Notes on Statistical Analyses of Data from a Cancer Mortality-Dietary Project

by

W.T. Federer, L.C. Clark, N. Miles-McDermott, and D.S. Robson

BU-917-M*

November 1986

Revised September 1987

1. INTRODUCTION

Many statistical considerations and judgments are involved when analyzing and interpreting the data from a large epidemiologic project such as the current one. Some of these are discussed below. In this project, there were 65 counties in the survey with two communes per county, resulting in 130 observations. Also, data were obtained for 50 males and 50 females per commune. In all, several hundred variables were observed. For some variables like cancer mortality on eight male and ten female cancers, only county data by sex and age are available. For other variables, responses are available on an individual basis, resulting in 6,500 observations. Since there were several hundred independent variables possibly related to only 65 observations on the dependent variables on age-truncated cancer mortality rates by sex, overparameterization of functional relationships and predictive equations is an immediate possibility. Caution needs to be practiced and steps need to be taken in order to deal with this difficulty. A number of statistical and practical problems arose prior to and during the course of analyzing the data. Statistical notes on a number of particular problems are given for the following items in the sections as numbered on the left:

---

* In the Technical Report Series of the Biometrics Unit, Cornell University, Ithaca, NY 14853.

2) confirmatory versus exploratory data analysis,

3) issues on pooling samples,

4) pooling and sample size bias,

5) effect of range and errors in variables on correlation coefficients,

6) validation of multiple correlation coefficients,

7) univariate versus multivariate analyses,

8) quality control checks in routine laboratory analyses,

9) data management checks on statistical analyses obtained from computer packages or otherwise,

10) implementing GLIM for quality control of a routine serological assay system,

11) validating exploratory data analyses,

12) asymmetric composites, and

13) use of linear combinations as predictor variables in regression.

2. CONFIRMATORY VERSUS EXPLORATORY DATA ANALYSIS

Hypotheses should be constructed prior to conducting an investigation or before studying the data if hypothesis testing or significance testing is to be appropriately applied and interpreted. This step is an essential part of a *confirmatory analysis*, where the objective is to determine the plausibility of a hypothesis and the degree of confidence in the proposed hypothesis. This is done by stating a level of significance, a size of the test, and/or a confidence level. If the data are identically and independently distributed and if the sample is a random one from the population for which inferences are being made, the probability values (p-values) obtained from standard statistical tables are appropriate. P-values are only meaningful if the above conditions are met or if a procedure has been devised to account for a violation of one or more of the required conditions.

Exploratory analysis, on the other hand, is not concerned with p-values. The data are studied to determine what patterns or relationships might exist in this particular set of data. These can then be used to construct possible hypotheses for future investigations. In special cases, the patterns or relationships existing in the data may direct the investigator's attention to theory in that particular subject matter field. This may indicate that no further investigation is required to reach a conclusion. In general, most exploratory data analyses will be used to define hypotheses for future investigation.

There appears to be considerable confusion about these two types of analysis, both of which can be appropriately used in any investigation. All too often an exploratory data analysis is made and then p-values are attached. The results of an exploratory data analysis should not be

assessed with p-values but rather with the strength of an association and with the theory associated with the subject matter in that field. The addition of p-values can mislead the reader and distort the results. (Note an exception in that one may use Scheffé's multiple comparisons procedure in an exploratory nature and still set p-values.)

Another condition that is often violated in investigations is the representativeness of the sample for the population about which inferences are being made. The sample may be and often is unrepresentative of the population being considered. One method of obtaining a representative sample is to take a random sample from the population for which inferences are desired. Many investigations use unrepresentative samples. The sample of counties selected in this geographic study is not representative of the population in China. It was selected to attain specific conditions, i.e., to maximize the range in six male and one female cancer mortality rate. Hence, it is not a representative survey of the Chinese population. Any interpretation of the data should be made with this in mind.

3. ISSUES ON POOLING SAMPLES

In the process of planning the project, one proposal was to sample 70-80 counties with two communes per county and with 100 individuals per county. This would have resulted in approximately 8000 samples of plasma, red blood samples, and urine. Laboratory work for 8,000 samples on 100-200 biochemical and other characters would be a tremendous undertaking. At several early meetings on the project, characters were divided into the A group (must be done), B group (highly desirable to be done), C group (some experimenters would like them done but not of high priority), and a D group (probably would not be done). Since the cancer mortality data is by county, by sex, and by age, results for individual biochemical analysis would need to be expressed on a county level as a mean value for continuous variables as a proportion for discrete variables. Therefore, it was proposed that the 100 samples be pooled into groups of 25. Pooling by sex and commune reduced the laboratory work to *four percent* of that originally proposed and allowed for more of the proposed biochemical analyses to be performed as well as to provide a larger volume for each sex by commune sample. The proposal was accepted and allowed all laboratory analyses to be performed since the sex by commune specific pools contained 50 ml of plasma. Using 260 pooled samples plus checks (see Section 8) did not unduly overload the laboratory faculties of cooperating organizations for a particular assay. In addition, the pools of blood samples then were quite large, allowing many laboratory analyses to be performed. This would not have been possible if the individual samples had been analyzed.

The rationale and justification for pooling needs to be scrutinized closely for each assay. Since the cancer mortality measures are by county, sex, and age and *not* by individual, any relationship or predictive studies

of biochemical or other characters with cancer mortality will necessarily have to use county sex and age groupings. Hence, the analyses by individuals is inefficient since these data must be pooled. The values obtained from the pools are appropriate provided the group mean is a sufficient summary statistic for purposes at hand. We should not blithely assume that the arithmetic mean is a statistically sufficient summary of the county distribution of the variate X. If disease is a threshold phenomenon, for example, then the mean is not sufficient. We would then need to know the percentiles of the distribution of X among individuals in the county. Similar problems arise with the questionnaire data. For many questions the mean by commune by sex is *not* sufficient. The goals and limitations of any study must be carefully considered prior to conducting laboratory analyses. This study definitely indicated that pools be used rather than individual analyses.

