MATHEMATICAL MODELING STRATEGIES FOR THE ANALYSIS OF EPIDEMIOLOGIC RESEARCH

Raymond S. Greenberg
Atlanta Cancer Surveillance Center and Department of Community Health and Biometry, Emory University, Decatur, Georgia 30030

David G. Kleinbaum
Department of Biostatistics and Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27514

INTRODUCTION

In the inaugural volume of the Annual Review of Public Health, Reuel Stallones suggested the following central axiom of epidemiology (27): “Disease does not distribute randomly in human populations.” As a corollary to this axiom, Professor Stallones stated: “Variations in the frequency of human disease occur in response to variations in the intensity of exposure to etiologic agents or other more remote causes, or to variations in the susceptibility of individuals to the operation of these causes.”

The principal objective of many epidemiologic studies is to evaluate the association between exposure to a single risk factor and the occurrence of a specific disease. In this type of investigation, it is essential to isolate the effect of interest from the effects of other risk factors. The effects of relevant covariates may be controlled in the design or analysis of epidemiologic studies.

Two approaches to the analytical control of covariates are stratification and mathematical modeling. The technique of stratified analysis was introduced in a seminal paper by Mantel & Haenszel (21). This procedure involves the categorization of the study population into discrete levels of the relevant
covariates. The association between the exposure and disease then is evaluated for each subset of the study population and, when appropriate, a summary measure of association and test statistic may be calculated.

Until quite recently, stratification was the predominant approach to the analysis of epidemiologic research. The appeal of stratification lies in its simplicity, clarity, and minimal statistical assumptions needed for making inferences. Unfortunately, stratification upon even a modest number of covariates may result in the allocation of only a few subjects to individual strata. As a consequence, the estimation of the main association of interest becomes imprecise and unreliable. Even when there are sufficient numbers of subjects in each stratum, it may be difficult to interpret the pattern of effect estimates across strata.

The growing recognition of the limitations of stratification has stimulated interest in alternative methods of analysis. In an earlier contribution to this Annual Review, Hanley (14) discussed a variety of multivariate statistical techniques. Among the mathematical models presented in that review were discriminant analysis, multiple logistic regression, and log-linear methods. Of these three approaches, logistic regression is the most widely utilized and is the focus of the present discussion.

In contrast to stratification, mathematical modeling can easily accommodate several covariates. Other features of mathematical models include the ability to "smooth out" or dampen variation attributable to unimportant factors, and the capacity to incorporate continuous independent variables. With the introduction of a number of computer statistical packages, multivariate techniques have become accessible to a wider audience of investigators (1, 8, 15).

Mathematical modeling has several potential disadvantages, however. First, these techniques have underlying statistical assumptions that may not be valid for the data under consideration (10). Second, a particular model may provide an inadequate description of the true relationship under investigation (11). Third, erroneous conclusions can be drawn from the results of statistical computer packages if the user is unfamiliar with the coding scheme employed for categorical variables (19).

Aside from these technical concerns about the use of mathematical models, several pragmatic issues have yet to be resolved. For example, how should the investigator decide which cofactors are eligible for inclusion into a model? Once the eligible covariates are determined, in what sequence should they be entered or deleted from the model? Should the inclusion of variables be dictated entirely by tests of statistical significance? On an even more fundamental level, should variables be considered in a forward selection procedure, or by backward elimination?

At present, there are no universal answers to these questions. Instead, the individual investigator must select the approach that will be used to model a set of data. This lack of uniformity has created a state of confusion, which results
in misunderstandings between epidemiologists and, more seriously, may produce misinterpretations of data.

The purpose of this paper is to clarify some of the current uncertainties about mathematical modeling of epidemiologic data. A specific approach to variable selection is recommended, with illustrations from a recently completed study. These guidelines are intentionally general, and alternative strategies may be preferred for specific applications. Only the basic principles of analysis are presented here; the reader interested in a more detailed discussion may consult several current textbooks (5, 18, 26).

GOALS OF MULTIVARIATE ANALYSIS

Two distinct goals may direct multivariate analyses. First, the investigator may be interested in predicting a dependent variable from a set of independent variables. For this purpose, it is desirable to obtain a model that conforms to the observed data as closely as possible. In other settings, the main objective is to assess the relationship between the dependent variable and one or more independent variables, as indicated by the respective regression coefficients. This goal of effect estimation is central to many epidemiologic studies and is the main focus of this discussion. In this regard, two methodologic concepts must be addressed: interaction and confounding. The following sections provide some background information on these two essential concepts.

INTERACTION

Interaction may be defined as a condition that exists when the relationship of interest varies according to the level (i.e. value) of one or more covariates. That is to say, the effect of the main exposure is altered by the presence of another factor. To appreciate this condition, it is helpful to think in terms of the usual measure of association employed in cohort studies, the risk ratio (RR). The RR corresponds to the risk of disease among persons who have the exposure of interest divided by the disease risk among the unexposed. When interaction exists, the RR for the exposure of interest varies according to the level of a cofactor. For case-control studies, the usual measure of association is the odds ratio (OR), which corresponds to the odds of exposure in cases, divided by the odds of exposure in controls. Interaction in case-control studies is present when the OR for the main exposure varies according to the level of a cofactor. The following example provides an illustration of interaction in epidemiologic data.

