Ann. Rev. Public Health. 1983. 4:155–80

APPROPRIATE USES OF MULTIVARIATE ANALYSIS

James A. Hanley

Department of Epidemiology and Health, McGill University, Montreal, Quebec, Canada H3A 2B4

INTRODUCTION

Comparison of the articles in today's biomedical literature with those of twenty years ago reveals many changes. In particular, there seem to have been large increases over time in three indices: the number of authors per article, the number of data-items considered, and the use of multivariate statistical methods. While cause and effect among these three indices is unclear, there is little doubt that the growth in a fourth factor, namely, computing power and resources, has made it much easier to assemble larger and larger amounts of data. Packaged collections of computer programs, driven by simple keywords and multiple options, allow investigators to manage, edit, transform, and summarize these data and fit them to a wide array of complicated multivariate statistical "models." In addition to making it easy for the investigator to include a larger number of variables in otherwise traditional methods of statistical analysis, the increased speed and capacity of computers have also been partly responsible for the new methods being developed by contemporary statisticians. For example, some of the survival analysis techniques discussed below can involve several million computations.

How do these trends in the availability and use of multivariate statistical methods affect the health researcher who must decide what data to collect and how to analyze and present them? How does the reader of the research report get some feeling for what the writer is attempting to do when he uses some of these complex-sounding statistical techniques? Are these methods helping or are they possibly confusing the issue?

Unfortunately one cannot look to one central source for guidance about these newer methods. Descriptions of many of them are still largely scat-

0163-7525/83/0510-0155$02.00

tered in the (often highly technical) statistical literature or else presented in monographs in which the connections to other related techniques may not be very evident. Moreover, the reader is often not interested in references to the technical intricacies of maximum likelihood equations, to the methods of solving them, or to the computer program or package used to perform the calculations; rather he is worried about what the technique is attempting to do, what the parameters mean, and whether the assumptions and conclusions are appropriate.

The plan of this chapter then is not so much to review all of the recent developments in statistical methodology, but rather to use examples from the literature (*a*) to give an overview of what multivariate analysis is all about, (*b*) to describe, in general terms, what it can and cannot be expected to do, and (*c*) to discuss in a little more detail some newer techniques, as well as some that were developed some time ago but are only now becoming popular, namely (*i*) logistic regression, (*ii*) log-linear models for multiway contingency tables, (*iii*) proportional hazards models for survival data, and (*iv*) discriminant analysis.

MULTIVARIATE ANALYSIS: AN OVERVIEW

Scope

The term multivariate analysis has come to describe a collection of statistical techniques for dealing with several data-items in a single analysis. Although authors differ about where to draw exact boundaries, for example whether multiple regression is a univariate or multivariate technique, it is more a matter of semantics than it is of substance. I follow here the convention of others (10, 28, 33, 43) and define any analysis that involves three or more variables simultaneously as "multivariate." As such, the term multivariate analysis encompasses everything except confidence intervals, chi-square tests for two-way contingency tables, t-tests (unpaired), one-way analysis of variance, and simple correlation and regression. It includes a huge variety of techniques, since even with just three variables, there are a large number of possibilities (Table 1). The method of analysis depends heavily on whether one is interested in interrelationships or in comparisons, and on whether variables are qualitative or quantitative. The most I can do in this short space is to give a brief roadmap, along with pointers to helpful descriptions or examples. In many situations there will not be one single best method of analysis. As Bishop et al (10) point out, multivariate analysis should be thought of as a "codification of techniques of analysis, regarded as attractive paths rather than straightjackets, which offer the scientist valuable directions to try."

Table 1 A taxonomy of parametric statistical methods

	Response variable(s)			
	Univariate		Multivariate	
Stimulus variable(s)	Discrete [1]	Continuous [2]	Discrete [3]	Continuous [4]
Univariate				
Discrete	Contingency table	t-test	Multi-dimensional contingency table	Discriminant analysis
		One-way analysis of variance (Anova)		Logistic regression
Continuous	Logistic regression	Correlation		Multivariate regression
	Discriminant analysis	Simple regression		
Multivariate				
Discrete	Multi-dimensional contingency table	Multi-way Anova	Multi-dimensional contingency table	Multivariate Anova (Manova)
Continuous	Logistic regression	Partial correlation		Multivariate regression
	Discriminant analysis	Multiple regression		Canonical analysis
Mixed	Logistic regression	Analysis of covariance (Ancova)		Multivariate regression
	Discriminant analysis			Canonical analysis

Types of Analyses

Multivariate statistical techniques may be conveniently divided into those in which the variables involved (*a*) are all of "equal status" or (*b*) fall naturally (or with some gentle pushing) into two sets, those which are influenced (response variables) and those which influence (stimulus variables).

In the first group of techniques, which includes Principal Components Analysis, Factor Analysis, and Cluster Analysis, the emphasis is on the internal structure of the data-items in a single sample.

Principal Components Analysis (PCA) asks whether a large number of quantitative data items on each subject can be combined and reduced to a single (or at most a few) new variables (principal components) without losing much of the original information. In other words, the aim is to describe the subjects in terms of their scores (weighted sums of the original variables) on a much smaller number of new variables. These new variables (components) are built to be uncorrelated with each other, so as to avoid any redundancy. Also, they are arranged in decreasing order of "information" so that subjects are furthest apart from each other on the first component, less far apart on the second, and so on. If the total information in the original variables is "compressible," the subjects will not vary very much

on the latter components, and these can be discarded as redundant. Theoretically, since there are as many principal components as there are original variables, retaining them all permits one to reproduce the original data. An example in which the first principal component captured 67% of phenotypic variance in a population and was then used as a (univariate) index of overall body size in all subsequent analyses can be found in (11).

Factor Analysis (FA) asks whether subjects' quantitative responses on a large number of items and the patterns or correlations among these responses are "explainable" by thinking of each item or variable as measuring or reflecting a different mix of a smaller number of underlying "factors" or "traits" or "dimensions." As originally conceived, it differs from PCA in a number of ways. Whereas PCA "constructs" new variables from already observed ones, FA goes in the other direction, "reconstructing" the observed variables from latent ones. This distinction may have been too subtle and has largely evaporated; moreover, most computer packages use principal components as one way of extracting factors. Second, FA usually assumes that although factors are translated into variables by a "mixing formula" that is common to all subjects, variables will also contain some variation that is unique to each subject. Third, whereas PCA is more a data-reduction technique, FA seeks actually to understand and label the various "factors." Fourth, unlike PCA, FA does not necessarily produce unique answers. Indeed, there are many methods of factor analysis.

FA techniques are used primarily to explore relationships and to reduce the dimensionality of a data set. They serve more for instrument building and index construction than as direct analytic tools. However, although they are closely associated in psychology with establishing construct validity, at least one author (40) considers them generally inappropriate for developing health indices. These techniques have been somewhat more useful when the context is of a physical nature, such as in studying air pollution patterns (35), but even then, there are difficulties (5). The few published examples of FA in epidemiology and public health have either concluded the obvious or concluded nothing at all. The same seems to hold true for their use in the medical literature (28).

By far the majority of the applications of multivariate statistical methods in the health sciences are of the second kind, where one or more variables serve as "outcomes" or "responses" or "target variables" (28), and others serve as "predictors" or "explanatory" or "carrier" (48) variables. These two sets of terms are gradually replacing the older and quite misleading terms, "dependent" and "independent" variables. Some authors subdivide the explanatory variables further into those of primary interest ("study variables") and those of a "disturbing" or "confounding" or "nuisance" nature; I return to this subdivision below.