Another potential temptation is to interpret the results of geographic studies as if they indicated individual risk. This is particularly true for what may appear to be important interactions in county level data and which may not be observed as interactions in epidemiologic studies conducted on individuals. Hence, interpreting the analyses for geographic studies as individual's risk is inappropriate. In other words this epidemiologic study should not be used as a clinical one.

4. POOLING AND SAMPLE SIZE BIAS

This proposal contends that plasma pooling is statistically acceptable if the distribution of individual values is normal (and the pool value equals the mean of the individual values). Suppose, therefore, that the plasma concentration X of a certain "risk factor" is normally distributed in each age (i) and sex (j) specific population,

$$f_{X_{ij}}(x) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_{ij}}{\sigma_{ij}}\right)^2} . \qquad (1)$$

By definition of "risk factor," this population frequency distribution is "size-biased"; i.e., (1) is the frequency distribution of X among the survivors, where survival probability is, itself, a monotonic function of x. We may, therefore, regard (1) as a conditional density function $f_X(x|\text{alive})$, and hence we have the model

$$f_{X_{ij}}(x|\text{alive}) = g_{X_{ij}}(x) P_{ij}(\text{alive}|X=x)/P_{ij}(\text{alive}) \qquad (2)$$

where

$$P_{ij}(\text{alive}) = \text{sex-specific (j) probability of surviving to the specified age (i)} \qquad (3)$$

and

$$g_{X_{ij}}(x) = \text{the form that the age-and-sex-specific distribution would have taken if this factor were not a "risk factor."} \qquad (4)$$

The "selection function"

$$\frac{P_{ij}(\text{alive}|X=x)}{P_{ij}(\text{alive})} = \frac{f_{X_{ij}}(x|\text{alive})}{g_{X_{ij}}(x)} = r_{ij}(x) \quad \text{(say)} \qquad (5)$$

is assumed to be a monotonic function of x where, typically, higher values of x are risky; i.e.,

$$\frac{dr_{ij}(x)}{dx} < 0 \qquad \forall x \quad . \tag{6}$$

If both the selected population (1) and the unselected (and unobservable) population (4) are normally distributed, then (6) can hold only if the variances in these two populations are equal, in which case

$$r_{ij}(x) = c_{ij} e^{-\Delta_{ij} x} \tag{7}$$

where $\mu_{ij} + \Delta_{ij}\sigma^2_{ij}$ is the mean value of X in the unselected population and

$$c_{ij} = e^{\Delta_{ij}(\mu_{ij} + \frac{1}{2}\Delta_{ij}\sigma^2_{ij})} P_{ij}(\text{alive}) \quad . \tag{8}$$

In this normally distributed population the hazard rate (Δ_{ij}) with respect to x is thus proportional to the difference between the mean value of X in the unselected and the selected population, the proportionality factor being the reciprocal of the common variance in these two populations. In the homoscedastic case $\sigma^2_{ij} \equiv \sigma^2$ (the usual concomitant to the normality assumption) this proportionality factor is then the same for all age and sex classes.) Note that this result extends to the multivariate normal model with a common covariance matrix in the selected and unselected populations; the exponent in (7) then becomes the linear discriminant function of the several variables, and the survival model is log-linear.

Since concentration in the plasma is a nonnegative random variable the normal model is only an approximation, at best, and the log-normal or gamma models might better serve the purpose. The gamma distribution has particularly convenient mathematical properties under pooling, while the log-normal has particularly inconvenient properties* with respect to pooling, so we consider the former case here:

---

* The distribution of a mean of log-normal random variables is not expressible in closed form. Note that (7) becomes $r(x) = C\chi^{-\Delta}$ in the log-normal case.

$$f_X(x|\text{alive}) = \beta^{\alpha} x^{\alpha-1} e^{-\beta x}/\Gamma(\alpha) \qquad (9)$$

where the distribution of $\bar{x}_n$ has this same form with parameters $n\alpha$ and $n\beta$. If the distribution $g_X(x)$ in the unselected population is also a gamma distribution and if the selection function is monotonic, then

$$g_X(x) = (\beta-\Delta)^{\alpha} x^{\alpha-1} e^{-(\beta-\Delta)x}/\Gamma(\alpha) \qquad (10)$$

and, again,

$$r(x) = ce^{-\Delta x} \quad . \qquad (11)$$

In this case the selected and unselected populations must have the same coefficient of variation rather than the same variance (i.e., the two gamma distributions share a common parameter α; and an assumption of homoscedasticity across age and sex classes would now be replaced by an assumption of constant coefficient of variation ($= 1/\sqrt{\alpha}$)).

The simple exponential form of (7) is, more generally, a property of the one-parameter exponential family

$$f(x;\theta) = \exp[\theta_f x + a(x) + b(\theta_f)] \ , \qquad (12)$$

where the function $a(x)$ may depend on unknown nuisance parameters that are *constant* within the family under consideration (as in the *homoscedastic* normal family or the gamma family with a *constant* coefficient of variation). If both $f_X(x|\text{alive})$ and $g_X(x)$ are members of the *same* exponential family, differing only in the value of θ, then (7) holds with $\Delta = \theta_f - \theta_g$ and $\&nc = b(\theta_f) - b(\theta_g)$. Other examples which may be mentioned are the discrete distribution families including the Poisson, binomial and negative binomial. The latter case which might apply, for example, to transient but potentially lethal parasite attacks would give

$$f_X(x|\text{alive}) = \frac{\Gamma(\alpha+x)}{x!\Gamma(\alpha)} \left(\frac{p}{\beta}\right)^x \left(1 - \frac{p}{\beta}\right)^{\alpha} \qquad (\beta>1) \qquad (13)$$

$$g_X(x) = \frac{\Gamma(\alpha+x)}{x!\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^x \left(1 - \frac{1}{\beta}\right)^\alpha \tag{14}$$

where p ($0 \leq p \leq 1$) is then the probability of surviving an attack. In this case the selection function r(x) is

$$r(x) = \left(\frac{\beta-p}{\beta-1}\right)^\alpha p^x \ . \tag{15}$$

Here $g_X(x)$ is the negative binomial distribution of number of exposures to the parasite (as if the parasite were strictly nonlethal), while $f_X(x|\text{alive})$ is the distribution of number of bouts with the parasite among living individuals.

