goes to zero. This limit is usually called the *hazard* or *intensity* of the outcome at time t. The Cox model is then defined as a model for these hazards,

$$\lambda(t; x_1) = \exp(\beta_1 x_1)\lambda_0(t) .$$

In epidemiologic studies, these hazards are purely theoretical quantities; thus, it is important to understand the approximate forms of the model given above and what those forms imply about observable rates.

The Cox model may be extended to allow X_1 to vary over time. Let us write $X_1(t)$ as an abbreviation for "the exposure as of time t" and $x_1(t)$ for the actual numerical value of $X_1(t)$ at time t. Then the *Cox model with time-dependent covariates* implies that the incidence rate at time t in the subpopulation that has exposure level $x_1(t)$ at time t is

$$I[t; x_1(t)] \approx \exp[\beta_1 x_1(t)]\lambda_0(t) . \tag{3.21}$$

This model may be the most widely used model for time-dependent exposures. Usually, a time-dependent exposure $X_1(t)$ is not defined as the actual amount at time t, but instead is some cumulative and lagged index of expo-

sure up to t. For example, if time is measured in months and exposure is cumulative tamoxifen lagged 3 months, $X_1(t)$ would mean "cumulative amount of tamoxifen taken up to month $t - 3$" and $x_1(t)$ would be a value for this variable.

There are biases that can arise in use of Cox models to estimate effects of time-dependent exposures. These biases and alternative models are described in Robins et al. (1992) and Robins and Greenland (1994).

Trend Models: Univariate Exposure Transforms 3.3.14

Consider again the linear risk model (3.7). If this model were correct, a plot of average risk across the subpopulations defined by X_1 (that is, a plot of risk against X_1) would yield a line. Ordinarily, however, there is no compelling reason to think the model is correct, and we might wish to entertain other possible models for the trend in risk across exposure levels. We can generate an unlimited variety of such models by *transforming* exposure, that is, by replacing X_1 in the model by some function of X_1.

To illustrate, we could replace years exposed in model (3.7) by its logarithm, to get

$$R(x_1) = \alpha + \beta_1 \ln(x_1)\,. \tag{3.22}$$

This is still called a linear risk model, because a plot of average risk against the new regressor $\ln(X_1)$ would yield a line. But it is a very different model from model (3.7) because if model (3.22) were correct, a plot of average risk against years exposed (X_1) would yield a *logarithmic curve* rather than a line. Such a curve starts off very steep for $X_1 < 1$, but levels off rapidly beyond $X_1 > 1$. One technical problem can arise in using the logarithmic transform: It is not defined if X_1 is negative or zero. If the original exposure measurement can be negative or zero, it is common practice to add a number c to X_1 that is big enough to insure $X_1 + c$ is always positive. The resulting model is

$$R(x_1) = \alpha + \beta_1 \ln(x_1 + c)\,. \tag{3.23}$$

The shape of the curve represented by this model (and hence results derived using the model) can be very sensitive to the value chosen for c, especially when the values of X_1 may be less than 1. Frequently, c is set equal to 1, although there is usually no compelling reason for this choice.

Among other possibilities for exposure transforms are simple power curves of the form

$$R(x_1) = \alpha + \beta_1 x_1^p\,, \tag{3.24}$$

where p is some number (typically 1/2 or 2) chosen in advance according to some desired property. For example, with X_1 as years exposed, use of $p = 1/2$ yields the *square-root* model

$$R(x_1) = \alpha + \beta_1 x_1^{1/2},$$

which produces a trend curve that levels off as X_1 increases above zero. In contrast, use of $p = 2$ yields the simple *quadratic* model

$$R(x_1) = \alpha + \beta_1 x_1^2,$$

which produces a trend that rises more and more steeply as X_1 increases above zero. One technical problem can arise when using the power model (3.24). It is not defined if p is fractional and X_1 can be negative. To get around this limitation, we may add some number c to X_1 that is big enough to insure $X_1 + c$ is never negative, and then use $(X_1 + c)^p$ in the model; again, however, the result may be sensitive to choice of c.

The trend implications of linear and exponential models are vastly different, and hence the implications of exposure transforms are also different. Consider again the exponential risk model (3.8). If this model were correct, a plot of average risk against X_1 would yield an exponential curve, rather than a line. If β_1 is positive, this curve starts out slowly but rises more and more rapidly as X_1 increases; it eventually rises more rapidly than does any power curve (3.24). Such rapid increase is often implausible and we might wish to use a slower-rising curve to model risk.

One means of moderating the trend implied by an exponential model is to replace x_1 by a fixed power x_1^p with $0 < p < 1$, for example

$$R(x_1) = \exp\left(\alpha + \beta_1 x_1^{1/2}\right).$$

Another approach is to take the logarithm of exposure. This transform produces a new model:

$$\begin{aligned} R(x_1) &= \exp[\alpha + \beta_1 \ln(x_1)] = \exp(\alpha)\exp[\beta_1 \ln(x_1)] \\ &= e^{\alpha}\exp[\ln(x_1)]^{\beta_1} = e^{\alpha} x_1^{\beta_1}. \end{aligned} \tag{3.25}$$

A graph of risk against exposure under this model produces a power curve, but now (unlike (3.24)), the power is the unspecified (unknown) coefficient β_1 instead of a prespecified value p, and the multiplier of the exposure power is e^{α} (which must be positive) instead of β_1. Model (3.25) might thus appear more appropriate than model (3.24) when we want the power of X_1 to appear as an unknown coefficient β_1 in the model, rather than as a pre-specified value p. As earlier, however, X_1 must always be positive in order to use model (3.25) otherwise, one must add a constant c to it such that $X_1 + c$ is always positive.

When β_1 is negative in model (3.25) risk declines more and more gradually across increasingly exposed subpopulations. For example, if $\beta_1 = -1$, then under model (3.25) $R(x_1) = e^{\alpha} x_1^{-1} = e^{\alpha}/x_1$, which would imply risk declines 50% (from

$e^{\alpha}/1$ to $e^{\alpha}/2$) when going from $X_1 = 1$ to $X_1 = 2$, but declines less than 10% (from $e^{\alpha}/10$ to $e^{\alpha}/11$) when going from $X_1 = 10$ to $X_1 = 11$.

The exposure transforms and implications just discussed carry over to the analogous models for odds and rates. For example, we can modify the logistic model (which is an exponential odds model) by substituting the odds $O(x_1)$ for the risk $R(x_1)$ in models (3.22) to (3.25). Similarly, we can modify the rate models by substituting the rate $I(x_1)$ for $R(x_1)$. Each model will have implications for the odds or rates analogous to those described above for the risk; because the risks, odds, and rates are functions of one another (see Rothman and Greenland 1998, Chap. 3), each model will have implications for other measures as well.

Any trend in the odds will appear more gradual when transformed into a risk trend. To see this, note that

$$R(x_1) = O(x_1)/[1 + O(x_1)] < O(x_1)\,,$$

and hence

$$O(x_1)/R(x_1) = 1 + O(x_1)\,.$$

This ratio of odds to risk grows as the odds (and the risks) get larger. Thus, the logistic risk model, which is an exponential odds model, implies a less-than-exponential trend in the risk. Conversely, any trend in the risks will appear steeper when transformed into an odds trend. Thus, the exponential risk model implies a greater-than-exponential trend in the odds, although when risks are uniformly low (under 10% for all possible X_1 values), the risks and odds will be close and so there will be little difference between the shape of the curves produced by analogous risk and odds models.

The relation of risk and odds trends to rate trends is more complex in general, but in typical applications follows the simple rule that rate trends tend to fall between the less steep risk and more steep odds trends. For example, an exponential rate model typically implies a less than exponential risk trend but more than exponential odds trend. To see why these relations can be reasonable to expect, recall that, if incidence is measured over a span of time Δt in a closed cohort, then $R(x_1) < I(x_1)\Delta t < O(x_1)$. When the risks are uniformly low, we obtain $R(x_1) \doteq I(x_1)\Delta t \doteq O(x_1)$ (see Rothman and Greenland 1998, Chap. 3), and so there will be little difference in the curves produced by analogous risk, rate, and odds models.

Interpreting Models After Transformation

3.3.15

One drawback of models with transformed regressors is that the interpretation of the coefficients depends on the transformation. As an example, consider the model (3.25) which has $\ln(x_1)$ in place of x_1. Under this model, the risk ratio for a one-unit increase in X_1 is

$$R(x_1 + 1)/R(x_1) = e^{\alpha}(x_1 + 1)^{\beta_1}/e^{\alpha}(x_1)^{\beta_1} = [(x_1 + 1)/x_1]^{\beta_1}\,.$$

which will depend on the value x_1 used as the reference level: If β_1 equals 1 and x_1 is 1, the risk ratio is 2, but if β_1 equals 1 and x_1 is 2, the ratio is 1.5. Here, β_1 is the power to which x_1 is raised, and so determines the shape of the trend. The interpretation of the intercept α is also altered by the transformation. Under model (3.25), $R(1) = e^{\alpha}1^{\beta_1} = e^{\alpha}$, thus, α is the log risk when $X_1 = 1$, rather than when $X_1 = 0$, and so is meaningful only if 1 is a possible value for X_1.

As a contrast, consider again the model $R(x_1) = \exp\left(\alpha + \beta_1 x_1^{1/2}\right)$. Use of $x_1^{1/2}$ rather than x_1 moderates the rapid increase in the slope of the exponential dose-response curve, but also leads to difficulties in coefficient interpretation. Under the model, the risk ratio for a one-unit increase in X_1 is

$$\exp\left[\alpha + \beta_1(x_1 + 1)^{1/2}\right] \Big/ \exp\left(\alpha + \beta_1 x_1^{1/2}\right) = \exp\left\{\beta_1\left[(x_1 + 1)^{1/2} - x_1^{1/2}\right]\right\} .$$

Here, β_1 is the log risk ratio per unit increase in the *square root* of X_1, which is rather obscure in meaning. Interpretation may better proceed by considering the shape of the curve implied by the model, for example, by plotting $\exp\left(\alpha + \beta_1 x_1^{1/2}\right)$ against possible values of X_1 for several values of β_1. (The intercept α is less important in this model, because it only determines the vertical scale of the curve, rather than its shape.) Such plotting is often needed to understand and compare different transforms.

3.4 Multiple Regression Models

Suppose now we wish to model the full multiple regression $E(Y|\boldsymbol{X} = \boldsymbol{x})$. Each of the previous models for the single regression $E(Y|X_1 = x_1)$ can be extended to handle this more general situation by using the following device: In any model for the single regression, replace $\beta_1 x_1$ by

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n . \tag{3.26}$$

To illustrate the idea, suppose we wish to model average risk of death by age 70 across female subpopulations defined by

$X_1 =$ years of tamoxifen therapy,

$X_2 =$ age at start of tamoxifen use, and

$X_3 =$ age at menarche,

with $\boldsymbol{X} = (X_1, X_2, X_3)$. Then the multiple linear risk model for $R(\boldsymbol{x})$ is

$$R(\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ,$$

while the multiple logistic risk model is

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\,.$$

If instead we wished to model the death rate, we could use the multiple linear rate model

$$I(\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

or a multiple exponential rate model

$$I(\boldsymbol{x}) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\,.$$

Because (3.26) can be clumsy to write out when there are three or more regressors ($n \geq 3$), several shorthand notations are in use. Let us write $\boldsymbol{\beta}$ for the vertical list (column vector) of coefficients $\beta_1, \ldots, \beta_n$. Recall that $\boldsymbol{x}$ stands for the horizontal list (row vector) of values $x_1, \ldots, x_n$. We will let $\boldsymbol{x\beta}$ stand for $\beta_1 x_1 + \ldots + \beta_n x_n$. We can then represent the multiple linear risk model by

$$R(\boldsymbol{x}) = \alpha + \boldsymbol{x\beta} = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n\,, \tag{3.27}$$

the multiple logistic model by

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \boldsymbol{x\beta})\,, \tag{3.28}$$

the multiple exponential rate model by

$$I(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x\beta})\,, \tag{3.29}$$

and so on for all the models discussed earlier.

Relations Among Multiple-Regression Models 3.4.1

The multiple-regression models (3.27)–(3.29) are not more general than the single-regression models given earlier, nor do they contain those models as special cases. This is because they refer to entirely different subclassifications of the target population: The single-regression models refer to variations in averages across subpopulations defined by levels of just one variable; in contrast, the multiple-regression models refer to variations across the much finer subdivisions defined by the levels of several variables. For example, it is possible for $R(x_1)$ to follow the single-logistic model (3.10) without $R(\boldsymbol{x})$ following the multiple-logistic model (3.28) conversely, it is possible for $R(\boldsymbol{x})$ to follow the multiple-logistic model without $R(x_1)$ following the single-logistic model.

The preceding point is often overlooked because the single-regression models are often confused with multiple-regression models in which all regressor coefficients but one are zero. The difference is, however, analogous to the differences discussed earlier between the vacuous models (3.1)–(3.3) (which are so general as to imply nothing) and the constant regression models (3.4)–(3.6) (which are so

restrictive as to be unbelievable in typical situations). To see this, consider the multiple-logistic model

$$R(\boldsymbol{x}) = \text{expit}(\alpha + \beta_1 x_1)\,. \tag{3.30}$$

The right side of this equation is the same as in the single-logistic model (3.10) but the left side is crucially different: It is the multiple-risk regression $R(\boldsymbol{x})$, instead of the single-regression $R(x_1)$. Unlike model (3.10) model (3.30) *is* a special case of the multiple-logistic model (3.28) the one in which $\beta_2 = \beta_3 = \ldots = \beta_n = 0$. Unlike model (3.10) model (3.30) asserts that risk *does not vary* across subpopulations defined by $X_1, X_2, \ldots, X_n$ *except* to the extent that X_1 varies. This is far more strict than model (3.28) which allows risk to vary with $X_2, \ldots, X_n$ as well as X_1 (albeit only in a logistic fashion). It is also far more strict than model (3.10) which says absolutely nothing about whether or how risk varies across subpopulations defined by $X_2, \ldots, X_n$ within specific levels of X_1.