The main types of techniques for dealing with stimulus-response studies are presented in Table 1, in the form of a multiway grid, according to whether the stimulus and response variable(s) (rows and columns, respectively) are one or many and according to whether they are all recorded on continuous measurement scales, or are all categorical (discrete), or a mixture of both.

It is worth dwelling for a moment on a number of contrasts between methods for analyzing a single (univariate) response that is "measured" on a continuous scale (*column 2*) and those for a corresponding response that is discrete (*column 1*).

1. Methods for analyzing a continuous response have been in existence for considerably longer (the principle of least squares for fitting a regression line dates back at least two centuries; the newest technique, analysis of covariance, is at least 50 years old).
2. These methods tend to choose parameters and judge the amount of variation explained by various factors using easily understood "distance" criteria such as least squares; in other words, they keep the analysis in the same scale or "metric" that the actual observations were measured on; by contrast, methods for analyzing a discrete response tend to measure "distance" and "fit" using a probability or "likelihood" scale (likelihood is defined as the probability, calculated after the fact, of observing the data values one did). Although the method of fitting parameters to maximize the likelihood is in no sense inferior (if anything it is generally superior from a technical standpoint), it is easier for readers to comprehend changes in R-squared than changes in a log-likelihood!
3. Regression equations for a continuous response are usually linear, involving additive terms, and can be fitted from simple summary statistics, whereas those for a discrete response are often nonlinear, and need to be fitted iteratively with several passes through the data.
4. Estimates from these nonlinear regressions tend to have skewed sampling distributions, giving rise to confidence intervals that are not symmetric. The odds ratio used in epidemiologic studies is a case in point. Fortunately, it is often possible to work in a scale (e.g. log) in which the confidence interval will be of a simpler, symmetric, shape and to change back to the desired scale at the finish.

As can be seen from Table 1, multiway contingency tables, logistic regression, and discriminant analysis all play dual functions: they can be used to analyze either a single response variable and several stimuli or several responses and a single stimulus. Indeed, as discussed below, this ability to reverse a "multiple response, single stimulus" situation and cast it into a more traditional and more workable "one response, multiple stimuli" regression framework is key to handling multiple response data.

As one proceeds to treat several response variables and several stimulus variables simultaneously, the level of complexity increases considerably: all but the few with n-dimensional vision are quickly lost. As a result, even though computer programs are available for them, the two "doubly-multivariate" techniques, multivariate regression and multivariate analysis of variance (*Column 4,* Table 1), are seldom used. Instead, investigators try first to construct a "univariate" response and then relate this to the several stimulus variables.

MULTIVARIATE ANALYSIS: PURPOSES

In this section I discuss the *Why* of multivariate techniques. Although there are many different techniques, they share a number of common aims and a common underlying philosophy. Of course, they also have many of the same pitfalls; I discuss some of these below.

It is difficult to discuss multivariate techniques without also discussing the concept of statistical "models." It sometimes helps to think of these models as comprising two parts, one that is deterministic (dealing with the expected structure, almost like a "law") and one that is stochastic (dealing with random variation). This first part will be of a more global nature, describing what should happen. It might describe how two chemical agents act together on a host or how a lung grows in volume as it grows in linear dimensions; it might be based on or summarize a psychological or sociological theory; or it might be a rough straight-line or curvilinear pattern seen in the data, and which one wants to follow up. This "structural" part of the overall statistical model can be thought of as describing the systematic variations or pattern one would expect in a body of data. Although it is usually described in explicit mathematical equations with coefficients, powers, and the like, it does not have to be so precise. For example, the model might be: "the dose response relationship has no threshold," or "the underlying curve is expected to be concave," or "the risk of cancer will vary with age and be different in exposed and nonexposed groups, but the risk of cancer among the exposed relative to that among the nonexposed will remain the same over all ages."

The other part of the model, which some would regard as the probabilistic element, deals with the deviation of the observed data from the postulated pattern. It is often difficult, however, to separate the two parts of the overall model, since it is not clear where prior knowledge (pattern) ends and ignorance (unexplained variation) begins, i.e. whether aberrations are observed because the postulated pattern is a poor one (lack of fit) or because of some other reason. Although this separation into systematic and random components, i.e. into signal and noise, is often used for responses that are

recorded on a continuous scale, it is done much less frequently for binary responses. One learns very early in linear regression to think of both the systematic (the straight line) and the random (the scatter of the individual points from the line). In a binary regression, one still thinks of a systematic line (possibly "s-shaped" such as a probit or logit curve) but seldom stops to think about the noise about this curve. Part of the reason for not doing so is that the curve is fitted using likelihood, rather than distance, as the metric and part is that the variation is binary, not continuous. The virtue of this "systematic plus random" paradigm has been recently illustrated in the Generalised Linear Interactive Modelling (GLIM) computer program (6): the program "generalizes" to a wide variety of continuous and binary response regressions by using different probabilistic models (Gaussian, Binomial, Poisson, etc) and different "link functions" for changing the systematic portion of the model from straight line to s-shaped and so on. GLIM points out that in fact there is a "distance" minimization intrinsic to the method of Maximum Likelihood.

With this preamble, I now go on to discuss, via examples where possible, the main aims and uses of multivariate statistical techniques and models. We see four main purposes:

1. to summarize, to smooth out, to see patterns
2. to make comparisons fair, to compare like with like
3. to make comparisons clear, to remove noise
4. to study many factors at once, to explain variation.

Purpose 1: To Smooth Out, to See the Forest From the Trees

How might one investigate whether and in what way breast cancer incidence rates have changed over time, using the available incidence data from 1935 to 1980 collected by the Connecticut tumor registry? This is an example of a single target variable, binary in nature (cancer or not), and the influence of two "stimulus" variables, age and year of birth. Suppose we know the numbers of cancers in each of nine five-year periods from 1935 to 1980 for each of 12 five-year age groups, along with numbers at risk in each of these $9 \times 12 = 108$ "cells."

As a first step, one could plot the 108 observed age specific incidence rates against age and use lines of different colors to connect together the data points to form age-specific incidence curves for the different birth cohorts. Some of these plots, derived from the data published in Reference (60), are given in Figure 1 (*left*); they show that although there seem to be cohort effects, it is difficult to measure them very precisely from these "raw" data points. Most would believe that the jagged pattern of straight-line segments has no special meaning, and would think of it only as noise that is obscuring

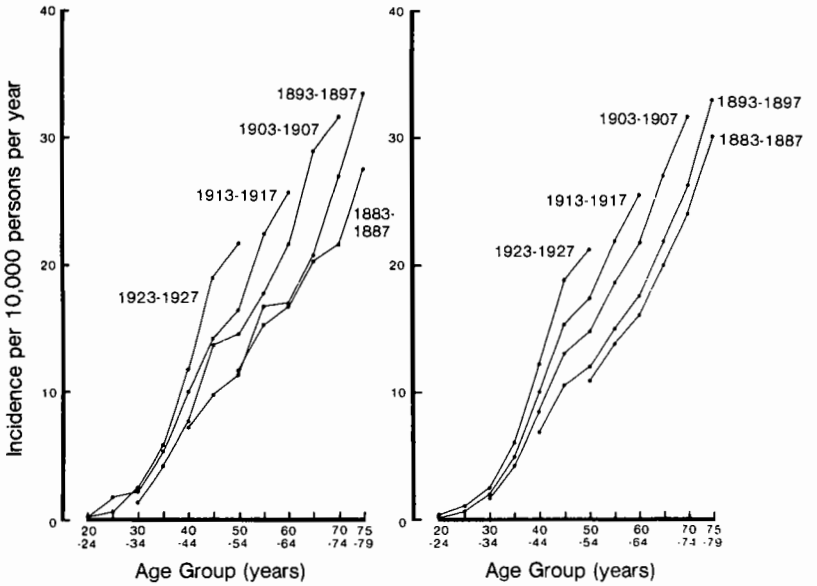

Figure 1 Age and cohort specific breast cancer incidence rates in Connecticut, 1935–1980. *Left:* observed rates in five selected cohorts. *Right:* smoothed rates obtained from a multiplicative model.

the "real" underlying pattern. They would prefer instead a series of "smoother" incidence plots, one for each birth cohort. These systematic "curves" could be produced by smoothing each one by eye, but doing so would ignore two considerations: first, the rates are calculated from numerators and denominators of varying stability (something the eye looking at a data point cannot see) and, second, if rates vary smoothly across age, they probably also do so across cohorts. Thus, one would need to smooth in two directions at once. This could be done by postulating a single "parent" plot, consisting of 12 points (left unsmoothed to begin with) and specifying that the plots for the separate cohorts are to be obtained by multiplying the parent plots by separate proportionality factors. Admittedly, the task is too complicated to perform manually, but that is hardly an obstacle. This "model-fitting" serves a number of purposes.