Example 1: Interaction

In a recent study, Hall (13) examined factors that influence the risk of depressive symptoms in low income mothers with young children. The study sample included 114 women, all with at least one child of kindergarten age.
The women were interviewed in their homes during the first four months of 1982. Information was collected on measures of everyday stressors, life events, the quality of the primary intimate relationship, the social network, as well as depressive and psychosomatic symptoms. In this example, we consider the observed relationship between the social network and the risk of depressive symptoms. The social network may be described as an individual's social and community ties. For this study, it was measured by the Berkman Social Network Index and scored with ordinal values between one and four (2). The level of depressive symptoms was measured as a continuous variable with the Center for Epidemiologic Studies--Depression Scale (24). Previous research has indicated that persons with a score of 16 or greater on this scale have a high level of depressive symptoms. Therefore, respondents in this study with scores below 16 were classified as having low depressive symptoms, whereas respondents with scores of 16 or higher were classified as having high depressive symptoms. The results of this analysis are summarized in Table 1.

Among employed women there was no relationship between the social network and the presence of depressive symptoms, since the OR was approximately the null value of one. In contrast, a lack of social network ties was a strong risk factor for depressive symptoms among unemployed women, as indicated by an OR significantly greater than one. Because the effect of the social network depends so heavily on employment status, it would be misleading to summarize these data with a single measure of effect. In particular, the Mantel-Haenszel summary measure (21) would be weighted toward the estimate in the subgroup with the largest number of observations. With this weighting scheme, the effects of strong risk factors may be obscured when the effects occur only in small subsets of the population (12). If heterogeneous stratum-specific estimates cannot be weighted in a meaningful fashion, then it is preferable to avoid summarization across levels of the covariate(s).

Models of Interaction

In mathematical modeling, interaction between two or more risk factors generally is depicted by the inclusion of a term that is the product of the interacting variables. For the linear model involving two risk factors, \( X_1 \) and \( X_2 \), interac-

<table>
<thead>
<tr>
<th>Employment status</th>
<th>Estimated OR</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>0.92</td>
<td>(0.05, 18.00)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>15.32</td>
<td>(5.58, 42.03)</td>
</tr>
</tbody>
</table>

*Adapted from Hall (13).
tion is depicted as a departure from the simple addition of the effects of the component risk factors:

\[ Y = \alpha + \beta X_1 + \gamma X_2 + \delta X_1 X_2, \]

where \( \alpha, \beta, \gamma, \) and \( \delta \) are regression coefficients and \( Y \) is the dependent variable. In other words, the presence of interaction is indicated by a nonzero coefficient for the cross-product term.

The inclusion of a cross-product term in a logistic model with two risk factors, \( X_1 \) and \( X_2 \), takes the following form:

\[ Y = 1/(1 + \exp[-(\alpha + \beta X_1 + \gamma X_2 + \delta X_1 X_2)]). \]

Again, the presence of interaction is denoted by a nonzero coefficient for the cross-product term. However, interaction with the logistic model represents a departure from the simple multiplication of the effects of the component risk factors (17).

A great deal has been written on the use of additive and multiplicative models in epidemiology. These arguments may be summarized with the general statement that the choice of a model depends upon the context and objectives of the research. If the only intent is to provide accurate predictions, then the simplest model that provides a close fit to the data is desirable (11). In other situations, the investigator may be interested in obtaining a model that reflects the pertinent biologic processes. Epidemiologic data often cannot provide insight into the detail of causal mechanisms (23). However, the structure of epidemiologic models may provide a general indication of the type of mechanism involved. For example, when several etiologic agents act interchangeably at a single step of a multistage pathway, an additive model is implied. In contrast, when etiologic agents act at different steps, a multiplicative model may be obtained (25).

A third application of mathematical models in epidemiology is to determine the health impact of exposures in a study population. Interaction in the public health domain reflects the extent to which risk factors for the disease tend to occur together in the same individuals. Interaction in the public health setting may be viewed as a departure from additivity of incidence rates (4).

From the preceding discussion it is evident that in some situations an additive model is appropriate, whereas in other circumstances a multiplicative model is preferred. The reader should keep this caveat in mind when selecting an analytical strategy. Because logistic regression involves multiplicative effects, it is not appropriate for certain types of analysis. As Greenland (11) has demonstrated, the complete analysis of epidemiologic data may involve examination with more than one candidate model.
CONFOUNDING

Confounding may be described as the mixing of two or more effects. Confounding is present when there is a meaningful difference in the estimated effect of interest, depending upon whether a covariable is included or eliminated from the analysis. In epidemiology, confounding usually is assessed by comparing crude and adjusted effect estimates (e.g. RRs). The term crude RR implies that the influences of covariables were not taken into account. When the effects of other risk factors are controlled, the RR is said to be adjusted. Therefore, confounding of the RR occurs if and only if

\[ \text{crude RR} \neq \text{adjusted RR}. \]

An analogous comparison of crude and adjusted ORs can be used to assess confounding in case-control studies.

**Example 2: Confounding**

Hall (13) used the population described in Example 1 to evaluate factors that are associated with psychosomatic symptoms in low income mothers with young children. Psychosomatic symptoms refer to sensations of bodily disorders that may arise from emotional factors. Symptoms were measured by a 15-item index derived from the work of MacMillan (20). Because the outcome variable in a logistic regression analysis must be a binary variable, the continuous range of psychosomatic scores was dichotomized at the seventy-fifth percentile into high and low symptom groups. In this example, we consider the observed relationship between the social network and the risk of psychosomatic symptoms. The social network was measured and scored as previously described. First, the data were examined for possible interactions. Unlike Example 1, no significant interactions were found in this analysis. Next, the crude relationship between the social network and psychosomatic symptoms was assessed. Then, the effect of the social network was evaluated, with adjustment for other risk factors, including level of income, employment status, marital status, and the type of primary intimate relationship. The results of this analysis are presented in Table 2.

