326-75

# INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

## SIMULTANEOUS STATISTICAL INFERENCE AND STATISTICAL POWER IN SURVEY RESEARCH APPLICATIONS OF THE GENERAL LINEAR MODEL

William T. Bielby
and
James R. Kluegel

UNIVERSITY OF WISCONSIN-MADISON

# Simultaneous Statistical Inference and Statistical Power in Survey Research Applications of the General Linear Model

William T. Bielby
Institute for Research on Poverty
University of Wisconsin-Madison

James R. Kluegel
Department of Sociology
University of California-Riverside

January 1976

During the preparation of this paper the authors were supported by the National Institute of General Medical Sciences Training Program in Methodology and Statistics Grant 5-T01-GM01526-08. The research reported here was supported in part by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison by the Office of Economic Opportunity pursuant to the Economic Opportunity Act of 1964. The opinions expressed are those of the authors.

ABSTRACT

In this paper we review neglected issues of simultaneous statistical inference and statistical power in survey research applications of the general linear model, and we find that classical hypothesis testing as it is currently applied, is inadequate for the purposes of social research. The intelligent use of statistical inference demands control over the overall level of Type I error and knowledge of the magnitude of effects one is likely to detect. We suggest techniques that can be used to routinely incorporate considerations of simultaneous inference and power into the statistical analysis of survey data. Several examples of applications of these techniques are presented.

## I. Introduction

Our purpose in this paper is to provide for a more informed use of statistical inference in tests of hypotheses in survey applications of the general linear model (GLM). This model, like any model, is comprised of a set of assumptions that permit the derivation of certain general principles. The assumptions of the GLM are of particular utility to the survey researcher in that they permit one to draw inferences about the structure of relationships among variables to larger populations on the basis of sample survey data. In this paper we suggest that when conducting multiple statistical tests of hypotheses within the GLM framework, our results will be more meaningful if we know the overall probability of rejecting a false null hypothesis and the probability of finding statistically significant results when substantively meaningful effects exist.

By following the current practice for doing multiple tests for hypotheses on parameters of linear models, researchers are inadequately controlling the probability of rejecting a true null hypothesis--the probability of making a Type I error. Inference considerations in situations where multiple tests of hypotheses are conducted are qualitatively different from procedures described for single hypothesis testing in most texts. Procedures currently employed yield Type I error rates that can be considerably lower than the true probability of rejecting a true null hypothesis when a number of hypotheses are tested. Scientific norms of parsimony that dictate researchers be conservative in their claims of the empirical effects of social variables are clearly violated as

social researchers systematically underestimate the likelihood that Type I errors are occurring in their analyses. Drawing on a considerable body of literature on simultaneous inference in the GLM, we argue that analysts of survey data must reconceptualize their approach to statistical inference. We discuss several techniques for treating the simultaneous inference problem in the GLM, and present, with examples, procedures for applying these techniques to survey data.

In current research practice little concern has been expressed for the power of GLM statistical tests. Analysts of large sample survey data often dismiss power considerations with the assertion that their tests have "more than enough power"--even trivial effects yield statistically significant results (Blau and Duncan, 1967, pp. 17-18). We present examples below to demonstrate that for many GLM hypotheses, whether or not the tests are characterized by "more than enough power" can be quite problematic. Indeed, most researchers are confronted with a situation where they must take as given two important determinants of the power of statistical tests, sample size and the configuration of independent variables.[1] Thus, we argue that it is imperative that the analyst compute the magnitudes of the effects that are likely to be detectable in a given set of data. As we shall demonstrate below, the power of GLM tests can be routinely calculated.

The importance of the consideration of issues in simultaneous statistical inference and power for the informed use of statistical tests of hypotheses requires that the survey researcher be aware of major issues and procedures pertaining to these two areas. In this paper,

we provide a critical examination of issues and procedures in simultaneous statistical inference and statistical power as they apply to the research. We begin our exegesis with a brief review of assumptions of the GLM and common hypothesis tests in survey research applications. Drawing on the body of literature on simultaneous inference in the GLM, we argue that analysts of survey data must reconceptualize their approach to statistical inference. We discuss several techniques for treating the simultaneous inference problem in the GLM, and present, with examples, procedures for applying these techniques to survey data. Following the treatment of simultaneous inference we examine factors that influence the ability to detect substantively meaningful effects--statistical power. A procedure for estimating the power of statistical tests is discussed and illustrative examples of the influence of various factors on statistical power are presented. We conclude with some suggestions for improving the use of statistical inference in making meaningful decisions about the merits of the hypotheses being tested.