1. It produces more realistic plots, and uses many fewer numbers or "parameters" to do so (for the entire dataset, there would be 20 cohort parameters and 12 age parameters).
2. It draws the eye away from the randomness (which should be binomial or Poisson around each fitted point) and toward the pattern, in the same way that an image becomes clearer the further away one stands from its rough grain.

3. The raw plots generated from the earliest and latest cohorts are based on fewer data points (age groups) and are the most difficult to judge, whereas the corresponding synthetic plots are generated from parameters that were estimated from the entire data set. This concept of borrowing strength from neighboring data points is a central one in multivariate analysis.

To some, the idea that it takes $20 + 12 = 32$ numbers to describe 20 plots is still unappealing. Surely, they might argue, the parent plot (12 parameters) is not in reality so complicated that it could not be described by a truly smooth, two or three parameter curve or possibly by separate curve segments for pre- and post-menopause. Likewise, they would consider it quite likely that the 20 proportionality factors by which this incidence curve changes from cohort to cohort themselves form a smoothly changing series that could be described by many fewer parameters. Others would argue that one should "leave well enough alone" and that any further smoothing or modeling might do more harm than good. In this example, with the relatively large amount of data, the additional reduction might indeed be unnecessary; however, had the data been scarcer, it is likely that the further smoothing would have been required.

There are two more serious objections to the approach just described. First, for any one cohort, the entire parent curve is multiplied through by the same value. This does not allow for cohort effects that are age-specific, e.g. changes in the age at which women in different cohorts completed their first full-term pregnancy might affect the risk of premenopausal breast cancer differently than they would the risk of postmenopausal cancer. This is an example of what statisticians call an interaction: an effect of one factor (age) that is not constant across different values or levels of another (year of birth). Second, the actual goodness of fit of the smoothed curves to the raw data points needs to be evaluated. Before it is, any other expected or suspected patterns can be built into the fitted curves (provided that there are not so many assumptions and exceptions that one ends up with almost as many parameters as data points) and their "fit" tested by examining whether in fact the fitted curves come closer to the raw data points than before, and whether the discrepancies (residuals) are more or less haphazard and unexplainable. See (51) for a nice account of the use of regression models in studying regional variations in cardiovascular mortality.

As already mentioned, the assumption of smoothness and of orderly patterns of change is a central one in multivariate analysis. It stems from the belief (or maybe just the hope) that nature is basically straightforward, and that if there are no good biologic or other reasons to the contrary, relationships tend to be linear rather than quadratic, quadratic rather than

cubic, etc. [For a description of this principle of "Occam's Razor," see Ref. (54).] In the breast cancer example just described, however, the changes in some possible risk factors have been "man-made" and more sudden, e.g. world wars, shifts in childbearing habits, oral contraceptives, etc, and it may indeed be some sudden changes in incidence (as it was with liver cancer) that alert us to newly introduced causative (or protective) agents.

Purpose 2: To Make Comparisons Fair

The majority of analytic studies involving humans are of an observational, rather than experimental, nature. As a result, when one compares responses of one group with those of another, the fundamental scientific principle of holding all other factors constant or equal may be violated. Consequently, differences (or nondifferences) in responses may be caused by differences (imbalances) in factors that cannot be controlled experimentally, rather than by the basic variable (groups) under study. Such variables, referred to as "confounding," "disturbing,' or "extraneous" by various authors, can, if ignored, have insidious effects. For example, male and female applicants had similar acceptance rates in each of the various faculties at Berkeley, yet the crude overall (schoolwide) acceptance rate for females was considerably lower (9) because females were more likely to apply to those faculties for which the acceptance rates were lower. This artifact is referred to as Simpson's Paradox, and is always a possibility in observational studies.

Although standardization for imbalances (e.g. in age or sex), used to put comparisons of rates on a fair footing, is one of the oldest epidemiologic tools, it is sometimes ignored. A particularly distressing example is the recent controversy in the US and Britain regarding possible cancer-causing effects of water fluoridation, based on findings that cancer rates had increased more in cities that had been fluoridated than in those that had not. As subsequent articles pointed out, these effects disappear if differences in the demographic structure of the two groups of cities are taken into account. [See Refs. (19, 20) for some recent British investigations and a guide to the earlier US studies.] One of the benefits (didactically speaking) was the helpful illustration of two methods of standardization (41).

Standardization was also used recently in a slightly different context (31). It showed that, although the crude infant mortality rate is much higher in Massachusetts than in Sweden, if infant mortality rates in the two areas were standardized for birthweight, Massachusetts would actually have a slightly lower one. The point of the analysis was not to explain away or hide the differences in mortality rates, but rather to show that it is an advantage in birthweight, and not the superiority of Swedish hospital care, that gives Swedish infants a survival advantage. Although the country of birth seems as if it is the main study variable and birthweight simply a "nuisance

factor," in reality, birthweight matters everything and country not at all. Luckily, as the accompanying editorial pointed out, of the two variables, birthweight (and through it, presumably the infant mortality rate) is the modifiable one.

To many, the term multivariate analysis has come to mean a statistical model that uses regression-type equations and distributional assumptions to link observed values of a response variable to values of various explanatory variables. Up to this point, the discussion in this section has centered around yes/no responses and explanatory variables that were either naturally discrete (sex, race, country, faculty) or forced to be discrete (age group, birthweight group). These types of data lend themselves to such straightforward tabulation and computation of standardized rates (a technique known as a *stratified analysis*) that one might rightly ask what is "multivariate" about the method other than the fact that it involves three or more variables. The answer is that by averaging results over a number of cells (strata), analysis techniques such as that of Mantel-Haenszel (used to combine data from several 2 X 2 tables into a single summary) do, at least implicitly, assume that all tables are measuring a common odds ratio. If the underlying odds ratios are not the same in each table, then the single odds ratio produced by the Mantel-Haenszel technique measures a weighted average of these separate ratios, and since the weighting is related to the relative sizes of the separate tables, the average will be somewhat arbitrary. The same is true of rates that are computed with reference to some standard population—they depend on the assumed mix of categories in the model population. This emphasizes a central issue in all multivariate analyses: One cannot adjust or standardize a comparison without making certain assumptions. Probably the best way to view statistical models is as "a series of approximations to the truth": one can realize that the assumptions (model) used to adjust a comparison may not be entirely correct but proceed as best one can, or one can forego any adjustment because one did not realize the need or was afraid to make assumptions. It is a choice between the results being approximately correct and being precisely wrong!