From the crude OR, women with low social network scores were estimated to have a fourfold excess risk of psychosomatic symptoms when compared to women with high social network levels. After adjustment for the other risk factors, the effect of a low social network was twice as strong as the crude effect. In other words, the association between low social network and the risk of psychosomatic symptoms was confounded by the other risk factors. For this example, the crude OR was an underestimate of the adjusted association. That is to say, the crude estimate was closer to the null value of unity than the fully
adjusted estimate. The direction of the bias therefore may be described as "toward the null." Consequently, failure to control for the effects of the covariates provided an inaccurate summary of the association between the social network and psychosomatic symptoms.

One also must distinguish between the issues of validity and statistical precision. The primary concern in this analysis is to obtain a valid (unconfounded) estimate for the effect of the social network. The most valid estimate from these data was the fully adjusted OR. At the same time, it should be recognized that adjustment produced a statistically imprecise result, as indicated by the broad 95% confidence intervals. When such a loss of precision occurs, a smaller number of control variables may produce a point estimate similar to the fully adjusted value, but with a greater degree of precision. For these data, control of only two covariables, marital status and income, provided the same OR as the fully adjusted value. However, no gain in statistical precision resulted with the smaller subset of control variables. Therefore, the data were summarized by the fully adjusted OR.

**Confounding in Epidemiologic Data**

The issue of confounding should be entertained only after interactions have been considered (18). We have already noted that when strong interactions are present, it usually is desirable to avoid summarization across levels of the covariate(s). Without summarization, an adjusted effect estimate cannot be calculated, and hence confounding usually cannot be assessed. One should be aware of the situation in which the crude effect estimate has a more extreme value than all of the stratum-specific estimates. For these data, any weighted average of the stratum-specific effect estimates will differ from the crude value. Thus, we recognize the presence of confounding, without the need to calculate an adjusted summary effect estimate. This leads us to the general conclusion that in the presence of strong interaction, stratum-specific estimates should be presented, regardless of whether there also is confounding in the data.

---

**Table 2** Crude and adjusted estimates of the odds ratio between low social network and psychosomatic symptoms

<table>
<thead>
<tr>
<th>Variables controlled</th>
<th>Estimated OR</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>4.13</td>
<td>(1.36, 12.53)</td>
</tr>
<tr>
<td>Income, employment, marital status, type of primary intimate</td>
<td>9.43</td>
<td>(1.72, 51.79)</td>
</tr>
</tbody>
</table>

*Adapted from Hall (13).*
A second practical point is that confounding should not be evaluated with a statistical test (7, 9, 22). With a large sample size, it is possible to detect statistically significant differences between crude and adjusted measures, even when the interpretations of the two measures are quite similar. However, important differences between crude and adjusted measures may not reach statistical significance when the sample size is small. Therefore, inferences about confounding should be based solely upon the magnitude of change in the effect estimate, not on the statistical significance of these differences.

Another consideration is the process by which the set of potential confounders is determined. It should be emphasized that adjustment for nonconfounders (overadjustment) can lower the statistical precision of the effect estimates. It is advisable to limit the list of potential confounders to previously documented risk factors for the disease. With this approach, the exposure of interest is evaluated only after taking into account the effects of all known determinants of the disease. When there is no previous information on risk factors for the disease, or when the prior work is inconclusive, the set of potential confounders may be limited to basic demographic characteristics.

One situation does occur in which a cofactor should not be controlled, even when it is a recognized determinant of the disease. If the exposure of interest is suspected to operate through an intervening variable, adjustment for the effect of the intervening variable may lead to an underestimate of the exposure-disease association. Typically, the precise causal pathway cannot be defined with epidemiologic data. However, when there is a clear temporal sequence of exposure to risk factors, the possibility of an intervening variable should be considered.

**THE LOGISTIC MODEL**

Multiple logistic regression is a popular approach to mathematical modeling in epidemiology. The general form of this model is

\[
P(X) = \frac{1}{1 + \exp[-(\alpha + \sum_{i=1}^{k} \beta_i X_i)]},
\]

where \(P(X)\) is the dependent variable, \(\alpha\) is a constant term, the \(\beta_i\) are regression coefficients, and the \(X_i\) are independent variables. In this model, the dependent variable may range between the values of zero and one. Because \(P(X)\) is constrained to these limits, it can be used to model probabilities. This model was introduced originally for the analysis of cohort studies (28), where the outcome is the probability (risk) of disease. A natural measure for this model, nevertheless, is the (adjusted) risk odds ratio, which will approximate the risk ratio under the rare disease assumption.
In the absence of interaction, the risk odds ratio ($\hat{OR}$) for comparing two levels ($E^*, E^0$) of exposure to a single risk factor is given by:

$$\hat{OR} = \exp[\beta(E^* - E^0)]$$ \hspace{1cm} (5)

where $\beta$ is the regression coefficient for the exposure variable. Readers interested in the derivation of Eq. 5 should consult one of several textbooks (5, 18, 26). In the situation where the exposure has only two levels ($E^* = 1$, $E^0 = 0$), then Eq. 5 simplifies to

$$\hat{OR} = \exp[\beta(1 - 0)] = \exp(\beta).$$ \hspace{1cm} (6)

From either Eqs. 5 or 6, the $\hat{OR}$ is equal to unity when the regression coefficient, $\beta$, is zero. In other words, a test of the null hypothesis that $\beta$ is zero is equivalent to a test of the null hypothesis that the $\hat{OR}$ is unity.

