

3. In an analysis of the effect of oral contraceptives on stroke based on the data in Table 11-2, suppose that you were interested in the oral-contraceptive effect and wished only to control for possible confounding by hypertension using stratification. What would be the stratum-specific risk ratio estimates for oral contraceptive use for the two strata of hypertension? In an ordinary stratified analysis, why is there a separate referent category in each stratum?

4. Show that if there is an excess over a multiplicative effect among those with joint exposure to two causes, there will also be an excess over an additive effect.

5. The investigators of the study described in Table 11-2 claimed that women who faced increased risk from one risk factor ought to avoid additional risk from another risk factor, regardless of whether the two factors interacted in the causation of the disease. Does this suggestion make sense? What would it imply about seat belt use for women who take oral contraceptives?

6. List reasons why the study of biologic interaction is more difficult than the study of the effects of single factors.

## REFERENCES

1. Ahlbom A, Alfredsson L. Interaction: a word with two meanings creates confusion. *Eur J Epidemiol.* 2005;20:563-564.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:Chapter 5, Concepts of Interaction.
3. Collaborative Group for the Study of Stroke in Young Women. Oral contraceptives and stroke in young women. *JAMA.* 1975;231:718-722.
4. Knol MJ, Vanderweele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.

## Using Regression Models in Epidemiologic Analysis

The straight line depicted in Figure 12-1 is an example of a simple mathematical model. It is a model because we use the mathematical equation for the straight line that is fitted to the data as a way of describing the relation between the two variables in the graph, in this case cigarette smoking and laryngeal cancer mortality. Models in epidemiology are used for various purposes, the two primary ones being to make predictions and to control for confounding. Prediction models are used to estimate risk (or other epidemiologic measures) based on information from risk predictors. For example, an equation can be used to estimate a person's risk of heart attack during a 10-year period based on information about the person's age, sex, family history, blood pressure, smoking history, weight, height, exercise habits, and medical history. Values for each of these predictors could be inserted into an equation that predicts the risk of heart attack from the combination of risk factors. The model must have terms in it for all the risk factors listed.

In contrast to the goal of risk prediction for specific people, much of epidemiologic research is aimed at learning about the causal role of specific factors for disease. In causal research, regression models are used to evaluate the causal role of one or more factors while simultaneously controlling for possible confounding effects of other factors. Because this use of multivariable regression models differs from the use of models to obtain estimates of risk for people, there are different considerations that apply to the construction of multivariable models for causal research. Unfortunately, many courses in statistics do not distinguish between the use of regression models for prediction of individual risk and the use of such models for causal inference.

The data in Figure 12-1 illustrate an almost perfect linear relation between the number of cigarettes smoked per day and the age-standardized mortality rate of laryngeal cancer. Seldom do epidemiologic data conform to such a striking linear pattern. The line drawn through the data points is a *regression line*, meaning that



3. In an analysis of the effect of oral contraceptives on stroke based on the data in Table 11-2, suppose that you were interested in the oral-contraceptive effect and wished only to control for possible confounding by hypertension using stratification. What would be the stratum-specific risk ratio estimates for oral contraceptive use for the two strata of hypertension? In an ordinary stratified analysis, why is there a separate referent category in each stratum?
4. Show that if there is an excess over a multiplicative effect among those with joint exposure to two causes, there will also be an excess over an additive effect.
5. The investigators of the study described in Table 11-2 claimed that women who faced increased risk from one risk factor ought to avoid additional risk from another risk factor, regardless of whether the two factors interacted in the causation of the disease. Does this suggestion make sense? What would it imply about seat belt use for women who take oral contraceptives?
6. List reasons why the study of biologic interaction is more difficult than the study of the effects of single factors.

## REFERENCES

1. Ahlbom A, Alfredsson L. Interaction: a word with two meanings creates confusion. *Eur J Epidemiol.* 2005;20:563-564.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:Chapter 5, Concepts of Interaction.
3. Collaborative Group for the Study of Stroke in Young Women. Oral contraceptives and stroke in young women. *JAMA.* 1975;231:718-722.
4. Knol MJ, Vanderweele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.

## Using Regression Models in Epidemiologic Analysis

The straight line depicted in Figure 12-1 is an example of a simple mathematical model. It is a model because we use the mathematical equation for the straight line that is fitted to the data as a way of describing the relation between the two variables in the graph, in this case cigarette smoking and laryngeal cancer mortality. Models in epidemiology are used for various purposes, the two primary ones being to make predictions and to control for confounding. Prediction models are used to estimate risk (or other epidemiologic measures) based on information from risk predictors. For example, an equation can be used to estimate a person's risk of heart attack during a 10-year period based on information about the person's age, sex, family history, blood pressure, smoking history, weight, height, exercise habits, and medical history. Values for each of these predictors could be inserted into an equation that predicts the risk of heart attack from the combination of risk factors. The model must have terms in it for all the risk factors listed.

In contrast to the goal of risk prediction for specific people, much of epidemiologic research is aimed at learning about the causal role of specific factors for disease. In causal research, regression models are used to evaluate the causal role of one or more factors while simultaneously controlling for possible confounding effects of other factors. Because this use of multivariable regression models differs from the use of models to obtain estimates of risk for people, there are different considerations that apply to the construction of multivariable models for causal research. Unfortunately, many courses in statistics do not distinguish between the use of regression models for prediction of individual risk and the use of such models for causal inference.

The data in Figure 12-1 illustrate an almost perfect linear relation between the number of cigarettes smoked per day and the age-standardized mortality rate of laryngeal cancer. Seldom do epidemiologic data conform to such a striking linear pattern. The line drawn through the data points is a *regression line*, meaning that



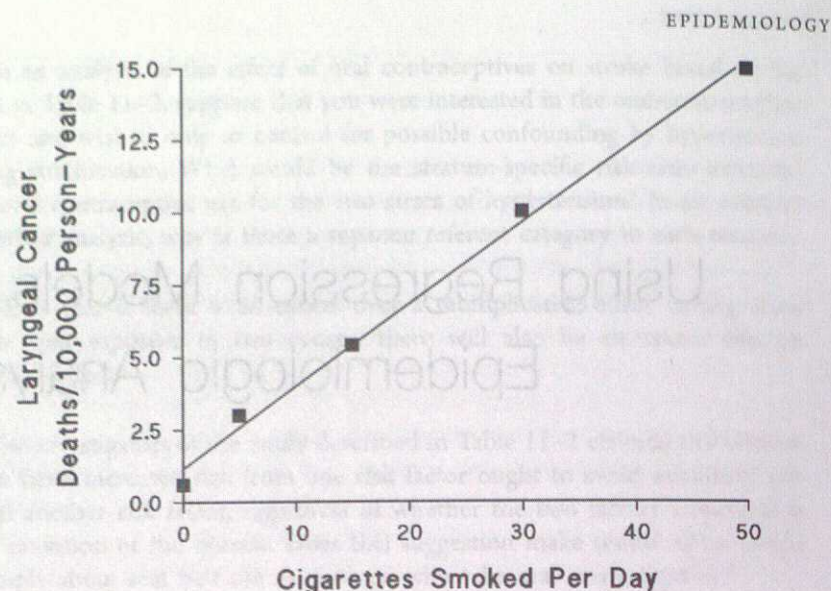


Figure 12-1 Age-standardized mortality from laryngeal cancer according to number of cigarettes smoked daily, derived from data of Kahn.<sup>1</sup> (Adapted from Rothman et al<sup>2</sup> with permission of the *American Journal of Epidemiology*.)

it estimates average values for the variable on the vertical scale ( $Y$ ) according to values of the variable on the horizontal scale ( $X$ ). In this case, it is a *simple regression* because it can be described as a single straight line in the following form:

$$\hat{Y} = a_0 + a_1X$$

$\hat{Y}$  (often called "Y-hat") is the estimated value of  $Y$  for any given value of  $X$ ;  $a_0$  is the intercept, or the value of  $\hat{Y}$  when  $X$  is zero; and  $a_1$  is the coefficient of  $X$ , which describes the slope of the line, or the number of units of change in  $\hat{Y}$  for every unit change in  $X$ . In the figure,  $\hat{Y}$  is the age-standardized mortality rate from laryngeal cancer, and  $X$  is the number of cigarettes smoked daily.

