Commentary

Modeling and Variable Selection in Epidemiologic Analysis

SANDER GREENLAND, MS, DrPH

Abstract: This paper provides an overview of problems in multivariate modeling of epidemiologic data, and examines some proposed solutions. Special attention is given to the task of model selection, which involves selection of the model form, selection of the variables to enter the model, and selection of the form of these variables in the model. Several conclusions are drawn, among them: a) model and variable forms should be selected based on regression diagnostic procedures, in addition to goodness-of-fit tests; b) variable-selection algorithms in current packaged programs, such as conventional stepwise regression, can easily lead to invalid estimates and tests of effect; and c) variable selection is better approached by direct estimation of the degree of confounding produced by each variable than by significance-testing algorithms. As a general rule, before using a model to estimate effects, one should evaluate the assumptions implied by the model against both the data and prior information. *(Am J Public Health 1989; 79:340-349.)*

Introduction

All statistical methods are based on mathematical models for the process generating the data. Statistical modeling may be defined as using the data to select an explicit mathematical model for the data-generating process. Model selection and model explicitness distinguish modeling from more basic statistical methods.

Although modeling can be useful for detecting and summarizing data patterns in a more efficient manner than simple stratified analyses, these advantages are purchased by a higher risk of bias. The assumptions used in most modeling procedures are more restrictive than those used in simple stratified analyses, if one or more of these assumptions are incorrect, the estimates and tests derived from modeling may be seriously compromised. Vandenbroucke1 proposed two complementary solutions to this problem: a) attempt to fit a complete family of models in order to see what conclusions seem to follow regardless of the model one selects; and b) validate modeling results against the results obtained from simple cross-tabulation. The latter suggestion is relatively straightforward to implement, and was in fact recommended by the forefathers of current modeling approaches in epidemiology.2 The first suggestion, however, is not quite as straightforward to implement, because any truly "complete" family of models will be too large to allow one to fully explore the consequences of using different models in the family. Neither suggestion gets at the problem of deciding which variables should be controlled (by stratfication or modeling) when estimating exposure effects.

The present paper provides an overview of the problems that arise in the selection of models and control variables, and examines some of the proposed solutions. Because current knowledge is so limited, the present paper is *not* intended to provide a definitive guide to modeling. Nevertheless, some conclusions and recommendations can be culled from the literature. One recommendation is worth stating at the outset: To maximize the validity of model-based estimates, an investigator should use both the data and prior information to critically evaluate the epidemiologic assumptions implied by the model and the statistical assumptions required by the model.

The Problem of Model Selection

Investigators who decide to use a multivariate model to estimate exposure effects face a bewildering variety of potential models. Consider a situation in which one wishes to estimate the effect of one exposure factor, and one has data containing seven potential "control" variables (covariates) with possible confounding or interactive effects. How many different models could one construct for the process that generated these data?

First, one has a virtually limitless number of model forms to choose from, since there is a model form corresponding to every way in which one can express the dependent variable (e.g., incidence rate) as a mathematical function of the exposure and control variables. Having selected a model form, one will have to choose a subset from the $2^7 = 128$ possible subsets of the variables to enter in the model. Having selected, say, four of the seven variables to enter in the model as control variables, one will have to choose among six possible two-way interactions of the four control variables, plus four possible two-way interactions of the control variables with the exposure, for a total of $2^{6+4} = 1{,}024$ possible combinations of two-way interactions to enter in the model. For every control variable measured on a continuous scale, one will also have to choose from the infinitude of forms in which to enter the variable in the model, e.g., it could be entered: a) as a categorical variable (in which case the number of categories and the appropriate cutpoints must be chosen); b) as a single continuous variable, possibly after some transformation (e.g., logarithmic); or c) as several polynomial terms, e.g., x and x^2 (in which case the number of variables in the model is increased).

All of the above choices are complicated by the fact that they are confounded, in that what appear to be the best model and variable forms will be influenced by the choice of variables and product terms, and vice-versa. Even restriction to a few common forms of models and transformations will still leave thousands of possible choices. Small wonder, then, that many investigators adopt an automated approach to model selection: Use the most convenient computer-packaged model, enter continuous variables "as is" (without considering transformations), and select variables for the model according to some mechanical algorithm. The fact that such approaches appear in textbooks, articles, and programs may provide one with a sense that such approaches are justfiable, if not the best available. Such a sense would, however, be false;3-9 in fact, little research has been done on the relative performance of different modeling strategies.

The Problems of Simple Stratified Analysis

Despite its problems, modeling can begin to look attractive when one confronts the limitations of simple stratified analyses (i.e., examining only stratum-specific and summary estimates and tests from multiway contingency tables).

First, one must confront the variable-selection problem in stratified analysis no less than in modeling, since most data sets will yield a sparse stratfication (i.e., a table with very small cells) if stratfication is based on all or most of the candidates for control. Unfortunately, both stratum-specific and standardized measures8-11 require a fair number of subjects in each cell for validity (five per cell is generally thought to be a safe minimum); thus use of such methods will almost always entail severe limits on the number of variables one controls.

Second, in order to ensure sufficient numbers in each cell, one must convert each continuous variable into a series of categories. One may also have to combine categories of other, discrete variables. The final categories must be broad enough to ensure sufficient numbers per cell, yet the use of broad categories increases the danger that confounding by the variable will take place within the categories. Furthermore, the most basic stratified methods make no use of the natural ordering of the categories formed from continuous variables. Such categorization problems are circumvented by entering the continuous variables directly in a model; doing so, however, brings us back to the problem of selecting the forms for the model and the variable.

Third, although there are a variety of basic methods that can be applied to sparse stratfications (such as Mantel-Haenszel techniques7-9 and "hybrid" standardized measures11), the validity of these and many other stratfication methods (such as the Woolf and maximum-likelihood estimators of a common odds ratio7,8) depends on the same assumptions as certain multivariate models. For example, the Mantel-Haenszel odds ratio is a valid effect estimate only under the assumption that the underlying stratum-specific odds ratios are constant across the strata.10,11 The latter is equivalent to the assumption that the joint effect of exposure and the stratfication variables is given by a logistic model containing one variable representing exposure status and a series of indicator variables (whose number is one less than the number of strata) which represent the stratum membership of each subject.12

Finally, stratified analysis methods are based on the same assumptions about sampling variability (i.e., the distribution of random error) as those used in modeling. These assumptions are further discussed below. First, however, we will examine the one problem that can be largely avoided in simple stratified analysis: selection of a mathematical form for the dependence of the outcome on the study variables.