Although the logistic model was introduced for the analysis of cohort studies, it is most commonly used for the analysis of case-control studies. The application of Eq. 4 to case-control research may appear to be inappropriate, because the “outcome” in a case-control study is exposure status rather than the occurrence of disease. Nevertheless, as has been demonstrated, with adjustment for covariates, similar results are obtained for modeling case-control data with the disease or exposure as the dependent variable (6).

**Logistic Model with Interaction**

The general form of the logistic model given in Eq. 4 can be adapted to include potential confounders and interaction terms:

$$P(X) = 1/\{1 + \exp[-(\alpha + \beta E + \sum_{i=1}^{m} \gamma_i V_i + \sum_{j=1}^{n} E \delta_j W_j)]\},$$ \hspace{1cm} (7)

where $\alpha$, $\beta$, $\gamma_i$, and $\delta_j$ are regression coefficients, $E$ is the exposure of interest, the $V_i$ are potential confounders, and the $EW_j$ are interaction terms (i.e. the $W_j$ are possible effect modifiers). If one of the interaction terms, say $EW_1$, is statistically significant, then its regression coefficient, $\delta_1$, will be significantly different from zero. In the presence of such an interaction, the $\hat{OR}$ for comparing two levels ($E^*, E^0$) of the exposure can be estimated from

$$\hat{OR} = \exp[\beta(E^* - E^0) + \delta_1(E^* - E^0)W_1].$$ \hspace{1cm} (8)

Of course, Eq. 8 can be generalized to include more than one interaction term by including all significant cross-product terms, with their respective regres-
sion coefficients. Also, in the situation where the exposure has only two levels \((E^* = 1, E^0 = 0)\), Eq. 8 simplifies to

\[
\hat{OR} = \exp(\beta + \delta_1W_1).
\]

When significant interactions are present, a single summary measure of effect may misrepresent the epidemiologic relationships involved. It can be seen from Eqs. 8 and 9 that the \(\hat{OR}\) will vary according to the value(s) of the covariate(s) in \(W_1\). Therefore \(\hat{OR}\)s can be calculated from the model as a function of the interacting covariate(s).

In modeling interactions, one must choose the components for cross-product terms to be evaluated. Even when there are a modest number of covariables, a large number of possible interactions may be envisioned. To simplify the selection process, we recommend that the two-factor interactions (exposure with single covariates) be considered. Three-factor interactions (exposure with two covariates) also may be tested in certain situations. However, evaluating all possible interactions with more than two components may require a large number of statistical tests. The performance of multiple statistical tests increases the likelihood of an interaction resulting from chance alone (Type I error). Furthermore, the interpretation of higher-order interactions may be extremely difficult. As higher-order interactions are added to the model, problems may arise from collinearity of the predictor variables. Unless there is an \textit{a priori} reason to suspect a higher-order interaction, exploratory analyses may be limited to the exposure, covariables, and up to two-factor cross-product terms.

When a significant interaction is found, the lower-order terms that are involved in that interaction should be retained in the model. This requirement derives from the hierarchy principle, as discussed in the work of Bishop et al (3). Thus, if a two-factor interaction involving \(E\) and \(W_1\) is statistically significant, both \(E\) and \(W_1\) also should be retained in the model.

\section*{Logistic Model Without Interaction}

When none of the tested cross-product terms are statistically significant, the logistic model in Eq. 7 may be simplified to the following:

\[
P(X) = \frac{1}{1 + \exp\left[-(\alpha + \beta E + \sum_{i=1}^{m} \gamma_i V_i)\right]}.
\]

The \(\hat{OR}\) estimated from this logistic model is the fully adjusted measure. An adjusted \(\hat{OR}\) comparing two levels of exposure \((E^*, E^0)\) is calculated from
adjusted \( \hat{OR} = \exp[\beta(E^* - E^0)] \).

Notice that the potential confounders, \( V_i \), do not enter directly into this formula. The influences of the confounders are manifested in the model as they affect the estimated value of \( \beta \), the regression coefficient for the exposure. Confounding is assessed by a comparison of the fully adjusted \( \hat{OR} \) with the crude \( \hat{OR} \). The logistic model for the crude effect of the exposure is

\[
P(X) = \frac{1}{1 + \exp[-(\alpha + \beta_0 E)]}.
\]

From this logistic model, the crude \( \hat{OR} \) comparing two levels of exposure \((E^*, E^0)\) is calculated from

\[
\text{crude} \ \hat{OR} = \exp[\beta_0(E^* - E^0)].
\]

Thus, confounding is assessed by a comparison of the adjusted \( \hat{OR} \) (estimated from Eq. 11) and the crude \( \hat{OR} \) (estimated from Eq. 13). If these effect estimates are meaningfully different, then confounding is present. Obviously, confounding relates to differences in the values of \( \beta \) and \( \beta_0 \). The assessment of confounding does not involve a test of statistical significance. Any meaningful difference in the crude and adjusted \( \hat{ORs} \) is a threat to validity, regardless of whether it arose by chance or some other cause.