The equation for the regression line in Figure 12-1 is  $\hat{Y} = 1.15 + 0.282X$ . These values refer to deaths per 10,000 person-years. The intercept, 1.15, represents the number of deaths per 10,000 person-years that are estimated to occur in the absence of cigarette smoking. There is also a direct observation for the rate at the zero level of smoking, which is 0.6 deaths per 10,000 person-years. The regression line estimates a slightly larger value, 1.15, than the value that was observed; this estimate is based not just on the zero level for smoking but on all five data points. The regression line slope, 0.282, indicates that the number of deaths per 10,000 person-years is estimated to increase by 0.282 for every additional cigarette smoked daily.

Assuming that confounding and other biases have been properly addressed, the slope value of 0.282 quantifies the effect of cigarette smoking on death from laryngeal cancer. The regression line also allows us to estimate mortality rate ratios at different smoking levels. For example, from the regression line, we can estimate the mortality rate among those who smoke 50 cigarettes daily (equivalent to 2.5 packs/day) to be 15.2 deaths per 10,000 person-years. Compared with

the estimated rate among nonsmokers of 1.15 deaths per 10,000 person-years, the estimated rate ratio for smoking 2.5 packs daily is  $15.2/1.15 = 13.3$ . Put in these terms, we can readily see that the regression coefficient indicates a strong effect of smoking on laryngeal cancer mortality.

## THE GENERAL LINEAR MODEL

Models that incorporate terms for more than one factor at a time can be used as an alternative to stratification to achieve control of confounding. These models succeed in controlling confounding because when several risk factors are included, the effect of each is unconfounded by the others. Let us consider an extension of the simple linear model in Figure 12-1 to a third variable.

$$\hat{Y} = a_0 + a_1X_1 + a_2X_2 \quad [12-1]$$

Equation 12-1, like the one for Figure 12-1, has the same outcome variable,  $\hat{Y}$  (also known as the *dependent* variable), but there are now two predictor variables,  $X_1$  and  $X_2$ , which are referred to as *independent* variables. Suppose that  $Y$  is the mortality rate from laryngeal cancer, as in Figure 12-1, and that  $X_1$ , as before, is the number of cigarettes smoked daily. The new variable,  $X_2$ , might be the number of grams of alcohol consumed daily (alcohol is also a risk factor for laryngeal cancer). With two independent variables and one dependent variable, the data points must now be visualized as being located within a three-dimensional space: two dimensions for the two independent variables and one dimension for the dependent variable. Imagine a room in which the edge of the floor against one wall is the axis for  $X_1$  and the edge where the adjacent wall meets the floor is the axis for  $X_2$ . The line from floor to ceiling where these two adjacent walls meet would be the  $Y$  axis. Equation 12-1 is a straight line through the three-dimensional space of this room.

What is the advantage of adding the term  $X_2$  to the model? Ordinarily, because cigarette smoking and alcohol consumption are correlated, we might expect that cigarette smoking and alcohol drinking would be mutually confounding risk factors for laryngeal cancer. A stratified analysis could remove that confounding, but the confounding can also be removed by fitting Equation 12-1 to the data. In a model such as Equation 12-1 with terms for two predictive factors, smoking ( $X_1$ ) and alcohol ( $X_2$ ), the coefficients for these terms,  $a_1$  and  $a_2$  respectively, provide estimates of the effects of cigarette smoking and alcohol drinking that are mutually unconfounded. Mathematically, there is no limit to the number of terms that could be included as independent variables in a model, although limitations of the data provide a practical limit. The general form of Equation 12-1 is referred to as the *general linear model*.

## TRANSFORMING THE GENERAL LINEAR MODEL

The dependent variable in a regression model is not constrained mathematically to any specific range of values. In actual epidemiologic applications, however, the



dependent variable might be constrained in various ways. For example, the dependent variable might be  $FEV_1$  (forced expiratory volume in 1 second), a measure of lung function that cannot be negative. As another example, the dependent variable might be the occurrence of disease, which is measured as either *no* or *yes* and is usually assigned a value of 0 or 1. This dichotomy is a highly constrained variable, because only two values are observable, and the estimates would theoretically be constrained to be within the range [0,1]. It is common when using constrained outcome variables to use a transformation to avoid getting impossible values for the dependent variable. For example, the straight line in Figure 12-1 has an intercept of 1.15 deaths per 10,000 person-years. With only slightly different data points, however, it would have been possible to have the line cross the Y-axis at a value less than 0, implying a negative mortality rate for nonsmokers. A negative mortality rate is impossible, but mathematically there is nothing in the fitting of a straight line that confines the line to positive territory.

How could we fit a model for rate data that avoids the possibility of the dependent variable taking negative values? We can transform the data to confine the line to positive territory. One way to achieve that is to fit the straight line to the logarithm of the mortality rate rather than to the mortality rate itself:

$$\ln(\hat{Y}) = a_0 + a_1X_1 + a_2X_2 \quad [12-2]$$

where  $\ln(\hat{Y})$  is the natural logarithm of  $\hat{Y}$ . In Equation 12-2, the left side can range from minus infinity to plus infinity, as can the right side, but  $\hat{Y}$  itself must always be positive because one cannot take the logarithm of a negative number. This equation can be solved for  $\hat{Y}$  by taking the antilogarithm of both sides, giving

$$\hat{Y} = e^{a_0 + a_1X_1 + a_2X_2} \quad [12-3]$$

Equation 12-3 allows only positive values for  $\hat{Y}$ . On the other hand, to achieve this nicety, we no longer have a simple linear model but an exponential model instead.

Having an exponential model has some implications for the interpretation of the coefficients. Consider again the simple linear model in Figure 12-1. The slope, 0.282, is a measure of the absolute amount of increase in the death rate from laryngeal cancer with each additional cigarette smoked per day. If a similar model were applied to an exposure that was measured on a dichotomous scale, with the "unexposed" condition assigned a value of 0 and the "exposed" condition assigned a value of 1, the coefficient  $a_1$  would correspond to the rate difference between the exposed and unexposed states, which can be determined by subtracting the equation given that a person is unexposed (when  $X = 0$ ) from the equation given that a person is exposed (when  $X = 1$ ).

$$\text{exposed } (X = 1): \quad \hat{Y}_e = a_0 + a_1X = a_0 + a_1$$

$$\text{unexposed } (X = 0): \quad \hat{Y}_u = a_0 + a_1X = a_0$$

$$\text{exposed-unexposed:} \quad \hat{Y}_e - \hat{Y}_u = a_1$$

Thus, without any transformation,  $a_1$  can be interpreted as an estimate of the rate difference between exposed and unexposed persons. If, however, we use the logarithmic transformation that is shown in Equations 12-2 and 12-3, we find that the coefficient  $a_1$  in that model is not interpretable as a rate difference:

$$\text{exposed:} \quad \ln(\hat{Y}_e) = a_0 + a_1X = a_0 + a_1$$

$$\text{unexposed:} \quad \ln(\hat{Y}_u) = a_0 + a_1X = a_0$$

$$\text{difference:} \quad \ln(\hat{Y}_e) - \ln(\hat{Y}_u) = a_1$$

$$\text{ratio:} \quad \frac{\hat{Y}_e}{\hat{Y}_u} = e^{a_1}$$

Rather, the antilogarithm of the coefficient (which is what you get when you raise the constant  $e$  to the power of the coefficient) is the rate ratio of exposed to unexposed persons. Thus, the transformation that provides for the good behavior of the predictions from the model with respect to avoiding negative rate estimates also has an implication for the interpretation of the coefficient. Without the transformation, the coefficient estimates rate differences; with the transformation, the coefficient estimates rate ratios (after exponentiating).

## THE LOGISTIC TRANSFORMATION

Suppose that we had data for which the dependent variable was a risk measure. Whereas rates are never negative but can go as high as infinity, risks are mathematically confined to the narrower range, [0,1]. For any straight line with nonzero slope,  $Y$  ranges from minus infinity to plus infinity rather than from 0 to 1. Consequently, a straight line model without transformation could lead to individual predicted risk values that are either negative or greater than 1. There is a commonly used transformation, however, the *logistic transformation*, that will confine the predicted risk values to the proper range.