Selection of Model Form

Nearly everyone finds it convenient to employ "off-the-shelf" model forms, i.e., whatever is available in packaged computer programs. For case-control data, this almost inevitably leads to the use of the logistic model for the probability (Pr) that a sample subject is a case, 7, 8, 13, 15

$\Pr(\text{case}|x,z_1,...,z_n) = 1/[1 -+ \exp(- a - bx - c_l z_l - ...\ C_n Z_n)]$,

where x represents the exposure of interest, and $z_1,...,z_n$ represent the n variables selected for control (or functions of them). For cohort data, one often sees the logistic model the exponential form of the Cox proportional-hazards model, 4, l4 or the exponential regression model for the person-time rate of disease, l3, 15

$\text{Rate}(x,z_1,...,z_n) = \exp(a + bx + c_1 z_1 + ... + c_n z_n)$.

Under each of these models, the approximate relative risk for the effect of exposure level x" versus exposure level x', adjusted for $z_1,...,z_n$ is exp[b(x" - x')].

In most epidemiologic applications, these three regression models (logistic, Cox, exponential) are based on similar assumptions, l4, 15 and have similar epidemiologic meanings. The two major epidemiologic assumptions implied by these models are:

a) Either the incidence odds or the incidence rate of disease increases exponentially with a simple linear increase in the exposure x or any of the control variables $z_1,...z_n$.

b) The joint effects of exposure and the control variables are multiplicative (for example, if a one-unit increase in x alone multiplies incidence two-fold and a one-unit increase in z_1 alone multiplies incidence three-fold, a simultaneous one-unit increase in both x and z_1 will multiply incidence by six-fold).

Neither assumption is reasonable in all applications.9 Although much research has been done examining alternative or more general model forms, 16-19 none of this research has directly addressed the issue of how much inferences about exposure effects will be distorted if the effects are estimated from a model whose assumptions are violated (i.e., a misspecified model). This is somewhat unfortunate, since in epidemiologic applications any model is at best a rough approximation to the "true" disease pattern. There is however a growing body of research on the effect of using an incorrect model (e.g., see White20), and several conclusions can be drawn from this work.

First, the usual standard-error estimates for the coefficient estimates will be biased if one uses an incorrect model to derive the estimates.20 This implies that model-based confidence intervals will not cover the true effect measure at the nominal (e.g., 95%) coverage rate if the model is incorrect.

Second, even if the general model form is correct, the power of a model-based test of exposure dose-response will be reduced if the form of exposure entered in the model is incorrect.21 For example, entering exposure untransformed into the model as bx can result in marked loss of power for the test of b = 0 if in the model the exposure effect is correctly represented by b(log x).21

Third, in many applications, estimates of effect for extreme and uncommon exposure levels will be more distorted by model misspecfication than estimates for less extreme, common levels. Thus, if one uses a logistic model to study x = cigarettes per day and myocardial infarction incidence, but the true effect of smoking is linear, then one should expect the logistic estimate of the effect of 50 cigarettes per day (versus no smoking) to be more biased than the logistic estimate of the effect of 20 cigarettes per day. Reverse situations (in which the estimated effect for less extreme exposure is more biased) can occur when the extreme observations unduly influence or "leverage" the estimates, or when the effect of exposure is not monotonic, but at least the former problem is detectable using regression diagnostics22 (see below). In either case, model misspecfication can also result in "off-center" confidence intervals, in that the model-based intervals will tend to fall more frequently off to one side of the true value.

Fourth, simply transforming the variables in the model (e.g., replacing x by log x) or adding product terms to the model will not always adequately cope with incorrectness of the basic model form. For example, if the true model is linear in the rate, i.e.,

$$\text{Rate}(x,z_1,...,z_n) = a + bx + c_1z_1 + ... + c_nz_n,$$

and x and some of the z are continuous variables (e.g., cigarettes per day, age), no variable transformations or product terms added to the exponential rate model will allow it to correctly replicate the true dependency of the rate on exposure and the control variables (unless one uses a "saturated" model with as many parameters as observed exposure-covariate combinations). Also, even though transformations and product terms will improve the approximation of the exponential model to the true (linear) form, the product terms will complicate both fitting the model and interpretation of the final results.

In the most severe instances of the above problems, it is possible for a modeling result to be entirely dependent on model assumptions, or to be an artifact of some violation of those assumptions. Suppose, for example, that the true effect of exposure follows a U-shaped dose-response curve, but one fits a model that allows only monotonic (nonreversing) dose-response, such as any of the models given above. Then the estimated exposure coefficient may indicate that no exposure effect exists, even if the exposure effect is strong.

Regression Diagnostics

Clearly, it would be of value to choose the correct model for analysis, or at least the best approximation among the available candidate models. Unfortunately, all model-selection methods are subject to error, and no optimal method for selecting the best model form is known. For example, it is now widely accepted that simple "global" tests of fit are inadequate for model selection: such tests often have low power,23 and so need to be supplemented by various regression diagnostic procedures (such as residual analysis) in order to assure that one has a reasonable chance of detecting important violations of model assumptions. 22, 24-26

Although the literature on regression diagnostics for the models common in epidemiology is still fairly technical, there are many simple graphical methods for checking the reasonableness of model assumptions. One can examine graphs of observed rates (or, in case-control studies, observed odds ratios) or transformations of these to aid in deciding which models and variable transformations would be reasonable to employ;9 more sophisticated exploratory graphing can be done via smoothing techniques,27 although the latter generally incorporate some modeling assumptions. Once a model is fitted, observed or smoothed values may be plotted against or with values predicted (fitted) by the model to detect systematic departures of the observations from the model form;24 for example, a plot of the observed or smoothed values against model-fitted values should tend to form a straight line if the model form is correct.

Because model assumptions may be violated in a fashion undetectable by formal tests or diagnostic procedures, model-based results should be subjected to sensitivity analysis, i.e., checking whether similar results are obtained when different models or assumptions are used for analysis. At a minimum, one should validate modeling results by seeing whether similar results are somehow apparent in the relatively nonparametric setting of simple stratified analysis.1 If they are not, the model-based results may well be more misleading than those from stratified analysis.