To end this section, I discuss briefly situations in which the response variable is continuous rather than discrete (I shall discuss more complicated methods for standarizing rates, below), and address issues of matching and of adjustment by regression. In some experimental studies, it is possible to compare responses to two or more maneuvers applied to the same individual. The advantage of having each subject serve as his own control is obvious: the comparison is immediately fair with respect to an infinity of variables that could otherwise theoretically bias it. When this is not possible, the next best thing, using balancing or randomization (or both), to equalize the two groups receiving the different maneuvers, is often difficult.

This is especially true if the numbers in the two groups are so small that it is impossible to balance them adequately, or if the study is an observational one and the groups have already been formed. For example, in a recent study (42) comparing the ventilatory function, as measured by forced expiratory volume (FEV), of workers who had worked in a vanadium factory for at least four months with that of an unexposed reference group, investigators matched the subjects for two variables known to influence lung function: age (to within two years) and cigarette smoking (to within five cigarettes daily). However, since the two groups differed by an average of 3.4 cm in height, a variable with a very strong relationship to FEV, some standardization or adjustment was required. The authors achieved this using the finding of Cole (17) that past age 20, the predicted FEV for a man of a certain age and height is approximately of the form

$$\text{FEV} = \text{height-squared} \times (a + b \times \text{age})$$

Both members of each matched pair were already concordant for age and smoking; thus, if one simply divided each man's recorded FEV by his squared height, the resulting paired values could be taken as FEV's that were adjusted for one member being taller or shorter than the other. Since the effect was as though the pairs had been also matched for height, the comparison was carried out using a straightforward paired t-test on the differences in the pairs of adjusted FEV's. Although the task will often be more difficult than in this elegant example, the principle generally remains the same: one calculates what each subject's response would be expected to be if all of the variables that distort or bias the comparison were held equal, say at the mean of each covariable. The term *analysis of covariance* (3, 4) has generally been applied to adjustments of a simple additive nature, but as we have just seen, if some other relationship more appropriately and more accurately describes the way in which the covariate(s) affect the response, and if it is easy to derive, it is certainly preferable. Usually this relationship between response and confounders is estimated "internally" from the data at hand, unless the study is small and some outside norms (e.g. weight and height charts, dental maturity curves) are deemed better. Researchers generally feel safer using internal standardization; by doing so, they avoid problems of different measurement techniques, inappropriate reference samples, etc. In the vanadium study just cited, one could actually test Cole's FEV internally in the group of nonexposed workers. If the study did not have a pure unexposed group, and relied instead on the within-group variation in the amount of exposure, one would probably treat the exposure more as a continuous variable and use a multiple regression approach.

Purpose 3: To Sharpen Comparisons

With the considerable emphasis on using multivariate techniques such as analysis of covariance to control bias, it is often forgotten that these methods may also be used to eliminate unwanted variability and thereby increase the signal to noise ratio. Users and readers alike often have the impression that if the subjects in two groups are balanced with respect to some major explanatory variable, there is no need to account for that variable in any analysis. This misconception is especially likely to arise in a large randomized trial in which the balance is expected, and seen, to be good. A recent example (50), dealing with a subject that may be more amusing than relevant from a public health viewpoint, illustrates the usefulness of analysis of covariance in increasing the precision of various comparisons.

Figure 2a shows the responses of the 25 subjects in each of the five groups. The considerable "within group" variation makes it difficult to judge whether, compared with this large source of "noise," any apparent systematic differences in longevity among the groups are more random than real. Some guidance is given by Figure 2b, which shows that much of the noise is due to the fact that larger subjects tend to live for longer and smaller ones

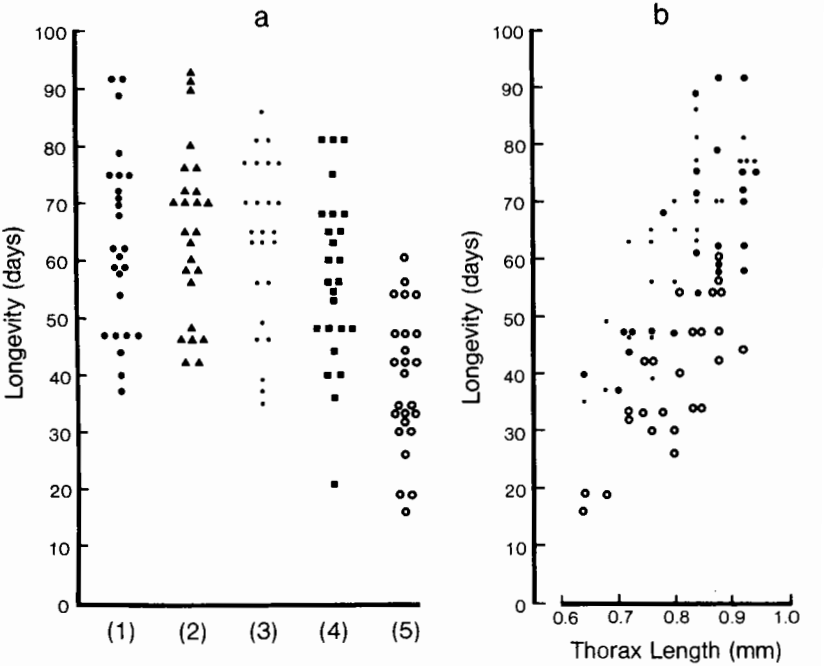

Figure 2 Longevity of male fruitflies in relation to amount of sexual activity: (*a*) observed lifetimes in each of five control and experimental groups; (*b*) observed lifetimes in relation to size (three groups shown).

for shorter lengths of time. Faced with this, it is clear that the smaller subjects should be compared with other smaller subjects and larger ones with other larger ones. This way, within each size category the within-group variation would be considerably less, thereby allowing systematic between-group differences to "shine through" more easily. Thus, the strong relationship between longevity and size would become irrelevant. Indeed, the experiment could have been planned very tightly by matching on size and analyzing the intergroup comparisons by paired t-tests or other techniques for matched subjects.

However, this would pose problems if subjects were to be individually matched, since it might not be possible to obtain perfect matches. Moreover, in human studies, with fewer cooperative subjects to subdivide along a wide scale, with many variables to match on, with the difficulty of obtaining all matching data before forming study groups or (in the observational study) with groups who had formed themselves well before any study was contemplated, the difficulties become formidable. To understand how a multivariate analysis can help to overcome these practical problems and allow the researcher to still benefit from a more tightly controlled study, imagine for the moment that the longevity study had been performed not with 25 but 10 subjects per group. Figure 3a illustrates one such possibility. At this point, any efforts at forming size categories, as in Figure 3b, would lead to a certain amount of "trading," i.e. it might be that a slight advantage for one group in the "small-size" category could be balanced off against a disadvantage for that group in the "next size up" category. However, one might not be so lucky, and in any case the within-group responses in the now broader size-categories will be larger. Intuitively, one would like to "homogenize" the subjects within each category by making them all the same size. One way to do this would be to forcibly "slide" the points laterally until they coincide on the size scale as in Figure 3c; to compensate for this change, one would likewise slide the responses vertically by corresponding increments, using an appropriate "exchange ratio" or slope. The slope could be estimated from the data by regression methods. This simple concept of equalization, which is the basis for analysis of covariance, is largely obscured by the all-in-one computational packages that fit the slope and calculate the between and within group variation in a single step. To perform an analysis of covariance for two extraneous variables $X1$ and $X2$, one might imagine responses plotted as vertical bars standing on a two-dimensional grid of $(X1, X2)$ points. To homogenize the responses with respect to $X1$ and $X2$, one would first slide the bars diagonally along the grid to a single $(X1, X2)$ point and adjust each vertical height (response) by the sum of $B1 \times$ shift in $X1$ and $B2 \times$ shift in $X2$, where $B1$ and $B2$ are regression coefficients describing how the response changes with each variable (while holding all other variables constant).