It is easier to understand the logistic transformation if we think of it as two transformations. The first converts the risk measure,  $R$ , to a transformed measure that ranges from zero to infinity instead of [0,1]. This transformation is accomplished by using  $R/(1-R)$  in place of  $R$ . For values of  $R$  near 0, the quantity  $R/(1-R)$  will be little different from  $R$ , but as  $R$  approaches 1, the denominator of the transformed value approaches 0 and the ratio  $R/(1-R)$  approaches infinity. Thus, this transformation raises the upper end of the range from 1 to infinity. The quantity  $R/(1-R)$  is called the *risk-odds* (any proportion divided by its complement is an odds). The second transformation converts the risk-odds to a measure that ranges all the way from minus infinity to plus infinity. That transformation is the same as the one used previously for incidence rates: one simply takes the logarithm of the risk-odds. The resulting measure, after both transformations, is  $\ln[R/(1-R)]$ , a quantity that is called a "logit." The two-step transformation is known as the logistic transformation.



The logistic model is one in which the logit is the dependent variable of a straight-line equation:

$$\ln \left[ \frac{R}{1-R} \right] = a_0 + a_1 X \quad [12-4]$$

Equation 12-4 shows only a single independent variable, but, just as in other linear models, it is possible to add additional independent variables to the model, making it a "multiple logistic" model. What is the interpretation of the coefficient  $a_1$  in this model? For an  $X$  that is dichotomous (ie,  $X = 1$  for exposed and  $X = 0$  for unexposed), the coefficient  $a_1$  is the ratio of logits for exposed relative to unexposed. This ratio is equal to the logarithm of the risk-odds ratio:

$$\ln \left[ \frac{R_1}{1-R_1} \right] - \ln \left[ \frac{R_0}{1-R_0} \right] = \ln \left[ \frac{\frac{R_1}{1-R_1}}{\frac{R_0}{1-R_0}} \right] = \ln \left[ \frac{R_1(1-R_0)}{R_0(1-R_1)} \right] = a_1 \quad [12-5]$$

This result means that, in the logistic model, the antilogarithm of the coefficient of a dichotomous exposure term estimates the odds ratio of risks.

$$\frac{R_1(1-R_0)}{R_0(1-R_1)} = e^{a_1}$$

As a consequence of this interpretation for the logistic coefficient, the logistic model has become a popular tool for the analysis of case-control studies, in which the odds ratio is the primary statistic of interest.

## CHOICES AMONG MODELS

From a mathematical perspective, the advantages of these transformations are tied to the mathematical behavior of the measures, ensuring that individual estimates from the models conform to the allowed range. From a practical standpoint, however, the transformations dictate what type of measure the coefficients in the model will estimate. If one has risk data and wishes to estimate risk difference, the logistic model will not conveniently provide it will provide odds ratios. If one is using a model to obtain risk estimates for people, it may be important to avoid getting estimates of risk that are negative or greater than 100%, because these are invalid estimates. On the other hand, if the model is being used primarily to assess an overall effect of the exposure from the coefficient in the fitted model, there may be less concern about whether all the individual estimates stay within their allowable ranges and more interest in which effect measure the model can provide. In many epidemiologic applications, it is the choice among effect measures that dictates the type of model the investigator ought to use.

Consider the data in Table 12-1, which describe the risk of acquiring a hypothetical disease over a 5-year period according to the subject's age at the start of the period. Twenty subjects were followed for 5 years, and each either did or did

Table 12-1 Risk  
OF DEVELOPING A  
HYPOTHETICAL DISEASE  
DURING A 5-YEAR PERIOD  
FOR 20 SUBJECTS

Subject No.	Age	Disease
1	18	0
2	21	0
3	22	0
4	25	0
5	26	0
6	28	0
7	33	0
8	34	0
9	35	0
10	37	0
11	42	1
12	47	1
13	55	0
14	56	1
15	58	0
16	61	1
17	65	1
18	68	1
19	75	1
20	77	1

not develop the disease. These data are plotted as a scatterplot in Figure 12-2. In a scatterplot with a binary outcome variable that takes values of either 0 or 1, all the observations fall either at 0 or at 1 on the vertical axis. Figure 12-2 also shows the linear regression line through the 20 data points and its equation. The intercept of the regression line is the estimated value of the risk for those with age 0. The value of the intercept is -0.49, an impossible value for a risk. In fact, the line estimates a negative risk for all ages less than 24 years and a risk greater than 100% for all ages greater than 74 years.

One can avoid the inadmissible risk estimates from the regression line in Figure 12-2 by fitting a logistic model instead of a straight line. The logistic model for the same data from Table 12-1 is illustrated in Figure 12-3. Its sigmoid shape is characteristic of the logistic curve. This shape keeps the curve within the range [0,1] for any age, preventing the impossible estimates that come from the linear model in Figure 12-2.

It might appear from a comparison of these two figures that the logistic model is always preferable for risk data, because it cannot result in estimates of risk that are inadmissible (ie, either less than zero or greater than one). This example is presented, however, to make the point that the logistic model is not always preferable. The age coefficient in the straight-line equation in Figure 12-2 is interpretable as a risk difference for each year of age: it indicates that the risk increases by an estimated 2% for each year of age. It is true that the fitted straight



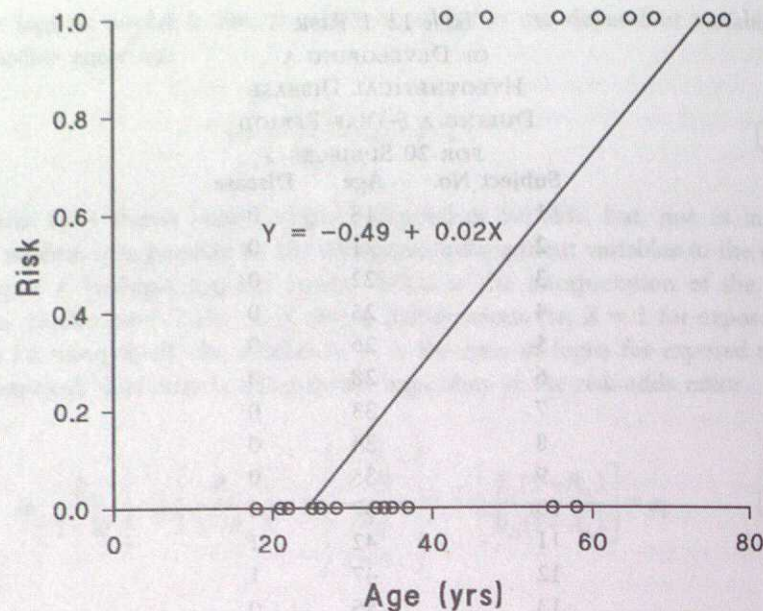


Figure 12-2 Scatterplot of risk data from Table 12-1 and a linear regression line fitted to the data.

line is not expected to fit the data well outside the central region of the graph. Nevertheless, for this central region in the middle of the age span, the straight line provides a simple and useful way of estimating the risk difference for each year of age. In contrast, the logistic model in Figure 12-3 does not permit direct estimation of a risk difference. Instead, it allows estimation of an odds ratio associated with a 1-year increase in age, from the antilogarithm of the logistic coefficient:  $e^{0.144} = 1.15$ , which is the risk-odds ratio for each 1-year increase in age. Although there is nothing fundamentally wrong with estimating the odds ratio, the straight-line model may be preferable if one wishes to estimate a risk difference. As mentioned earlier, the logistic model is particularly appropriate for the analysis of case-control studies because the odds ratio can be obtained from it, and the odds ratio is the statistic of central interest for estimating rate ratios in case-control studies.

#### CONTROL OF CONFOUNDING WITH REGRESSION MODELS

One of the principal advantages of multivariable regression models for epidemiologic analysis is the ease with which several confounding variables can be controlled simultaneously. In a multivariable regression model, the inclusion of several variables results in a model in which each term is unconfounded by the other terms in the model. This approach makes it easy and efficient to control confounding by several variables at once, something that might be difficult to achieve through a stratified analysis.

For example, as described in Chapter 10, suppose that you were conducting an analysis with five confounding variables, each of which had three categories.