Dose-Response Modeling

Dose-response assumptions regarding an ordered variable can easily be checked by converting to an unordered categorical form (a minimum of five categories if possible), and entering this categorical form into the model as a series of unordered indicator variables.8, 9 One can then plot the fitted coefficients of the indicators against the central values of their corresponding categories. This plot should approximate a straight line if the original variable can be entered directly into the model, without transformation. The shape of the plot may also suggest an appropriate form in which to enter the variable; for example, a J-shaped plot, if biologically plausible (as, e.g., with alcohol use and heart disease), would suggest that the variable be entered using a quadratic as well as linear term. Walter, et al,61 present an alternative approach for categorical dose-response analysis.

The preceding approaches are subject to sample-size limitations (see below). In particular, there must be a sufficient number of subjects in each category to produce valid estimates of the outcome or effect measure for that category (which at best would imply no less than five cases expected per category for person-time rate data, plus five noncases expected per category for count data). On the other hand, one should avoid forming exposure categories so broad that there could be important variations in exposure effect within the categories, or covariate categories so broad that there could be important confounding by the covariate within the categories. In particular, it is usually desirable to assign definite boundaries to end categories, even if this means that some outlying observations are omitted from the analysis. For example, "76-80 years" might be suitable for the highest age category, but use of "76+ years" would run the risk of producing distortion if the category included many subjects over 80.

The preceding problems of category construction need not arise in the final analysis, provided that one restores categorized variables to their continuous form after exploratory dose-response analysis; such restoration can also be recommended on efficiency grounds.21 The problems also are avoided if one uses smoothing27 instead of categorization for dose-response analysis.

Almost inevitably, several different forms of a doseresponse curve will appear compatible with the data. In some cases there may be extensive external (prior) information indicating what sort of curve to expect. Prominent examples occur in modeling rates of epithelial tumors, such as bronchial carcinoma; for many such tumors, epidemiologic data and biological theory indicate that within birth cohorts the smoking-specific age-incidence curves should be well approximated by certain "log-log" curves.28 There are, however, several limitations on the utility of prior information, especially in case-control analyses. For example, previous findings may not apply to the particular population being studied, or within levels of the variables chosen for control; case-control matching on a variable can greatly distort the dose-response curve for the variable (usually obliterating the dose response); and, when dealing with age-incidence curves, one must distinguish cohort from cross-sectional curves, the latter being what one would observe in typical case-control studies (the two types of curves may have very different shapes if distributions of other, uncontrolled risk factors vary across birth cohorts or calendar time8).

A special problem in multivariate modeling is the potential for one or a few observations to have exaggerated influence on the final results. For example, it is possible for the data from one subject or category to entirely determine whether a particular model fits well, or whether a particular variable appears to have an effect. Although similar problems can occur in basic analyses (for example, the mean of a set of observations can be strongly influenced by a single outlier), they may be less obvious in multivariate analysis: Influential observations need not have an unusual value for any single variable, nor need they exhibit large residuals.22,24 As a result, an extensive technical literature on detecting and measuring influence has developed,22 and an increasing number of programs incorporate influence measures. When these are not available, one can search scatterplots and multiway cross tabulations for subjects with unusual *combinations* of variable values. The influence of a subject identified in this search can then be assessed by refitting the model after excluding that subject from the data set.

Distributional Assumptions

The models given above (formulas 1-3) specify only the dependence of the expected outcome (risk or rate) on the independent variables in a model. Such models are often termed *structural models,79* since such models specify the (fixed) structural features of the sampled populations which are of central epidemiologic interest. Until recently, discussions of model selection in epidemiologic analysis focused exclusively on specfication of the best structural model form in a restricted class of candidate forms. Specfication of the structural form is not, however, sufficient to fit epidemiologic models (such as formulas 1-3) by the usual methods (e.g., maximum likelihood).

It is usually left to the computer package to specify the other necessary component for model fitting, namely the *sampling model* for the repeated-sampling distribution of the observations about their expectations (i.e., the model for the random or unexplained component of variation in observations). Models for probabilities or expected proportions (such as formula I above) are almost always fit by assuming a binomial distribution for the number of cases observed at each exposure-covariate combination, while models for rates—such as (2) and (3) above—are almost always fit by assuming a Poisson distribution for this number. The same sampling models are used to derive tests and confidence intervals from simple stratified analyses.

Most sampling models are based on the assumptions that individual response probabilities are homogeneous within exposure and covariate levels, and responses are independent across individuals. If these assumptions are violated, the usual sampling models will no longer hold. For example, if in a closed cohort the outcome of one subject affects the outcome of another subject (as with contagious diseases), the distribution of cases will not in general be binomial. If the assumed sampling model is in fact incorrect, the standard errors of both stratified estimates and model-based estimates may be seriously underestimated, and a structural model may appear to fit poorly (according to goodness-of-fit statistics) even if it is perfectly correct.

To circumvent such problems, methods have been developed that employ more general sampling models than the usual binomial and Poisson models. Examples include models for over-dispersion (i.e., excess variation) of discrete data24,30; such models can be fit using the GLIM package,30, 31 and some special cases are preprogrammed in the EGRET package.32

There is also an enormous literature on modeling of contagious diseases (for a recent example see Haber, *et* al33).

Selection of Variables

Even if one ignores (for the moment) the problem of selecting a model form, one still faces the task of selecting the variables to enter in the model. This task includes not only choosing which primary variables to enter, but also which product terms to enter.

Any variable-selecting strategy should begin by screening out recorded variables that would be inappropriate candidates for control. For example, in a study of the effect of an exposure on disease, it is widely recognized that variables influenced by the exposure or disease are inappropriate for control, since control of such variables may lead to considerable bias.6-9 (Special methods for control of confounding by intermediate variables have only recently been developed.34) It is thus essential to exclude such variables from the pool of candidates for control.