When confounding is present, the adjusted \( \hat{OR} \) should be interpreted as the best estimate of the effect under consideration. In Example 2, we note that simultaneous adjustment for all covariates may produce a statistically imprecise result. In some situations, modeling a smaller subset of the covariates may yield the same estimate of the OR as the fully adjusted model, with a greater degree of precision. Again, we emphasize that obtaining a valid effect estimate should be the primary consideration. Therefore, models with reduced numbers of covariates should be entertained only if the regression coefficient for the exposure is equivalent to the fully adjusted \( \beta \).

**Estimation of Parameters**

For logistic modeling, the preferred approach to estimation of regression coefficients is through maximum likelihood (ML) methods. With these procedures, a likelihood function is constructed that represents the probability of observing the data obtained as a function of the unknown parameters (regression coefficients). The values of the parameters in the likelihood function are selected to maximize the value of the function. The details of ML estimation may be found in current textbooks (5, 18, 26).
Hypothesis Testing in Logistic Regression

A straightforward procedure may be used to assess the statistical significance of terms in a logistic regression. To develop this test, we first note that each model has a maximized likelihood value, which is the value of the likelihood function when the unknown parameters are replaced by their ML estimates. Symbolically we represent the maximized likelihood value by \( \hat{L} \), and from it we can calculate a statistic referred to as the log likelihood:

\[
\log \text{likelihood} = -2 \ln(\hat{L}),
\]

where \( \ln(\hat{L}) \) is the natural logarithm of \( \hat{L} \). The log likelihood statistic is provided in the computer output from standard packages for logistic regression with ML estimation. The use of the log likelihood statistic for hypothesis testing is illustrated with the following general models.

Consider the comparison of two logistic models that have all but one independent variable in common:

Model 1: \( P(X) = 1/(1 + \exp[-(\alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 E V_1)]) \), 15a.

Model 2: \( P(X) = 1/(1 + \exp[-(\alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2)]) \). 15b.

The independent variables in Model 2 include all of the independent variables in Model 1 except for the interaction term. The difference in the log likelihood statistics of these models is distributed as a chi-square statistic, and is calculated in the following manner:

\[
-2 \ln(\hat{L}_2) - [-2 \ln(\hat{L}_1)] = -2 \ln(\hat{L}_2/\hat{L}_1). \tag{16}
\]

This procedure is referred to as a likelihood ratio test and corresponds to a test of the null hypothesis that Model 1 does not fit the data better than Model 2. Another way of stating this null hypothesis is that the regression coefficient of the interaction term is zero. In general, the degrees of freedom for this type of test statistic is equal to the number of parameters that are deleted from the larger model to obtain the reduced model. Since one parameter is deleted from Model 1 to obtain Model 2, the likelihood ratio test for this comparison is distributed as a chi-square with one degree of freedom.

Consider the situation in which the larger model has two parameters more than the smaller model:

Model 3: \( P(X) = 1/(1 + \exp[-(\alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 E V_1 + \delta_2 E V_2)]) \), 17a.

Model 2: \( P(X) = 1/(1 + \exp[-(\alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2)]) \). 17b.
In this comparison, two interaction terms are tested simultaneously for statistical significance. The likelihood ratio test therefore is distributed as a chi-square with two degrees of freedom. The null hypothesis for this test is that the regression coefficients for both interaction terms are zero. If this null hypothesis is rejected, then one may conclude that at least one of the interactions has a nonzero coefficient.

Other approaches to hypothesis testing can be employed in logistic regression. The various testing procedures should yield similar results for a given application. Therefore, the alternative tests are not presented in this review, but may be found in other references (18, 26).

Interval Estimation in Logistic Regression

In the absence of significant interactions, a large sample $100(1-\alpha)\%$ confidence interval for the OR of a dichotomous exposure ($1=$ exposed, $0=$ unexposed) is given by:

$$100(1-\alpha)\% CI = \exp\left[\hat{\beta} \pm Z_{1-\alpha/2}\sqrt{\text{Var}(\hat{\beta})}\right].$$

An extension of this formula can be used to determine large sample confidence limits for the OR of a polychotomous or continuous variable (26). The calculation of confidence limits for the OR in the presence of significant interactions is more complicated. The approach to interval estimation of models with interaction is presented in a recent textbook (18).

Features of the Logistic Model

The present discussion has focused on the logistic model. The logistic function is a reasonable model for the analysis of many epidemiologic relationships. Specifically, the dependent variable in a logistic function has a range of possible values between zero and one, which is convenient for modeling risks. Also, dose-dependent effects in epidemiology often adhere to the sigmoidal shape of a logistic function. The ML estimation procedures used in logistic regression can accommodate nominal, ordinal, or continuous independent variables. Finally, as demonstrated in this review, conventional epidemiologic measures of association can be calculated from the logistic model.

It should not be assumed that the logistic model is appropriate for the analysis of all epidemiologic relationships. Several limitations of the logistic model must be recognized. First, if the relationship(s) between the predictor variable(s) and the outcome is (are) not monotonic, then a logistic function may not provide a close fit to the data. Quadratic terms may be included in the model to improve the fit, but a simpler approach may be to use an alternative model form (10). A logistic model also is less desirable when the effects of two or more independent variables are additive. When additivity of effects is observed, a linear regression model is likely to involve fewer and less complex terms than a
corresponding logistic regression analysis. A third limitation of the logistic
function occurs when two or more highly correlated independent variables are
included. Under these conditions, the parameter estimates become unstable and
may not have a useful epidemiologic interpretation. If the goal of the analysis is
to estimate individual effects, then it may be necessary to model the correlated
factors separately (10).