Another "size-biased" *normal* distribution of plasma concentration X may be derived from a nonselected *normal* distribution $g_X(x) = \phi[(x-\mu_g)/\sigma_g]/\sigma_g$ by assuming that there is an optimal concentration μ^* for survival, and that deviations above or below this optimum are equally risky, and in particular that

$$P(\text{alive}|X=x) \sim e^{-\left(\frac{x-\mu^*}{2\sigma_{A^*}}\right)^2} \ . \tag{16}$$

Under these conditions (2) becomes

$$f_X(x|\text{alive}) = \sqrt{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_*^2}}\ \phi\left[\frac{x - \left(\frac{\mu_g}{\sigma_g^2} + \frac{\mu^*}{\sigma_A^2}\right)\Big/\left(\frac{1}{\sigma_g^2} + \frac{1}{\sigma_A^2}\right)}{\sqrt{\frac{1}{\frac{1}{\sigma_g^2} + \frac{1}{\sigma_{A^*}^2}}}}\right] . \tag{17}$$

The mean value of X in the surviving population is thus a weighted average of the optimum concentration μ^* and the mean μ_g of the unselected population.

Note that X might represent some linear combination of plasma variables (giving more credence to the normality assumptions) arrived at by multivariate analysis methods. Plant ecologists have been using normality in this manner to model density (abundance) of a plant species as a function of a linear environmental variable bearing a label such as "elevation."

5. EFFECT OF RANGE AND ERRORS IN VARIABLES ON CORRELATION COEFFICIENTS

Many sets of data represent a selected subpopulation of the entire population. Experimenters then compute correlation coefficients and are surprised by the smallness of the coefficient or even of the sign in some cases. The type of sample or subpopulation selection can easily account for this. To illustrate, consider a graph like Figure 5.1 which represents an envelope of scatter-points for two correlated variables, X and Y. The cigar-shaped envelope would indicate a high positive correlation between X and Y. Suppose that the data set obtained for a particular study was similar to that represented by the shaded portion of the figure. Such a set would indicate a negative value for the computed correlation whereas it is high and positive in the entire population.

As a second example, consider that the sample collected is represented by the area to the right of the dashed line of Figure 5.1. Here the correlation is close to zero. A situation similar to this is encountered every year by colleges when students are selected (screened) for admission into college. There may be a high correlation between measures of pre-college performance, e.g., SAT scores, American College Board scores, high school grade average, etc., and actual success (grade point average) in college. When only a small fraction of the high school population is selected for admission, as at Cornell University for example, the correlations of pre-college measurements and grade point average in college are dramatically reduced.

In the bivariate normal case, for example, with selection of the top $100\alpha\%$ with respect to X, the correlation (ρ_α) between X and Y in this selected fraction of the population is expressible in terms of the correlation $\rho = (=\rho_1)$ in the entire population as

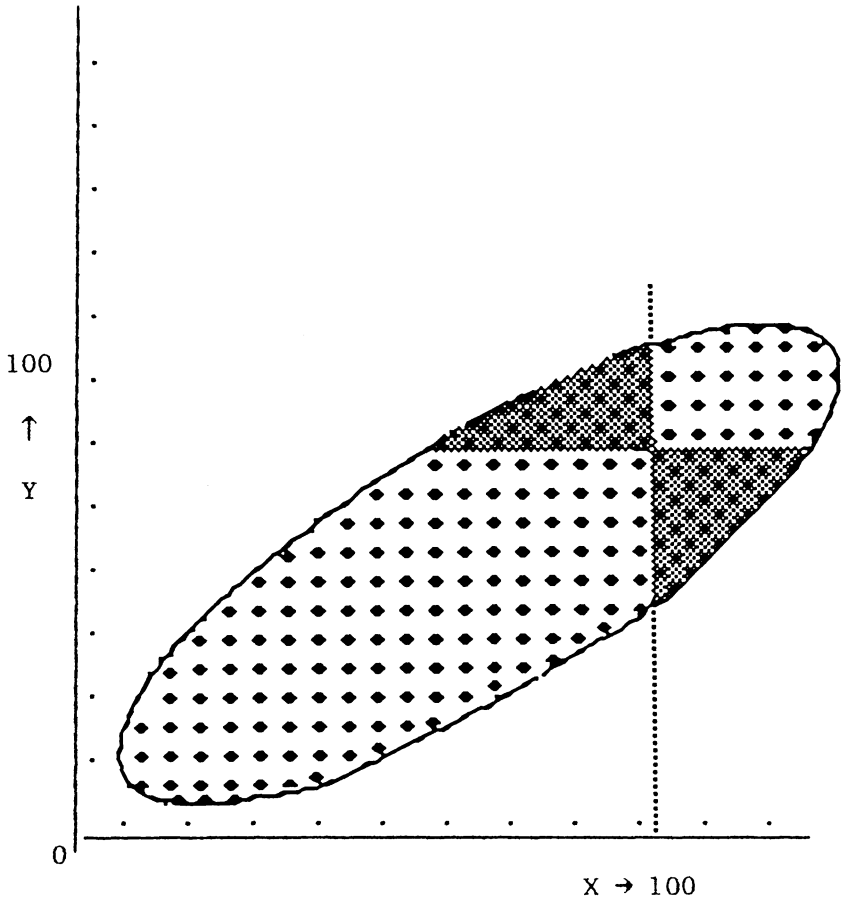


Figure 5.1. Scatter diagram example.

$$\rho_\alpha = \rho\sqrt{\frac{1-\delta_\alpha}{1-\rho^2\delta_\alpha}} = \mathrm{sgn}(\rho)\sqrt{\frac{\rho^2-\rho^2\delta_\alpha}{1-\rho^2\delta_\alpha}} \tag{18}$$

where

$$0 \le \delta_\alpha = \frac{\phi(z_{1-\alpha})}{\alpha}\left[\frac{\phi(z_{1-\alpha})}{\alpha} - z_{1-\alpha}\right] \le 1 \tag{19}$$

and

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \qquad 1-\alpha = \int_{-\infty}^{z_{1-\alpha}} \phi(z)dz \quad . \tag{20}$$

Some examples of the magnitude of the reduction in correlation due to selection are given in Figure 2.