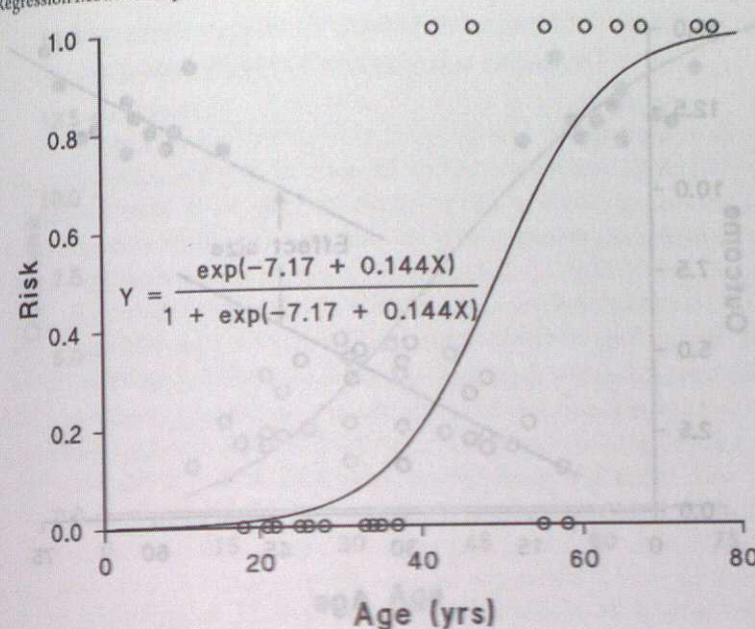


Figure 12-3 Scatterplot of risk data from Table 12-1 and a logistic regression line fitted to the data.

To control for these variables in a stratified analysis, you would need to create three strata for the first variable, then divide each of those three strata into three more substrata for the second variable, giving nine strata, and so on until there are  $3 \times 3 \times 3 \times 3 \times 3 = 243$  strata. If any of the variables required more than three categories, or if there were more than five variables to control, the number of strata would rise accordingly. With such a large number of strata required for a stratified analysis to control several variables, the data could easily become uninformative because there are too few subjects within each stratum to give useful estimates. Multivariable regression modeling solves this problem by allowing a much more efficient way to control for several variables simultaneously. Everything has its price, however, and so it is for multivariable regression analysis. The price is that the results from the regression model are readily susceptible to bias if the model is not a good fit to the data.

To illustrate the problem, consider the hypothetical data in Figure 12-4, with data points for exposed and unexposed people by age and by some unspecified but continuous outcome measure. These data show an unfortunate situation in which there is no overlap in the age distributions between the exposed population and unexposed population. If a stratified analysis were undertaken to control for age, there would be little or no overlap in any age category between exposed and unexposed groups, and the stratified analysis would produce no estimate of effect. In essence, a stratified analysis attempting to control for age would give the result that there is no information in the data.

In contrast, in a multiple regression with both age and exposure terms, the model will essentially fit two parallel straight lines through the data, one relating age to the outcome for unexposed people and the other relating age to the outcome for exposed people. If the dichotomous exposure term is coded 0 or 1, at



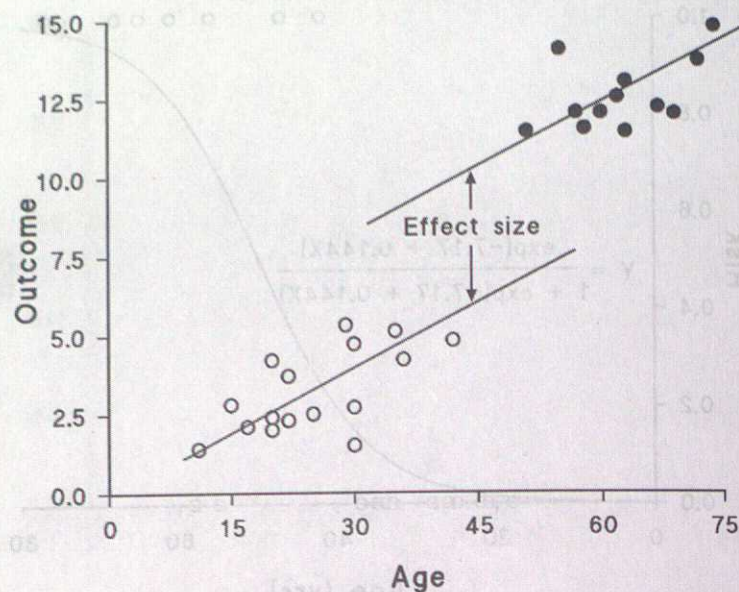


Figure 12-4 Hypothetical example of a multivariable linear regression of outcome data involving a dichotomous exposure variable (exposed = solid circles, unexposed = open circles) and age.

any age the difference in the outcome between exposed and unexposed is equal to the coefficient for the exposure, which measures the exposure effect:

$$\text{Outcome} = a_0 + a_1 \cdot \text{Exposure} + a_2 \cdot \text{Age}$$

Thus the regression model produces a statistically stable estimate from the nonoverlapping sets of data points. Basically, the model extrapolates the age relation for the unexposed and exposed persons and estimates the effect from the extrapolated lines, as indicated in Figure 12-4. This estimation process is much more efficient than a stratified analysis, which for these data would not produce any effect estimate at all.

But what if the actual relation between age and the outcome were as pictured in Figure 12-5? If age has the curvilinear relation pictured there, there is no effect of exposure on the outcome, and the effect estimated from the model depicted in Figure 12-4 would be incorrect. The apparent effect in Figure 12-4 is simply a bias introduced by the model and its inappropriate extrapolation beyond the observations. Because we cannot know whether the model in Figure 12-4 is appropriate or whether the relation is actually like the pattern depicted in Figure 12-5, the lack of results from a stratified analysis of these data begins to look good compared with the regression analysis, which might produce an incorrect result. Saying nothing is better than saying something incorrect.

Stratified analysis has other advantages over regression analysis. With stratified analysis, both the investigator and the reader (if the stratified data are presented) are aware of the distribution of the data by the key study variables. When examining the output from a multiple regression analysis, on the other hand, the reader

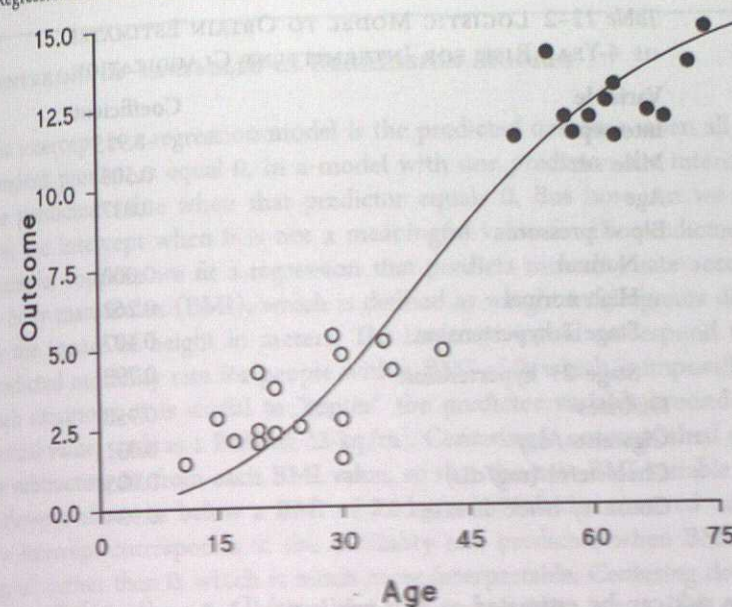


Figure 12-5 A possible age-outcome curve for the data in Figure 12-4 (exposed = solid circles, unexposed = open circles).

is typically in the dark, and the researcher may not be much better off than the reader. For this reason, a multivariable regression analysis should be used only as a supplement to a stratified analysis, rather than as the primary analytic tool. Unfortunately, many researchers tend to leap into regression modeling without arming themselves first with the knowledge that would come from a stratified analysis. Typically, the reader is also deprived and is presented with no more than the coefficients from the regression model. This approach has the lure of seeming sophisticated, but it is a mistake to plunge into regression modeling until one has viewed the distribution of the data and analyzed them according to the methods presented in Chapter 10.

## PREDICTING RISK FOR A PERSON

Much of the advice on how to construct a regression model is crafted for models aimed at making individual predictions regarding the outcome of interest. For example, Murabito et al<sup>3</sup> published a logistic model that provided 4-year risk estimates for intermittent claudication (the symptomatic expression of atherosclerosis in the lower extremities). The model is given in Table 12-2.