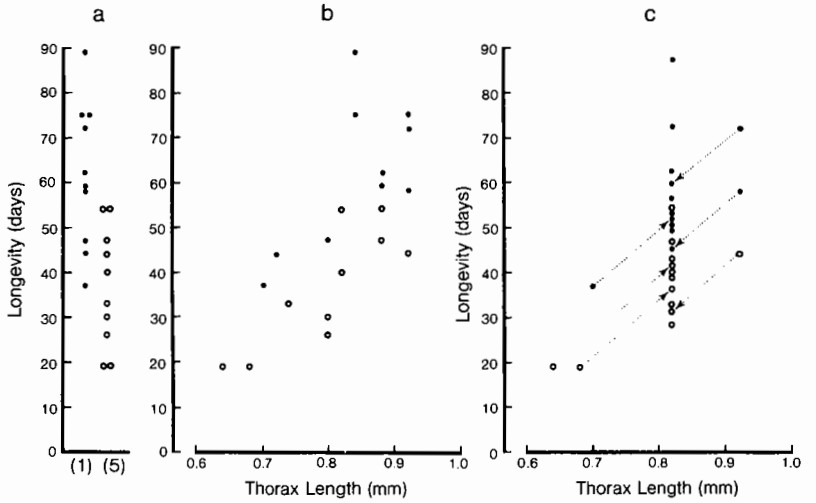

Figure 3 Longevity of ten fruitflies in each of two groups: (*a*) longevity shows wide within-group variation; (*b*) subjects cannot be easily matched on thorax size; (*c*) "matching" produced by analysis of covariance; lifetimes are adjusted to what would have been expected had each subject's thorax length been 0.82 mm (adjustment process shown for six subjects). Analysis (i) corrects imbalance of 0.40 mm in average thorax lengths of two groups and (ii) reduces within-group variation.

Provided that a large fraction of the observations ("degrees of freedom") do not need to be expended in estimating what the form of the adjustment should be, this analysis of covariance technique can be extended to several extraneous variables.

Purpose 4: To Study Several Factors

In many health studies, there will be several stimulus variables of primary interest. For example, one might investigate what characteristics of school-children and their environment are associated with their caries experience. Even when the stimulus variables are categorical, the classical multiway analysis of variance is rarely appropriate for such observational studies, since the cells will be of varying sizes (the "design" will be unbalanced). Instead, one usually analyzes such survey data by multiple regression methods, using indicator ("dummy") variables for factors that are categorical (e.g. gender). It is this flexibility that makes multiple regression so attractive. Indeed, if one had to choose between becoming familiar with classical analysis of variance or with regression techniques, one should probably choose the latter: it can accommodate a mixture of categorical and continuous variables and can evaluate these factors in the presence of other variables that are of a disturbing nature rather than of any direct interest. The key to understanding both its strength and at the same time its synthetic

nature is realizing that it produces an estimate of the effect of a factor even though there may be no two individuals in the data set for whom all other relevant factors are in fact equal. I comment below on the opportunities for misinterpretation of multiple regression analyses; however, there are three points that are specifically related to "risk-factor" studies.

The first concerns the situation in which the distributions of the different risk factors are not independent of each other in a fairly small data set, that is, if risk factor B was present in different proportions in those individuals who had risk factor A and in those who did not. Here, even if the two factors truly contribute independently in an additive way to the response being studied, it is still not possible to obtain independent estimates of these two effects from the sample. The two estimates will be correlated, and each estimated effect will have to be presented "adjusted for the other." This problem, addressed under "collinearity" in statistics textbooks, can become quite serious in health studies if one cannot obtain a good spread of one factor, such as amount of chronic exposure to loud noise, across each level of another factor, such as age. In such situations, one may have to adjust the response (hearing loss) through the use of some outside age-specific norms for hearing loss in unexposed individuals.

The second concerns how to deal with the variable "age" in the following hypothetical stepwise multiple regression analysis of caries experience.

Factor	*Multiple R-squared*	*Change*
Age of child	43%	—
Education of mother	50%	7%
Intake of fluoride	55%	5%
Frequency of toothbrushing	59%	4%
Consumption of soft drinks	62%	3%

It is mistaken to interpret this kind of output as evidence that the last four factors account for "only 19%" of the variance, when in fact they account for 19 out of the 57 percentage points (100 minus 43) that remain after age has already been accounted for. Because the crude or total variation in caries in this study could have been arbitrarily widened or narrowed by simply studying a wider or narrower age range, and because the real interest is in why two individuals, of the same age, have had different caries experience, the variation introduced by studying children of different ages is quite irrelevant. It can be removed either by actually subtracting from each response an amount attributable to age and analyzing the residuals or, as was indicated above, by a conceptual subtraction in which age is left in the analysis of variance table but all further explanations of variance are measured out of 57 rather than out of 100. A formal statistical test of whether

these latter variables are really explaining any variation does in fact judge their contribution relative to what is left to explain, rather than to what has already been explained. [See Reference (15) for a useful discussion of the appropriate terminology for variables such as age and sex.]

The third point deals with submitting our caries study, with its multitude of explanatory variables, some of them demographic, such as language group, race and place of residence, and some that are more "basic" (including life style characteristics such as diet and quality of dental care) to a multiple regression. Because either set of variables, or a combination of variables from the two sets, might do well in explaining the observed variation in caries, one needs to be careful and be guided by the purpose of the analysis. Broad demographic labels, e.g. language spoken at home, that are only predictive through their association with more causal variables, are more relevant for using the results locally to identify those with greater dental care needs. However, the results of an analysis that focuses on direct or proximal variables, e.g. mother's knowledge of oral hygiene practice, are more likely to be transportable to other settings and to uncover mechanisms governing caries. If one does not separate these two sets of variables, but instead submits them all to a regression analysis, the resulting picture may be quite blurred: part of the variance associated with a certain factor may be correctly credited to that factor, whereas part of it may be credited to some demographic variable that is only a proxy for the factor. For the results to make sense, the variables offered to a regression must first make sense.

SELECTED MULTIVARIATE TECHNIQUES

In this section I discuss a number of multivariate techniques for analyzing discrete responses, techniques that have become popular in the last ten years.

Discriminant Analysis

Discriminant Analysis (4, 47) began as a method of predicting to which of several categories an individual belonged, using several pieces of information collected about him and similar information collected about past individuals known to belong to the various categories. It has come to have three main uses (see Table 1): (*a*) as a way of carrying out a multivariate t-test comparing two or more samples on several continuous-type responses simultaneously and as a means of controlling the false-positive results associated with separate analyses (33); (*b*) more in its original spirit, in screening, diagnosis and prognosis (32, 64); (*c*) as a form of multiple regression for categorical responses (43).

If Discriminant Analysis is used in the second way, to simply construct a one-dimensional score from many variables, and if the scores one obtains are used as though they were the result of a single test (25), few distributional assumptions are needed regarding either the discriminating variables (indicants) or the resulting scores. Further, if one has sufficient numbers of proven cases one can use the empirical distributions of scores to construct score-specific predictions (25, 53). With fewer cases, one will need to fit some distribution to either the scores or to the discriminating variables. The third use, to adjust for disturbing variables before comparing proportions, or to study the effects of several variables on the probability of a certain yes/no outcome, is best discussed in the context of multiple logistic regression.

Multiple Logistic Regression

Logit and probit curves (21) have been used for several years to study a binary response to a single stimulus variable. However, it was only in the early 1970s after the publication of three signal articles (2, 62, 65) and a comprehensive monograph (21) that the "logistic model" began to be used for studying multiple stimuli. It was not until the 1980s that the technique was integrated into biostatistics textbooks (3) and took its place as the primary method for analyzing the relationship between a binary response and several discrete or continuous stimulus variables. It now stands in the same relation to binary response data as classical regression does to continuous response data.