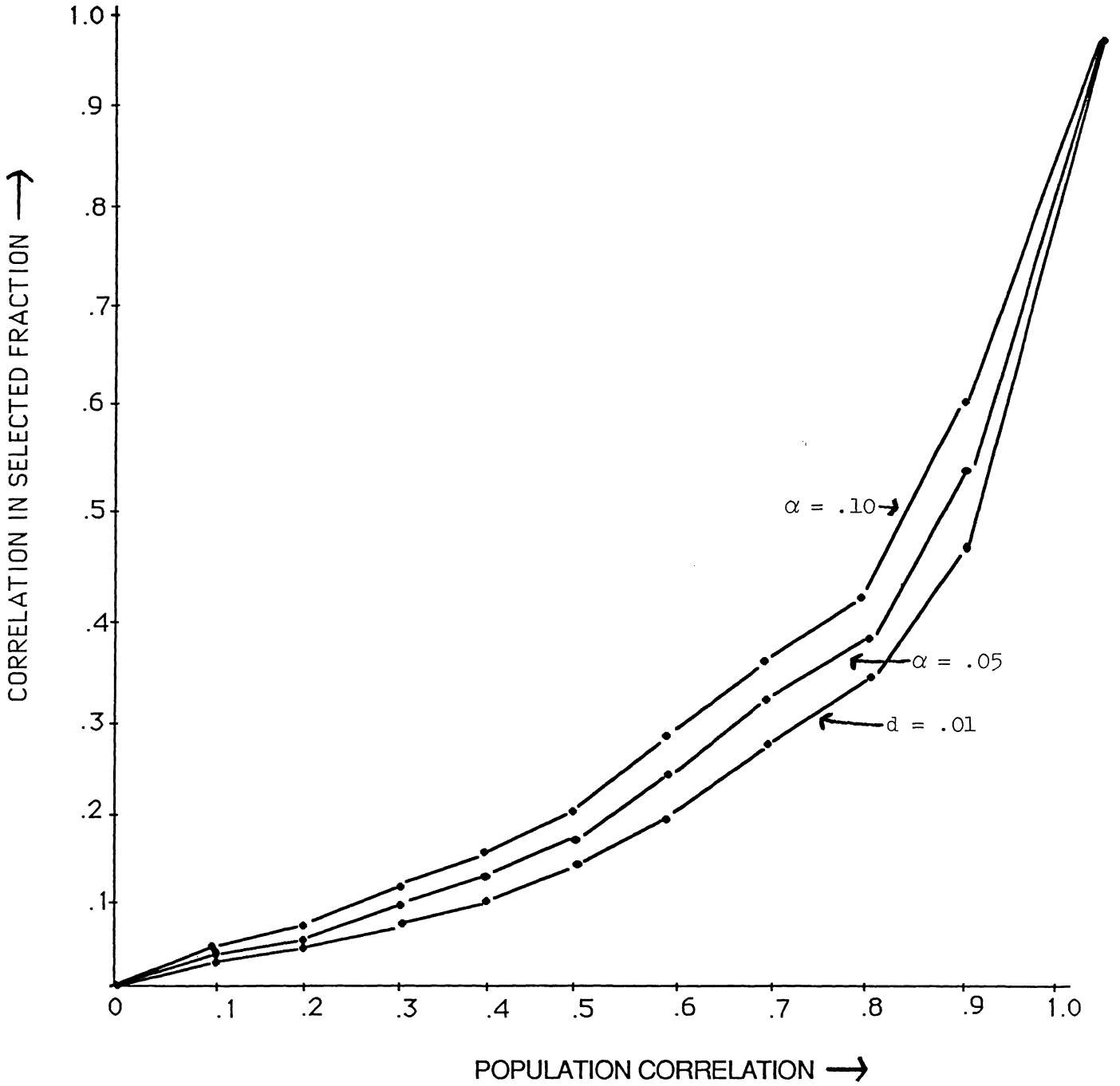


Figure 5.2. Values for the correlation between X and Y in a selected fraction (α) from a bivariate normal population when selection is based only on large values of X.

As can be observed from the above, any restriction in the range of variables can greatly reduce positive correlations and vice versa for large negative correlations, in the population. For the Project, the range in variables as measured by within-county relative to between-county mean squares was quite large for most of the variables studied. With such a relatively large range in cancer mortality, variables and between-biochemical variables, any relationships present should be detectable in this study. If the F values for between-county to within-county mean squares has been low, this would indicate a small range in county means for a variable, and hence the variable would not be of much use for relationships or predictive studies.

Another item that escapes the attention of many investigators and which does not receive appropriate discussion in statistical literature is measurement error in variables and their effect on correlations. To illustrate, suppose that the variable Y has measurement errors which are uncorrelated with a variable X which has no measurement errors. Further, suppose that the true values of Y are highly correlated with X. The sum of squares for the X variate is S_X, say, the sum of squares for the Y variate is $S_Y + S_{MY}$ where S_Y is the sum of squares among the true Y values and S_{MY} is the sum of squares due to measurement error in Y, and the cross-product of the measured Y values and the X values is $C_{XY} + C_{XM} = C_{XY}$, since X and measurement error in Y has zero correlation. The correlation coefficient would ordinarily be computed as:

$$r_{XY} = C_{XY}/\sqrt{S_X(S_Y+S_{MY})} \quad , \tag{21}$$

whereas the correct correlation between X and Y is $r_{XY} = C_{XY}/\sqrt{S_X S_Y}$. The measurement error in Y can reduce the computed value of the correlation

coefficient. The situation can be even more pronounced when both X and Y have large measurement errors and these are uncorrelated with each other and with the true values of X and Y. A correlation coefficient computed as

$$r_{XY} = C_{XY}/\sqrt{(S_X+S_{MX})(S_Y+S_{MY})} \qquad (22)$$

would result in lower correlations than there should be. The reduction depends upon the relative sizes of the X measurement errors, S_{MX}, and/or the Y measurement errors, S_{MY}, to S_X and S_Y. If S_{MX}/S_X and S_{MY}/S_Y are near zero, the correlation is essentially correct. However, if these ratios deviate far from zero, apparent correlations will be very small when computed as above. Experimenters often consider correlations small or nonexistent in this situation when in fact the correlation may be quite high. One method of circumventing this problem is to obtain estimates of S_{MX} and S_{MY} and subtract these values from the sum of squares for the X and Y values. This procedure can produce correlations which are greater than one. A better method, if possible, is to reduce S_{MX} and S_{MY} to zero or near zero by improvement of experimental techniques.