To obtain individual risk estimates from this model, one would multiply the coefficients for each variable in the table by the values for a given person for each variable, which gives the logit for a given person. Because the exponentiated logit equals the risk-odds,  $\exp(\text{logit}) = [R/(1-R)]$ , the logit can be converted to a risk estimate by taking into account the relation between risk and risk-odds:

$$\text{Odds} = \frac{\text{Risk}}{1 - \text{Risk}} \quad \text{or} \quad \text{Risk} = \frac{\text{Odds}}{1 + \text{Odds}}$$



Table 12-2 LOGISTIC MODEL TO OBTAIN ESTIMATES OF 4-YEAR RISK FOR INTERMITTENT CLAUDICATION

Variable	Coefficient
Intercept	-8.915
Male sex	0.503
Age	0.037
Blood pressure	
Normal	0.000
High-normal	0.262
Stage 1 hypertension	0.407
Stage 2+ hypertension	0.798
Diabetes	0.950
Cigarettes/day	0.031
Cholesterol (mg/dL)	0.005
Coronary heart disease	0.994

Thus, the risk can be estimated as  $R = \exp(\text{logit}) / [1 + \exp(\text{logit})]$ . For example, suppose we wish to estimate the 4-year risk of intermittent claudication for a 70-year-old nonsmoking man who has normal blood pressure, diabetes, coronary heart disease, and a cholesterol level of 250 mg/dL. The logit would be  $-8.915 + 1 \cdot 0.503 + 70 \cdot 0.037 + 0 \cdot 0.000 + 1 \cdot 0.950 + 0 \cdot 0.031 + 250 \cdot 0.005 + 1 \cdot 0.994 = -2.628$ , and the risk estimate over the next 4 years for intermittent claudication to develop would be  $\exp(-2.628) / [1 + \exp(-2.628)] = 6.7\%$ . If the man had stage 2 hypertension instead of normal blood pressure, the logit would be  $-1.830$  and the risk estimate would be 13.8%.

In a model such as the one in Table 12-2, the purpose of including each individual term in the model is to improve the estimate of risk. To produce a useful risk estimate, it does not matter whether any of the predictor terms is causally related to the outcome. In the model in Table 12-2, some of the predictors cannot be viewed as causes: age is an example of a noncausal predictor, as is heart disease, which presumably does not cause intermittent claudication, although the two diseases may have causes in common. Nevertheless, despite not being causes of intermittent claudication, both age and the presence of coronary heart disease are good predictors of the risk of developing intermittent claudication; therefore, it makes sense to include them in the prediction model. Other predictors, such as cigarette smoking, hypertension, and diabetes, may be causes of intermittent claudication.

## STRATEGY FOR CONSTRUCTING REGRESSION MODELS FOR EPIDEMIOLOGIC ANALYSIS

Although a detailed discussion of the strategy for constructing multivariable regression models for epidemiologic analysis goes beyond the scope of this book, I outline here some basic principles for the use of these models in causal research.

## CENTERING OF VARIABLES IN REGRESSION MODELS

The intercept in a regression model is the predicted outcome when all independent predictors equal 0. In a model with one predictor, the intercept is the predicted value when that predictor equals 0. But how can we interpret the intercept when 0 is not a meaningful value for the predictor? For example, suppose we fit a regression that predicts mortality rate according to body mass index (BMI), which is defined as weight in kilograms divided by the square of height in meters. The intercept would correspond to the predicted mortality rate for people with a BMI of 0, which is impossible. In such situations, it is useful to "center" the predictor variable around some central value, such as a BMI of 22 kg/m<sup>2</sup>. Centering is accomplished simply by subtracting 22 from each BMI value, so that the new BMI variable is the difference above or below a BMI of 22 kg/m<sup>2</sup>. With the centered variable, the intercept corresponds to the mortality rate predicted when BMI is 22 kg/m<sup>2</sup>, rather than 0, which is much more interpretable. Centering does not affect the basic interrelationships of the study variables, but it does make it easier to interpret the coefficients, especially when there are product terms in the model that would further complicate the interpretations.

## DO A STRATIFIED ANALYSIS FIRST

The first step should always be a stratified analysis. The main contribution of a multivariable regression analysis to causal research is to enable the simultaneous control of several confounding factors. In accomplishing this goal, multivariable modeling ought to be thought of as a supplement to stratified analysis, to be used in situations in which there are too many confounders to be handled comfortably in a stratified analysis. Even in those circumstances, it is common that most of the confounding stems from one or two variables and a multivariable regression model will give essentially the same result as a properly conducted stratified analysis.

## DETERMINE WHICH CONFOUNDERS TO INCLUDE IN THE MODEL

Start with a set of predictors of the outcome based on the strength of their relation to the outcome, as indicated from analyses of each factor separately or from a preliminary model in which all potential confounders are included. Then, build a model by introducing predictor variables one at a time. After each term is introduced, examine the amount of change in the coefficient of the exposure term. If the exposure coefficient changes considerably (most investigators look for a 10% change), then the variable just added to the model is a confounder (provided that it also meets the conditions for a confounder described in Chapter 7); if not, it is not an important confounder. To judge the confounding effect in this way, it is essential for the exposure to be included in the model as a single term. For example, if the exposure is cigarette smoking, one might enter a single term that quantifies the amount of cigarette smoking rather than several terms for levels of cigarette smoking. It is likewise important to avoid any product terms that



### STEPWISE MODELS IN EPIDEMIOLOGIC ANALYSIS

Stepwise construction of regression models is accomplished by an algorithm that automatically selects which terms to include in the final model. The algorithm typically selects terms based on the level of statistical significance of the coefficient for each term. Stepwise modeling makes much more sense for the construction of a prediction model than for a causal model. As discussed in Chapter 8, statistical testing does not allow us to grasp either the strength of a relation or the precision of an estimate in isolation; it mixes the two. Using statistical significance levels to determine which potential confounders to include in a model is a bad idea, whether it is part of an automatic stepwise algorithm or not, for several reasons. First, the amount of confounding depends on two associations—the relation between the potential confounder and the exposure, and the relation between the potential confounder and the outcome. The coefficient that is tested for significance in a stepwise algorithm evaluates only the latter relation; it ignores the relation between the potential confounder and the exposure. Therefore, this method can result in the inclusion of variables that are not confounding. It can also omit variables that are confounding but for which the relation with the outcome is not “statistically significant.”

involve cigarette smoking (or whatever the exposure variable is) at this stage of the analysis.

### ESTIMATE THE SHAPE OF THE EXPOSURE-DISEASE RELATION

If the exposure is a simple dichotomy, one can estimate the exposure effect directly from the coefficient of the exposure term after the confounders have been entered into the model. If, however, the exposure is a continuous variable, such as the number of cigarettes smoked daily, the exposure term needs to be redefined after the confounders are entered into the model. The reason for redefining the exposure term is that the single exposure term that was in the model for the purpose of evaluating confounding will not reveal the shape of the exposure-disease relation for a continuous exposure variable. If the model involves a logarithmic transformation, as do most of the models commonly used in epidemiologic analysis, a single term for a continuous exposure variable will be mathematically constrained to take the shape dictated by the model. In a logistic model, the exposure coefficient is the logarithm of the odds ratio for a unit change in the exposure variable. If the exposure is the number of cigarettes smoked daily, the coefficient of a single term that corresponds to the number of cigarettes smoked daily would be the logarithm of the odds ratio for each single cigarette smoked. Because there is only a single term, the model dictates that the effect of each cigarette multiplies the odds ratio by a constant amount. The result is an exponential dose-response curve between exposure and disease (Fig. 12-6).