After this preliminary screening, there still may be more candidate variables than observations, in which case one cannot fit a model containing all the candidate variables. Even if one can enter all or most of the candidate variables in a model, the estimate of exposure effect derived from the resulting model may be highly unstable or biased; this is especially likely to occur if the set of variables entered in the model has a high multiple correlation with the study exposure. It thus can be valuable to select for control only a limited subset of the candidate variables. In doing so, however, one will confront the following dilemma: No matter what selection procedure is used, the validity of the estimates and tests obtained from the final model depends on the assumption that the omitted variables are *not* in fact confounders (conditional on control of the included variables); but, in practice, there will always be uncertainty regarding the validity of this assumption (were there no uncertainty, no variable-selection algorithms would be needed).

In a sense, variable selection parallels screening for disease: a variable-selection algorithm is a screening device, and a good algorithm will be highly sensitive for detecting confounders and highly specific for screening out nonconfounders. Nevertheless, the performance of an algorithm must be evaluated primarily in terms of the validity of the estimates and tests of effect that it yields, rather than in terms of sensitivity and specificity.

Two major types of selection algorithms are in use in health research: conventional stepwise regression, and charge-in-estimate methods. The following sections describe these methods, and focus on the shortcomings of conventional stepwise regression as a confounder screening tool.

Conventional Stepwise Regression

The most commonly implemented "canned" variable-selection routines are stepwise regression algorithms.35,36 These proceed as follows. One first divides all candidate variables into two classes: forced-in variables, which are entered into the initial model and *not* subject to deletion from the model; and the remaining nonforced variables, which are subject to selection and deletion by a stepwise algorithm. One chooses an initial model, and then subjects it to sequential modfication by repeated application of a selection-deletion cycle, such as the following:

Selection: Among all candidate variables not in the current model, find the one that would have the most sign)ificant coefficient if entered in the model; if its coefficient would be "statistically sign)ificant" (e.g., p < 0.05), enter it into the model.

Deletion: Among all variables in the current model, find the one with the least sign)ificant coefficient; if its coefficient is not statistically sign)ificant, delete it from the model.

The algorithm continues until either no further model changes occur over one complete cycle, or a preset number of variable selections and deletions occur. *Forward-selection* algorithms start with an initial model that includes only forced-in variables, and execute only the selection step. *Backward-deletion* (or backward-elimination) algorithms start with an initial model that includes all potential control variables, and execute only the deletion step (note, however, that a model with all potential control variables may be numerically impossible to fit).

Many variants of the basic stepwise algorithms just outlined are possible. There is, however, absolutely no reason to expect that any conventional selection strategy (such as stepwise or all-possible-subsets regression35 38) will select the best possible subset of variables or product terms to control. In fact, many theoretical reasons have been put forth for expecting conventional algorithms to do poorly in confounder selection.3-9 37 For example, if conventional signficance levels are used, it is likely that some important confounders and product terms will *not* end up in the final model;23,38 in other words, as conventionally implemented the methods have poor sensitivity. In addition, both the sensitivity and specificity of the methods are impaired by the fact that they ignore covariate-exposure associates (which can be as crucial as covariate-disease associations in determining bias). Use of collapsibility tests39-41 to select control variables addresses only the last objection, and so in theory would only be a little better for variable selection than the usual stepwise algorithms.

More generally, in deciding whether to control a variable it is not at all clear that the null hypothesis (of no covariate effect or of no confounding) is the appropriate hypothesis to test. Logically, one can justify failing to control a variable only if one can reject the alternative hypothesis that the covariate is an important confounder.42 Thus, if one wishes to conduct a statistical test of a confounder, a logically sound approach would employ equivalence testing in place of significance testing (see ref. 43 for an introduction to the concept of equivalence testing).

No matter what algorithm is adopted, it is essential that exposure be forced in the model for, if it is not, the probability of inappropriately selecting nonconfounders and omitting confounders will be greatly increased.3,8,37

Change-in-Estimate Methods

Perhaps the most commonly proposed alternative to conventional algorithms is to select variables (and product terms) using a "change-in-estimate" method, in which variables are selected based on relative or absolute changes in the estimated exposure effect.3-9 Such a method logically requires that the exposure variable be present in every fitted model.

As with conventional stepwise algorithms, the "chang-in-estimate" approach can be implemented in a stepwise fashion,8 in which for example the cycle becomes

Selection: Among all candidate variables not in the current model, find the one which if entered would result in the largest change in the estimated exposure effect; if this change exceeds a certain amount (e.g., a 10 per cent change in the risk ratio), enter the variable in the model.

Deletion: Among all variables in the current model, find the one for which its deletion from the model would result in the smallest change in the estimated exposure effect; if this change is less than a certain amount, delete the variable from the model.

Again, many variants are possible. For example, selection may be based on changes in the confidence limits for the exposure effect, rather than changes in the point estimate,44 this approach will more adequately account for collinearity of the study exposure with the candidate variables than will a change-in-point-estimate approach.37,44

Prior Information and Variable Selection

Automatic variable selection algorithms (such as those given above) do not take advantage of prior information concerning effects of factors not forced into the model. For this reason, many authors recommend that prior information take precedence in determining order of entry, with extensive forcing of a *priori* important variables into the model3-6 37 (e.g., age would invariably be forced in when analyzing a cohort study of lung cancer). Several authors further note that when one has reliable prior information concerning the magnitude of a control-variable coefficient, it can be worthwhile to use this information in the construction of the final coefficient estimate for that variable.37 44-46 Nevertheless, such use carries the risk of increasing error if the prior information is in fact incorrect.37-44

Evaluations of the Methods

Informal theoretical arguments have tended to favor the change-in-estimate approach over signficance-testing approaches such as conventional stepwise selection and all possible-subsets regression.3-9 Furthermore, it is not difficult to find real examples in which "canned" stepwise and other signficance-testing approaches give palpably absurd results, but the change-in-estimate approach does not;6 in contrast, no examples of a reverse situation (real data for which ordinary stepwise estimates appear superior) have been published.

Although there have been very few simulation or mathematical studies of variable-selection criteria for epidemiologic situations, those studies that have been done have tended to support earlier criticisms of signficance testing approaches. For example, Robins44 has given some formal justfications for previous intuitive arguments against significance-tests for confounder selection; he found that such tests can indeed lead to confidence intervals for the exposure effect that are invalid (i.e., cover the true exposure coefficient at less than the nominal rate) in typical nonrandomized studies. Some of these findings are illustrated in the nontechnical discussion by Robins and Greenland.38 Other authors have found that variable selection based on signficance tests can lead to nonnormality of coefficient estimators;47,48 a consequence of such nonnormality is that tests and intervals computed from the estimated coefficient and its standard error will not be valid.