Variation in any of the several variables studied can be affected by measurement error. Thus, in studying relationships, it is imperative that measurement error of a variable, e.g., selenium, calcium, cancer mortality, etc., is small or nonexistent relative to the variance of a variable. Note that the biochemical method could be perfect but an unskilled and/or uncareful technician or analyst could introduce considerable measurement error.

6. VALIDATION OF MULTIPLE CORRELATION COEFFICIENTS

Since most studies are multivariate in nature, a multiple correlation coefficient rather than simple two variate correlations is appropriate. After obtaining a correlation, one is always faced with interpreting the results. A first step in interpretation is to consider whether or not the size of the computed multiple correlation is too high, too low, or reasonable for the particular set of data under consideration. Overparameterization can easily result, when the Y variates are not independent and there are fewer degrees of freedom than what the experimenter presumes. Also, too many X-variates may be included because the experimenter assumes more independent observations than are actually present. In all cases, the goal should be to maximize R^2 while minimizing the number v of X variates used. The fewer the number of X variates, the easier will be the interpretation. If the experimenter has n independent observations and uses *any* n-1 X noncollinear variates, a value of $R^2 = 1$ will always result.

One method of validating high R^2 values for a set of data is the following. For the multiple correlation desired, the experimenter has selected a Y-variate and v X-variates for which it is believed that there is a relationship. The experimenter has hypothesized that Y is related to the v X-variates. This is the confirmatory part of the analysis. For the same Y values, there may be a second set of v X-variates which are unrelated to the Y-variate values. If not, a set of v random variates can be used. If a high value of R^2 is obtained near that for the hypothesized R^2 for this set of data, then overparameterization appears to be present. Note that the assumed set of noncorrelated v variables could actually be correlated with Y, i.e., the assumption was incorrect. However, this would not arise if the v uncorrelated variates were each a random sample from a single population.

For the project data, if one type of cancer mortality rate is highly related to v biochemical measurements, this can be validated by selecting v other biochemical variables which should be unrelated with this particular cancer mortality rate. Then, if a high R^2 is obtained in the former case and a low R^2 in the latter case, the investigator could have confidence of a real relationship. Note that under the null hypotheses of no relation between Y and the v X variates, the mean value of R^2 is not zero but is $v/(n-1)$ where v is the number of X variates and n is the sample size.

Collinearity among the v independent variables (X) can affect the value of R^2. For example, consider that there are two X variables which have correlations of $r_{Y1} = .2$ and $r_{Y2} = .8$. R^2 can be unity if r_{12} equals either 0.75 or -0.42. The minimum value of R^2 is attained when $r_{12} = r_{Y1}/r_{Y2} = 0.25$ and is 0.64. For a more detailed study see Federer (1961). Thus, the nature and extent of the relationship among the independent variables can have a significant effect on the size of R^2. This should always be considered when interpreting relationships of dependent and independent variables.

7. UNIVARIATE VERSUS MULTIVARIATE ANALYSES

Seldom, if ever, is a variable Y explained by a single X-variable. Most variables are correlated to some extent. In explaining variation in a variable Y, several X variables will be required in most cases. Since this is true, investigators will be concerned with a multivariate situation rather than a univariate one. Simple correlation coefficients will usually not be of much value in explaining relations. Rather, a multiple correlation will be more useful. Also, when considering a set of variables, either Y or X, it may be useful to try multivariate techniques.

In considering eight different male cancers and ten different female cancers, it may be useful to use a technique from multivariate analysis known as principal components. If the male cancer mortality rates are not independent then some may be highly correlated. In this situation it may be appropriate to combine those cancer sites which lack independence in this study. If this is not done, a false sense of consistency may result. The same thing may be true for some of the other variables measured, e.g., biochemical characters. This type of study was made on the eight different male cancers, on the ten different female cancers (see Federer *et al.*, 1986), and is planned for a number of biochemical characters. The purpose of the study will be to determine if fewer variables can be used in showing relationships between the cancer mortalities and the biochemical variables or other variables.

In studying only simple correlation coefficients, apparent relations may be misleading or spurious in that the correlation observed is not a relationship between the two variables but is caused because both variables are correlated with a third and/or other variables. The use of partial correlations may be helpful in overcoming this problem. In a study of the magnitude of this project, simple correlation coefficients are considered to be of little primary value.

8. QUALITY CONTROL CHECKS IN ROUTINE LABORATORY ANALYSES

In conducting routine laboratory analyses, a variety of checks should always be used. For the Project, it was suggested that blind checks be randomly distributed in with test samples, that n/3 checks at the low level of a component, n/3 checks at a medium level, and n/3 checks at a high level be used, and that n be approximately the square root of the number of samples being processed. If 289 samples were being processed, then $n = \sqrt{289} = 17$ check samples should be used. This would result in approximately six samples at each level, low, medium, and high.

If the actual amount of a component in a check sample is known, then the difference between the level obtained from routine laboratory analysis and the actual amount in the sample will be a measure of the bias in the method and/or analyst. If the actual amount is not known, a bias in the method and/or analyst may be difficult to find. Also, if the amount of bias changes with the amount of the component in the sample, problems arise in the interpretation of results obtained from routine laboratory analyses. In order to have some information on levels in the check samples, the high check sample may be diluted by a factor of two (or four) to obtain the medium level sample and by a factor of four (or eight) to obtain the low sample. The laboratory analyses should then give the amount in the medium level check samples as one-half of that obtained for the high samples. Likewise, the low level check samples should contain only one-fourth that obtained for the high level samples. If this result obtains, then any bias there is, is consistent over all levels.

In cases where the actual amount of a component in the check sample needs to be determined, the experimenter could conduct many analyses on the check sample and essentially determine the exact amount. If the level were

high, then the sample could be diluted to obtain the medium and low check samples with essentially known values.