This exponential shape will be fit to the data regardless of the actual shape of the relation between exposure and disease, as long as the exposure variable is continuous and confined to a single term in a model that uses a logarithmic

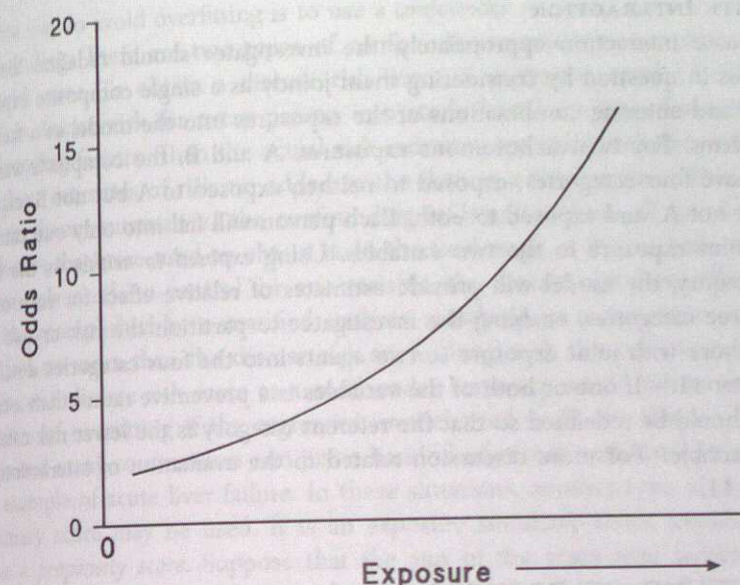


Figure 12-6 Shape of a positive dose-response relation between exposure and disease from a logistic model with a single continuous exposure term.

transformation. In linear models, a linear relation, rather than an exponential relation, will be guaranteed. The problem is that the actual relation set by nature may be nothing like the shape of the relation posed by the model. Indeed, few dose-response relations in nature look like the curve in Figure 12-6.

To avoid this difficulty of having the model dictate the shape of the dose-response relation, the investigator can allow the data, rather than the model, to determine the shape. To accomplish this goal, the investigator must redefine the exposure term in the model. One popular approach is *factoring* the exposure. Factoring here refers to categorizing the exposure into ranges and then creating a separate term for each subrange of exposure, except for one category that becomes by default the reference value. For example, cigarette smoking could be categorized as zero cigarettes per day, 1 to 9 per day, 10 to 19 per day, and so on. The model would have a term for each cigarette-smoking category except 0, which is the reference category. The variable corresponding to each term would be dichotomous, simply revealing in which category a given person fell. Each smoker would have a value of 1 for the smoking category that applied and a value of 0 for every other category; a nonsmoker would have a 0 for all smoking categories. The resulting set of coefficients in the fitted model indicate a separately estimated effect for each level of smoking, determined by the data and not by the mathematics of the model.

Another approach for estimating the shape of the dose-response trend is to use curve-smoothing methods, such as *spline regression*, which allow a different fitted curve to apply in different ranges of the exposure variable.

The important point in evaluating dose-response relations is to avoid letting the model determine the shape of the relation between exposure and disease. Whether one uses factoring, splines, or other smoothing methods, it is desirable to allow the data, not the model, to define the shape of the dose-response curve.



### EVALUATE INTERACTION

To evaluate interaction appropriately, the investigator should redefine the two exposures in question by considering them jointly as a single composite exposure variable and entering combinations of the exposures into the model as a factored set of terms. For two dichotomous exposures, A and B, the composite variable would have four categories: exposed to neither, exposed to A but not B, exposed to B but not A, and exposed to both. Each person will fall into only one category of the joint exposure to the two variables. Using *exposed to neither* as the reference category, the model will provide estimates of relative effect for each of the other three categories, enabling the investigator to partition the risk or risk ratio among those with joint exposure to two agents into the four categories described in Chapter 11.<sup>4,5</sup> If one or both of the variables is a preventive rather than a cause, then it should be redefined so that the referent category is the lower risk category of the variable.<sup>6</sup> For more discussion related to the evaluation of interaction, see Chapter 11.

### OVERFITTING OF REGRESSION MODELS AND SUMMARY CONFOUNDER SCORES

The great advantage of regression models for epidemiologic analysis is the ability to control simultaneously for several confounding variables. Very often, there is little confounding from most of the potential confounders in a body of data, apart from one or two that exhibit moderate or strong confounding. When it is not necessary to control for more than one or two variables, it is often advisable to present the results of a stratified analysis as the primary findings. With several confounders that all contribute at least moderate amounts of confounding, however, some multivariable regression approach is preferred. In extreme cases, there may be a large number of variables that all contribute substantial confounding, and in some of these situations, it may not be feasible to fit a regression model. For example, in a cohort study with a small number of outcomes, the sparsity of outcomes may pose a problem. One rule of thumb that is commonly cited is that there should be at least 10 or 15 observations for every term in a regression model. With fewer observations than that, the fitted model may be *overfitted*, which means that the model will be too heavily influenced by random error in the data.<sup>7,8</sup>

In most epidemiologic studies, "observations" are equivalent to people, but in many studies, it is not the total number of people that serves as the limiting factor but the number of people with the study outcome. Suppose that you conducted a cohort study in 10,000,000 people aimed at learning about the causes of acute liver failure. This is a large cohort, but because acute liver failure is rare, in 1 year you might observe only 70 cases. This small number of cases is the limiting number for determining how many terms could be included in a regression model without overfitting. With 70 cases, the maximum that the model can reasonably accommodate will be about 7 to 10 terms. The number of variables may be less than the number of terms. For example, if a variable such as age is categorized into 10-year age categories, age alone may require several terms in the model, one for each category apart from the referent category. Therefore, the number of variables that can be accommodated may be quite limited if overfitting is to be avoided.

One way to avoid overfitting is to use a *confounder summary score* to control for confounding. There are two types of confounder summary scores. One is a *disease risk score*. To obtain a disease risk score, a regression model is fitted that predicts disease risk for every person in a study based on the information from confounding factors. Then the actual risk estimates are calculated for each person, and these estimates of risk are added to the data as a new variable. This new variable in theory summarizes the confounding information from all the disease risk predictors that were used to obtain it. In the final stage of the analysis, the investigator only needs to control for one variable, the disease risk score. This control can be accomplished by a stratified analysis, matching, or a regression model that contains no more than the disease risk score along with the exposure variable.<sup>9</sup>

Use of a disease risk score as a confounder summary would not overcome the problem of overfitting if the regression model used to fit the disease risk score contained many confounders and just a handful of people with the outcome, as in the example of acute liver failure. In these situations, another type of confounder summary score may be used. It is an exposure summary score, usually referred to as a *propensity score*. Suppose that the aim of the acute liver failure study is to examine the relation between administration of antibiotics and the development of acute liver failure. To get a summary exposure score, one could fit a regression model that predicts whether each person would receive an antibiotic. Overfitting that model is less of a problem than overfitting a model that predicts risk of acute liver failure: although few people get acute liver failure, many people use antibiotics. From the regression model that predicts antibiotic exposure, one would calculate a propensity score for each person in the study and treat that score as a summary confounder. Then, in the final analysis, the investigator would need to control for only a single variable, the propensity score, which effectively will control for all the component variables used to estimate the propensity score. Because exposures typically are not as rare as many disease outcomes, propensity scores can often be used to control for a large number of confounding factors in a regression model without the difficulty of overfitting the model.

Using a summary confounding score amounts to more work than fitting the typical multivariable regression model, and despite the greater effort it does not always appear to result in substantially better control of confounding.<sup>10</sup> Nevertheless, the use of summary confounder scores, especially propensity scores, is increasing. One reason, as just described, is the ability to control for numerous confounders when there are few outcome events in the data. Another value in using a confounder summary score is the ability to examine the range of confounder scores for all subjects and to exclude outlier subjects who are outside the range common to both exposed and unexposed subjects (a procedure called *trimming*). Suppose one looked at the age distribution of exposed and unexposed subjects and discovered that exposed subjects tend to be older, with the oldest exposed subjects having ages well above those of the oldest unexposed subjects. On the other end of the age distribution, there may be unexposed subjects younger than any exposed subjects. It would be good practice to restrict the age distributions of exposed and unexposed subjects to the range common to both. Doing so would reduce residual confounding by age. The same approach could be used with a confounder summary score, likewise reducing residual confounding (Fig. 12-7). The advantage of trimming the distributions based on a summary confounder score is that one needs to



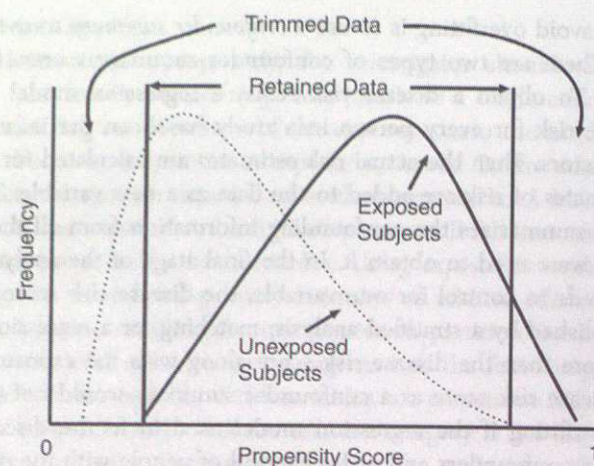


Figure 12-7 Trimming of subjects outside the range of propensity scores that is common to both exposed and unexposed subjects.