Using a simulation study of case-control data, Mickey and Greenland38 compared signficance-testing algorithms with a change-in-point-estimate algorithm. Their study was limited to one binary exposure and one interval-scaled candidate variable (potential confounder); 48 situations were simulated, with 1900 trials per situation. They found that when conventional 5 per cent signficance levels were used in signficance-testing algorithms, important confounders were underselected; consequently, when confounding was present, the algorithms produced models that yielded biased estimates of exposure effect and invalid confidence intervals. The extent of confidence-interval invalidity (under-coverage) produced by signficance-testing algorithms was severe in some cases, and was directly proportional to the strength of the candidate variable as a confounder; the "double-testing" algorithm of Fleiss49 did especially poorly. These results support earlier arguments3-9 37 that conventional significance-testing algorithms (such as stepwise regression) underselect confounders and so yield invalid point and interval estimates. Other simulation findings23 indicate that product terms will be underselected by signficance-testing algorithms.

Mickey and Greenland38 also found that much of the bias in the signficance-testing algorithms could be prevented by increasing the nominal signficance level of the tests beyond 20 per cent (as recommended by Dales and Ury5). Nevertheless, among the variable-selection algorithms examined, the 10 per cent change-in-estimate method produced the most valid point and interval estimates.

It thus appears that the evidence to date tends to favor the change-in-estimate approach over signficance-testing approaches (such as conventional stepwise regression). It also appears that the major cause of the inferior performance of signficance-testing approaches is the low power (insensitivity) of the tests for detecting true confounders. This inferiority will be most pronounced in small samples, and can be mitigated by markedly increasing the signficance levels of the tests.

It seems likely that algorithms better than any in current use could be developed. For example, a potential source of difficulty for change-in-point-estimate methods is that they take no account of random variability in the coefficient estimates being compared, whereas signficance-testing algorithms may depend on such variability in an inappropriate fashion. Empirical Bayes estimation50 of the compared coefficients would introduce some consideration of random variability into a change-in-estimate algorithm, as would use of confidence-limit changes instead of point-estimate changes.

General Cautions in the Use of Algorithms

As noted earlier, it is essential to keep exposure forced in the model in order to avoid bias in the selection of confounders.3 X 37 Also, no matter how good its performance under repeated-sampling (long-run) evaluations, any algorithm may make what are obviously terrible choices in particular data sets;37 44 as a result, the user should monitor the results of each selection and deletion step. As a final caution, note that if (as in all the procedures discussed thus far) one uses the data to guide model and variable selection, the standard error of the resulting effect estimate is not correctly estimated by simply taking the standard error computed from the final model (as is always done).37, 44, 46,47 The error introduced by using the standard error from the final model to compute tests and confidence limits has not been thoroughly studied, although it appeared small in the special cases considered by Mickey and Greenland.38

Product Terms

Special distinction needs to be drawn between product terms involving exposure (such as x_1z_2) and not involving exposure (such as z_1z_2). In theory, at least, the criterion for entering product terms not involving exposure should be no different than for any basic covariate: terms which produce an important change in the estimated exposure effect when entered should be entered. Nevertheless, the need for such terms indicates that the chosen model form may be a poor one for adequate control of confounding by the basic covariates.

The entry of product terms involving exposure will often dramatically alter the exposure coefficient, simply because the *meaning* of that coefficient is changed when product terms are entered. As a consequence, the change in the exposure coefficient b should not be used to decide whether a product term involving exposure should be entered in the model. To illustrate, consider a logistic model with one covariate z: Without product terms, the odds ratio from exposure level x" versus exposure level x' is exp[b(x" - x')]; if the product term xz is entered and has coefficient c_{xz}, this odds ratio becomes

$$\exp[(b + c_{xz}z)(x'' - x')],$$

which is a function of z. Note that if an arbitrary number, say 20, was added to z, the value of b in the model without product terms would not change but the value of b in the second model (with xz) would be changed by $-20c_{xz}$ to compensate for the constant increase in the values of z. Such examples show that when a product term involving the exposure and a covariate is entered, the exposure coefficient b can change simply because it no longer represents the exposure effect at nonzero values of the covariate. Note also that the exposure effect may vary dramatically across z (so that the xz term is essential), but if z has a zero mean the exposure coefficient b may change very little upon entering the product term.

In most situations, if one includes the product ("interaction") of two variables in a model, one should also include the single terms ("main effects") for each variable X this rule is sometimes called the hierarchy principle.29 Models containing a product xz but not both components x and z depend on the coding of x and z for proper interpretation, and are usually implausible or nonsensical in subject-matter terms. For example, if x and z are binary indicators of the presence of two conditions (I = present, 0 = absent), a model with xz but neither x nor z entered implies that x and z have an effect only in one another's presence.

The decision to enter a product term involving exposure in the model may be based in part on the magnitude of the variation in effect (effect modfication) that is apparent when the term is in the model. In the preceding example, if z" and z' represent high and low observed values of z, the magnitude of variation in the effect of exposure level x" versus exposure level x' can be measured by the ratio of the odds ratios:

$$\exp[(b + c_{xz}z'')(x'' - x')] / \exp[(b + c_{xz}z')(x'' - x')] = \exp[c_{xz}(z'' - z')((x'' - x')]$$

If x and z are binary (zero-one) indicator variables, the last expression simplifies to $\exp(c_{xz})$. Nevertheless, while one may wish to include all product terms that correspond to large variation in the chosen effect measure, some consideration of statistical error is essential: such product terms form an integral component of effect estimates, and thus errors in their coefficient estimates will directly contribute to errors in the estimated effects. In particular, one must determine whether such effect variation as exists can be accurately estimated from the available data.