To summarize results from check samples, the means, variances, and ranges should be computed. The range of values can be taken to be limits of error for the n/3 samples at each level. As the number of check samples n is increased, it will be possible to obtain a frequency distribution for the results. The nature and form of the distribution may be an important item of information, and it may be possible to describe the distribution mathematically. One should carefully scrutinize the check data from the analyst to ascertain that they stay in control from day-to-day and that their results are reliable. Outliers should be studied to ascertain the nature of measurement errors occurring for any particular set of laboratory data.

The basis for suggesting $\sqrt{n}$, where n = number of samples processed, follows. Let k = n+c be the block size where n is the number of treatments (samples) to be compared with a control and c is the number of replicates for the control in each block. Further, let N = rk be the total number of samples processed, and let r be the number of replicates for each of the n treatments being compared with the control. Then, the standard error variance of a difference between a treatment mean and the control mean is

$$\frac{\sigma^2}{r}\left(1 + \frac{1}{c}\right) = \frac{k\sigma^2}{N}\left(1 + \frac{1}{c}\right) = \frac{(n+c)}{N}\sigma^2\left(1 + \frac{1}{c}\right) = \frac{\sigma^2}{N}\left(n + c + \frac{n}{c} + 1\right) \quad . \qquad (23)$$

The value of c minimizing the above is obtained by taking the derivative of the above and equating to zero, which is

$$1 - \frac{n}{c^2} = 0 \qquad \text{or} \qquad c = \sqrt{n} \quad . \qquad (24)$$

A somewhat more complicated version of the above is described by Yates (1936).

In order to provide evidence on the validity of the laboratory analyses performed for the project, it was suggested that a detailed analysis be performed on the results obtained for the check samples. In particular, the mean, the variance, the range, the coefficient of variation, and the frequency distribution should be obtained for the n/3 checks at the low level, the medium level, and the high level. If actual values are known for a level, the bias in the mean for a particular analysis can be obtained. Low and stable variances at the three levels is desirable. This would indicate high precision for a particular laboratory analysis. Such a study as this would add considerable credence to the laboratory analyses for test samples being reported for the project. Without such evidence as the above, the reader of such reported results would have nothing but faith to go on in accepting the results. Also, the frequency distribution would help in locating gross errors and outliers for a particular laboratory analysis.

9. DATA MANAGEMENT CHECK ON STATISTICAL ANALYSES OBTAINED FROM COMPUTER PACKAGES AND OTHERWISE

In the course of statistical consulting, it has become apparent that investigators make too many assumptions and do too little checking of statistical computer packages. Some of the packages simply give wrong answers, and others may appear to give what is wanted, but the statistical analysis given is for the wrong problem. In the latter case, investigators often assume that the program is giving the desired answers when it is not. One simple check that should always be used whenever a computer package is employed is to include a set of data for which the statistical analysis interpretation, and results, are known. If the package gives the desired result on the known example then it can usually be assumed that the desired statistical analysis will be obtained for the new data. The known data set can be obtained from a textbook, can be computed with a pocket or desk calculator, can be constructed by the investigator, or can be obtained from a previous investigation. Note that such examples are not foolproof for a variety of reasons. Also, the effect of missing data on the results and the effect of missing data as handled by a package, needs to be known in order to interpret results. In addition, when the known data set is run through the computer and a print-out is obtained, the print-out should be annotated by describing the exact nature of each calculation. This annotated computer output (ACO) can become quite useful for interpreting output from investigations.

Although some documentations of statistical computer packages are excellent, others are inadequate and actually wrong in some cases. Some of the packages where the above difficulties have been encountered were in polynomial regression, covariance analyses, and multivariate analyses. It is well to remember that humans write programs and humans, any human, can

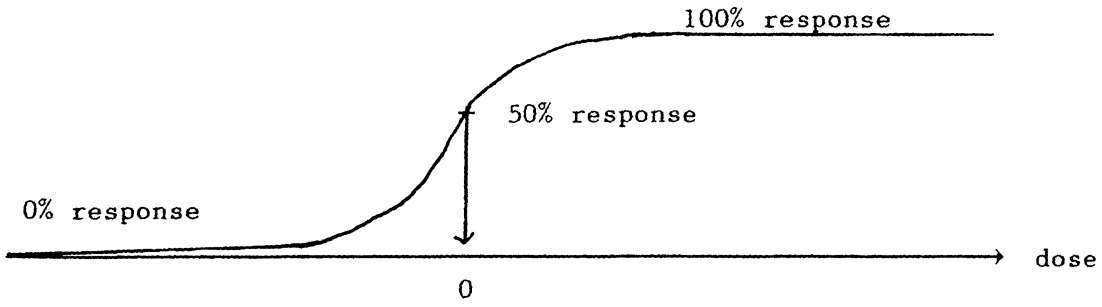
err. The more of a novice the user of computer programs is the more likely she/he is to misuse or misunderstand the package. The continual use of examples for which analyses and results are known, can often be of great help and comfort to the less experienced.

Another method of checking results is to consider the plausibility of the summarized results from an investigation. To illustrate, a chemical laboratory was running monthly analyses on sugarcane samples. The results were in the 13.0 to 16.0 range. In month ten, the results came back in the range 1.3 to 1.6 and in month eleven the range was 13.0 to 16.0 again. It is obvious that the laboratory personnel misplaced the decimal point even if they refused to concede this. There was no way that sugarcane plants could have been in the 1.3 to 1.6 range and still be alive. The biochemical and other variables for the Project should be closely scrutinized to ascertain plausibility and credibility.