trim for only that one variable. Trimming for age may reduce residual confounding by age, but it does not affect residual confounding by other variables. Trimming for several variables may control residual confounding at the price of losing much of the data. Trimming the data once according to a summary confounder score reduces residual confounding much more effectively than trimming by a single variable, and it conserves data by making the trimming process efficient.<sup>11</sup>

There is an important caution to keep in mind when selecting variables for propensity score models. When regression modeling is taught, students are often told that the aim is to put into the model the best predictors of the outcome. That may be true for regression models that are used to make individual predictions. When the purpose of the model is to control confounding, however, as it is in propensity score models, putting in the best predictors of the outcome may be undesirable. Very strong predictors of exposure may or may not be confounders, depending on their relation to the disease. If a strong predictor of exposure is unrelated to the disease, including it in a propensity score regression model will not accomplish any control of confounding. Including such a variable will, however, lead to separation of the exposed and unexposed distributions, as depicted in Figure 12-7, resulting in loss of data that are useful for direct comparison and ultimately resulting in wider confidence intervals. Therefore, it is important to avoid including strong predictors of exposure in a propensity score regression model unless they are also confounders, which means that they are also predictors of disease and that they are not in the causal pathway between exposure and disease.

#### EXAMPLE OF USING PROPENSITY SCORES: ARE DRUG-ELUTING STENTS BETTER THAN BARE-METAL STENTS?

In pharmacoepidemiology, the use of propensity scores to model the probability of treatment as a summary confounder score has become a dominant method for the control of confounding. This method was employed by Mauri et al<sup>13</sup> in a 2008

#### VARIABLE MATCHING RATIOS, CONFOUNDING, AND TRIMMING

Suppose you were conducting a cohort study of treated (exposed) and untreated (unexposed) patients in which there was expected to be substantial confounding by indication. Accordingly, you have decided to compute a propensity score for each subject to control for confounding. One option in the data analysis is to match an untreated patient to every treated patient, ending up with two cohorts that are matched by their propensity scores. All of the variables that went into the propensity score model should be adequately controlled in the comparison between the treated cohort and the individually matched untreated cohort. In fact, one can usually demonstrate that the compared cohorts show a balance for all the risk factors that went into the propensity score model, even if they were severely unbalanced before the matching process.<sup>12</sup> Such a demonstration can persuade skeptics that a propensity score model can indeed control effectively for many confounding variables simultaneously. Another advantage of matching is that it automatically achieves the trimming illustrated in Figure 12-7.

Matching one untreated subject to each treated subject comes at a cost, however. Many subjects may have propensity scores that put them in the range to be matched, but simply do not get matched. They are omitted from the analysis, resulting in a loss of information and leading to a wider confidence interval. These subjects could be retained in the study if either a stratified analysis or a regression model were used to analyze the data. Alternatively, the matching could be expanded so that instead of matching only one unexposed person with each exposed person, all unexposed persons who have approximately the same propensity score as the exposed person would be included. Including as many subjects as possible that match on propensity score produces a variable matching ratio rather than a fixed ratio of untreated patients for every treated patient. A variable matching ratio simply means that the ratio of untreated to treated subjects varies across matched sets. A drawback to a variable matching ratio is that one cannot display a simple table that shows the desired balance between all treated and untreated subjects for each of the variables that goes into the propensity score. Such overall balance will result only if the matching ratio is fixed, yielding the same number of unexposed people for each exposed person. A table that demonstrates balance is a useful way to persuade skeptics that use of the propensity score has successfully controlled for all its component variables that could be confounding, but this persuasive demonstration is not possible when the matching ratio varies across matched sets.

Rather than exclude subjects from the study who could have been included simply to obtain a table showing a balance of potential confounders, one might consider using a two-step process: first, select matched pairs (ie, using a fixed matching ratio of 1) to produce a table that shows balance for the individual variables in the propensity score model; second, add back into the data those subjects who could have been matched but were excluded to keep the matching ratio to a value of 1. The first step shows



that the propensity score achieves balance between the cohorts, because it uses a fixed matching ratio. The second step expands the comparison group by allowing a variable matching ratio, which avoids the loss of information from those who would have been omitted from the matched-pair analysis.

The above discussion of a cohort study does not apply to a case-control study. For case-control studies, matching on any variable related to exposure induces a selection bias that must be controlled in the analysis (see Chapter 7). Showing balance between cases and controls is no reassurance that the bias has been removed. The different behavior of matching in cohort and case-control studies derives from the fact that the matching in cohort studies is between exposed and unexposed subjects, whereas in case-control studies it is between subjects who have disease (the cases) and those at risk for disease (the controls)—an entirely different phenomenon.

study on the comparative safety of two different kinds of stents. Stents are tubular wire cages used to keep arteries patent after narrowed vessels have been widened by angioplasty. The authors identified adults undergoing stenting at Massachusetts hospitals for acute myocardial infarction during an 18-month period. They then measured the risk of death over the 2 years following insertion of the stent, comparing the risk for patients receiving bare-metal stents and those receiving drug-eluting stents. The latter stents are coated with medication to prevent scar tissue formation within the artery walls. Some previous studies, but not all, had indicated that patients with drug-eluting stents had a greater risk of eventual cardiac complications and death than patients receiving bare-metal stents.

The study by Mauri et al.<sup>13</sup> showed some differences in the characteristics of patients receiving the two kinds of stents. These differences were consistent with the explanation that a greater proportion of those receiving bare-metal stents were treated under emergency conditions. It is possible that drug-eluting stents were more widely used in big referral centers, whereas patients with an immediate cardiac crisis were more likely to be treated at local hospitals that were more likely to use bare-metal stents. Of course, this difference in prognosis would bias the study results unless it could be controlled adequately in the analysis. To attempt to control for these confounding differences, the authors calculated the propensity for each patient to receive a drug-eluting stent rather than a bare-metal stent, and then they matched patients receiving one kind of stent to patients receiving the other kind according to their propensity score.

No analytic method is perfect in controlling for confounding, because of unmeasured confounders or imperfect measurement of identified confounders. Mauri et al.<sup>13</sup> proposed an ingenious way to gauge the effectiveness of control of confounding in their study. The risk of death after a myocardial infarction is highest during the first few days and then declines gradually over time. Based on the mechanism of action, drug elution from a stent is thought to have no effect during the first days after placement of the stent. Therefore, the researchers compared the risk of death among patients receiving the two types of stents during the 2 days following insertion of the stent, a period in which the type of stent should make no difference. If the propensity score model were effective in controlling for

confounding, then one would expect to see the same risk of death over the 2 days following insertion for patients receiving drug-eluting versus bare-metal stents. In fact, however, the 2-day risk for those receiving a bare-metal stent was almost double the 2-day risk for those receiving a drug-eluting stent: 1.2% versus 0.7%. This difference indicates that matching on the propensity score did not balance the risk factors that predict death between the two groups of patients.

Unfortunately, Mauri et al. incorrectly focused on the lack of statistical significance of the difference in 2-day risk of death between the two study cohorts. The *P* value was 0.06, but using statistical significance to assess a difference is a poor approach to interpreting the data, as explained in Chapter 8. In this situation, the authors' mistake was even more profound, because the question at hand was not whether the difference between the cohorts might have been compatible with chance. Rather, the question was the size of the imbalance in risk factors and how much it biased the final results of the study. The authors claimed that their study showed a lower risk of death over 2 years for patients receiving drug-eluting stents. After control of confounding, they found that the 2-year risk of death was 10.7% among patients receiving a drug-eluting stent and 12.8% (20% greater) among those receiving a bare-metal stent. These conclusions, however, ignored the analysis that demonstrated residual confounding, which indicated almost double the risk for short-term death among the bare-metal stent group. If the confounding that was evident during the mortality experience of the first 2 days stemmed from risk factors that persisted over the next 2 years, then we would expect that, from confounding alone, the risk over 2 years would be almost twice as high for patients receiving bare-metal stents as for those receiving drug-eluting stents. Because the risk of death for patients receiving bare-metal stents was in fact much less than twice as high (only 20% higher), one can conclude that these data indicate that it was actually much safer to receive a bare-metal stent, a conclusion opposite to that drawn by the authors.