More generally, we must be careful to distinguish between models that refer to different multiple regressions. For example, compare the two exponential rate models:

$$I(x_1, x_2) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2) \tag{3.31}$$

and

$$I(x_1, x_2, x_3) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)\,. \tag{3.32}$$

These are different models. The first is a model for the regression of rates on X_1 and X_2 only, while the second is a model for the regression of rates on X_1, X_2, and X_3. The first model in no way refers to X_3, while the second asserts that rates do not vary across levels of X_3 if one looks within levels of X_1 and X_2. Model (3.32) is the special case of

$$I(x_1, x_2, x_3) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

(the case in which $\beta_3 = 0$), while model (3.31) is not, and this special case implies model (3.31).

Many textbooks and software manuals fail to distinguish between models such as models (3.31) and (3.32), and instead focus only on the appearance of the right-hand side of the models. Most software fits the model that ignores other covariates ((3.31) in the above example) rather than the more restrictive model (3.32) when requested to fit a model with only X_1 and X_2 as regressors. Note that if the less restrictive model is inadequate, then the more restrictive model must also be inadequate.

Unfortunately, if the less restrictive model appears adequate, it does *not* follow that the more restrictive model is also adequate. For example, it is possible for the model form $\exp(\alpha + \beta_1 x_1 + \beta_2 x_2)$ to describe adequately the double regression $I(x_1, x_2)$ (which means it describes adequately rate variation across X_1 and X_2 when X_3 is ignored), and yet at the same time describe poorly the triple regression

$I(x_1, x_2, x_3)$ (which means that it describes inadequately rate variation across X_1, X_2, and X_3). That is, a model may describe poorly the rate variation across X_1, X_2, and X_3 even if it describes adequately the rate variation across X_1 and X_2 when X_3 is ignored. The decision as to whether the model is acceptable should depend on whether rate variation across X_3 is relevant to the analysis objectives. For example, if the objective is to estimate the effect of changes in X_1 on the death rate, and X_2 and X_3 are both potential confounders (as in the tamoxifen example), we would want the model to describe adequately rate variation across all three variables. But if X_3 is instead affected by the study exposure X_1 (as when X_1 is past estrogen exposure and X_3 is an indicator of current uterine bleeding), we would ordinarily not want to include X_3 in the regression model (because we would not want to adjust our exposure effect estimate for X_3).

Product Terms (Statistical Interactions) 3.4.2

Each model form described above has differing implications for measures of association derived from the models. Consider again the linear risk model with three regressors X_1, X_2, and X_3, and let x_1^* and x_1 be any two values for X_1. Under the model, the risks at $X_1 = x_1^*$ and $X_1 = x_1$ and their difference RD when $X_2 = x_2$ and $X_3 = x_3$ are

$$R\left(x_1^*, x_2, x_3\right) = \alpha + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 ,$$

$$R\left(x_1, x_2, x_3\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ,$$

$$RD = \beta_1 \left(x_1^* - x_1\right) .$$

Thus, the model implies that the risk difference between two subpopulations with the same X_2 and X_3 levels depends only on the difference in their X_1 levels. In other words, the model implies that the risk differences for X_1 within levels of X_2 and X_3 will not vary across levels of X_2 and X_3. Such an implication may be unacceptable, in which case we can either modify the linear model or switch to another model. A simple way to modify a model is to add *product terms*. For example, suppose we want to allow the risk differences for X_1 to vary across levels of X_2. We then may add the product of X_1 and X_2 to the model as a fourth variable. The risks and their differences will then be

$$R\left(x_1^*, x_2, x_3\right) = \alpha + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1^* x_2 ,$$

$$R(x_1, x_2, x_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2 , \tag{3.33}$$

$$RD = \beta_1 (x_1^* - x_1) + \gamma_{12} (x_1^* - x_1) x_2 = (\beta_1 + \gamma_{12} x_2)(x_1^* - x_1) . \tag{3.34}$$

Under model (3.33), the risk difference for $X_1 = x_1^*$ versus $X_1 = x_1$ is given by (3.34), which depends on X_2.

A model (e.g., (3.33)), that allows variation of the risk difference for X_1 across levels of X_2 will also allow variation in the risk difference for X_2 across levels of X_1.

As an example, let x_2^* and x_2 be any two possible values for X_2. Under model (3.33) the risks at $X_2 = x_2^*$ and $X_2 = x_2$ and their difference RD when $X_1 = x_1$, $X_3 = x_3$ are

$$R\left(x_1, x_2^*, x_3\right) = \alpha + \beta_1 x_1 + \beta_2 x_2^* + \beta_3 x_3 + \gamma_{12} x_1 x_2^* ,$$

$$R(x_1, x_2, x_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{12} x_1 x_2 ,$$

$$RD = \beta_2 \left(x_2^* - x_2\right) + \gamma_{12} x_1 \left(x_2^* - x_2\right) = \left(\beta_2 + \gamma_{12} x_1\right) \left(x_2^* - x_2\right) . \qquad (3.35)$$

Thus, under the model, the risk difference for $X_2 = x_2^*$ versus $X_2 = x_2$ is given by (3.35), which depends on X_1. (3.34) and (3.35) illustrate how product terms modify a model in a symmetric way. The term $\gamma_{12} x_1 x_2$ allows the risk differences for X_1 to vary with X_2 and the risk differences for X_2 to vary with X_1.

If we have three regressors in a model, we have three unique two-way regressor products (x_1x_2, x_1x_3, x_2x_3) that we can put in the model. More generally, with n regressors, there are $\binom{n}{2}$ pairs and hence $\binom{n}{2}$ two-way products we can use. It is also possible to add triple products (e.g., $x_1x_2x_3$) or even more complex combinations to the model, but such additions are rare in practice; notable exceptions are body mass indices, such as kg/m^2 (Michels et al. 1998). A model without product terms is sometimes called a "main-effects only" model, and can be viewed as the special case of a model with product terms (the special case in which all the product coefficients γ_{ij} are zero).

Consider next an exponential-risk model with the above three variables. Under this model, the risks at $X_1 = x_1^*$ and $X_1 = x_1$ and their ratio RR when $X_2 = x_2$, $X_3 = x_3$ are

$$R\left(x_1^*, x_2, x_3\right) = \exp\left(\alpha + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3\right) ,$$

$$R(x_1, x_2, x_3) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) ,$$

$$RR = \exp\left[\beta_1 \left(x_1^* - x_1\right)\right] . \qquad (3.36)$$

Thus, the model implies that the risk ratio comparing two subpopulations with the same X_2 and X_3 levels depends only on the difference in their X_1 levels. In other words, the model implies that the risk ratios for X_1 will be constant across levels of X_2 and X_3. If this implication is unacceptable, product terms can be inserted, as with the linear model. These terms allow the risk ratios to vary in a limited fashion across levels of other variables. The preceding discussion of product terms can be applied to linear and exponential models in which the odds or rate replace the risk. For example, without product terms, the logistic model implies that the odds ratios for each regressor are constant across levels of the other regressors (because the logistic model is an exponential odds model); we can add product terms to allow the odds ratios to vary. Likewise, without product terms, the exponential rate model implies that the rate ratios for each regressor are constant across levels

of the other regressors; we can add product terms to allow the rate ratios to vary.

Although product terms can greatly increase the flexibility of a model, the type of variation allowed by product terms can be very limited. For example, model (3.33) implies that raising X_2 by one unit (i.e., comparing subpopulations that have $X_2 = x_2 + 1$ instead of $X_2 = x_2$) will yield a risk difference for X_1 of

$$[\beta_1 + \gamma_{12}(x_2 + 1)](x_1^* - x_1) = (\beta_1 + \gamma_{12}x_2)(x_1^* - x_1) + \gamma_{12}(x_1^* - x_1)\,.$$

In other words, the model implies that shifting our comparison to subpopulations that are one unit higher in X_2 will change the risk difference for X_1 in a linear fashion, by an amount $\gamma_{12}(x_1^* - x_1)$, regardless of the reference values x_1, x_2, x_3 of X_1, X_2, X_3.

Trends and Product Terms 3.4.3

Each of the above models forces or assumes a particular shape for the graph obtained when average outcome (regression) is plotted against the regressors. Consider again the tamoxifen example. Suppose we wished to plot how the risk varies across subpopulations with different number of years exposed but with the same age at start of exposure and the same age at menarche. Under the linear risk model, this would involve plotting the average risk

$$R(x_1, x_2, x_3) = \alpha + \beta x_1 + \beta_2 x_2 + \beta_3 x_3$$

against X_1, while keeping X_2 and X_3 fixed at some values x_2 and x_3. In doing so, we would obtain a line with an intercept equal to $\alpha + \beta_2 x_2 + \beta_3 x_3$ and a slope equal to β_1. Whenever we changed X_2 and X_3 and replotted $R(\boldsymbol{x})$ against X_1, the intercept would change (unless $\beta_2 = \beta_3 = 0$), but the slope would remain β_1. Because lines with the same slope are parallel, we can say that the linear risk model given above implies *parallel linear* trends in risk with increasing tamoxifen (X_1) as one moves across subpopulations of different starting age (X_2) and menarche age (X_3). This means that each change in X_2 and X_3 adds some constant (possibly negative) amount to the X_1 curve. For this reason, the linear risk model is sometimes called an *additive* risk model.

If we next plotted risks against X_2, we would get analogous results: The linear risk model given above implies parallel linear relations between average risk and X_2 as one moves across levels of X_1 and X_3. Likewise, the model implies parallel linear relations between average risk and X_3 across levels of X_1 and X_2. Thus, the linear model implies additive (parallel) relations among all the variables.

If we are unsatisfied with the linearity assumption but we wish to retain the additivity (parallel-trend) assumption, we could transform the regressors. If we are unsatisfied with the parallel-trend assumption, we can allow the trends to vary across levels of other regressors by adding product terms to the model. For

example, adding the product of X_1 and X_2 to the model yields model (3.33), which can be rewritten

$$R(x_1, x_2, x_3) = \alpha + (\beta_1 + \gamma_{12}x_2)x_1 + \beta_2 x_2 + \beta_3 x_3 .$$

From this reformulation, we see that the slope for the line obtained by plotting average risk against X_1 while keeping X_2, X_3 fixed at x_2, x_3 would be $\beta_1 + \gamma_{12}x_2$. Thus, the slope of the trend in risk across X_1 would vary across levels of X_2 (if $\gamma_{12} \neq 0$), and so the trend lines for X_1 would not be parallel. We also see that γ_{12} is the difference in the X_1-trend slopes between subpopulations with the same X_3-value but one unit apart in their X_2-value.

An entirely different approach to producing nonparallel trends begins with an exponential model. For example, under the exponential risk model (3.36) a plot of average risk against X_1 while keeping X_2 and X_3 fixed at x_2 and x_3 would produce an *exponential* curve rather than a line. This exponential curve would have intercept $\exp(\alpha + \beta_2 x_2 + \beta_3 x_3)$. If, however, we changed the value of X_2 or X_3 and replotted risk against X_1, we would *not* obtain a parallel risk curve. Instead, the new curve would be *proportional* to the old: A change in X_2 or X_3 *multiplies* the entire X_1 curve by some amount. For this reason, the exponential model is sometimes called a *multiplicative risk* model. If we were unsatisfied with this proportionality-of-trends assumption, we could insert product terms into the model, which would allow for certain types of nonproportional trends. Proportional trends in risk appear parallel when plotted on a logarithmic vertical scale; when product terms with nonzero coefficients are present, logarithmic trends appear nonparallel.

Analogous comments and definitions apply if we substitute odds or rates for risks in the above arguments. For example, consider the multiple-logistic model in the exponential-odds form:

$$O(\boldsymbol{x}) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) .$$

A plot of the disease odds $O(\boldsymbol{x})$ against X_1 while keeping X_2 and X_3 fixed would produce an exponential curve; a plot of the log odds (logit) against X_1 while keeping X_2 and X_3 fixed would produce a line. If we changed the value of X_2 or X_3 and replotted the odds against X_1, we would obtain a new curve proportional to the old; that is, the new odds curve would equal the old multiplied by some constant amount. Thus, the logistic model is sometimes called a multiplicative-odds model. For analogous reasons, the exponential rate model is sometimes called a multiplicative-rate model. In both these models, inserting product terms into the model allows certain types of departures from proportional trends.

3.4.4 Interpreting Product-Term Models

Several important cautions should be highlighted when attempting to build models with product terms and interpret coefficients in models with product terms. First, the so-called "main-effect" coefficient β_j will be meaningless when considered alone if its regressor X_j appears in a product with another variable X_k that cannot

be zero. In the tamoxifen example, X_1 is years of exposure, which can be zero, while X_3 is age at menarche (in years), which is always above zero. Consider the model

$$\begin{aligned} R(x_1, x_2, x_3) &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_{13} x_1 x_3 \\ &= \alpha + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 + \gamma_{13} x_1) x_3 \\ &= \alpha + (\beta_1 + \gamma_{13} x_3) x_1 + \beta_2 x_2 + \beta_3 x_3 \,. \end{aligned} \tag{3.37}$$

Under this model, $\beta_1 + \gamma_{13} x_3$ is the slope for the trend in risks across X_1 given $X_2 = x_2$ and $X_3 = x_3$. Thus, if X_3 was 0, this slope would be $\beta_1 + (\gamma_{13} \times 0) = \beta_1$, and so β_1 could be interpreted as the slope for X_1 in subpopulations of a given X_2 and with $X_3 = 0$. But X_3 is age at menarche and so cannot be zero; thus, β_1 has no simple epidemiologic interpretation. In contrast, because X_1 is years exposed and so can be zero, β_3 does have a simple interpretation: Under model (3.37) $\beta_3 + \gamma_{13} x_1$ is the slope for X_3 given $X_1 = x_1$; hence, $\beta_3 + \gamma_{13} \times 0 = \beta_3$ is the slope for X_3 in subpopulations with no tamoxifen exposure ($X_1 = 0$).