10. IMPLEMENTING GLIM FOR QUALITY CONTROL OF A ROUTINE SEROLOGICAL ASSAY SYSTEM

10.1 Introduction

If a serum antibody assay system is under control then each assay theoretically produces points on the *same* dose-response curve which, when centered at zero, has the appearance:

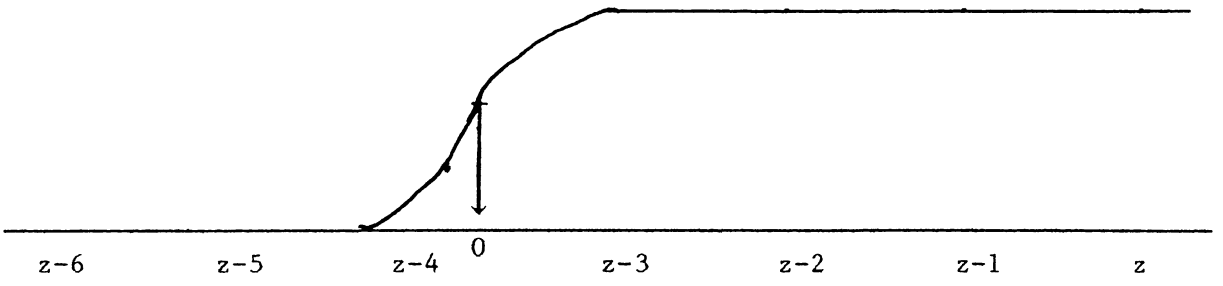


where "dose is measured in (base 2) log units of the amount of antibody per unit volume of serum. If the system is under control, then both the location and the shape of this curve remain constant between tests.

The statistical computing package GLIM (General Linear Interactive Modeling) offers a convenient, real time data analysis system for calculating titers, testing constancy of shape and, if a standard serum is assayed, testing constancy of location of the dose-response curve. As data from successive assays are accumulated in memory, the shape of this curve will be estimated with ever-increasing precision, and eventually a shift from GLIM to a custom programmed microcomputer could be implemented to reduce data analysis costs.

10.2 Conceptual Model of a Valid Assay

Each serum dilution series produces a different array of equally spaced points along the dose-response curve centered at zero, each starting from the right with the serum concentration of Z itself (or some standard dilution thereof), and proceeding to the left by two-fold dilution steps of unit length on the $\log_2$ scale.



The (fractional) number of steps required to reach the point of 50% response determines for us the location (z) of the starting point, which is reported as the titer in $\log_2$ units, or serum titer = 2^z (in the diagram, $z \approx 3.8$, or titer ≈ 14).

The dose response curve in theory represents the *expected* proportion of positive responses at each dilution while the *observed* proportion will depend on the number (n) of replicates tested per dilution. The expected proportion decreases continuously from 1 to 0 as antibody concentration decreases, while the observed proportion varies discretely over the fractions n/n, (n-1)/n, ···, 1/n,0. Interpolation among such observed responses to estimate that dilution level producing the theoretical 50% response may be accomplished in any of a variety of ways described in a variety of textbooks treating the topic of quantal bioassay. Most of these estimation procedures treat the data of each serum independently, and fail

to exploit the *a priori* information that in a controlled system all assays obey the same theoretical dose-response relation.

Estimation procedures which utilize these key assumptions must, in principle, simultaneously analyze the data from all serum samples which have been assayed under the fixed set of controlled conditions, since all of these assays contribute equal information concerning the common but unknown shape of the dose-response curve. Only recently has such a cumbersome task of data manipulation and analysis become a realistic option for routine assay systems demanding prompt output of results. Once a computer data file is updated by entering the current raw data, however, the output of results from an interactive system such as GLIM is virtually instantaneous.

10.3 GLIM Bioassay Analysis Capabilities

The existing and most readily usable statistical methodology capable of exploiting the assumption of a common dose-response curve does require specification of the parametric form of the curve, and the use of GLIM for such purposes imposes restrictions on this specification. Probit, logit and Weibull curves, the most widely used quantal response curves in bioassay analysis, are programmed in GLIM in a manner which permits a combined analysis to estimate a common shape parameter while estimating the separate titer of each serum. In the case of the probit model, for example, where the dose-response curve is specified to have the shape of a cumulative normal probability distribution, the shape parameter (β) is the reciprocal of the standard deviation (σ) of that normal distribution ($\beta=1/\sigma$). The assumption of constant shape of this response function is thus analogous to the homoscedasticity assumption in analyses of variance.

GLIM also provides statistical tests of homogeneity of shape or location, or both. If past data are stored in a computer file, GLIM can compare the shape parameter estimated from previous data and the shape parameter estimating from current data to test whether any change has occurred in the *precision* of the assay system. If a standard serum is assayed in the current and previous runs, then GLIM can compare titers estimated in current and previous assays of the standard to test whether *accuracy* has been maintained.

11. VALIDATING EXPLORATORY DATA ANALYSES

If possible, all hypotheses should be formulated prior to studying data. However, there are situations for which the data point to consideration of some previously unformulated hypotheses. Thus, if one is doing an exploratory data analysis with the idea of validating it, it is suggested that a fraction of the data, say 25 of the 65 counties, be used to set up hypotheses from exploratory data analyses. Then, the hypotheses can be tested on the remaining part of data, say on the 40 counties. For such a procedure the statistical analyses on the 40 counties can then be associated with p-values for constructing confidence intervals and making tests of significance. In any exploratory data analysis approach, it is suggested that only a fraction of the data be used in the exploratory stage. Then, any interesting analyses and hypotheses can be made on the remaining data. In larger studies an investigator may wish to use a random sample of half of the total data set for exploratory purposes and to use the second half of the data set for the confirmatory part of an analysis.

12. ASYMMETRIC COMPOSITES

Let $A_1, \cdots, A_k$ denote the k samples which are candidates for compositing at a site, and let these same symbols $A_1, \cdots, A_k$ denote readings that might be obtained from these samples if they were measured individually. The composite of all k samples will be denoted by C and referred to as a regular composite:

$$\text{Regular composite:} \quad \frac{1}{k}(A_1 \cup \cdots \cup A_k) = C \quad . \tag{25}$$

We also introduce irregular or asymmetric composites:

$$\text{Composite } A_1 \text{ with } A_2 \cup \cdots \cup A_k: \quad pA_1 \cup \frac{1-p}{k-1}(A_2 \cup \cdots \cup A_k) = C_1$$

$$= pA_1 \cup (1-p)/(k-1) \bigcup_{i=2}^{k} A \tag{26}$$

where asymmetry holds if $p \neq 1/k$ (if $p = 1/k$ then $C_1 = C$).