Here, signficance testing is useful. A small p value for the product-term coefficient indicates that at least the *direction* of variation in effect (e.g., increasing or decreasing with z) will be correctly estimated if the term is entered. On the other hand, a large p value indicates that such variation as exists cannot be estimated with any assurance that the apparent direction of variation is correct. Furthermore, if the coefficient of the product term cannot be accurately estimated, a more accurate estimate of the average effect of exposure may result from omitting the term. Thus there is some justfication for omitting from the model product terms involving exposure with "nonsignificant" p values, as long as one recognizes that this omission is more out of statistical necessity than actual lack of effect variation, and that the criterion for signficance may be best set much higher than *0.05.* In particular, one should not conclude from a large p value that no variation is present, since a large p value may also result if important variation is present but there is insufficient power to detect it. On the other hand, one must be aware that the need for such product terms indicates that the model form may be a poor one for representing exposure effects.

There has been considerable research on the effects of measurement error on effect estimates. Most of this literature is highly technical, but since measurement error is usually a major source of bias in estimates, the analyst should be aware of its potential effects.

Errors of measurement in a variable are *nondifferential* if they are not associated with the true value of any variables in the problem; otherwise the errors are *differential.* The errors in two variables are said to be *independent* if the errors in one variable are not associated with the errors in the other; otherwise the errors are said to be *dependent.* One important result is that if the errors in the outcome variable and a covariate are independent and nondifferential, the bias in the covariate coefficient estimate produced by these errors will be "towards the null," i.e. toward zero.5' If the covariate represents an exposure of interest, this means that the errors will contribute towards underestimation of the exposure

effect. If the covariate represents a confounder, the errors will result in an underestimation of the confounding effect of the covariate; as a consequence, entering the covariate in the model will not remove all the confounding by the covariate.

The effects of dependent or differential errors are much less predictable, and there are some exceptions to the aforementioned observations about independent nondifferential errors. In any case, it may be possible to correct for the effects of measurement error if sufficient background information on the error is available.8, 51

Interpretation of Modeling Results

Biological processes are not identifiable from epidemiologic data. This means that for any statistical model for epidemiologic data, numerous biological models will predict the data should follow that statistical model (e.g., see references *52* or *53).* Thus multivariate modeling of epidemiologic data cannot offer any qualitative improvement in our ability to test particular theories of biological action. Indeed, one could reasonably argue that no advances in chronic-disease epidemiology have involved multivariate modeling as an indispensable component.

One could further argue that multivariate modeling has increased the prevalence of statistical fallacies in biological inference. An example is the naive belief that product terms in a model correspond to or reflect biological interactions, such as synergy or antagonism. This fallacy has been discussed at length;8,9 it seems to stem in part from the fact that product terms are often referred to in the statistical literature as "interaction terms" (as in "tests for interaction"). Under certain restrictive assumptions there is a special correspondence between the need for product terms in an additive incidence model and the existence of certain types of biological interactions.54 In general, however, there is no simple relationship between product terms and biological interactions, or for that matter between statistical model forms and biological models.53

Such interpretational problems need not arise if the primary objective of modeling is to provide an efficient means of estimating epidemiologically meaningful measures (such as exposure-induced changes in average risk or life expectancy). This objective will, however, be better served if one reports the final modeling results in terms of familiar epidemiologic measures (e.g., model-fitted risks, risk differences or risk ratios), rather than in terms of the parameter estimates from the fitted model. To illustrate, suppose one fits a logistic model to estimate dependence of myocardial infarction risk on coffee use while controlling for several covariates, and that the final fitted coefficient for coffee use (measured in cups per day) is 0.098, with an estimated standard error of 0.050. One could then report this coefficient and standard error as the modeling result (perhaps along with the coefficient p value). Nevertheless, one could additionally provide a more intelligible presentation of the same results by transforming these statistics. If, for example, one was interested in an approximate relative-risk estimate, one could present the fitted odds ratio for the effect of five cups per day, $\exp[(0.098)5] = 1.6$, and its corresponding 95 per cent confidence interval, $\exp[(0.098 + 1.96[0.050])5] = (1.0, 2.7)$.

The preceding example was particularly simple, in that it required only an exponential transform of a single coefficient. More complex methods will be required in other situations. Such methods have been illustrated for a number of special cases, including estimation of standardized incidence from logistic models for cohort data,55 and estimation of exposure-specific incidence from case-control data.56

Sample-Size Requirements

As discussed earlier, model-based estimates may be rendered invalid by violations of model assumptions (specification error). However, even if all the assumptions are correct, bias can arise from inadequacies of sample size; I will refer to this problem as *finite-sample bias.* Such bias is also a threat to simple stratified analyses.

The fitting procedures used for most models of incidence and prevalence (including those discussed earlier) are *largesample* or *asymptotic* procedures. Examples include maximum likelihood (ML), Mantel-Haenszel (MH), and weighted least squares (WLS or inverse-variance weighted) methods for obtaining summary measures and for fitting risk or rate models. 7-9, 12-20, 24,57,58 The estimates from models fitted by such methods are *not* unbiased in typical applications, but for large enough samples their bias will be negligible.59 Unfortunately, it is difficult to specify exactly what "large enough" is, since this is highly dependent on the model and procedure under discussion. There have been many mathematical and simulation studies of a few special cases; only a cursory overview of major points can be given here.

For modeling purposes, the most critical point is that the degree of finite-sample bias present in modeling results tends to be directly proportional to the number of variables entered in the model. (This is obvious in the context of modeling purely categorical factors, in which every added variable reduces average cell size by a factor of two or more.) Consequently, attempts to control confounding by simply entering all possible confounders in a model can in some situations lead to more bias than they remove. Viewed another way, the threat of finite-sample bias further emphasizes the importance of valid confounder-selection methods. Categorical Factors

If only categorical factors are employed, it appears that the finite-sample bias of point estimates obtained from ML and WLS methods will be negligible if all the cells in the cross-classfication of all the variables in the model, including the outcome, are "large"; this means that one would require about five observations per cell before applying WLS procedures, and somewhat less for ML procedures.29 While this "large-cell" criterion is appropriate for simple standardization methods,8-11 it may be too conservative for modeling (especially when using ML estimation). A less conservative criterion, applicable to log-linear and logistic modeling of categorical factors, requires only that the marginal totals composing the sufficient statistics of the model (the "configurations" of the model29) be large. Unfortunately, this criterion does not readily extend to non-multiplicative models.