As mentioned earlier, if a regressor X_j cannot be zero, one can insure a simple interpretation of the intercept α by recentering the regressor, that is, by subtracting a reference value from the regressor before entering it in the model. Such recentering also helps provide a simple interpretation for the coefficients of variables that appear with X_j in product terms. In the example, we could recenter by redefining X_3 to be age at menarche minus 13 years. With this change, β_1 in model (3.37) would now be the slope for X_1 (years of tamoxifen) in subpopulations of a given X_2 (age at start of tamoxifen) in which this new X_3 was 0 (that is, in which the age at menarche was 13).

Rescaling can also be important for interpretation of product-term coefficients. As an example, suppose X_1 is serum cholesterol in mg/dl and X_2 is diastolic blood pressure (DBP) in mm Hg, and that the product of X_1 and X_2 is entered into the model without rescaling, say as $\gamma_{12} x_1 x_2$ in an exponential rate model. Then γ_{12} would represent the difference in the log rate ratio for a 1 mg/dl increase in cholesterol when comparing sub-populations 1 mm Hg apart in DBP. Even if this term was important, it would appear very small in magnitude because of the small units used to measure cholesterol and DBP. To avoid such deceptive appearances, we could rescale X_1 and X_2 so that their units now represented important increases in cholesterol and DBP. For example, we could redefine X_1 as cholesterol divided by 20 and X_2 as DBP divided by 10. With this rescaling, γ_{12} would represent the difference in the log rate ratio for a 20 mg/dl increase in cholesterol when comparing subpopulations 10 mm Hg apart in DBP.

Another caution is that, in most situations, a product term in a model should be accompanied by terms for all variables and products contained within that product. For example, if one enters $\gamma_{12} x_1 x_2$ in a model, $\beta_1 x_1$ and $\beta_2 x_2$ should also be included in that model; and if one enters $\delta_{123} x_1 x_2 x_3$ in a model, all of $\beta_1 x_1, \beta_2 x_2, \beta_3 x_3, \gamma_{12} x_1 x_2, \gamma_{13} x_1 x_3$, and $\gamma_{23} x_2 x_3$ should be included in that model. This rule, sometimes called the "hierarchy principle" (Bishop et al. 1975), is useful in avoiding models with bizarre implications. As an example, suppose X_1 is serum-

lead concentration and X_2 is age minus 50 years. If $\gamma_{12} > 0$, the 1-year mortality-risk model

$$R(x_1, x_2) = \exp(\alpha + \beta_2 x_2 + \gamma_{12} x_1 x_2)$$

implies that serum-lead is positively related to risk among persons above age 50 ($X_2 > 0$), is unrelated to risk among persons of age 50 ($X_2 = 0$), and is negatively related to risk among persons below age 50 ($X_2 < 0$); if $\gamma_{12} < 0$, it implies a negative relation over 50 and a positive relation below 50. Rarely (if ever) would we have grounds for assuming such unusual relations hold. To prevent use of absurd models, many regression programs automatically enter all terms contained within a product when the user instructs the program to enter the product into the model.

Models violating the hierarchy principle often arise when one variable is not defined for all subjects. As an example, suppose in a study of breast cancer in women that X_1 is age at first birth (AFB) and X_2 is parity. Because X_1 is undefined for nulliparous women ($X_2 = 0$), one sometimes sees the breast-cancer rate modeled by a function in which age at first birth appears only in a product term with parity, such as $\exp(\alpha + \beta_2 x_2 + \gamma_1 x_1 x_2)$. The rationale for this model is that the rate will remain defined even when age at first birth (X_1) is undefined, because $x_1 x_2$ will be zero when parity (X_2) is zero.

One can sometimes avoid violating the hierarchy principle if there is a reasonable way to extend variable definitions to all subjects. Thus, in the tamoxifen example, age at start of tamoxifen was extended to the untreated by setting it to age 70 (end of follow-up) for those subjects, and for those subjects who started at age 70 or later. The rationale for this extension is that, within the age interval under study, untreated subjects and subjects starting tamoxifen at age 70 or later would have identical exposures.

Our final caution is that product terms are commonly labeled "interaction terms" or "statistical interactions". We avoid these labels because they may inappropriately suggest the presence of biologic (mechanical) interactions between the variables in a product term. In practice, regression models are applied in many situations in which there is no effect of the regressors on the regressand (outcome). Even in causal analyses, the connections between product terms and biologic interactions can be very indirect, and can depend on many biologic assumptions. For descriptions of these connections see Greenland (1993) and Rothman and Greenland (1998, pp 386–387).

3.4.5 Categorical Regressors

Consider a regressor whose possible values are discrete and few, and perhaps purely nominal (that is, with no natural ordering or quantitative meaning). An example is marital status (never married, currently married, formerly married). Such regressors may be entered into a multiple-regression model using *category indicator variables*. To use this approach, we first choose one level of the regressor

as the *reference level*, against which we want to compare risks or rates. For each of the remaining levels (the *index* levels), we create a binary variable that indicates whether a person is at that level (1 if at the level, 0 if not). We then enter these indicators into the regression model.

The entire set of indicators is called the *coding* of the original regressor. To code marital status, we could take "currently married" as the reference level and define

$X_1 = 1$ if formerly married, 0 if currently or never married,

$X_2 = 1$ if never married, 0 if ever married

(i.e., currently or formerly married) .

There are $2 \times 2 = 4$ possible numerical combinations of values for X_1 and X_2, but only three of them are logically possible. The impossible combination is $X_1 = 1$ (formerly married) and $X_2 = 1$ (never married). Note, however, that we need two indicators to distinguish the three levels of marital status, because one indicator can only distinguish two levels.

In general, we need $J - 1$ indicators to code a variable with J levels. Although these indicators will have 2^{J-1} possible numerical combinations, only J of these combinations will be logically possible. For example, we will need four indicators to code a variable with five levels. These indicators will have $2^4 = 16$ numerical combinations, but only five of the 16 combinations will be logically possible.

Interpretation of the indicator coefficients depends on the model form and the chosen coding. For example, in the logistic model

$$R(x_1, x_2) = \text{expit}(\alpha + \beta_1 x_1 + \beta_2 x_2) , \tag{3.38}$$

$\exp(\beta_2)$ is the odds ratio comparing $X_2 = 1$ persons (never married) to $X_2 = 0$ persons (ever married) within levels of X_1. Because one cannot have $X_2 = 1$ (never married) and $X_1 = 1$ (formerly married), the only level of X_1 within which we can compare $X_2 = 1$ to $X_2 = 0$ is the zero level (never or currently married). Thus, $\exp(\beta_2)$ is the odds ratio comparing never married ($X_2 = 1$) to currently married ($X_2 = 0$) people among those never or currently married ($X_1 = 0$). In a similar fashion, $\exp(\beta_1)$ compares those formerly married to those currently married among those ever married.

In general, the type of indicator coding described above, called *disjoint category coding*, results in coefficients that compare each index category to the reference category. With this coding, for a given person no more than one indicator in the set can equal 1; all the indicators are zero for persons in the reference category. Another kind of coding is *nested indicator coding*. In this coding, levels of the regressor are grouped, and then codes are created to facilitate comparisons both within and across groups. For example, suppose we wish to compare those not currently married (never or formerly married) to those currently married, and also compare those never married to those formerly married. We can then use the

indicators

$Z_1 = 1$ if never or formerly married (i.e., not currently married),

0 otherwise (currently married);

$Z_2 = 1$ if never married, 0 if ever married.

Z_2 is the same as the X_2 used above, but Z_1 is different from X_1. The combination $Z_1 = 0$ (currently married), $Z_2 = 1$ (never married) is impossible; $Z_1 = Z_2 = 1$ for people who never married. In the logistic model

$$R(z_1, z_2) = \text{expit}(\alpha + \beta_1 z_1 + \beta_2 z_2)\,, \tag{3.39}$$

$\exp(\beta_2)$ is now the odds ratio comparing those never married ($Z_2 = 1$) to those ever married ($Z_2 = 0$) within levels of Z_1. Note that the only level of Z_1 in which this comparison can be made is $Z_1 = 1$ (never or formerly married). Similarly, $\exp(\beta_1)$ is now the odds ratio comparing those formerly married ($Z_1 = 1$) among those never married ($Z_2 = 0$).

There can be quite a large number of options for coding category indicators. The choice among these options may be dictated by which comparisons are of most interest. As long as each level of the regressor can be uniquely represented by the indicator coding, the choice of coding will not alter the assumptions represented by the model. There is, however, one technical point to consider in choosing codes. The precision of the estimated coefficient for an indicator will directly depend on the numbers of subjects at each indicator level. For example, suppose in the data there were 1000 currently married subjects, 200 formerly married subjects, and only 10 never married subjects. Then any indicator that had "never married" as one of its levels (0 or 1) would have a much less precise coefficient estimate than other indicators. If "never married" were chosen as the reference level for a disjoint coding scheme, all the indicators would have that level as its zero level, and so all would have very imprecise coefficient estimates. To maximize precision, many analysts prefer to use disjoint coding in which the largest category (currently married in the above example) is taken as the reference level.

In choosing a coding scheme, one need not let precision concerns dominate if they get in the way of interesting comparisons. Coding schemes that distinguish among the same categories produce equivalent models. Therefore, one may fit a model repeatedly using different but equivalent coding schemes, in order to easily examine all comparisons of interest. For example, one could fit model (3.38) to compare those never or formerly married with those currently married, then fit model (3.39) to compare the never with formerly married.

Although indicator coding is essential for purely nominal regressors, it can also be used to study quantitative regressors as well, especially when one expects qualitative differences between persons at different levels. Consider number of marriages as a regressor. We might suspect that people of a given age who have had one marriage tend to be qualitatively distinct from people of the same age who have

had no marriage or two marriages, and that people who have had several marriages are even more distinctive. We thus might want to code number of marriages in a manner that allowed qualitative distinctions among its levels. If "one marriage" was the most common level, we might take it as the reference level and use

$X_1 = 1$ if never married, 0 otherwise;

$X_2 = 1$ if two marriages, 0 otherwise;

$X_3 = 1$ if three or more marriages, 0 otherwise.

We use one variable to represent "three or more" because there might be too few subjects with three or more marriages to produce acceptably precise coefficients for a finer division of levels. The coding just given would provide comparisons of those never married, twice married, and more-than-twice married to those once married. Other codings could be used to make other comparisons.

Trend Models in Multiple Regression 3.5

Multiple regression models can be extended to produce much more flexible trend models than those provided by simple transformations. The latter restrict trends to follow basic shapes, such as quadratic or logarithmic curves. The use of multiple terms for each exposure and confounder allows more detailed assessment of trends and more complete control of confounding than possible with simple transformations.

Categorical Trends 3.5.1

One way to extend trend models is to categorize the regressor and then use a category-indicator coding such as discussed above. The resulting analysis may then parallel the categorical (tabular) trend methods discussed for example in Chap 17 of Rothman and Greenland (1998). Much of the advice given there also applies here. To the extent allowed by the data numbers and background information, the categories should represent scientifically meaningful constructs within which risk is not expected to change dramatically. Purely mathematical categorization methods such as percentiles (quantiles) can do very poorly in this regard and so are best avoided when such information is available. On the other hand, the choices of categories should *not* be dictated by the results produced; for example, manipulation of category boundaries to maximize the effect estimate will produce an estimate biased away from the null, while manipulation of boundaries to minimize a ***P***-value will produce a downwardly biased ***P***-value. Similarly, manipulation to minimize the estimate or maximize the ***P***-value will produce a null-biased estimate or an upwardly biased ***P***-value.

There are two common types of category codes used in trend models. *Disjoint coding* produces estimates that compare each index category (level) to the reference level. Consider coding weekly servings of fruits and vegetables with

$$\begin{aligned} X_1 &= 1 \quad \text{for } < 15, \quad 0 \text{ otherwise;} \\ X_2 &= 1 \quad \text{for } 36\text{–}42, \ 0 \text{ otherwise;} \\ X_3 &= 1 \quad \text{for } > 42, \quad 0 \text{ otherwise.} \end{aligned}$$

In the rate model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,, \tag{3.40}$$

$\exp(\beta_1)$ is the rate ratio comparing the "< 15" category with the "15–35" category (which is the referent), and so on, while $\exp(\alpha)$ is the rate in the "15–35" category (the category for which all the X_j are zero). When model (3.40) is fit, we can plot the fitted rates on a graph as a step function. This plot provides a crude impression of the trends across (but not within) categories.

Confounders may be added to the model in order to control confounding, and these too may be coded using multiple indicators or any of the methods described below. We may plot the model-adjusted trends by fixing each confounder at a reference level and allowing the exposure level to vary.