An asymmetric composite alone is useless, but if accompanied by either A_1 alone or the regular composite C, the pair of readings may be combined to form estimates of A_1 or C. Use A_1 and C_1 to estimate C:

$$\hat{C} = \frac{(k-1)C_1 - (kp-1)A_1}{k(1-p)} \tag{27}$$

or use C_1 and C to estimate A_1:

$$A_1 = \frac{(k-1)C_1 - k(1-p)C}{kp-1} \quad . \tag{28}$$

If A_1, C_1 and C are all available then weighted estimates of both C and A_1 can be formed, measurement error variance can be estimated, and composite validity can be tested.

$$\tilde{C} = \frac{k^2(1-p)^2\hat{C} + [(k-1)^2 + (kp-1)^2]C}{k^2(1-p)^2 + (k-1)^2 + (kp-1)^2} \tag{29}$$

$$V_{meas.}(\tilde{C}) = \frac{(k-1)^2 + (kp-1)^2}{k^2(1-p)^2 + (k-1)^2 + (kp-1)^2}\,\sigma^2_{meas.} \tag{30}$$

$$\tilde{A}_1 = \frac{[k^2(1-p)^2 + (k-1)^2]A_1 + (kp-1)^2\hat{A}_1}{k^2(1-p)^2 + (k-1)^2 + (kp-1)^2} \tag{31}$$

$$V_{meas.}(\tilde{A}_1) = \frac{k^2(1-p)^2+(k-1)^2}{k^2(1-p)^2+(k-1)^2+(kp-1)^2}\,\sigma^2_{meas.} \tag{32}$$

$$D = \frac{(k-1)C_1-(kp-1)A_1-k(1-p)C}{\sqrt{k^2(1-p)^2+(k-1)^2+(kp-1)^2}} \tag{33}$$

$$V_{meas.}(D) = \sigma^2_{meas.} \quad . \tag{34}$$

If such a compositing scheme is followed at n sites with A_1 being randomly selected from the k individuals in the i'th composite then the test statistic

$$Z = \frac{\sum_1^n D}{\sqrt{\sum_1^n D^2}} \tag{35}$$

is approximately (asymptotically) $N(0,1)$ under the null hypothesis of composite validity, and $\sum D^2/n$ then estimates $\sigma^2_{meas.}$.

Note that additional asymmetric composites might be formed, as

$$C_k = pA_k \cup \frac{1-p}{k-1}(A_1 \cup \cdots \cup A_{k-1}) \tag{36}$$

to provide additional degrees of freedom and enhanced power in the test of composite validity while also providing an estimate of the within-site variance component as, for example,

$$\frac{(A_1-A_k)^2}{2} - \frac{D_1^2+D_k^2}{2} \quad . \tag{37}$$

The number of asymmetric composites so constructed, and the choice of p in these composites might depend upon whether the study is purely cross sectional or mixed cross sectional and longitudinal. In the latter case there might be equal interest in the site mean C and the individual A_1, suggesting the choice

$$p = \frac{k+1}{2k} \tag{38}$$

giving

$$V_{meas.}(\tilde{C}) = V_{meas.}(\tilde{A}_1) = \frac{5}{6}\sigma^2_{meas.} \tag{39}$$

when only one asymmetric composite is formed. This 20 percent increase in efficiency of $\tilde{A}$ as compared to A_1 alone would serve to enhance the precision of any estimated longitudinal change in this individual while correspondingly reducing the measurement error in the cross-sectional estimate $\tilde{C}$ as compared to C. Note that for this choice of p, the asymmetric composite C_1 is formed as

$$C_1 = \tfrac{1}{2}A_1 + \tfrac{1}{2}C \tag{40}$$

so that

$$\hat{A}_1 = 2C_1 - C \qquad \text{and} \qquad \tilde{A}_1 = \frac{5A_1+2C_1-C}{6} \tag{41}$$

$$\hat{C} = 2C_1 - A_1 \qquad \text{and} \qquad \tilde{C} = \frac{5C+2C_1-A_1}{6} \tag{42}$$

$$D = \frac{2C_1-A_1-C}{\sqrt{6}} \quad . \tag{43}$$

13. USE OF LINEAR COMBINATIONS AS PREDICTOR VARIABLES IN REGRESSION

In a principal components analysis, the various principal components are linear combinations of all the variates. For example, let us consider a principal components analysis on 64 biochemical variables. Suppose that the last 40 or more principal components added little or nothing to explaining variance. This could and has led investigators to considering the first few principal components as regressor variables in a regression analysis. This must be done with extreme care and caution. For example, let Y be the mortality rates for the 65 counties and let Z_1 be the first principal component on 64 biochemical characters. A simple regression of Y on Z_1 would give predicted values of

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} Z_{1i} = \hat{\alpha} + \hat{\beta} \sum_{j=1}^{64} X_{ij} a_j \tag{44}$$

where $\hat{\alpha}$ is the estimated intercept, $\hat{\beta}$ is the estimated slope, a_j are coefficients obtained from a principal components analysis, for example, and $i = 1, \cdots, 65$ pairs of observations (Y_i, Z_i). Note that Y is *not* a function of one regressor variable but is a function of 64 X variables. The dimensionality has *not* been reduced since Y is still a function of 64 Xs.

As regressor variables, which are principal components, are added, R^2 has to approach unity. Also, if the individual X_{ij} are used as regressor variables, as the number of Xs approach 64, R^2 will approach unity whether or not any of the Xs are correlated with the Y variable. This example amply illustrates that any statistical technique can be misused and/or misinterpreted. The investigator should have some knowledge of the nature and the assumptions of a statistical procedure before using and before making interpretations of the results obtained.

14. LITERATURE CITED

Federer, W. T. (1961). A note on partial regression coefficients. BU-139-M in the Mimeo Series of the Biometrics Unit, Cornell University.

Federer, W. T., L. C. Clark, C. E. McCulloch, and N. J. Miles-McDermott. (1987). Principal components analysis of age-truncated cancer mortality data from a cancer mortality-dietary study. BU-915-M in the Technical Report Series of the Biometrics Unit, Cornell University.

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *J. Agric. Science* **26**, 424-455.