In many epidemiologic studies the cross-classfied data will be sparse, in that most of the cells will be very small; for example, in a matched-pair study, stratification on the matching factors will yield strata with only zero or one observation per cell.7~9 Methods for fitting models to such sparse data have been developed, the most common being conditional maximum likelihood (CML)7,8 and Mantel-Haenszel methods.57,58 The estimates derived from such methods will have negligible finite-sample bias if *either* all the cell numbers are large *or* if the number of informative strata are large (where "informative strata" can be roughly interpreted as those in which there is some variation in both the exposure

variable and the outcome variable). Like the earlier criteria this criterion may be conservative, but more accurate criteria have yet to be developed.

Sparse-data methods should not be confused with smallsample (exact) methods: the validity of sparse-data methods still requires that the total number of subjects be large, whereas the validity of exact methods does not depend on sample size.

If some of the variables are continuous and entered in the model as such (e.g., if x is cigarettes per day and entered as bx, instead of a series of category indicators), the above criteria are no longer relevant, and the total sample-size requirement will be reduced. In fact, as long as the number of parameters in the model is small relative to the total number of cases (and, for probability models such as the logistic model, the total number of noncases), effects may be validly estimated using incidence models (such as models 13 above), even if there is only one person at each observed exposure-covariate combination.24

Unfortunately, the reduction in sample-size requirement for continuous-covariate models is purchased by additional dose-response assumptions (e.g., that bx gives the correct form of dose response), and thus by additional risk of bias due to model misspecification. Furthermore, exactly what number of parameters is "small" depends not only on the sample size, but also on the model form and the joint distribution of exposure and covariates in the observed data. As a consequence, practical sample-size criteria for continuous-data models have yet to be developed.

Sample-Size Requirements for Tests and Intervals

The simplest and most common method of computing tests and confidence intervals from model results is the *Wald method,* in which for example the test statistic for the null hypothesis b = 0 (no exposure effect) is computed from the estimate b of b and estimated standard error s of b as b/s, and the 95 per cent confidence interval is computed from b _ 1.96s. For multiplicative models (such as the logistic), the Wald method appears adequate under the same large-sample conditions discussed above for point estimates. For other models, however, the sample-size required for the Wald method to be valid may be impractically large. In particular, ordinary Wald tests and intervals for parameters in additive relative-risk models can be grossly invalid, even at very large sample sizes.19,23

Moolgavkar and Venzon19 recommend that, for nonmultiplicative models, one employ likelihood-ratio tests and profile-likelihood confidence intervals in place of the usual Wald procedures. Such likelihood-based procedures appear to be valid for both multiplicative and additive models under the "large-sample" conditions discussed above. Unfortunately, profile-likelihood intervals cannot be conveniently computed with ordinary packaged programs.

Practical methods for computing exact tests and confidence intervals for logistic model parameters have recently been developed 60 and are available in the EGRET package.32 Although present versions of the program have rather severe upper limits on sample size, future developments in computing should allow these limits to be raised to the point that they surpass the minimum sample sizes necessary for valid use of asymptotic procedures. (This is already the case for analyses involving only three binary covariates.)

Conclusions

Both stratified and modeling analyses have limitations: validity of both stratified and modeling results depends on distributional assumptions which, on occasion, may be violated; both stratified and modeling analyses will have to confront the problem of control-variable selection; and both types of analyses will ultimately be limited by the size and quality of the data set, for neither approach can compensate for methodological shortcomings of the study (such as misclassification, selection bias, or lack of power to address the questions of interest).

The chief advantages of modeling are that it allows one to control more potential confounders than one could in simple stratified analysis, and it allows one to more precisely estimate variation in the effect of exposure across levels of other factors (effect modfication). These advantages come at the price of stronger assumptions about effects of exposure and covariates on outcome measures. This need for stronger assumptions implies that proper use of modeling will be more laborious than stratified analyses, since the assumptions should be carefully checked. More effort will also be required to correctly interpret and intelligibly present modeling results.

One can view the problems discussed here as consequences of a more general problem of causal inference in epidemiology: Because confounding is a complex and unknown function of many covariates, there is rarely enough information in nonexperimental data to allow one to construct an acceptably accurate estimate of exposure's effect from the data alone.3' One can obtain an effect estimate of acceptable precision only by making certain assumptions (e.g., that the effects of all the continuous covariates follow a linear dose-response curve); if, however, these additional assumptions are incorrect, then the estimate based on the assumptions will be biased to an unknown degree.37 A particular consequence of this dilemma is that one should not expect a model-based estimate to incorporate lower total error than a simple stratified estimate which is adjusted for the same covariates, unless some of the additional implications of the model (e.g., linearity of dose response) are approximately correct. This caution applies even if the model is only used to construct a risk score for stratification (e.g., as in reference 3).

Vandenbroucke' proposed a set of sensible guidelines for employing stratified and modeling methods in a complementary fashion. These can be summarized as stating that one should first perform a series of stratified and modeling analyses involving the same variables, in order to validate model choices and results against the stratified data; only then should one embark on modeling exercises that cannot be directly validated against parallel stratified analyses. To these guidelines, I would add that one should always check the assumptions underlying any model or estimation method used to derive summary statistics.

ACKNOWLEDGMENTS

I am grateful for the comments on this paper supplied by Alexander Walker, Joseph Fleiss, and the students of Public Health 223, Spring 1988, UCLA. This work was supported by Grant No. HS 05753 from the National Center for Health Senices Research and Health Care Technology Assessment.

REFERENCES

1. Vandenbroucke JP: Should we abandon statistical modeling altogether?
Am J Epidemiol 1987,126:10-13.

Gordon T: Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies. J Chronic Dis 1974; 27: 97-102.

3. Miettinen OS: Stratification by a multivariate confounder score. Am J Epidemiol 1976, 104:609-620.

4.Rothman KJ: Epidemiologic methods in clinical trials. Cancer 1977; 39: 1771-1775.

5. Dales LG, Un HK: An improper use of statistical signficance testing in studying covariables. Int J Epidemiol 1978; 4:373-375.

6. Greenland S, Neutra RR: Control of confounding in the assessment of medical technology. Int J Epidemiol 1980; 9:361-367.

7. Breslow NE, Day NE: Statistical Methods in Cancer Research. 1: The Analysis of Case-Control Studies. Lyon: IARC, 1980.