Incremental coding (nested coding) can be useful when one wishes to compare each category against its immediate predecessor (Maclure and Greenland 1992). For "Number of servings per week", we could use

$$\begin{aligned} Z_1 &= 1 \quad \text{for } > 14, \ 0 \text{ otherwise;} \\ Z_2 &= 1 \quad \text{for } > 35, \ 0 \text{ otherwise;} \\ Z_3 &= 1 \quad \text{for } > 42, \ 0 \text{ otherwise.} \end{aligned}$$

Note that if $Z_2 = 1$, then $Z_1 = 1$, and if $Z_3 = 1$, then $Z_1 = Z_2 = 1$. In the model

$$\ln[I(z_1, z_2, z_3)] = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 \,, \tag{3.41}$$

$\exp(\beta_1)$ is the rate ratio comparing the 15–35 category ($Z_1 = 1$ and $Z_2 = Z_3 = 0$) to the < 15 category ($Z_1 = Z_2 = Z_3 = 0$). Similarly, $\exp(\beta_2)$ is the rate ratio comparing the 36–42 category ($Z_1 = Z_2 = 1$ and $Z_3 = 0$) to the 15–35 category ($Z_1 = 1$ and $Z_2 = Z_3 = 0$). Finally, $\exp(\beta_3)$ compares the > 42 category ($Z_1 = Z_2 = Z_3 = 1$) to the 36–42 category ($Z_1 = Z_2 = 1$ and $Z_3 = 0$). Thus, $\exp(\beta_1)$, $\exp(\beta_2)$, and $\exp(\beta_3)$ are the incremental rate ratios across adjacent categories. Again, we may add confounders to the model and plot adjusted trends.

3.5.2 Regression with Category Scores

A common practice in epidemiology is to divide each covariate into categories, assign a score to each category, and enter scores into the model instead of the original variable values. Ordinal scores or codes (e.g., 1, 2, 3, 4, 5 for a series of five

categories) should be avoided, as they can yield quantitatively meaningless dose-response curves and harm the power and precision of the results (Lagakos 1988; Greenland 1995b, c; Rothman and Greenland 1998, pp 311–312). Category midpoints can be much less distortive but are not defined for open-ended categories; category means or medians can be even less distortive and are defined for open-ended categories. Unfortunately, if there are important nonlinear effects within categories, no simple scoring method will yield an undistorted dose-response curve, nor will it achieve the power and precision obtainable by entering the uncategorized covariates into the model (Greenland 1995b, c). We thus recommend that categories be kept narrow and that scores be derived from category means or medians, rather than category scores. We further recommend that one examine models with uncategorized covariates whenever effects are clearly present.

Power Models 3.5.3

Another approach to trend analysis and confounder control is to use multiple power terms for each regressor. Such an approach does not require categorization, but does require care in selection of terms. Traditionally, the powers used are positive integers (e.g., x_1, x_1^2, x_1^3), but fractional powers may also be used (Royston and Altman 1994). As an illustration, suppose X_1 represents the actual number of servings per week (instead of an indicator). We could model trends across this regressor by using X_1 in the model along with the following powers of X_1:

$$X_2 = X_1^{1/2} = \text{ square root of } X_1 ,$$

$$X_3 = X_1^2 = \text{ square of } X_1 .$$

The multiple-regression model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ,$$

is now just another way of writing the power model

$$\ln[I(x_1)] = \alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2 . \tag{3.42}$$

We can plot fitted rates from this model using very fine spacings to produce a *smooth curve* as an estimate of rate trends across X_1. As always, we may also include confounders in the model and plot model-adjusted trends.

Power models have several advantages over categorical models. Most importantly, they make use of information about differences within categories, which is ignored by categorical models and categorical analyses (Greenland 1995a, b, c). Thus, they can provide a more complete picture of trends across exposure and more thorough control of confounders. They also provide a smoother picture of trends. One disadvantage of power models is a potentially greater sensitivity of estimates to *outliers*, that is, persons with unusual values or unusual *combinations* of values for the regressors. This problem can be addressed by performing delta-beta analysis, as discussed below.

3.5.4 Regression Splines

Often it is possible to combine the advantages of categorical and power models through the use of *spline models*. Such models can be defined in a number of equivalent ways, and we present only the simplest. In all approaches, one first categorizes the regressor, as in categorical analysis (although fewer, broader categories may be sufficient in a spline model). The boundaries between these categories are called the *knots* or *join points* of the spline. Next, one chooses the *power* (or order) of the spline, according to the flexibility one desires within the categories (higher powers allow more flexibility).

Use of category indicators corresponds to a zero-power spline, in which the trend is flat within categories but may jump suddenly at the knots; thus, category-indicator models are just special and unrealistic types of spline models. In a first-power or *linear* spline, the trend is modeled by a series of connected line segments. The trend within each category corresponds to a line segment; the slope of the trend may change only at the knots, and no sudden jump in risk (discontinuity in trend) can occur.

To illustrate how a linear spline may be represented, let X_1 again be "Number of servings per week" but now define

$$\begin{aligned} X_2 &= X_1 - 14 \quad \text{if } X_1 > 14,\ 0 \text{ otherwise;} \\ X_3 &= X_1 - 35 \quad \text{if } X_1 > 35,\ 0 \text{ otherwise.} \end{aligned}$$

Then the log-linear rate model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{3.43}$$

will produce a log-rate trend that is a series of three line segments that are connected at the knots (category boundaries) of 14 and 35. To see this, note that when X_1 is less than 14, X_2 and X_3 are zero, so the model simplifies to a line with slope β_1:

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1$$

in this range. When X_1 is greater than 14 but less than 35, the model simplifies to a line with slope $\beta_1 + \beta_2$:

$$\begin{aligned} \ln[I(x_1, x_2, x_3)] &= \alpha + \beta_1 x_1 + \beta_2 x_2 = \alpha + \beta_1 x_1 + \beta_2 (x_1 - 14) \\ &= \alpha - 14\beta_2 + (\beta_1 + \beta_2) x_1 . \end{aligned}$$

Finally, when X_1 is greater than 35, the model becomes a line with slope $\beta_1 + \beta_2 + \beta_3$:

$$\begin{aligned} \ln[I(x_1, x_2, x_3)] &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= \alpha + \beta_1 x_1 + \beta_2 (x_1 - 14) + \beta_3 (x_1 - 35) \\ &= a - 14\beta_2 - 35\beta_3 + (\beta_1 + \beta_2 + \beta_3) x_1 . \end{aligned}$$

Thus, β_1 is the slope of the spline in the first category, β_2 is the change in slope in going from the first to second category, and β_3 is the change in slope in going from the second to third category.

The trend produced by a linear spline is generally more realistic than a categorical trend, but can suddenly change its slope at the knots. To smooth out such sudden changes, we may increase the order of the spline. Increasing the power to 2 produces a second-power or *quadratic* spline, which comprises a series of parabolic curve segments smoothly joined together at the knots. To illustrate how such a trend may be represented, let X_1, X_2, and X_3 be as just defined. Then the model

$$\ln[I(x_1, x_2, x_3)] = \alpha + \beta_1 x_1 + \gamma_1 x_1^2 + \gamma_2 x_2^2 + \gamma_3 x_3^2 \tag{3.44}$$

will produce a log-rate trend that is a series of three parabolic segments smoothly connected at the knots of 14 and 35. The coefficient γ_1 corresponds to the curvature of the trend in the first category, while γ_2 and γ_3 correspond to the changes in curvature when going from the first to second and second to third category. A still smoother curve could be fit by using a third-power or *cubic* spline, but for epidemiologic purposes the quadratic spline is often smooth and flexible enough.

One disadvantage of quadratic and cubic splines is that the curves in the end categories (tails) may become very unstable, especially if the category is open-ended. This instability may be reduced by *restricting* one or both of the end categories to be a line segment rather than a curve. To restrict the lower category to be linear in a quadratic spline, we need only drop the *first* quadratic term $\gamma_1 x_1^2$ from the model; to restrict the upper category, we must subtract the *last* quadratic term from all the quadratic terms, and drop the last term out of the model. To illustrate an upper category restriction, suppose we wish to restrict the above quadratic spline model for log rates (3.44) so that it is linear in the upper category only. Define

$$Z_1 = X_1 = \text{number of servings per week},$$

$$Z_2 = X_1^2 - X_3^2,$$

$$Z_3 = X_2^2 - X_3^2.$$

Then the model

$$\ln[I(z_1, z_2, z_3)] = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 \tag{3.45}$$

will produce a log-rate trend that comprises smoothly connected parabolic segments in the first two categories ("< 14" and "15–35"), and a line segment in the last category ("> 35") that is smoothly connected to the parabolic segment in the second category. (If we also wanted to force the log-rate curve in the first category to follow a line, we would drop Z_2 from the model.)

To plot or tabulate a spline curve from a given spline model, we select a set of X_1 values spaced across the range of interest, compute the set of spline terms

for each X_1 value, combine these terms with the coefficients in the model to get the model-predicted outcomes, and plot these predictions. To illustrate, suppose X_1 is servings per week and we wish to plot model (3.45) with $\alpha = -6.00, \beta_1 = -0.010, \beta_2 = -0.001$, and $\beta_3 = 0.001$ over the range 0–50 servings per week in 5-serving increments. We then compute Z_1, Z_2, Z_3 at $0, 5, 10, \ldots, 50$ servings per week, and compute the predicted rate

$$\exp(-6.00 - 0.010z_1 - 0.001z_2 + 0.001z_3)$$

at each set of Z_1, Z_2, Z_3 values and plot these predictions against the corresponding X_1 values $0, 5, 10, \ldots, 50$. For example, at $X_1 = 40$ we get $Z_1 = 40, Z_2 = 40^2 - (40 - 35)^2 = 1575$, and $Z_3 = (40 - 14)^2 - (40 - 35)^2 = 651$, for a predicted rate of

$$\exp[-6.00 - 0.010(40) - 0.001(1575) + 0.001(651)] = 2/1000 \text{ year}.$$

As with other trend models, we may obtain model-adjusted trends by adding confounder terms to our spline models. The confounder terms may be splines or any other form we prefer; spline plotting will be simplified, however, if the confounders are centered before they are entered into the analysis, for then the above plotting method may be used without modification. For further discussion of splines and their application, as well as more general nonparametric regression techniques, see Hastie and Tibshirani (1990), Green and Silverman (1994), and Greenland (1995a).

3.5.5 Models for Trend Variation

We may allow trends to vary across regressor levels by entering products among regressor terms. For example, suppose X_1, X_2, X_3 are power terms for fruit and vegetable intake, while W_1, W_2, W_3, W_4 are spline terms for age. To allow the fruit-vegetable trend in log rates to vary with age, we could enter into the model all $3 \times 4 = 12$ products of the X_j and W_k, along with the X_j and W_k. If in addition there was an indicator $Z_1 = 1$ for female, 0 for males, the resulting model would be

$$\begin{aligned}&\ln[R(x_1, x_2, x_3, w_1, w_2, w_3, w_4, z_1)]\\ &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 w_1 + \beta_5 w_2 + \beta_6 w_3 + \beta_7 w_4 + \beta_8 z_1\\ &\quad + \gamma_{11} x_1 w_1 + \gamma_{12} x_1 w_2 + \ldots + \gamma_{33} x_3 w_3 + \gamma_{34} x_3 w_4 .\end{aligned}$$

The same model form may be used if X_1, X_2, X_3 and W_1, W_2, W_3, W_4 represent category indicators or other terms for fruit-vegetable intake and age.

Models with products among multiple trend terms can be difficult to fit and may yield quite unstable results unless large numbers of cases are observed. Given enough data, however, such models can provide more realistic pictures of dose-response relations than can simpler models. Results from such models may be easily interpreted by plotting or tabulating the fitted trends for the key exposures of interest at various levels of the "modifying" regressors. In the above example, this

process would involve plotting the model-fitted rates against fruit and vegetable intake for each of several ages (e.g., for ages evenly spaced within the range of case ages).

Extensions of Logistic Models 3.6

Outcomes that are polytomous or continuous are often analyzed by reducing them to just two categories and applying a logistic model. For example, CD_4 counts might be reduced to the dichotomy ≤ 200, > 200; cancer outcomes might be reduced to cancer and no cancer. Alternatively, multiple categories may be created with one designated as a referent, and the other categories compared one at a time to the referent using separate logistic models for each comparison. While not necessarily invalid, these approaches disregard the information contained in differences within categories, in differences between non-reference categories, and in ordering among the categories. As a result, models specifically designed for polytomous or continuous outcomes can yield more precision and power than simple dichotomous-outcome analyses.

This section briefly describes several extensions of the multiple logistic model (3.28) to polytomous and ordinal outcomes. Analogous extensions of other models are possible.

Polytomous Logistic Models 3.6.1

Suppose an outcome variable Y has $I + 1$ mutually exclusive outcome categories or levels $y_0, \ldots, y_I$, where category y_0 is considered the reference category. For example, in a case-control study of relations of exposures to types of cancer, Y is a disease outcome variable, with y_0 = all control as the reference category, and I other categories $y_1, \ldots, y_I$, which correspond to the cancer outcomes (leukemia, lymphoma, lung cancer, etc.). Let $R_i(\boldsymbol{x})$ denote the average risk of falling in outcome category $Y_i (i = 1, \ldots, I)$ given that the regressors $\boldsymbol{X}$ equal $\boldsymbol{x}$; that is, let

$$R_i(\boldsymbol{x}) = \Pr(Y = y_i | \boldsymbol{X} = \boldsymbol{x}) .$$

The *polytomous logistic* model for this risk is then

$$R_i(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\alpha}_1 + \boldsymbol{x}\boldsymbol{\beta}_1\right)}{1 + \sum_{j=1}^{I} \exp(\boldsymbol{\alpha}_j + \boldsymbol{x}\boldsymbol{\beta}_j)} \tag{3.46}$$

This is a model for the risk of falling in cancer category y_i. When Y has only two levels, I equals 1, and so formula (3.46) simplifies to the binary multiple logistic model (3.28).