To these descriptions of the "logic" of logistic regression, I add one point dealing with its historical evolution. If one works with the odds (rather than the probability) of a yes/no event in relation to a series of explanatory variables $X1, X2, \ldots$, the logistic model implies that the logarithm of this odds can be written as

$$\log(\text{odds of yes/no}) = B0 + B1.X1 + B2.X2 + \ldots$$

If one thinks of the right-hand side of the equation as a score S, then it will have different distributions in the "yes" and "no" groups, just as in a discriminant analysis. The first justification for the multiple logistic model was that if the Xs in the "yes" and "no" populations follow two multivariate normal distributions, then the Ss will have univariate normal distributions. Then, if these two univariate normal distributions have equal variances, one obtains the logistic curve (62). It is still not well recognized that although these conditions are indeed sufficient to produce the logistic relationship, they are not necessary. First, one does not need multivariate normal Xs in

order for the Ss to be approximately normal; if there are sufficiently many of them to add together, if they are reasonably uncorrelated, and if they do not have highly skewed distributions, the central limit theorem guarantees distributions of Ss that are close to normal. Second, one does not even need the Ss to have normal distributions: several other pairs of distributions of scores will also generate the logistic relationship. The interested reader can verify this for himself, using as an example the data in Table 1 of Reference (14), which shows two Poisson-like distributions with the score (number of symptoms) averaging 0.5 per individual in the "no" group and 2.7 in the "yes" group. The important point is that even though logistic regression is now regarded as simply a convenient functional form for linking probabilities to explanatory variables, it does have some historical and statistical basis.

Epidemiologic studies, and their use of risk ratios (also called relative risks) to report comparisons from prospective (cohort) studies, have done much to popularize logistic regression (indeed one could say that the technique began with the Framingham Study). Studies involving a binary response and multiple stimuli do not need to force the stimulus variables into discrete categories required for a Mantel-Haenszel analysis but can use all the information in every variable: the coefficient for the main exposure of interest leads immediately to the odds ratio and the relative risk. In one recent study (34), the results were also presented as observed and expected numbers of cases, in much the same spirit as is done for comparisons of mortality rates.

Logistic regression has also become quite popular for analyzing case-control studies, as a result of some very significant insights into the logical connections with corresponding methods for cohort studies (12, 13, 56). Furthermore, as computing becomes cheaper, it probably will largely replace the traditional two-group linear discriminant analysis. It is a little more difficult to know how useful logistic regression will become for multi-category responses ("polychotomous logistic regression"), since there are several ways one might contrast the categories (29). Recent work, performed in the context of trying to place patients into one of several diagnostic categories on the basis of a number of binary indicants (symptoms, findings, test results etc), suggests that some of these methods are at least feasible (A. Wijesinha, unpublished information).

The arguments of Dawid (23) add further theoretical justification for choosing a more robust prospective model, such as logistic regression, over a retrospective one, such as discriminant analysis. By "prospective" Dawid means predicting responses from the given indicants, and by "retrospective" he means predicting the distribution of indicants from knowledge of the response.

In spite of these theoretical advantages, however, some direct comparisons of various discrimination techniques have not always shown a definitive advantage for logistic regression (30, 61). As Fienberg (29) points out, however, there is a difference between using these competing methods for discrimination (where it is the overlapping part of the score distribution that contributes to misclassification rates) and using them to make accurate probability predictions or adjustments across the entire probability scale. The fact that discriminant analysis can hold its own in the task for which it was first designed is no guarantee that it will be equally good for other purposes. Nevertheless, since it is inexpensive, it will probably continue to be used to screen for possible influential confounding variables before undertaking a logistic regression.

A disadvantage of logistic regression is that results are often presented as odds or log odds, or worse still, as unitless coefficients rather than using the more familiar probabilities. To aid with these nonlinear concepts, it is often appropriate to translate to log-odds back into the more familiar probability scale. Recent articles that used graphical methods (36, 38) or expected numbers of events (34) to describe the fitted models have been especially helpful in this regard.

Log-Linear Models for Multiway Tables

If the stimulus variables can all be considered categorical, binary response data can also be assembled into multiway contingency tables and analyzed using multiplicative models (the same one used to compute an expected cell entry in the simple 2 X 2 table), which become additive when transformed to a log scale. The logic behind these models and how they are fitted (almost always by computer iteration) is well described in recent textbooks (3, 10, 29). The attractiveness of log-linear models for multiway tables lies in their parallels with classical analysis of variance models, in their use as a way of standardizing comparisons of rates in complex data sets, and in the ease with which interactions and confounding variables can be identified. They agree with logistic models if one fits as many parameters as there are cells. The fits to the breast cancer incidence data discussed above are examples of a log-linear approach: the simplest curves involved points that were products of an average age-specific curve and different proportionality factors for the different cohorts. The best-fitting parameters (32 in the first "model" considered) could be fit by a variety of techniques, such as logistic regression of the 108 numerators and denominators on 32 dummy variables or a 20 by 12 by 2 contingency table analysis (with a number of cells missing because the cohorts were too young or the cancers occurring early in life to the furthest back cohorts were not in the registry). Some drawbacks to analyzing a binary response by a contingency table, rather than general

log-linear regression, approach include the fact that it tends to treat the response variable the same way as the stimulus variables, that it worries about reproducing the interrelationships among the stimulus variables, and that variables that are not categorical have to be made so.

Regression Methods for Life-Table Analysis

Although the life-table (used in the broad sense for techniques that analyze the time until events happen) has long been an essential epidemiologic tool, it is only in the last decade that it has been adapted into a multivariate method (22, 46). As are most of the other methods described in this section, it is log-linear, with the log of the time-specific "mortality" rate (hazard) linked to the "average" hazard and to the explanatory variables through a linear regression. The main differences from logistic regression are that the "average" hazard is not a single quantity but a function of time and that it is estimated nonparametrically. In the simplest case, the relationship between the hazard and the explanatory variables is assumed to remain constant over time. This constancy does not seem to hold always (45, 55) and statistical tests based on this "proportional hazards" model (58) can be quite misleading. Fortunately, some work has emerged (44, 57) and more is under way to produce diagnostic tests for checking the appropriateness of the assumed model, and suggesting when effects of variables should be allowed to vary over time.

POSSIBLE PITFALLS IN MULTIVARIATE ANALYSIS

This section deals with potential risks in the use of multivariate analyses. I do not discuss the risks of specific techniques, details of which will be found in the appropriate textbooks, but rather the issues that cut across techniques, and that arise simply because data are multivariate. Indeed the main message is that the more multivariate the data, the greater the opportunities for problems.

Adding Noise

Although I stress above that including other variables in the analysis of a comparative study can sharpen a comparison, it can also dull it, especially if the user allows a stepwise regression to decide which of many other variables are important. The gain or loss in precision will depend on how strongly these other variables influence the response being studied. For example, including the last digit of each individual's telephone number in a multiple regression will waste one degree of freedom or the equivalent of one individual. Worse still, if the average value of this variable is not equal in the groups being compared (and in any one study with small groups, it

almost certainly will not), any "adjustments" to the responses on the basis of this variable will actually add unwanted variation. Although users try to guard against such occurrences by first testing whether the slope of the observed relationship is real rather than random, they often use a lax criterion (e.g. a p-value less than, say, 0.20). This, together with the often large numbers of "possibly explanatory" variables "offered" to a regression, adds to the chances of decreasing rather than increasing the precision of a comparison. One way to avoid this artifact of chance is first to split one's data set into two or more smaller sets and retain only those variables that are influential in each subset.