CHOICE OF A MODEL

The thorough analysis of a set of epidemiologic data should not be limited to a
single model form. In some situations, the ultimate preference for a particular
model is based upon the hypothesized underlying biological process. In other
circumstances, a final model form is chosen to allow certain types of inference
(e.g. public health impact). In still other settings, the analytic approach may be
governed by statistical considerations, such as the closeness of fit to the
observed data, or the number and complexity of terms. For any epidemiologic
analysis, the investigator should consider a variety of different models. This
type of comparison is feasible because of the widespread availability of statisti­
cal software. The investigator should avoid the temptation to select a model
simply because it is in vogue at that time.

A Strategy for Modeling

The process of modeling epidemiologic data is most meaningful when it is
conducted through a logical sequence of decisions. Unfortunately,
epidemiologists do not agree on the sequence in which these decisions should
be made, or even on how to make such decisions. This lack of uniformity is
complicated further by the abbreviated manner in which analytic methods are
described in published research. The methods section of a published epidemio­
logic study usually does not indicate the criteria that were used to select
covariates, the number and types of interactions tested, or the approach that
was used to assess confounding. The results of modeling typically are summa­
rized by presenting the final model, with estimated regression coefficients,
measures of effect, and test statistics. The reader usually cannot verify any of
these calculations, or even determine whether the final model form was
appropriate for the data.

The following strategy is intended to provide some uniformity for estimating
measures of association with mathematical models. This approach is shaped by
the current state of knowledge about interaction and confounding. Because
these concepts are evolving rapidly, it may be necessary to modify the model­
ing strategy in the future. For the present, we note that this approach has been
applied successfully for a variety of research issues. In these applications, the
results of modeling have been consistent with the results of other analytic
approaches, such as stratified analysis.
The reader is forewarned that “cookbook” approaches to multivariable analysis have inherent dangers. The use of any general strategy may be impaired by small sample sizes, the constraints of statistical procedures, and the nature of the process under consideration. With these caveats in mind, we recommend the following steps for mathematical modeling:

**STEP 1: SPECIFY VARIABLES**  The disease and main exposure(s) should be defined, along with the independent risk factors to be assessed as potential confounders and the interactions that will be evaluated.

**STEP 2: CONSTRUCT FULL MODEL**  The full model should include the disease, exposure, potential confounders, and pertinent two-factor interactions.

**STEP 3: (OPTIONAL) ASSESS HIGHER ORDER INTERACTIONS**  The investigator may use a stepwise backward elimination procedure to assess the significance of three-factor interactions. In some situations it may be impossible to model multiple three-factor interactions simultaneously. Therefore, the investigator may be forced to assess three-factor terms by forward addition. If any of these higher-order interactions are retained, the component factors also must be retained in the model.

**STEP 4: ASSESS TWO-FACTOR INTERACTIONS**  The two-factor interactions that are not required by the results of Step 3 are tested for elimination by stepwise backward elimination. If no significant interactions are found, the investigator should proceed to Step 5. Otherwise, a generalization of Eq. 8 is used to estimate ORs as a function of the interaction(s).

**STEP 5: ASSESS CONFOUNDING**  The adjusted OR is estimated with Eq. 11. If the crude and adjusted measures are meaningfully different, then confounding is present. The investigator may either present the adjusted OR, or proceed to Step 6. If the crude and adjusted measures are not meaningfully different, then either OR can be presented as an unconfounded estimate of effect.

**STEP 6: (OPTIONAL) DETERMINE A MINIMAL SUBSET OF CONFOUNDERS**  When confounding is present, the investigator may be motivated to identify an unconfounded model with a more precise effect estimate than is obtained from the fully adjusted model. This is accomplished by backward elimination of covariables from the model, avoiding any eliminations that alter the regression coefficient of the exposure term(s).

The application of this strategy to the multivariate analysis of epidemiologic data is illustrated in the following example.
Example 3: Modeling Data with a Significant Interaction

For this example, we consider the relationship between daily stresses and the presence of depressive symptoms in low-income mothers with young children. This hypothesis was evaluated in the population studied by Hall (13). The daily stresses were defined as routine problems that worry, upset, or bother an individual. An instrument with 22 questions was developed to address the problems that low-income mothers of preschool-aged children might encounter on a daily basis, including financial concerns, role overload, parenting worries, work outside the home, and interpersonal problems. Each item was scored by the respondent on a four-point severity scale, from which a cumulative score was derived. The instrument is similar to the "Hassles Scale" developed by Kanner et al (16). For simplicity of classification, the range of stress scores was divided into quartiles. The specific outcome under consideration was depressive symptoms and was measured as described in Example 1.

STEP 1: VARIABLE SPECIFICATION The dependent variable was depressive symptoms, the exposure was daily stresses, and the potential confounders included income, employment status, marital status, and type of primary intimate relationship. The tested interactions included four two-factor cross-product terms, each of which was comprised of the exposure and one of the above covariates.

STEP 2: CONSTRUCTION OF FULL MODEL The full model was constructed in the form of Eq. 7.