Model (3.46) represents I separate risk equations, one for each nonreference outcome level $y_1, \ldots, y_I$. Each equation has its own intercept $\boldsymbol{\alpha}_i$ and vector of

coefficients $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{in})$, so that there is a distinct coefficient β_{ik} corresponding to every combination of a regressor X_k and nonreference outcome level y_i ($i = 1, \ldots, I$). Thus, with n regressors in $\boldsymbol{X}$, the polytomous logistic model involves I intercepts and $I \times n$ regressor coefficients. For example, with seven nonreference outcome levels and three regressors, the model would involve seven intercepts and $7 \times 3 = 21$ regressor coefficients, for a total of 28 model parameters.

The polytomous logistic model can be written more simply as a model for the odds. To see this, note that the risk of falling in the reference category must equal one minus the sum of the risks of falling in the nonreference categories:

$$R_0(\boldsymbol{x}) = \Pr(Y = y_0 | \boldsymbol{X} = \boldsymbol{x}) = 1 - \frac{\sum_{i=1}^{I} \exp\left(\alpha_i + \boldsymbol{x}\boldsymbol{\beta}_i\right)}{1 + \sum_{j=1}^{I} \exp\left(\alpha_j + \boldsymbol{x}\boldsymbol{\beta}_j\right)}$$

$$= 1 \Big/ \left[1 + \sum_{j=1}^{I} \exp\left(\alpha_j + \boldsymbol{x}\boldsymbol{\beta}_j\right)\right] . \tag{3.47}$$

Dividing (3.47) into (3.46), we get a model for $O_i(\boldsymbol{x}) = R_i(\boldsymbol{x})/R_0(\boldsymbol{x})$ = the odds of falling in outcome category y_i versus category y_0:

$$O_i(\boldsymbol{x}) = \frac{\exp(\alpha_i + \boldsymbol{x}\boldsymbol{\beta}_i)/[1 + \sum_j \exp(\alpha_j + \boldsymbol{x}\boldsymbol{\beta}_j)]}{1/[1 + \sum_j \exp(\alpha_j + \boldsymbol{x}\boldsymbol{\beta}_j)]} = \exp(\alpha_i + \boldsymbol{x}\boldsymbol{\beta}_i) . \tag{3.48}$$

This form of the model provides a familiar interpretation for the coefficients. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ are two different vectors of values for the regressors $\boldsymbol{X}$. Then the ratio of the odds of falling in category y_i versus y_0 when $\boldsymbol{X} = \boldsymbol{x}_1$ and $\boldsymbol{X} = \boldsymbol{x}_0$ is

$$\frac{O_i(\boldsymbol{x}_1)}{O_i(\boldsymbol{x}_0)} = \frac{\exp(\alpha_i + \boldsymbol{x}_1\boldsymbol{\beta}_i)}{\exp(\alpha_i + \boldsymbol{x}_0\boldsymbol{\beta}_i)} = \exp\left[(\boldsymbol{x}_1 - \boldsymbol{x}_0)\boldsymbol{\beta}_i\right] .$$

From this equation, we see that the antilog $\exp(\beta_{ik})$ of a coefficient β_{ik} corresponds to the proportionate change in the odds of outcome i when the regressor $\boldsymbol{X}_k$ increases by one unit.

The polytomous logistic model is most useful when the levels of Y have no meaningful order, as with the cancer types. For further reading about the model, see McCullagh and Nelder (1989) and Hosmer and Lemeshow (2000).

3.6.2 Ordinal Logistic Models

Suppose that the levels $y_0, \ldots, y_I$ of Y follow a natural order. Order arises, for example, when Y is a clinical scale, such as y_0 = normal, y_1 = dysplasia, y_2 = neoplasia, rather than just a cancer indicator; Y is a count, such as number of malformations found in an individual; or the Y levels represent categories of a physical quantity, such as CD_4 count (e.g., > 500, 200–500, < 200). There are at least four different ways to extend the logistic model to such outcomes.

Recall that the logistic model is equivalent to an exponential odds model. The first extension uses an exponential model to represent the odds of falling in outcome category y_i versus falling in category y_{i-1} (the next lowest category):

$$\frac{R_i(\boldsymbol{x})}{R_{i-1}(\boldsymbol{x})} = \frac{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y = y_{i-1}|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x}\beta^*) \tag{3.49}$$

for $i = 1, \ldots, I$. This may be called the *adjacent-category logistic model*, because taking logarithms of both sides yields the equivalent *adjacent-category logit* model (Agresti 2002). It is a special case of the polytomous logistic model: From (3.48), the polytomous logistic model implies that

$$\frac{R_i(\boldsymbol{x})}{R_{i-1}(\boldsymbol{x})} = \frac{R_i(\boldsymbol{x})/R_0(\boldsymbol{x})}{R_{i-1}(\boldsymbol{x})/R_0(\boldsymbol{x})} = \frac{\exp\left(\alpha_i + \boldsymbol{x}\beta_i\right)}{\exp\left(\alpha_{i-1} + \boldsymbol{x}\beta_{i-1}\right)} = \exp[(\alpha_i - \alpha_{i-1}) + \boldsymbol{x}(\beta_i - \beta_{i-1})]\,.$$

The adjacent-category logistic model sets $\alpha_i^* = \alpha_i - \alpha_{i-1}$, and forces the I coefficient differences $\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i-1} (i = 1, \ldots I)$ to equal a common value $\boldsymbol{\beta}^*$. If there is a natural distance d_i between adjacent outcome categories y_i and y_{i-1} (such as the difference between the category means), the model can be modified to use these distances as follows:

$$R_i(\boldsymbol{x})/R_{i-1}(\boldsymbol{x}) = \exp(\alpha_i^* + \boldsymbol{x}\boldsymbol{\beta}^* d_i) \tag{3.50}$$

for $i = 1, \ldots, I$. This model allows the coefficient differences $\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i-1}$ to vary with the distances d_i between categories. Further information on adjacent-category models may be found in Greenland (1994) and Agresti (2002).

The second extension uses an exponential model to represent the odds of falling *above category* y_i versus *falling in or below* category y_i:

$$\frac{\Pr(Y > y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y \leq y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x}\boldsymbol{\beta}^*)\,, \tag{3.51}$$

where $i = 0, \ldots, I$. This is called the *cumulative-odds* or *proportional-odds* model. It can be derived by assuming that Y was obtained by categorizing a special type of continuous variable; for more details about this and other aspects of the model, see McCullagh and Nelder (1989).

The third extension uses an exponential model to represent the odds of falling *above outcome* category y_i versus *in* category y_i:

$$\frac{\Pr(Y > y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x}\boldsymbol{\beta}^*)\,, \tag{3.52}$$

where $i = 0, \ldots, I$. This is called the *continuation-ratio* model. The fourth extension uses an exponential model to represent the odds of falling *in* category y_i versus falling *below* y_i:

$$\frac{\Pr(Y = y_i|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y < y_i|\boldsymbol{X} = \boldsymbol{x})} = \exp(\alpha_i^* + \boldsymbol{x}\boldsymbol{\beta}^*)\,, \tag{3.53}$$

where $i = 1, \ldots, I$. This model may be called the *reverse continuation-ratio* model. It can be derived by reversing the order of the Y levels in model (3.52) but in any given application it is not equivalent to model (3.52) (Greenland 1994).

How does one choose from the above variety of ordinal models? Certain guidelines may be of use, although none is compelling. First, the adjacent-category and cumulative-odds models are *reversible*, in that only the signs of the coefficients change if the order of the Y levels is reversed. In contrast, the two continuation-ratio models are not reversible. This observation suggests that the continuation-ratio models may be more appropriate for modeling irreversible disease stages (e.g., osteoarthritic severity), whereas the adjacent-category and cumulative-odds models may be more appropriate for potentially reversible outcomes (e.g., blood pressure, cell counts) (Greenland 1994). Second, because the coefficients of adjacent-category models contrast pairs of categories, the model appears best suited for discrete outcomes with few levels (e.g., cell types along a normal-dysplastic-neoplastic scale). Third, because the cumulative-odds model can be derived from categorizing certain special types of continuous outcomes, it is often considered most appropriate when the outcome under study is derived by categorizing a single underlying continuum (e.g., blood pressure) (McCullagh and Nelder 1989). For a more detailed comparative discussion of ordinal logistic models and guidelines for their use, see Greenland (1994).

All the above ordinal models simplify to the ordinary logistic model when there are only two outcome categories ($I = 2$). One advantage of the continuation-ratio models over their competitors is of special importance: Estimation of the coefficients $\boldsymbol{\beta}^*$ in those models can be carried out if the levels of Y are numerous and sparse; Y may even be continuous. Thus, one can apply the continuation-ratio models without any categorization of Y. This advantage can be important because results from all the above models (including the cumulative-odds model) may be affected by the choice of the Y categories (Greenland 1994; Strömberg 1996). The only caution is that conditional (as opposed to unconditional) maximum likelihood must be used to fit the continuation-ratio model if the observed outcomes are sparsely scattered across the levels of Y (as would be inevitable if Y were continuous). See Greenland (1994) for further details, and Cole and Ananth (2001) for futher extensions of the model.

3.7 Generalized Linear Models

Consider again the general form of the exponential risk and rate models, $R(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x}\boldsymbol{\beta})$ and $I(\boldsymbol{x}) = \exp(\alpha + \boldsymbol{x}\boldsymbol{\beta})$ and the logistic risk model $R(\boldsymbol{x}) = \text{expit}(\alpha + \boldsymbol{x}\boldsymbol{\beta})$. There is no reason why we cannot replace the "exp" in the exponential models or the "expit" in the logistic model by some other reasonable function. In fact, each of these models is of the general form

$$E(Y|\boldsymbol{x}) = f(\alpha + \boldsymbol{x}\boldsymbol{\beta}), \tag{3.54}$$

where f is some function that is smooth and strictly increasing (i.e., as $\boldsymbol{\alpha} + \boldsymbol{x\beta}$ gets larger, $f(\boldsymbol{\alpha} + \boldsymbol{x\beta})$ gets larger, but never jumps or bends suddenly).

For any such function f, there is always an inverse function g that "undoes" f, in the sense that $g[f(u)] = u$ whenever $f(u)$ is defined. Hence, a general form equivalent to (3.54) is

$$g[E(Y|\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{x\beta}\,. \tag{3.55}$$

A model of the form (3.55) is called a *generalized linear* model. The function g is called the *link function* for the model; thus, the link function is ln for the log-linear model and logit for the logit-linear model. The term $\boldsymbol{\alpha} + \boldsymbol{x\beta}$ in it is called the *linear predictor* for the model and is often abbreviated η; that is, $\eta = \boldsymbol{\alpha} + \boldsymbol{x\beta}$ by definition.

All the models we have discussed are generalized linear models. Ordinary linear models (such as the linear risk model) are the simplest examples, in which f and g are both the identity function $f(u) = g(u) = u$, so that

$$E(Y|\boldsymbol{x}) = \boldsymbol{\alpha} + \boldsymbol{x\beta}\,.$$

The inverse of the exponential function exp is the natural log function $\ln(u)$. Hence, the generalized-linear forms of the exponential risk and rate models are the log-linear risk and rate models

$$\ln[R(\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{x\beta} \quad \text{and} \quad \ln[I(\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{x\beta}\,;$$

that is, the exponential risk and rate models correspond to a natural-log link function, because $\ln[\exp(u)] = u$. Similarly, the inverse of expit, the logistic function, is the logit function $\text{logit}(u)$. Hence, the generalized-linear form of the logistic-risk model is the logit-linear risk model

$$\text{logit}[R(\boldsymbol{x})] = \boldsymbol{\alpha} + \boldsymbol{x\beta}\,;$$

that is, the logistic model corresponds to the logit link function, because $\text{logit}[\text{expit}(u)] = u$.

The choices for f and g are virtually unlimited. In epidemiology, however, only the logit link $g(u) = \text{logit}(u)$ is in common use for risks, and only the log link $g(u) = \ln(u)$ is in common use for rates. In practice, these link functions are almost always the default, and are sometimes the only options in commercial software for risk and rate modeling. Some packages, however, allow easy selection of linear risk, rate, and odds models, which use the identity link. Some software (e.g., GLIM) allows the user to define their own link function.

The choice of link function can have a profound impact on the shape of the trend or dose-response surface allowed by the model, especially if exposure is represented by only one or two terms. For example, if exposure is represented by a single term $\beta_1 x_1$ in a risk model, use of the identity link results in a linear risk model and a linear trend for risk; use of the log link results in an exponential (log-linear) risk model and an exponential trend for risk; and use of a logit link results in a logistic model and an exponential trend for the odds. Gen-

eralized linear models encompass a broader range than the linear, log-linear, and logistic forms, however. One example is the complementary log-log risk model,

$$R(\boldsymbol{x}) = 1 - \exp[-\exp(\boldsymbol{\alpha} + \boldsymbol{x\beta})]\,,$$

which translates to the generalized-linear form

$$\ln[-\ln(1 - R(\boldsymbol{x}))] = \boldsymbol{\alpha} + \boldsymbol{x\beta}\,.$$

This model corresponds to the link function $\ln[-\ln(l-u)]$ and arises naturally in certain biology experiments. For further reading on this and other generalized linear models, see McCullagh and Nelder (1989).