Overoptimism Regarding Future Performance

The performance of discriminant functions or prediction equations constructed from a data-set is often judged by "resimulation" or by seeing how well the system "would have done" if it were used to classify the individuals in the data set. The results are generally overoptimistic for two reasons. First, because the weights were chosen on the very basis of doing well in this data set, they may well have "chased" or been fooled by any data patterns that were peculiar to that dataset. The random variation in a new dataset is unlikely to match the random peculiarities of the "training" dataset. As a result, knowing only a finite sample, but thinking of it as a universe, the system will be surprised a little more (16). Second, if one has enough candidate predictors to choose from, one is bound to find some coincidences. Similarly, if one builds an equation with enough variables, one will also get an irreproducibly good fit. There are a number of techniques for obtaining less optimistically biased estimates of future misclassification rates without actually doing a prospective test (27). However, they do not apply to the second bias mentioned above. In this latter situation, one needs to evaluate the system on a separate dataset. A number of studies that claimed high prediction accuracy solely on the basis of resimulation have "regressed toward the mean" (8, 24, 49, 63). Others have recognized this danger and have included the validation as an integral part of the task (53); one has even subjected the prediction system, which incidentally was constructed by logistic regression, to a comparative trial (52).

There has been speculation that there is some "natural law" that no matter how many variables are available for prediction, only four or five will finally remain in any stepwise regression (18). This claim would need to be examined more carefully, especially with regard to the influence of typical sample sizes. It does emphasize one point, namely that prediction of binary outcomes is a considerable task, given the considerable nonreducible uncertainty inherent in an all or nothing event. A method of measuring the attainable discrimination in a dataset and of deciding whether the search for predictors might be worth the effort is given in (32).

One Model for All

A recent example points up the serious inadequacy in a common approach to statistical predictions. The study asked whether two different types of gallstones could be distinguished on the basis of the features seen in a radiograph (26). Univariate analyses revealed that regardless of any other features, stones that appeared to be buoyant were invariably of one type; those that were not buoyant were sometimes of one type, sometimes the other. In spite of this, buoyancy ranked only third in the linear discriminant analysis which tried to predict the variation in types. This is clearly a situation in which buoyant cases could have been classified immediately, removed from the dataset, and discriminant analysis applied to the remaining cases. The unconditional "one model for all" approach is simplistic and possibly even misleading. Technically, the discriminant model could be made conditional through the use of interaction terms, provided one could anticipate which ones to include. An alternative, and more natural approach, which first partitions subjects on the most important variable, then partitions each of these subgroups separately, and so on in a branching fashion, is provided by recursive partitioning (also called Automatic Interaction Detection), a recent nonparametric classification system for use with larger data sets (25, 37). For smaller ones, the "kernel method" (1) seems to hold some promise.

Explaining Away a Difference

In the dental caries survey mentioned above, one would probably collect information on the frequency of visits to a dentist, and one might be tempted to take this variable into account in a multiple regression, when studying the effects of other risk factors on caries. If more caries result in more visits, then including the number of visits as an "explanatory" variable will lessen the observed impact of the other (real) risk factors: it will be one of the first variables to enter the regression equation and will thus "explain away" whatever variance might have been more appropriately accounted for by the risk factors being studied. Similar misinterpretations can arise if one includes as an explanatory variable one which is intermediate in the stimulus-response chain, as for example if one allowed for the amounts of medication given in a study comparing the lengths of stay following an operation performed in two different ways. Although it probably draws the correct conclusion, a recent study (39) shows just how easy it is to adjust away a difference, especially if other factors are not held constant. The authors state that the "data are in agreement with the hypothesis" that differences in weight, rather than in pO2 (Partial Oxygen Pressure), explain most if not all of the observed differences in blood pressure between children of the same age living at different altitudes. What is alarming is that the data might

also be in agreement with a similarly worded hypothesis stated in terms of family income, education, or any other variable that may be associated in a noncausal way with blood pressure, and on which high altitude children score lower than the comparison group.

CONCLUDING REMARKS

Investigation in the health sciences will continue to be of a multivariate nature. The statistical tools for dealing with the data generated by these studies are now largely in place; the challenge and the obligation will be to use them prudently (7, 59). Even though a number of lines of enquiry have become decidedly more complex in the past few decades (witness for example the current thinking on cholesterol and heart disease), by and large, questions still tend to be posed one dimension at a time. The same remains true in multivariate analysis, where even though the computations may sound high-dimensional, the statistical tests are univariate in spirit.

ACKNOWLEDGMENTS

I would like to thank my colleagues for their help with this article.

Literature Cited

1. Aitchinson, J., Aitken, C. G. G. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63: 413–20
2. Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* 59: 19–35
3. Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., Weisberg, H. I. 1980. *Statistical Methods for Comparative Studies.* New York: Wiley. 289 pp.
4. Armitage, P. 1971. *Statistical Methods in Medical Research.* Oxford/ Edinburgh: Blackwell. 504 pp.
5. Armstrong, J. S. 1967. Deviation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *Am. Stat.* 21:17–21
6. Baker, R. J., Nelder, J. A. 1978. *Manual for the GLIM system of generalized linear interactive modelling.* Oxford, GB: Numerical Algorithms Group
7. Barrett-Connor, E. 1979. Infectious and chronic disease epidemiology: Separate and unequal? *Am. J. Epidemiol.* 109: 245–49
8. Bell, R. S., Loop, J. W. 1971. The utility and futility of radiographic skull examination for trauma. *N. Engl. J. Med.* 284:236–39
9. Bickel, P. J., Hammel, E. A., O'Connell, J. W. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187:398–404
10. Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. 1976. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: MIT Press. 557 pp.
11. Boag, P. T., Grant, P. R. 1981. Intense natural selection in a population of Darwin's finches (Geospizinae) in the Galapagos. *Science* 214:82–84
12. Breslow, N. E. 1982. Design and analysis of case-control studies. *Ann. Rev. Public Health* 3:29–54
13. Breslow, N. E., Day, N. E. 1980. *Statistical Methods in Cancer Research. I. The Analysis of Case-Control Studies.* Lyon: Intl. Agency Res. Cancer. 338 pp.
14. Carpenter, R. G., Gardner, A., Pursall, E., McWeeny, P. M., Emery, J. L. 1979. Indentification of some infants at immediate risk of dying unexpectedly and justifying intensive study. *Lancet* 2: 343–46
15. Clark, D. W. 1981. A vocabulary for preventive and community medicine. In *Preventive and Community Medicine,* ed. D. W. Clark, B. MacMahon, pp.