One problem in this interesting case study was the authors' focus on the *P* value instead of the magnitude of the risk imbalance that they reported. Another problem surfaced in later published correspondence,<sup>14,15</sup> when the authors suggested that the differences in 2-day risks were unimportant for another reason. They dismissed the greater risk of death over the first 2 days for patients getting bare-metal stents because the risk difference for death over 2 days was 0.5%, small in relation to the difference of 2.1% in the other direction seen over 2 years. The value of a risk, however, is cumulative over the time period for which the risk is measured (see Chapter 4). The risks and the corresponding risk differences for the first 2 days are necessarily smaller than the risks that accumulate over 2 years. Suppose that the patients receiving bare-metal stents had a greater risk over 2 days because they were older than those receiving drug-eluting stents. The age difference would continue to contribute to the risk difference between the two cohorts over the next 2 years, with the cumulative effect being roughly proportional to the period of time. Differences in propensity score would likely work similarly.

To adjust the study findings for the discrepancy seen in the 2-day risk, one would need to project the difference in risk over 2 days to the full 2 years, which requires using the proportionality of the risks as an adjustment factor. Over 2 days, those getting bare-metal stents had a risk that was 73% higher than the



risk for those getting drug-eluting stents. Over 2 years, the risk observed in the bare-metal stent group was 20% higher than in the drug-eluting stent group. If the confounding alone would have led to a risk that was 73% higher, it seems that these data indicate that getting the bare-metal stent was considerably safer than getting the drug-eluting stent. After using the ratio of risks over the first 2 days to adjust the risk ratio at 2 years, one can then convert that 2-year risk ratio to a risk difference with some simple assumptions that we will not elaborate here. Using such methods, one can infer that based on the study of Mauri et al.<sup>13</sup> the bare-metal stent patients had an *absolute* risk of death that was actually 4.4% lower over 2 years than that of the drug-eluting stent patients, rather than the 2.1% greater risk the authors reported.

## SUMMARY

Regression models are extremely useful in epidemiologic analysis, both for predicting disease risk and for controlling confounding, especially when there are several important confounders that must be controlled simultaneously. Many advanced techniques, which are beyond the scope of this text, rely on different types of regression models in various ways. Nevertheless, there is a case to be made that epidemiologists have demonstrated an overreliance on regression models. As suggested earlier in this chapter, in almost all situations a stratified analysis should be undertaken as a first approach to analyzing the data in an epidemiologic study. If there are several important confounders, it may not be possible for all of them to be controlled simultaneously, but it is rare that there are more than one or two substantial confounders. Even when there are, a stratified analysis can be conducted that will measure and control confounding for each variable singly, a process that informs the multivariable regressions done subsequently. Furthermore, it may be feasible to control for two or three confounders with simple stratification and to get a good estimate of the exposure effect from the tables used for this analysis.

A great strength of a stratified analysis is that the data are revealed for all to peruse. To capitalize on that advantage, when stratified analyses are presented, the researcher should include more than just the summary results. The tables with the stratified data should also be presented. In this way, readers will have access to the key data from which unconfounded effect estimates can be calculated, and this approach keeps everyone well informed about the data. Presenting these tables will lead to fewer mistakes.

Another strength of stratified analysis is that for cohort studies it lends itself more readily than regression models to presentation of exposure-specific rates or risks. For example, one can use standardization to obtain exposure-specific rates of disease from a cohort. From those standardized rates, it is easy to calculate standardized rate differences and rate ratios. In contrast, most regression models are limited to one effect measure and do not offer estimation of exposure-specific rates or risks. Thus, although epidemiologists are taught that an advantage of cohort studies is the ability to measure absolute rates or risks rather than just relative measures, many cohort studies are analyzed and reported using regression models that provide estimates solely of ratio measures, with no information

on difference measures or actual rates or risks. This approach negates one of the important advantages of cohort studies. Using stratified analysis as the primary analysis will avert this problem.

In many cases, there will be enough recorded information on variables that have confounding effects so that fitting a regression model will ultimately be useful, after a thorough stratified analysis. The results from the regression model in most situations should be well anticipated by a preliminary stratified analysis. The regression results should be presented in the published work or final report only to the extent that they represent an important refinement of the findings. Rather than being the first analysis and often the only analysis presented, regression models should ordinarily reinforce what has already been shown.

## QUESTIONS

1. In a multivariable regression model with a nominal scale variable that has three categories, how many indicator terms would need to be included? In general, for a variable with  $n$  categories, what is the expression for the number of terms that would need to be included in the model?
2. The analysis depicted in Figure 12-4 is more efficient than a stratified analysis but also more biased. Why is it more biased?
3. Why is an exponential curve, such as the one in Figure 12-6, not a reasonable model for the shape of a dose-response trend? What would be the biologic implication of a dose-response curve that had the shape of the curve in Figure 12-6?
4. If the age term is removed from the model shown in Table 12-2, what would happen to the coefficients for blood pressure? Why?
5. In a regression model with a continuous exposure variable, why is it desirable to have a single exposure term in the model when evaluating confounding?
6. If we have a continuous exposure variable and use a single exposure term to evaluate confounding, the shape of the dose-response curve for that term will be implied by the model. That imposition can be avoided by factoring the exposure into several terms defined by categories of the exposure. The use of several exposure terms, however, will make it difficult to evaluate confounding. How can we evaluate confounding and also avoid the model-imposed restrictions on the shape of the dose-response curve?

## REFERENCES

1. Kahn HA. The Dorn study of smoking and mortality among U.S. Veterans: report on eight and one-half years of observation. National Cancer Institute Monograph 19. US DHEW, Public Health Service, 1966.



2. Rothman KJ, Cann CI, Flanders D, Fried MP. Epidemiology of laryngeal cancer. *Epidemiol Rev.* 1980;2:195-209.
3. Murabito JM, D'Agostino RB, Sibershatz H, Wilson PWF. Intermittent claudication: a risk profile from the Framingham Heart Study. *Circulation.* 1997;96:44-49.
4. Rothman KJ. *Modern Epidemiology.* 1st ed. Boston: Little, Brown; 1986.
5. Assman SE, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology.* 1996;7:286-290.
6. Knol MJ, Vanderweele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.
7. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med.* 2004;66:411-421.
8. Green SB. How many subjects does it take to do a regression analysis? *Multivar Behav Res.* 1991;26:499-510.
9. Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol.* 1976;104:609-620.
10. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol.* 2006;59:437-447.
11. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol.* 2010;172:843-854.
12. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.* 2005;14:465-476.
13. Mauri L, Silbaugh TS, Garg P, et al. Drug-eluting or bare-metal stents for acute myocardial infarction. *N Engl J Med.* 2008;359:1330-1342.
14. Rothman KJ. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301.
15. Mauri L, Normand S-LT. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301-302.

## Epidemiology in Clinical Settings

Clinical epidemiology focuses the application of epidemiologic principles on questions that relate to diagnosis, prognosis, and therapy. It also encompasses screening and other aspects of preventive medicine at both the population and the individual level. Therapeutic thinking has been greatly affected by advances in pharmacoepidemiology, an area that has extended the reach of epidemiologic research from the study of drug benefits to that of adverse effects and has led to the burgeoning fields of *outcomes research* and *comparative effectiveness*. Outcomes research marries epidemiologic methods with clinical decision theory to determine which therapeutic approaches are the most cost-effective, whereas comparative effectiveness aims to evaluate the effect of different interventions against one another in a variety of settings.

### DIAGNOSIS

Assigning a diagnosis is both crucial and subtle. To a large extent, the process of diagnosis may appear to involve intuition, conviction, and guesswork, processes that are opaque to quantification and analysis. Nevertheless, formal approaches to understanding and refining the steps in assigning a diagnosis have helped to clarify the thinking and solidify the foundation for diagnostic decision making. The basis for formulating a diagnosis comprises the data from signs, symptoms, and diagnostic test results that distinguish those with a specific disease from those who do not have that disease.

### The Gold Standard

Diagnosis cannot be a perfect process. Rarely does any sign or symptom, or any combination of them, distinguish completely between those with and those without a disease. Often a diagnosis is considered established when a specific combination of signs and symptoms that has been posed as the criterion for disease