3.8 Model Searching

How do we find a model or set of models acceptable for our purposes? There are far too many model forms to allow us to examine most or even much of the total realm of possibilities. There are several systematic, mechanical, and traditional algorithms for finding models (such as stepwise and best-subset regression) that lack logical or statistical justification and that perform poorly in theoretical and simulation studies; see Sclove et al. (1972), Bancroft and Han (1977), Freedman (1983), Flack and Chang (1987), Hurvich and Tsai (1990), and Weiss (1995). For example, the ***P***-values and standard-error (SE) estimates obtained when variables are selected using significance-testing criteria (such as "*F*-to-enter" and "*F*-to-remove") will be downwardly biased. In particular, the SE estimates obtained from the selected model underestimate the standard deviations (SDs) of the point estimates obtained by applying the algorithms across different random samples. As a result, the algorithms will tend to yield ***P***-values that are too small and confidence intervals that are too narrow (and hence fail to cover the true coefficient values with the stated frequency). Unfortunately, significance-testing criteria are the basis for most variable-selection procedures in standard packaged software.

Other criteria for selecting variables, such as "change-in-point-estimate" criteria, do not necessarily perform better than significance testing (Maldonado and Greenland 1993a). Viable alternatives to significance testing in model selection have emerged only gradually with recent advances in computing and with deeper insights into the problem of model selection. We first outline the traditional approaches after reinforcing one of the most essential and neglected starting points for good modeling: laying out existing information in a manner that can help the search avoid models in conflict with established facts. A powerful alternative to model selection is provided by hierarchical regression, also known as multilevel, mixed-model, or random-coefficient regression (Rothman and Greenland 1998, pp 427–432; Greenland 2000a, b).

Role of Prior Information

3.8.1

The dependence of regression results on the chosen model can be either an advantage or a drawback. The advantage comes from the fact that use of a model structure capable of reasonably approximating reality can elevate the accuracy of the estimates over those from the corresponding tabular analysis. The drawback comes from the fact that use of a model incapable of even approximating reality can decrease estimation accuracy below that of tabular analysis.

This duality underscores the desirability of using flexible (and possibly complex) models. One should take care to avoid models that are entirely unsupported by background knowledge. For example, in a cohort study of lung cancer, it is reasonable to restrict rates to increase with age, because there is enormous background literature documenting that this trend is found in all human populations. In contrast, one would want to avoid restricting cardiovascular disease (CVD) rates to strictly increase with alcohol consumption, because there are considerable data to suggest the alcohol-CVD relation is not strictly increasing (Maclure 1993).

Prior knowledge about most epidemiologic relations is usually too limited to provide much guidance in model selection. A natural response might be to use models as flexible as possible (a flexible model can reproduce a wide variety of curves and surfaces). Unfortunately, flexible models have limitations. The more flexible the model, the larger the sample needed for the usual estimation methods (such as maximum likelihood) to provide approximately unbiased coefficient estimates. Also, after a certain point, increasing flexibility may increase variability of estimates so much that the accuracy of the estimates is decreased relative to estimates from simpler models, despite the greater faithfulness of the flexible model to reality. As a result, it is usual practice to employ models that are severely restrictive in arbitrary ways, such as models without product terms (Robins and Greenland 1986). Hierarchical methods can help alleviate some of these problems by allowing one to fit larger models than one can with ordinary methods (Greenland 2000b).

Fortunately, estimates obtained from the most common epidemiologic regression models, exponential (log-linear) and logistic models, retain some interpretability even when the underlying (true) regression function is not particularly close to those forms (Maldonado and Greenland 1993b, 1994). For example, under reasonably common conditions, rate-ratio or risk-ratio estimates obtained from those models can be interpreted as approximate estimates of standardized rate or risk ratios, using the total source population as the standard (Greenland and Maldonado 1994). To ensure such interpretations are reasonable, the model used should at least be able to replicate qualitative features of the underlying regression function. For example, if the underlying regression may have a reversal in the slope of the exposure-response curve, we should want to use a model capable of exhibiting such reversal (even if it cannot replicate the exact shape of the true curve).

A major problem for epidemiology is that key variables may be unmeasured or poorly measured. No conventional method can account for these problems. Unmeasured variables may be modeled using prior information on their relation to

measured variables, but the results will be entirely dependent on that information (Leamer 1978; Greenland 2003a). Occasionally, measurement-error information may be in the form of data that can be used in special correction techniques (Carroll et al. 1995; Chap. II.5 of this handbook); otherwise, sensitivity analyses will be needed (Rothman and Greenland 1998, Chap. 19; Lash and Fink 2003).

3.8.2 Selection Strategies

Even with ample prior information, there will always be an overwhelming number of model choices, and so model search strategies will be needed. Many strategies have been proposed, although none has been fully justified.

Some strategies begin by specifying a minimal model form that is among the most simple credible forms. Here "credible" means "compatible with available information". Thus, we start with a model of minimal computational or conceptual complexity that does not conflict with background information. There may be many such models; in order to help insure that our analysis is credible to the intended audience, however, the starting model form should be one that most researchers would view as a reasonable possibility.

To specify a simple yet credible model form, one needs some knowledge of the background scientific literature on the relations under study. This knowledge would include information about relations of potential confounders to the study exposures and study diseases, as well as relations of study exposures to the study diseases. Thus, specification of a simple yet credible model can demand much more initial effort than is routinely used in model specification.

Once we have specified our minimal starting model, we can add complexities that seem necessary (by some criteria) in light of the data. Such a search process is sometimes called an *expanding* search (Leamer 1978). Its chief drawback is that often there are too many possible expansions to consider within a reasonable length of time. If, however, one neglects to consider any possible expansion, one risks missing an important shortcoming of the initial model. For example, if our minimal model involves only single "first-order" terms ("main effects") for 12 variables, we would have $\binom{12}{2} = 66$ possible two-way products among these variables to consider, as well as 12 quadratic terms, for a total of 78 possible expansions with just one second-order term. An analyst may not have the time, patience, or resources to examine all the possibilities in detail; this predicament usually leads to use of automatic significance-testing procedures to select additional terms, which (as referenced above) can lead to distorted statistics.

Some strategies begin by specifying an initial model form that is flexible enough to approximate any credible model form. A flexible starting point can be less demanding than a simple one in terms of need for background information. For example, rather than concern ourselves with what the literature suggests about the shape of a dose-response curve, we can employ a starting model form that can approximate a wide range of curves. Similarly, rather than concern ourselves

with what the literature suggests about joint effects, we can employ a form that can approximate a wide range of joint effects. We can then search for a simpler but adequate model by removing from the flexible model any complexities that appear unnecessary in light of the data. Such a search process, based on simplifying a complex model, is sometimes called a *contracting* or simplifying search (Leamer 1978).

The chief drawback of a purely contracting search is that a sufficiently flexible prior model may be too complex to fit to the available data. This is because more complex models generally involve more parameters; with more parameters in a model, more data are needed to produce trustworthy point and interval estimates. Standard model-fitting methods may yield biased estimates or may completely fail to yield any estimates (e.g., not converge) if the fitted model is too complex. For example, if our flexible model for 12 variables contains all first and second-order terms, there will be 12 first-order plus 12 quadratic plus 66 product terms, for a total of 90 coefficients. Fitting this model may be well beyond what our data or computing resources can support.

Because of potential fitting problems, contracting searches begin with something much less than a fully flexible model. Some begin with a model as flexible as can be fit, or maximal model. As with minimal models, maximal models are not unique. In order to produce a model that can be fit, one may have to limit flexibility of dose-response, flexibility of joint effects, or both. It is also possible to start a model search anywhere in between the extremes of minimal and maximal models, and proceed by expanding as seems necessary and contracting as seems reasonable based on the data (although again, resource limitations usually lead to mechanical use of significance tests for this process). Unsurprisingly, such *stepwise* searches share some advantages and disadvantages with purely expanding and purely contracting searches. Like other searches, care should be taken to insure that the starting and ending points do not conflict with prior information.

The results obtained from a model search can be very sensitive to the choice of starting model. One may check for this problem by conducting several searches, starting at different models. However, there are always too many possible starting models to check them all. Thus, if one has many variables (and hence many possible models) to consider, model search strategies will always risk producing a misleading conclusion.

Model Fitting 3.9

Residual Distributions 3.9.1

Different fitting methods can lead to different estimates; thus, in presenting results one should specify the method used to derive the estimates. The vast majority of programs for risk and rate modeling use *maximum-likelihood* (ML) estimation, which is based on very specific assumptions about how the observed values of Y

tend to distribute (vary) when the vector of regressors $\boldsymbol{X}$ is fixed at a given value $\boldsymbol{x}$. This distribution is called the error distribution or *residual distribution* of Y.

If Y is the person-time rate observed at a given level $\boldsymbol{x}$ of $\boldsymbol{X}$, and T is the corresponding observed person-time, it is conventionally assumed that the number of cases observed, $A = YT$, would tend to vary according to a Poisson distribution if the person-time were fixed at its observed value. Hence, conventional ML regression analysis of person-time rates is usually called *Poisson regression*. If, on the other hand, Y is the proportion of cases observed at a given level $\boldsymbol{x}$ of $\boldsymbol{X}$ out of a person-count total N, it is conventionally assumed that the number of cases observed, $A = YN$, would tend to vary according to a *binomial distribution* if the number of persons (person count) N was fixed at its observed value. Hence, conventional ML regression analysis of prevalence or incidence proportions (average risks) is sometimes called *binomial regression*. Note that if $N = 1$, the proportion diseased Y can be only 0 or 1; in this situation, $A = YN$ can be only 0 or 14 and is said to have a Bernoulli distribution (which is just a binomial distribution with $N = 1$). The binomial distribution can be deduced from the homogeneity and independence assumptions discussed for example in Rothman and Greenland (1998, pp 232–233). As noted there, its use is inadvisable if there are important violations of either assumption, e.g., if the disease is contagious over the study period.

If Y is the number of exposed cases in a 2×2 table, the conventionally assumed distribution for Y is the *hypergeometric*; ML fitting in this situation is usually referred to as conditional maximum likelihood (CML). CML fitting is closely related to partial-likelihood methods, which are used for fitting Cox models in survival analysis.

More details on maximum-likelihood model fitting in epidemiology can be found in Breslow and Day (1980, 1987), Hosmer and Lemeshow (2000), and Clayton and Hills (1993). More general and advanced treatments of maximum likelihood can be found in many books, including Cox and Hinkley (1974) and McCullagh and Nelder (1989).

3.9.2 Overdispersion

What if the residual distribution of the observed Y does *not* follow the conventionally assumed residual distribution? Under a broad range of conditions, it can be shown that the resulting ML fitted values (ML estimates) will remain approximately unbiased if no other source of bias is present (White 1994). Nonetheless, the estimated SDs obtained from the program will be biased. In particular, if the actual variance of Y given $\boldsymbol{X} = \boldsymbol{x}$ (the *residual variance*) is larger than that implied by the conventional distribution, Y is said to suffer from *overdispersion* or *extravariation*, and the estimated standard deviations and $\boldsymbol{P}$-values obtained from an ordinary maximum-likelihood regression program will be too small.

In Poisson regression, overdispersion is sometimes called "extra-Poisson variation"; in binomial regression, overdispersion is sometimes called "extra-binomial

variation". Typically, such overdispersion arises when there is dependence among the recorded outcomes, as when the outcome Y is the number infected in a group, or Y is the number of times a person gets a disease. As an example, suppose Y is the number of eyes affected by glaucoma in an individual. In a natural population, $Y = 0$ for most people and $Y = 2$ for most of the remainder, with $Y = 1$ very infrequently. In other words, the Y values would be largely limited to the extremes of 0 and 2. In contrast, a binomially distributed variable with the same possible values (0, 1, or 2) and the same mean as Y would have a higher probability of 1 than 2, and hence a smaller variance than Y.

Two major approaches have been developed to cope with potential overdispersion, both of which are based on modeling the residual distribution. One approach is to use maximum likelihood, but with a residual distribution that allows a broader range of variation for Y, such as the negative binomial in place of the Poisson or the beta-binomial in place of the binomial (McCullagh and Nelder 1989). Such approaches can be computationally intensive, but have been implemented in some software. The second and simpler approach is to model only the residual variance of Y, rather than completely specify the residual distribution. Fitting methods that employ this approach are discussed by various authors under the topics of quasi-likelihood, pseudo-likelihood, and generalized estimating-equation (GEE) methods; see McCullagh and Nelder (1989), McCullagh (1991), and Diggle et al. (2002) for descriptions of these methods. GEE methods are often used for logitudinal data analysis (Diggle et al. 2002), but have some serious limitations in that role (Robins et al. 1999).

Sample-Size Considerations 3.9.3

One drawback of all the above fitting methods is that they depend on "large-sample" (asymptotic) approximations, which usually require that the number of parameters in the model is much less than (roughly, not more than 10% of) the number of cases observed. Methods that do not use large-sample approximations (exact methods) can also be used to fit certain models. These methods require the same strong distributional assumptions as maximum-likelihood methods. An example is exact logistic regression (Cytel 2003).

Unfortunately, exact fitting methods for incidence and prevalence models are so computationally demanding that, at the time of this writing, they can be used to fit only a narrow range of models, and do not address all the problems arising from coefficient instability in small samples (Greenland et al. 2000). *Penalized likelihood estimation* and the related methods of Stein estimation and ridge regression address these problems and permit fitting of incidence and prevalence models while retaining acceptably (though still only approximately) valid small-sample results (Efron and Morris 1975; Copas 1983; Titterington 1985; Le Cessie and van Houwelingen 1992; Greenland 1997; Rothman and Greenland 1998, pp 429–430; Greenland 2001, Greenland 2003b, Greenland and Christensen 2001).

3.10 Model Checking

It is important to check a fitted model against the data. The extent of these checks may depend on what purpose we wish the model to serve. At one extreme, we may only wish the fitted model to provide approximately valid *summary* estimates or trends for a few key relationships. For example, we might wish only to estimate the average increment in risk produced by a unit increase in exposure. At the other extreme, we may want the model to provide approximately valid *regressor-specific* predictions of outcomes, such as exposure-specific risks by age, sex, and ethnicity. The latter goal is more demanding and requires more detailed scrutiny of results, sometimes on a subject-by-subject basis.