STEP 3: ASSESSMENT OF HIGHER ORDER INTERACTIONS No three-factor interactions were tested because there was no a priori reason to include such terms. Moreover, the sample size was insufficient to obtain stable estimates of higher-order interactions.

STEP 4: ASSESSMENT OF TWO-FACTOR INTERACTIONS The logistic regression of the full model is summarized in Table 3. We begin by focusing on the four interaction terms: stress with income, stress with employment, stress with marital status, and stress with type of primary intimate. The approach to exclusion of two-factor interaction terms was a stepwise backward elimination based upon statistical significance (p < .1).

From the results of regression on the full model, the least predictive interaction term was stress with type of primary intimate (p=0.87) and so it was eliminated, leading to the reduced model, which is depicted in Table 4. At this stage, the stress-income interaction term was eliminated (p=0.83), resulting in a further reduced model (Table 5).

The next elimination was the stress-employment interaction (p=0.55), which resulted in the model presented in Table 6. At this stage, the remaining
interaction term, stress-marital status, was statistically significant and therefore it was retained in the model.

**STEPS 5 AND 6: ASSESSMENT OF CONFOUNDING AND DETERMINATION OF A MINIMAL SUBSET OF CONFOUNDERS**  Because a significant interaction was present, it was elected not to attempt to determine a minimal subset of confounders. Alternatively, one or more of the following covariates could have been considered for elimination: income, employment status, and/or type of primary intimate. However, marital status must be retained in the model according to the hierarchy principle.

The effect of perceived everyday stressors was estimated by ORs comparing women with various levels of stress to women with the lowest amounts of stress. The pattern of point estimates is illustrated in Table 7. From these point estimates, the following conclusions were drawn:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Regression coefficient</th>
<th>Standard error</th>
<th>Chi square</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.616</td>
<td>1.428</td>
<td>3.35</td>
<td>.07</td>
</tr>
<tr>
<td>Stressors (str)</td>
<td>0.139</td>
<td>0.060</td>
<td>5.37</td>
<td>.02</td>
</tr>
<tr>
<td>Income (inc)</td>
<td>0.074</td>
<td>0.479</td>
<td>0.02</td>
<td>.88</td>
</tr>
<tr>
<td>Employment (emp)</td>
<td>0.186</td>
<td>1.224</td>
<td>0.02</td>
<td>.88</td>
</tr>
<tr>
<td>Marital status (mar)</td>
<td>1.924</td>
<td>1.732</td>
<td>1.23</td>
<td>.27</td>
</tr>
<tr>
<td>Intimate (int)</td>
<td>-0.449</td>
<td>0.482</td>
<td>0.87</td>
<td>.35</td>
</tr>
<tr>
<td>(Str) × (inc)</td>
<td>0.005</td>
<td>0.023</td>
<td>0.04</td>
<td>.83</td>
</tr>
<tr>
<td>(Str) × (emp)</td>
<td>-0.035</td>
<td>0.056</td>
<td>0.40</td>
<td>.53</td>
</tr>
<tr>
<td>(Str) × (mar)</td>
<td>-0.108</td>
<td>0.074</td>
<td>2.13</td>
<td>.14</td>
</tr>
</tbody>
</table>

*Adapted from Hall (13).*
estimates, we conclude that a strong positive relationship exists between perceived stress and depressive symptoms in unmarried women, whereas a weak positive association occurs in married women.

**FURTHER CONSIDERATIONS IN MODELING**

Regardless of the strategy used for mathematical modeling, certain statistical issues must be considered. Two of these issues, multiple hypothesis tests and collinearity of the exposure and covariates, are introduced in the following sections.

**Multiple Hypothesis Tests**

Even when an epidemiologic study is undertaken to evaluate a single exposure-disease relationship, the analysis often involves many tests of statistical significance. As advocated in this chapter, potential interactions must be assessed before the main effect of the exposure is considered. In the usual situation, at least five or six covariates might modify the effect of the exposure. If
higher-order interactions are entertained, the full model may contain over a dozen interaction terms. In principle, the statistical significance of each of these terms may be assessed by a separate test. As the number of tests increases, the likelihood that a null hypotheses will be erroneously rejected increases correspondingly (Type I error). The problems of multiple hypothesis testing are compounded when a variety of exposure-disease relationships are evaluated in a single study.

Several approaches are possible to reducing the chance of committing a Type I error in mathematical modeling of epidemiologic data. First, one may limit the number of tests performed. Because much of the testing involves interaction terms, it may be necessary to restrict the analysis to two-factor cross-product terms. As already noted, three-factor interactions and those of higher order may be difficult to interpret and also they may introduce collinearity into the model. Unless there is an a priori reason to anticipate a higher-order interaction, it is reasonable to limit the analysis to the exposure, covariables, and up to two-factor cross-product terms.

The total number of tests performed also may be reduced by assessing several variables in a single test. Equations 17a,b represented the full and reduced models for a likelihood ratio test of the significance of two interaction terms. The null hypothesis for this test was that the regression coefficients of both interaction terms were equal to zero. Since two terms were assessed simultaneously, the test statistic was distributed as a chi-square with two degrees of freedom. This type of aggregate or “chunk” test is especially useful when a large number of interactions must be evaluated. By testing several variables at the same time, the total number of tests is reduced. In the general case of \( n \) interactions, all \( n \) or some subset of these terms may be tested together. It must be recognized, nevertheless, that the conclusions reached by a “chunk” test may differ from those reached by separate tests of the individual variables.