8. Kleinbaum DG, Kupper LL, Morgenstern H: Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Lifetime Learning Publications, 1982.

9. Rothman KJ: Modern Epidemiology. Boston: Little, Brown, 1986.

10.Miettinen OS: Standardization of risk ratios. Am J Epidemiol 1982 96: 383-388.

11.Greenland S: Interpretation and estimation of summary ratios under heterogeneity. Stat Med 1982; 1:217-227.

12. Gart JJ: The comparison of proportions: a review of signficance tests, confidence intervals, and adjustments for stratfication. Rev Int Stat Inst 1971; 39:148-169.

13.Checkoway H, Pearce N, Crawford-Brown D: Research Methods in Occupational Epidemiology. New York: Oxford, 1989.

14.Cox DR, Oakes D: Analysis of Survival Data. New York: Chapman and Hall, 1984.

15.Breslow NE, Day NE: The Design and Analysis of Cohort Studies. New York: Oxford, 1988.

16.Thomas DC: General relative risk models for survival time and matched case-control data. Biometrics 1981, 29:276-281.

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

18.Breslow NE, Storer BE: General relative risk functions for case-control studies. Am J Epidemiol 1985; 122:149-162.

19.Moolgavkar SH, Venzon DJ: General relative risk regression models for epidemiologic studies. Am J Epidemiol 1987;126:949-961.

20.White H: Estimation, Inference, and Specfication Analysis. New York: Cambridge University Press, 1989.

21. Lagakos SW: Effects of mismodeling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 1988; 7:257-274.

22.Cook RD, Weisberg S: Residuals and Influence in Regression. New York: Chapman and Hall, 1986.

23.Greenland S: Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983; 2:243-256.

24.McCullagh P, Nelder JA: Generalized Linear Models. New York: Chap man and Hall, 1983.

25.Pregibon D: Data analytic methods for matched case-control studies. Biometrics 1984; 40:639-651.

26.Lustbader ED: Relative risk regression diagnostics. *In:* Moolgavkar SH, Prentice RL (eds): Modern Statistical Methods in Chronic Disease Epidemiology. New York: Wiley, 1986.

27.Hastie T, Tibshirani R: Generalized additive models. Stat Sci 1986; 1:297 318.

28.Doll R: An epidemiological perspective on the biology of cancer. Cancer Res 1978; 38:3573-3583.

29.Bishop YMM, Fienberg SE, Holland PW: Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press, 1975.

30.Bennett S: An extension of Williams' methods for overdispersion models. GLIM newsletter (to appear).

31.Numerical Algorithms Group: The Generalized Linear Interactive Mod eling System (GLIMtm). London: Royal Statistical Society, 1987.

32.Statistics and Epidemiology Research Corporation. EGRET Statistical SoRware. Seattle: SERC Inc, 1988.

33.Haber M, Longini IM, Cotsonis GA: Models for the statistical analysis of infectious disease data. Biometrics 1988; 44:163-173.

34. Robins JM: A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor eflfect. Math Model 1986, 7:1393-1512.

35.Dixon WJ (ed): BMDP Statistical Software. Berkeley, CA: University of California Press, 1985.

36.SAS Institute Inc: SAS Guide for Personal Computers. Gary, NC: SAS Institute Inc, 1985.

37.Robins JM, Greenland S: The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986; 123:392-402.

38.Mickey RM, Greenland S: A study of the impact of confounder-selection criteria on effect estimation. Am J Epidemiol 1989, 129.125-137.

39.Whittemore AS: Collapsibility of multidimensional contingency tables. J R Stat Soc B 1978; 40:328-340.

40.Ducharme GR, Lepage Y: Testing collapsibility in contingency tables. J R Statist Soc B 1986; 48:197-205.

41. Greenland S, Mickey RM: Closed-form and dually consistent methods for inference on collapsibility in 2x2xK and 2xJxK tables. Appl Stat 1988; 37:335-343.

42.Greenland S: Cautions in the use of preliminary-test estimators. Stat Med 1989, in press.

43.Hauck WW, Anderson S: A proposal for interpreting and reporting negative studies. Stat Med 1986; 5:203-209.

44. Robins JM: The statistical foundations of confounding in epidemiology. Technical Report No. 2. Boston, MA: Occupational Health Program, Harvard School of Public Health, 1983.

45.Miettinen OS, Cook EF: Confounding: essence and detection. Am J Epidemiol 1981, 114:593-603.

46.Leamer EE: Specfication Searches. New York: Wiley, 1978.

47.Sen PK: Asymptotic properties of maximum likelihood estimators based on conditional specfication. Ann Stat 1979; 7:1019-1033.

48.Griffiths WE, Hill RC, Pope PJ: Small sample properties of probit models. J Am Stat Assoc 1987; 82:929-937.

49.Fleiss JL: Significance tests have a role in epidemiologic research: reactions to A.M. Walker. Am J Public Health 1986; 76:559-560.

50.Cox DR, Hinkley DV: Theoretical Statistics. New York: Chapman and

Hall, 1974.

51.Fuller WA: Measurement Error Models. New York: Wiley, 1987.

52.Robins JM, Greenland S: Estimability and estimation of excess and etiologic fractions. Stat Med 1989, in press.

53. Siemiatycki J, Thomas DC: Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 1981; 10:383387.

54.Greenland S, Poole C: Invariants and noninvariantsin the concept of interdependent effects. Scand J Work Environ Health 1988; 14:125-129.

55. Flanders WD, Rhodes PH: Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. J Chronic Dis 1987; 40:697-704.

56.Greenland S: Multivariate estimation of exposure-specific incidence from case-control studies. J Chronic Dis 1981, 34:445-453.

57.Davis LJ: Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. Biometrics 1985; 41:487-495.

58.Liang K-Y: Extended Mantel-Haenszel estimating procedure for mulb variate logistic regression models. Biometrics 1987; 43:289-299.

59.Bickel PJ, Doksum KA: Mathematical Statistics. Oakland: Holden-Day, 1977.

60.Hirji KF, Mehta CR, Patel NR: Computing distributions for exact logistic regression. J Am Stat Assoc 1987, 82:1110-1117.

61.Walter SD, Feinstein AR, Wells CK: Coding ordinal dependent variables in multiple regression programs. Am J Epidemiol 1987;125:319-323.