Model diagnostics can detect discrepancies between data and a model only within the range of the data, and then only where there are enough observations to provide adequate diagnostic power. For example, there is much controversy concerning the health effects of low-dose radiation exposure (exposures that are only modestly in excess of natural background levels). This controversy arises because the natural incidence of key outcomes (such as leukemia) is low, and few cases have been observed in low-dose cohorts. As a result, several proposed dose-response models "fit the data adequately" in the low-dose region, in that each model passes the standard battery of diagnostic checks. Nonetheless, the health effects predicted by these models conflict to an important extent.

More generally, one should bear in mind that a good-fitting model is not the same as a correct model. In particular, a model may appear correct in the central range of the data, but produce grossly misleading predictions for combinations of covariate values that are poorly represented or absent in the data.

3.10.1 Tabular Checks

Both tabular methods (such as Mantel–Haenszel, Mantel and Haenszel (1959)) and regression methods produce estimates by merging assumptions about population structure (such as that of a common odds ratio or of an explicit regression model) with observed data. When an estimate is derived using a regression model, especially one with many regressors, it may become difficult to judge how much the estimate reflects the data and how much it reflects the model.

To investigate the source of results, we recommend one compare model-based results to the corresponding tabular (categorical-analysis) results. As an illustration, suppose we wish to check a logistic model in which X_1 is the exposure under study, and four other regressors X_2, X_3, X_4, X_5 appear in the model, with X_1, X_2, X_3 continuous, X_4, X_5 binary, and products among X_1, X_2, and X_4 in the model. Any regressor in a model must appear in the corresponding tabular analysis. Because X_2 and X_4 appear in products with X_1 and the model is logistic, they should be treated as modifiers of the X_1 odds ratio in the corresponding tabular analysis. X_3 and X_5 do not appear in products with X_1 and so should be treated as pure confounders (adjustment variables) in the corresponding tabular analysis. Because X_1, X_2, X_3

are continuous in the model, they must have at least three levels in the tabular analysis, so that the results can at least crudely reflect trends seen with the model. If all three of these regressors were categorized into four levels, the resulting table of disease (two levels) by all regressors would have $2 \times 4^3 \times 2^2 = 512$ cells, and perhaps many zero cells.

From this table, we would attempt to compute 3 (for exposure strata 1, 2, 3, versus 0) adjusted odds ratios (e.g., Mantel–Haenszel) for each of the $4 \times 2 = 8$ combinations of X_2 and X_4, adjusting all $3 \times 8 = 24$ odds ratios for the $4 \times 2 = 8$ pure-confounder levels. Some of these 24 adjusted odds ratios might be infinite or undefined due to small numbers, which would indicate that the corresponding regression estimates are largely model projections. Similarly, the tabular estimates might not exhibit a pattern seen in the regression estimates, which would suggest that the pattern was induced by the regression model rather than the data. For example, the regression estimates might exhibit a monotone trend with increasing exposure even if the tabular estimates did not. Interpretation of such a conflict would depend on the context: If we were certain that dose-response was monotone (e.g., smoking and esophageal cancer), the monotonicity of the regression estimates would favor their use over the tabular results; in contrast, doubts about monotonicity (e.g., as with alcohol and coronary heart disease) would lead us to use the tabular results or search for a model that did not impose monotonicity.

Tests of Regression and R^2 3.10.2

Most programs supply a "test of regression" or "test of model", which is a test of the hypothesis that all the regression coefficients (except the intercept $\boldsymbol{\alpha}$) are zero. For instance, in the exponential rate model

$$I(x) = \exp(\boldsymbol{\alpha} + \boldsymbol{x\beta}),$$

the "test of regression" provides a ***P***-value for the null hypothesis that all the components of $\boldsymbol{\beta}$ are zero, that is, that $\beta_1 = \ldots = \beta_n = 0$. Similarly, the "test of R^2" provided by linear regression programs is just a test that all the regressor coefficients are zero. A small ***P***-value from these tests suggests that the variation in outcomes observed across regressor values appears improbably large under the hypothesis that the regressors are unrelated to the outcome. Such a result suggests that at least one of the regressors is related to the outcome. It does *not*, however, imply that the model fits well or is adequate in any way.

To understand the latter point, suppose that $\boldsymbol{X}$ comprises the single indicator $X_1 = 1$ for smokers, 0 for nonsmokers, and the outcome Y is average year risk of lung cancer. In most any study of reasonable size and validity, "the test of regression" (which here is just a test of $\beta_1 = 0$) would yield a small ***P***-value. Nonetheless, the model would be inadequate to describe variation in risk, because it neglects amount smoked, age at start, and sex. More generally, a small ***P***-value from the test of regression only tells us that at least one of the regressors in the model should be included in some form or another; it does not tell us which regressor or what form

to use, nor does it tell us anything about what was left out of the model. Conversely, a large ***P***-value from the "test of regression" does not imply that all the regressors in the model are unimportant or that the model fits well. It is always possible that transformations of those regressors would result in a small ***P***-value, or that their importance cannot be discerned given the random error in the data.

A closely related mistake is interpreting the squared multiple-correlation coefficient R^2 for a regression as a goodness-of-fit measure. R^2 only indicates the proportion of Y variance that is attributable to variation in the fitted mean of Y. While $R^2 = 1$ (the largest possible value) does correspond to a perfect fit, R^2 can also be close to zero under a correct model if the residual variance of Y (i.e., the variance of Y around the true regression curve) is always close to the total variance of Y.

The preceding limitations of R^2 apply in general. Correlational measures such as R^2 can become patently absurd measures of fit or association when the regressors and regressand are discrete or bounded (Rosenthal and Rubin 1979; Greenland et al. 1986; Cox and Wermuth 1992; Greenland 1996). As an example, consider Table 3.1 showing a large association of a factor with a rare disease. The logistic model $R(x) = \text{expit}(\alpha + \beta_x)$ fits these data perfectly because it uses two parameters to describe only two proportions. Furthermore, $X = 1$ is associated with a 19-fold increase in risk. Yet the correlation coefficient for X and Y (derived using standard formulas) is only 0.09, and the R^2 for the regression is only 0.008.

Correlation coefficients and R^2 can give even more distorted impressions when multiple regressors are present (Greenland et al. 1986, 1991). For this reason, we strongly recommend against their use as measures of association or effect when modeling incidence or prevalence.

3.10.3 Tests of Fit

Tests of model fit check for nonrandom incompatibilities between the fitted regression model and the data. To do so, however, these tests must assume that the fitting method used was appropriate; in particular, test validity may be sensitive to assumptions about the residual distribution that were used in fitting. Conversely, it is possible to test assumptions about the residual distribution, but these tests usually have little power to detect violations unless a parametric regression model is assumed. Thus, useful model tests cannot be performed without making some assumptions.

Many tests of regression models are *relative*, in that they test the fit of an index model by assuming the validity of a more elaborate *reference* model that contains

Table 3.1. Hypothetical cohort data illustrating inappropriateness of R^2 for binary outcomes (see text)

	$X = 1$	$X = 0$
$Y = 1$	1900	100
Total	100,000	100,000

Risk ratio = 19, $R^2 = 0.008$.

it. A test that assumes a relatively simple reference model (i.e., one that has only a few more coefficients than the index model) will tend to have better power than a test that assumes a more complex reference model, although it will be valid only under narrower conditions.

When models are fit by maximum likelihood (ML), a standard method for testing the fit of a simpler model against a more complex model is the *deviance test*, also known as the likelihood-ratio test. Suppose that X_1 represents cumulative dose of an exposure, and that the index model we wish to test is

$$R(x_1) = \text{expit}(\alpha + \beta_1 x_1)\,,$$

a simple linear-logistic model. When we fit this model, an ML program should supply either a "residual deviance statistic" $D(\tilde{\alpha}, \tilde{\beta}_1)$, or a "model log-likelihood" $L(\tilde{\alpha}, \tilde{\beta}_1)$, where $\tilde{\alpha}, \tilde{\beta}_1$ are the ML estimates for this simple model. Suppose we wish to test the fit of the index model taking as the reference the fractional-polynomial logistic model

$$R(x_1) = \text{expit}\left(\alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2\right)\,.$$

We then fit this model and get either the residual deviance $D(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ or the log-likelihood $L(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ for the model, where $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are the ML estimates for this power model. The deviance statistic for testing the linear-logistic model against the power-logistic model (that is, for testing $\beta_2 = \beta_3 = 0$) is then

$$\Delta D(\beta_2, \beta_3) = D\left(\tilde{\alpha}, \tilde{\beta}_1\right) - D\left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)\,.$$

This statistic is related to the model log-likelihoods by the equation

$$\Delta D(\beta_2, \beta_3) = -2\left[L\left(\tilde{\alpha}, \tilde{\beta}_1\right) - L\left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)\right]$$

(McCullagh and Nelder 1989; Clayton and Hills 1993). If the linear-logistic model is correct (so that $\beta_2 = \beta_3 = 0$) and the sample is large enough, this statistic has an approximate χ^2 distribution with 2 degrees of freedom, which is the difference in the number of parameters in the two models.

A small ***P***-value from this statistic suggests that the linear-logistic model is inadequate or fits poorly; in some way, either or both the terms $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ capture deviations of the true regression from the linear-logistic model. A large ***P***-value does *not*, however, imply that the linear-logistic model is adequate or fits well; it means only that no need for the terms $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ was detected by the test. In particular, a large ***P***-value from this test leaves open the possibility that $\beta_2 x_1^{1/2}$ and $\beta_3 x_1^2$ are important for describing the true regression function, but the test failed to detect this condition; it also leaves open the possibility that some other terms not present in the reference model may be important in the same sense. These unexamined terms may involve X_1 or other regressors.

Now consider a more general description. Suppose that we wish to test an index model against a reference model in which it is nested (contained) and that this reference model contains p more unknown parameters (coefficients) than the index model. We fit both models and obtain either residual deviances of D_i and D_r for the index and reference models, or log-likelihoods L_i and L_r. If the sample is large enough and the index model is correct, the deviance statistic

$$\Delta D = D_i - D_r = -2(L_i - L_r) \tag{3.56}$$

will have an approximate χ^2 distribution with p degrees of freedom. Again, a small ***P***-value suggests that the index model does not fit well, but a large ***P***-value does not mean the index model fits well, except in the very narrow sense that the test did not detect a need for the extra terms in the reference model.

Whatever the size of the deviance ***P***-value, its validity depends on three assumptions (in addition to absence of the usual biases). First, it assumes that ML fitting of the models is appropriate; in particular, there must be enough subjects to justify use of ML to fit the reference model, and the assumed residual distribution must be correct. Second, it assumes that the reference regression model is approximately correct. Third, it assumes that the index model being tested is nested within the reference model. The third is the only assumption that is easy to check: In the previous example, we can see that the linear-logistic model is just the special case of the power-logistic model in which $\beta_2 = \beta_3 = 0$. In contrast, if we used the linear-logistic model as the index model (as above) but used the power-linear model

$$R(x_1) = \alpha + \beta_1 x_1 + \beta_2 x_1^{1/2} + \beta_3 x_1^2$$

as the reference model, the resulting deviance difference would be meaningless, because the latter model does *not* contain the linear-logistic model as a special case.

Comparison of non-nested models is a more difficult task unless the compared models have the same number of parameters. In the latter case, it has been suggested that (absent other considerations) one should choose the model with the highest loglikelihood (Walker and Rothman 1982).

3.10.4 Global Tests of Fit

One special type of deviance test of fit can be performed when Y is a proportion or rate. Suppose that, for every distinct regressor level $\boldsymbol{x}$, at least four cases would be expected if the index model were correct; also, if Y is a proportion, suppose at least four noncases would be expected if the index model were correct. (This criterion, while somewhat arbitrary, originated because it ensures that the chance of a cell count being zero is less than 2% if the cell variation is Poisson and the

index model is correct.) We can then test our index model against the *saturated* regression model

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \alpha_{\boldsymbol{x}},$$

where $\alpha_{\boldsymbol{x}}$ is a distinct parameter for every distinct observed level $\boldsymbol{x}$ of $\boldsymbol{X}$; that is, α_x may represent a different number for every level of $\boldsymbol{X}$ and may vary in any fashion as $\boldsymbol{X}$ varies. This model is so general that it contains all other regression models as special cases.

The degrees of freedom for the test of the index model against the saturated model is the number of distinct $\boldsymbol{X}$-levels (which is the number of parameters in the saturated model) minus the number of parameters in the index model, and is often called the *residual degrees of freedom* for the model. This *residual* deviance test is sometimes called a "global test of fit" because it has some power to detect any systematic incompatibility between the index model and the data. Another well-known global test of fit is the *Pearson* χ^2 *test*, which has the same degrees of freedom and sample-size requirements as the saturated-model deviance test.