Aside from limiting the number of tests, the investigator may reduce the chance of committing a Type I error by adjusting the significance level (6a). In other words, the significance level for each test is lowered so that the overall significance level does not exceed some nominal value (e.g. .05). When the tests are not independent, the usual approach to adjusting the significance level
is referred to as the Bonferroni correction. With this method, the nominal significance level is divided by the total number of tests to obtain the corrected significance level. For instance, if the nominal level is .05 and ten separate tests are performed, then the corrected significance level is .05/10 = .005. Each of the ten tests is performed with the adjusted significance level of .005. When the number of hypothesis tests is large, the corrected significance level may be so small that it is difficult to reject any null hypothesis. That is, an excessive adjustment for a Type I error will increase the chance of committing a Type II error. Therefore, it is advisable to use the Bonferroni correction only when there are a limited number of tests, or else not to use it at all.

A third and less commonplace approach to reducing the chance of committing a Type I error is to design a study in which the subjects are divided into an exploratory analysis subsample and a confirmatory analysis subsample (22a,b). The exploratory analysis subsample is randomly selected from the source population and includes not more than one-third of the total subjects. With this subsample, the investigator performs an exploratory analysis to evaluate the potential confounders and interactions. The covariates that appear important, although not necessarily statistically significant, then are included with the exposure and disease in the full model for the remaining subjects (i.e. the confirmatory analysis subsample). With this sequence, the exploratory analysis is used to screen out the covariables that do not appear to be either confounders or effect modifiers. This technique is likely to prove useful only in situations in which there are large numbers of subjects, since splitting the study population may result in an appreciable loss of statistical power in the confirmatory analysis. In other words, the division of the sample may produce an inflated Type II error rate.

The investigator may employ any of the preceding methods alone or in combination to achieve adequate control of the Type I error rate. Consider, as an example, the modeling of an exposure-disease relationship with five covariates. At the outset, the investigator can limit the modeled interactions to two-factor cross-product terms. Then, a “chunk” test can be performed on all five interactions together. In the absence of significant interaction, the main effect of the exposure can be tested, adjusting for the effects of the covariates. With this approach, the total number of hypothesis tests has been reduced to two. For this small number of anticipated tests a Bonferroni correction can be justified, so the nominal significance level of .05 is modified. Although more complicated analyses may arise (e.g. multiple significant interactions), the investigator may still successfully limit the number of hypothesis tests and, when it is appropriate, adjust the significance levels.

**Collinearity of Effects**

A further statistical problem arises when two or more highly related predictor variables are included in a model. In this situation, the coefficient estimates
become unstable, thereby reducing confidence in the results. When one of the correlated variables is the exposure, the estimate of the exposure-disease relationship may be greatly distorted. Although there are diagnostic techniques available to detect collinearity between the predictor variables, such methods are beyond the scope of this review. From an epidemiologic perspective, three types of related variables may be distinguished.

1. The collinear terms are exposures of interest, as illustrated by the use of systolic and diastolic blood pressures as risk factors for coronary heart disease. For correlated exposures, one solution is to create a new variable that combines information from each of the related exposures. This summary exposure variable can be substituted for the component exposures in the mathematical model. When the related exposures are diastolic and systolic blood pressure, a summary measure might be the mean arterial pressure, which is calculated as a weighted average of the diastolic and systolic pressures. Another approach to modeling two or more related exposures is to construct separate models for each exposure. If one of the correlated exposures is of primary interest, then several different models for that exposure-disease relationship may be constructed after stratification of the sample by levels of the other exposure(s).

2. In some situations, the effect of exposure to a single factor is related to one or more of the potential confounders. For example, when modeling an index of cumulative lifetime exposure to an environmental factor, a possible correlated control variable is the age at the time of investigation. If age is included as a control variable, the effect estimate for cumulative exposure can become highly unstable. Therefore, it may be necessary to delete the control variable from the model.

3. A third type of collinearity may arise between the exposure of interest and an interaction term involving the exposure. It should be recognized that the attempt to include higher-order interactions in a model will increase the likelihood that collinearity will occur. The simplest, and perhaps most feasible, solution to this problem is to delete the correlated interaction term from the model. When such a deletion is performed, it must be recognized that the variable is not eliminated because its effect is insignificant. Rather, the variable is removed because there is not enough information available to assess the statistical significance of its effect.

CONCLUSION

In this chapter, several aspects of mathematical modeling of epidemiologic data have been addressed. First, the selection of a particular model form (e.g. logistic model) should be based upon at least one of the following criteria: the nature of the research question, the biological process involved, or statistical considerations. Second, the approach to variable specification and selection should follow a logical sequence of decisions. Third, the concepts of interac-
tion and confounding must be incorporated into the modeling strategy. Fourth, statistical issues such as multiple hypothesis tests and collinearity of predictor variables must be addressed by the investigator.

The use of mathematical modeling in epidemiologic research has overcome many of the limitations of other analytic methods. Nevertheless, the effective use of mathematical models requires an appreciation of the underlying epidemiologic and statistical principles. The development of analytic strategies, such as the one presented here, will help to clarify these principles, and ultimately may simplify the analytic process.

ACKNOWLEDGMENTS

The authors thank Dr. Lynne A. Hall for allowing publication of her research findings. Dr. Greenberg was supported in part by contract N01-CN-61027 from the National Cancer Institute. The manuscript was prepared by Ms. Carol Settle and Ms. Vickie Thomas.

Literature Cited