3–15. Boston: Little, Brown. 794 pp. 2nd ed.
16. Cochran, W. G., Hopkins, C. E. 1961. Some classification problems with multivariate qualitative data. *Biometrics* 17:10–32
17. Cole, T. J. 1975. Linear and proportional regression models in the prediction of ventilatory function. *J. R. Statist. Soc. A* 138:297–337
18. Coles, L. S., Brown, B. W., Engelhard, C., Halpern, J., Fries, J. F. 1980. Determining the most valuable clinical variables: A stepwise multiple logistic regression program. *Meth. Inform. Med.* 19:42–49
19. Cook-Mozaffari, P., Bulusu, L., Doll, R. 1981. Fluoridation of water supplies and cancer mortality. I. A search for an effect in the UK on risk of death from cancer. *J. Epidemiol. Community Health* 35:227–32
20. Cook-Mozaffari, P., Doll, R. 1981. Fluoridation of water supplies and cancer mortality. II. Mortality trends after fluoridation. *J. Epidemiol. Community Health* 35:233–38
21. Cox, D. R. 1970. *The Analysis of Binary Data.* London: Methuen. 142 pp.
22. Cox, D. R. 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc. B* 34:187–202
23. Dawid, A. P. 1976. Properties of diagnostic data distributions. *Biometrics* 32:647–58
24. DeSmet, A. A., Fryback, D. G., Thornbury, J. R. 1979. A second look at the utility of radiographic skull examination for trauma. *Am. J. Roentgen.* 132:95–99
25. Diehr, P., Wood, R. W., Barr, V., Wolcott, B., Slay, L., Tompkins, R. K. 1981. Ocult headache: Presenting symptoms and diagnostic rules to identify patients with tension and migraine headache. *J. Chron. Dis.* 34:147–58
26. Dolgin, S. M., Schwartz, J. S., Kressel, H. Y., Soloway, R. D., Miller, W. T., Trotman, B., Soloway, A. S., Good, L. I. 1981. Identification of patients with cholesterol or pigment gallstones by discriminant analysis of radiologic features. *N. Engl. J. Med.* 304:808–11
27. Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 71:1–26
28. Feinstein, A. R. 1977. *Clinical Biostatistics.* St. Louis: Mosby. 468 pp.
29. Fienberg, S. E. 1980. *The Analysis of Cross-Classified Categorical Data.* Cambridge: MIT Press. 198 pp. 2nd ed.
30. Gardner, M. J., Barker, D. J. P. 1975. A case study in techniques of allocation. *Biometrics* 31:931–42
31. Guyer, B., Wallach, L. A., Rosen, S. L. 1982. Birth-weight-standardized neonatal mortality rates and the prevention of low birth weight: How does Massachusetts compare with Sweden? *N. Engl. J. Med.* 306:1230–33
32. Hanley, J. A., McNeil, B. J. 1982. Maximum attainable discrimination and the utilization of radiologic examinations. *J. Chron. Dis.* 35:601–11
33. Harris, R. J. 1975. *A Primer of Multivariate Statistics.* New York: Academic. 332 pp.
34. Heinonen, O. P., Stone, D., Monson, R. R., Hook, E. B., Shapiro, S. 1977. Cardiovascular birth defects and antenatal exposure to female sex hormones. *N. Engl. J. Med.* 296:67–70
35. Henry, R. C., Hidy, G. M. 1979. Multivariate analysis of particulate sulfate and other air quality variables by principal components—pt. I. Annual data from Los Angeles and New York. *Atmosph. Environ.* 13:1581–96
36. Higgins, M. W., Keller, J. B., Becher, M., Howatt, W., Landis, J. R., et al 1982. An index of risk for obstructive airways disease. *Am. Rev. Respir. Dis.* 125:144–51
37. Hooton, T. M., Haley, R. W., Culver, D. H., Morgan, W. M. 1981. The joint associations of multiple risk factors with the occurrence of nosocomial infection. *Am. J. Med.* 70:960–70
38. Horning, S. J., Hoppe, R. T., Kaplan, H. S., Rosenberg, S. A. 1981. Female reproductive potential after treatment for Hodgkin's disease. *N. Engl. J. Med.* 304:1377–82
39. Jongbloed, L. S., Hofman, A. 1983. Altitude and blood pressure in children. *J. Chron. Dis.* In press
40. Kaplan, R. M., Bush, J. W., Berry, C. C. 1976. Health status: Types of validity and the index of well-being. *Health Serv. Res.* 478–505
41. Kinlen, L., Doll, R. 1981. Fluoridation of water supplies and cancer mortality. III. A re-examination of mortality in cities in the USA. *J. Epidemiol. Community Health* 35:239–44
42. Kiviluoto, M. 1980. Observations on the lungs of vanadium workers. *Br. J. Indust. Med.* 37:363–66
43. Kleinbaum, D. G., Kupper, L. L. 1978. *Applied Regression Analysis and Other Multivariable Methods.* North Scituate, Mass: Duxbury. 556 pp.

44. Lagakos, S. W. 1981. The graphical evaluation of explanatory variables in proportional hazard models. *Biometrika* 68:93
45. Lagakos, S. W., Mosteller, C. F. 1981. A case study of statistics in the regulatory process the FD & C Red No. 40 experiments. *J. Nat. Can. Inst.* 66:197–212
46. Lee, E. T. 1980. *Statistical Methods for Survival Data Analysis.* Belmont: Lifetime Learning. 557 pp.
47. McNeil, B. J., Hanley, J. A. 1981. Statistical approaches to clinical predictions. *N. Engl. J. Med.* 304:1292–94
48. Mosteller, F., Tukey, J. W. 1977. *Data Analysis and Regression.* Reading: Addison-Wesley. 588 pp.
49. Murphy, J. F., Newcombe, R. G., Sibert, J. R. 1982. The epidemiology of sudden infant death syndrome. *J. Epidemiol. Community Health* 36: 17–21
50. Partridge, L., Farquhar, M. 1981. Sexual activity reduces lifespan of male fruitflies. *Nature* 294:580–81
51. Pocock, S. J., Shaper, A. G., Cook, D. G., Packham, R. F., Lacey, R. F., Powell, P., Russel, P. F. 1980. British Regional Heart Study: Geographic variations in cardiovascular mortality, and the role of water quality. *Br. Med. J.* 280:1243–49
52. Pozen, M. W., D'Agostino, R. B., Mitchell, J. B., Rosenfeld, D. M., Guglielmino, J. T., et al 1980. The usefulness of predictive instrument to reduce inappropriate admissions to the coronary care unit. *Ann. Intern. Med.* 92:238–42
53. Ramsdale, D. R., Faragher, E. B., Bennett, D. H., Bray, C. L., Ward, C., Beton, D. C. 1982. Preoperative prediction of significant coronary heart disease in patients with valvular heart disease. *Br. Med. J.* 284:223–26
54. Rothman, K. J. 1978. Occam's razor pares the choice among statistical models. *Am. J. Epidemiol.* 108:347–49
55. Sather, H., Coccia, P., Nesbit, M., Level, C., Hammond, D. 1981. Disappearance of the predictive value of prognostic variables in childhood acute lymphoblastic leukemia. *Cancer* 48:370–76
56. Schlesselman, J. J., Stolley, P. D. 1981. *Case Control Studies Design, Conduct, Analysis.* Oxford Univ. Press
57. Schoenfeld, D. A. 1979. Chi-squared goodness of fit tests for the proportional hazards regression model. *Biometrika* 67:145–53
58. Stablein, D. M., Carter, W. H., Novak, J. W. 1981. Analysis of survival data with nonproportional hazard functions. *Controlled Clin. Trials* 2:149–59
59. Stallones, R. A. 1980. To advance epidemiology. *Ann. Rev. Public Health* 1:69–82
60. Stevens, R. G., Moolgavkar, S. H., Lee, J. A. H. 1982. Temporal trends in breast cancer. *Am. J. Epidemiol.* 115:759–77
61. Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Dkene, A. M., Habbema, J. D. F., Gelpke, G. J. 1981. Comparison of discrimination techniques applied to a complex data set of head injured patients. *J. R. Statist. Soc. A* 144:145–75
62. Truett, J., Cornfield, J., Kannel, W. B. 1967. A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.* 20:511–24
63. Valdes-Dapena, M. A. 1980. Sudden infant death syndrome: A review of the medical literature 1974–1979. *Pediatrics* 66:597–614
64. Wagner, G., Tautu, P., Wolber, U. 1978. Problems of medical diagnosis—a bibliography. *Methods Inform. Med.* 17:55–74
65. Walker, S. H., Duncan, D. B. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54:167–79