Suppose we observe K distinct regressor values and we list them in some order, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K$. The statistic used for the Pearson test has the form of a residual sum-of-squares:

$$\mathrm{RSS}_{\mathrm{Pearson}} = \sum_k \left(Y_k - \hat{Y}_k\right)^2 / \hat{V}_k = \sum_k \left[\left(Y_k - \hat{Y}_k\right) / \hat{S}_k\right]^2,$$

where the sum is over all observed values $1, \ldots, K$, Y_k is the rate or risk observed at level $\boldsymbol{x}_k$, $\hat{Y}_k$ is the rate or risk predicted (fitted) at $\boldsymbol{x}_k$ by the model, $\hat{V}_k$ is the estimated variance of $\hat{Y}_k$ when $\boldsymbol{X} = \boldsymbol{x}_k$, and $\hat{S}_k = \hat{V}_k{}^{1/2}$ is the estimated standard deviation of Y_k under the model. In Poisson regression, $\hat{Y}_k = \exp(\hat{\alpha} + \boldsymbol{x}_k\hat{\boldsymbol{\beta}})$ and $\hat{V}_k = \hat{Y}_k/T_k$, where T_k is the person-time observed at $\boldsymbol{x}_k$; in binomial regression, $\hat{Y}_k = \mathrm{expit}(\hat{\alpha} + \boldsymbol{x}_k\hat{\boldsymbol{\beta}})$ and $\hat{V}_k = \hat{Y}_k(1 - \hat{Y}_k)/N_k$, where N_k is the number of persons observed at $\boldsymbol{x}_k$. The quantity $(Y_k - \hat{Y}_k)/\hat{S}_k$ is sometimes called the *standardized residual* at level $\boldsymbol{x}_k$; it is the distance between Y_k and $\hat{Y}_k$ expressed in units of the estimated standard deviation of Y_k under the model.

Other global tests have been proposed that have fewer degrees of freedom and less restrictive sample-size requirements than the deviance and Pearson tests (Hosmer and Lemeshow 2000). A major drawback of all global tests of fit, however, is their low power to detect model problems (Hosmer et al. 1997). If any of the tests yields a low $\boldsymbol{P}$-value, we can be confident the tested (index) model is unsatisfactory and needs modification or replacement (albeit the tests provide no clue as to how to proceed). If, however, they all yield a high $\boldsymbol{P}$-value, it does not mean the model is satisfactory. In fact, the tests are unlikely to detect any but the most gross conflicts between the fitted model and the data. Therefore, global tests should be regarded as crude preliminary screening tests only, to allow quick rejection of grossly unsatisfactory models.

The deviance and Pearson statistics are sometimes used directly as measures of distance between the data and the model. Such use is most easily seen for the

Pearson statistic. The second form of the Pearson statistic shows that it is the sum of squared standardized residuals; in other words, it is a sum of squared distances between data values and model-fitted values of Y. The deviance and Pearson global test statistics can also be transformed into measures of prediction error under the model; for example, see McCullagh and Nelder (1989) and Hosmer and Lemeshow (2000).

3.10.5 Model Diagnostics

Suppose now we have found a model that has passed preliminary checks such as tests for additional terms and global tests of fit. Before adopting this model as a source of estimates, it is wise to further check the model against the basic data, and assess the trustworthiness of any model-based inferences we wish to draw. Such activity is subsumed under the topic of *model diagnostics*, and its subsidiary topics of residual analysis, influence analysis, and model-sensitivity analysis. These topics are vast, and we can only mention a few approaches here. In particular, we neglect the classical topic of residual analysis, largely because its proper usage involves a number of technical complexities when dealing with the censored data and nonlinear models predominant in epidemiology (McCullagh and Nelder 1989). Detailed treatments of diagnostics for such models can be found in Breslow and Day (1987), Hosmer and Lemeshow (2000), and McCullagh and Nelder (1989).

3.10.6 Delta-Beta Analysis

One important and simple diagnostic tool available in some packaged software is *delta-beta* ($\Delta\beta$) *analysis*. For a data set with N subjects total, estimated model coefficients (or approximations to them) are recomputed N times over, each time deleting exactly one of the subjects from the model fitting. Alternatively, for individually-matched data comprising N matched sets, the delta-beta analysis may be done deleting one set at a time. In either approach, the output is N different sets of coefficients estimates: These sets are then examined to see if anyone subject or matched set influences the resulting estimates to an unusual extent.

To illustrate, suppose our objective is to estimate the rate-ratio per unit increase in an exposure X_1, to be measured by $\exp(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the estimated exposure coefficient in an exponential-rate model. For each subject, the entire model (confounders included) is re-fit without that subject. Let $\hat{\beta}_{1(-i)}$ be the estimate of $\hat{\beta}_1$ obtained when subject i is excluded from the data. The difference $\hat{\beta}_{1(-i)} - \hat{\beta}_1 \equiv \Delta\hat{\beta}_{1(-i)}$ is called the *delta-beta* for β_1 for subject i. The influence of subject i on the results can be assessed in several ways. One way is to examine the impact on the rate-ratio estimate. The proportionate change in the estimate from dropping subject i is

$$\exp\left(\hat{\beta}_{1(-i)}\right) / \exp\left(\hat{\beta}_1\right) = \exp\left(\hat{\beta}_{1(-i)} - \hat{\beta}_1\right) = \exp\left(\Delta\hat{\beta}_{1(-i)}\right) ,$$

for which a value of 1.30 indicates dropping subject i increases the estimate by 30%, and a value of 0.90 indicates dropping subject i decreases the estimate by 10%. One

can also assess the impact of dropping the subject on confidence limits, ***P***-values, or any other quantity of interest.

Some packages compute "standardized" delta-betas, $\Delta\hat{\beta}_{1(-i)}/\hat{s}_1$ where $\hat{s}_1$ is the estimated standard deviation for $\hat{\beta}_1$. By analogy with Z-statistics, any standardized delta-beta below −1.96 or above 1.96 is sometimes interpreted as being unusual. This interpretation can be misleading, however, because the standard deviation used in the denominator is not that of the delta-beta. A standardized delta-beta is only a measure of the influence of an observation expressed in SE units.

It is possible that one or a few subjects or matched sets are so influential that deleting them alters the conclusions of the study, even when N is in the hundreds (Pregibon 1981). In such situations, comparison of the records of those subjects to others may reveal unusual combinations of regressor values among those subjects. Such unusual combinations may arise from previously undetected data errors, and should at least lead to enhanced caution in interpretation. For instance, it may be only mildly unusual to see a woman who reports having had a child at age 45 or a woman who reports natural menopause at age 45. The combination in one subject, however, may arouse suspicion of a data error in one or both regressors, a suspicion worth the labor of further data scrutiny if that woman or her matched set disproportionately influences the results.

Delta-beta analysis must be replaced by a more complex analysis if the exposure of interest appears in multiple model terms, such as indicator terms, power terms, product terms, or spline terms. In that situation, one must focus on changes in estimates of specific effects or summaries, for example, changes in estimated risk ratios.

Conclusions

3.11

This chapter has reviewed basic principles and forms of parametric regression models and model fitting. Regression analysis is a vast subject, however, and many topics and details have been omitted. For further reading on fundamentals of parametric modeling a standard text is McCullagh and Nelder (1989). A standard introduction to nonparametric regression is Hastie and Tibshirani (1990). Nonparametric methods are connected to algorithmic modeling (machine learning) methods; for a comparison of parametric and algorithmic approaches see Breiman (2001). For an integrated coverage of parametric, nonparametric, and algorithmic methods see Hastie et al. (2001).

References

Agresti A (2002) Categorical data analysis. Wiley, New York

Bancroft TA, Han C-P (1977) Inference based on conditional specification: A note and a bibliography. Int Stat Rev 45:117–127

Berk R (2004) Regression analysis: A constructive critique. Sage publications, Thousand Oaks, CA
Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA
Breiman L (2001) Statistical modeling: The two cultures (with discussion). Statistical Science 16:199–231
Breslow NE, Day NE (1980) Statistical methods in cancer research. Vol I: the analysis of case-control data. IARC, Lyon
Breslow NE, Day NE (1987) Statistical methods in cancer research. Vol II: the design and analysis of cohort studies. IARC, Lyon
Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman and Hall, New York
Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, New York
Cole SR, Ananth CV (2001) Regression models for unconstrained, partially or fully constrained continuation odds ratios. Int J Epidemiol 30:1379–1382
Copas JB (1983) Regression, prediction, and shrinkage (with discussion). J Royal Stat Soc B 45:311–354
Cox DR (1972) Regression models and life tables (with discussions). J Royal Stat Soc B 34:187–220
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, New York
Cox DR, Oakes D (1984) Analysis of survival data. Chapman and Hall, New York
Cox DR, Wermuth N (1992) A comment on the coefficient of determination for binary responses. Am Statist 46:1–4
Cytel Corporation. LogXact Version 5 (software) (2003) Cytel Corp., Cambridge, MA
Diggle PJ, Heagerty P, Liang KY, Zeger SL (2002) The analysis of longitudinal data, 2nd edn. Oxford University Press, New York
Efron B, Morris CN (1975) Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70:311–319
Flack VF, Chang PC (1987) Frequency of selecting noise variables in subset regression analysis: a simulation study. Am Statist 41:84–86
Freedman DA (1983) A note on screening regression equations. Am Statist 37:152–155
Good IJ (1983) Good thinking: The foundations of probability and its applications. University of Minnesota Press, Minneapolis, MN
Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: A roughness penalty approach. Chapman and Hall, New York
Greenland S (1993) Basic problems in interaction assessment. Environ Health Perspect 101(suppl 4):59–66
Greenland S (1994) Alternative models for ordinal logistic regression. Stat Med 13:1665–1677
Greenland S (1995a) Dose-response and trend analysis: Alternatives to categorical analysis. Epidemiology 6:356–365

Greenland S (1995b) Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 6:450–454
Greenland S (1995c) Problems in the average-risk interpretation of categorical dose-response analyses. Epidemiology 6:563–565
Greenland S (1996) A lower bound for the correlation of exponentiated bivariate normal pairs. Am Statist 50:163–164
Greenland S (1997) Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses. Stat Med 16:515–526
Greenland S (2000a) Principles of multilevel modeling. Int J Epidemiol 29:158–167
Greenland S (2000b) When should epidemiologic regressions use random coefficients? Biometrics 56:915–921
Greenland S (2001) Putting background information about relative risks into conjugate prior distributions. Biometrics 57:663–70
Greenland S (2003a) The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields towards childhood leukemia. J Am Stat Assoc 98:47–54
Greenland S (2003b) Generalized conjugate priors for Bayesian analysis of risk and survival regressions. Biometrics 59:92–99
Greenland S, Christensen R (2001) Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. Stat Med 20:2421–2428
Greenland S, Maldonado G (1994) The interpretation of multiplicative-model parameters as standardized parameters. Stat Med 13:989–999
Greenland S, Schlesselman JJ, Criqui MH (1986) The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am J Epidemiol 123:203–208
Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H (1991) Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology 2:387–392
Greenland S, Schwartbaum JA, Finkle WD (2000) Problems from small samples and sparse data in conditional logistic regression. Am J Epidemiol 151:531–539
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, New York
Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York
Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16:965–980
Hurvich DM, Tsai CL (1990) The impact of model selection on inference in linear regression. Am Statist 44:214–217
Lagakos SW (1988) Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 7:257–274
Lash TL, Fink AK (2003) Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology 14:451–458

Le Cessie S, van Houwelingen HC (1992) Ridge estimators in logistic regression. Appl Stat 41:191–201
Leamer EE (1978) Specification searches: Ad hoc interference with nonexperimental data. Wiley, New York
Maclure M (1993) Demonstration of deductive meta-analysis: Ethanol intake and risk of myocardial infarction. Epidemiol Rev 15:328–351
Maclure M, Greenland S (1992) Tests for trend and dose response: Misinterpretations and alternatives. Am J Epidemiol 135:96–104
Maldonado G, Greenland S (1993a) Interpreting model coefficients when the true model form is unknown. Epidemiology 4:310–318
Maldonado G, Greenland S (1993b) Simulation study of confounder-selection strategies. Am J Epidemiol 138:923–936
Maldonado G, Greenland S (1994) A comparison of the performance of model-based confidence intervals when the correct model form is unknown: coverage of asymptotic means. Epidemiology 5:171–182
Mantel N, Haenszel WH (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719–748
McCullagh P (1991) Quasi-likelihood and estimating functions. In: Hinkley DV, Reid NM, Snell EJ (eds) Statistical theory and modelling. Chapman and Hall, London, Chap. 11
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York
Michels KB, Greenland S, Rosner BA (1998) Does body mass index adequately capture the relation of body composition and body size to health outcomes? Am J Epidemiol 147:167–172
Moolgavkar SH, Venzon DJ (1987) General relative risk regression models for epidemiologic studies. Am J Epidemiol 126:949–961
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710
Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9:705–724
Robins JM, Greenland S (1986) The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 123:392–402
Robins JM, Greenland S (1994) Adjusting for differential rates of prophylaxis therapy for PCP in high versus low dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc 89:737–749
Robins JM, Blevins D, Ritter G, Wulfsohn M (1992) G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3:319–336. Errata: Epidemiology (1993) 4:189
Robins JM, Greenland S, Hu FC (1999) Estimation of the causal effect of time-varying exposure on the marginal mean of a repeated binary outcome. J Am Stat Assoc 94:687–712
Rosenthal R, Rubin DB (1979) A note on percent variance explained as a measure of importance of effects. J Appl Psychol 9:395–396
Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott, Philadelphia

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Stat 43:425–467

Sclove SL, Morris C, Radhakrishna R (1972) Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. Ann Math Stat 43:1481–1490

Sheehe P (1962) Dynamic risk analysis of matched-pair studies of disease. Biometrics 18:323–341

Strömberg U (1996) Collapsing ordered outcome categories: a note of concern. Am J Epidemiol 144:421–424

Titterington DM (1985) Common structure of smoothing techniques in statistics. Int Stat Rev 53:141–170

Walker AM, Rothman KJ (1982) Models of varying parametric form in case-referent studies. Am J Epidemiol 115:129–137

Weiss RE (1995) The influence of variable selection: A Bayesian diagnostic perspective. J Am Stat Assoc 90:619–625

White H (1994) Estimation, inference, and specification analysis. Cambridge University Press, New York