## II. The General Linear Model: Assumptions and Common Hypothesis Tests

### A. GLM Assumptions

We shall concern ourselves with tests of hypotheses about the parameters of the GLM, the classical model stated in matrix terms as follows:

$$(1) \quad \underset{(Nx1)}{y} = \underset{(NxK)}{X}\,\underset{(Kx1)}{\beta} + \underset{(Nx1)}{\varepsilon}$$

(2) $E(\underline{\varepsilon}) = \underline{0}$

(3) $E(\underline{\varepsilon}\underline{\varepsilon}') = \sigma^2\underline{I}$

(4) $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\underline{I})$

(5) $\underline{X}$ is fixed (nonstochastic) and of full column rank.[2]

While the theory of statistical inference for the GLM was originally developed for the above model, assumption (5), fixed $\underline{X}$, is clearly untenable in the application of the model to survey data. It requires that the sampling design include *a priori* stratification on all independent variables, i.e. *a priori* specification of cell sizes for each combination of the levels of the independent variables. A modification of the above model allows for the typical survey design of multivariate sampling from the joint distribution of $\underline{y}$ and $\underline{X}$. We replace (2) through (5) above with the following assumptions:

(2a) $E(\underline{\varepsilon}|\underline{X}) = 0$

(3a) $E(\underline{\varepsilon}\underline{\varepsilon}'|\underline{X}) = \sigma^2\underline{I}$

(4a) $\underline{\varepsilon}|\underline{X} \sim N(\underline{0},\sigma^2\underline{I})$.

Thus it is required that the classical assumption holds *conditionally* on $\underline{X}$. The disturbance must be mean independent of the independent variables[3] and be conditionally independently normally distributed with zero mean and constant variance. While this more appropriate conditional GLM presents no differences in the treatment of Type I error, it does complicate the treatment of power. While our results with respect to Type I error hold unconditionally, the procedures for power calculations presented in this paper give results *conditional* upon the values of $\underline{X}$ realized in a particular sample (Graybill, 1961, pp. 204-205; Sampson, 1974). The

conditional power calculations presented herein must be considered upper bounds upon the unconditional power of the tests.[4]

B. Hypothesis Testing in Survey Applications of the GLM

In Table 1 we present an outline of the types of GLM hypotheses commonly tested in survey applications and the statistical tests applied to those hypotheses. In (1) we have the test of an individual coefficient, $\beta_i$. The t-test is just $b_i - \beta_i^*$ divided by the standard error of $b_i$, the usual t-ratio computed in regression programs. The one degree of freedom F-test is merely the square of the t-test.

Hypothesis (2) is the test that a subset of J coefficients are jointly equal to a set of J specified values. When $\underline{\beta}^*_{(2)}$ is specified to be a vector of zeros and $J = K - 1$, it is the common "overall" F-test of no regression. When $J < K - 1$, and $\underline{\beta}^*_{(2)}$ is a vector of zeros, it is the "increment to $R^2$" F-test for a subset of variables.

Hypotheses (1) and (2) comprise the majority of hypotheses tested in survey applications of the GLM. Although seldom conducted in non-experimental applications of the GLM, a researcher may want to test whether linear combinations of the coefficients are equal to some specified zero or nonzero values. Paralleling hypotheses (1) and (2), one can test a single linear combination with a t-test[5] or one degree of freedom F-test, or jointly test J linearly independent linear conbinations of coefficients. Indeed, hypotheses (1) and (2) are special cases of (3) and (4).

Finally, each of the F-tests for the hypotheses can be considered "increment to $R^2$" tests with J numerator and N - K denominator degrees of freedom as given in the equation for the u statistic found in Table 1.

Table 1. General Linear Hypothesis Statistical Tests

| NULL HYPOTHESIS | TEST |
|---|---|
| (1) Individual coefficient (J = 1): $H_o: \beta_i = \beta_i^*$ . | 1 df F-test or t-test: $t = (b_i - \beta_i^*)/s^2((\underline{X}'\underline{X})^{-1})^{\frac{1}{2}}$ . |
| (2) Set of J coefficients: $H_o: \underline{\beta}_{(2)} = \underline{\beta}^*_{(2)}$, where $\underline{\beta}' = (\underline{\beta}'_{(1)}\ \underline{\beta}'_{(2)})$. | J df F-test on increment to $R^2$ |
| (3) Linear combination of coefficients (J = 1): $H_o: \underline{a}'\underline{\beta} = a'\underline{\beta}^*$. | 1 df F-test or t-test: $t = \left[ \frac{(\underline{a}'\underline{b} - \underline{a}'\underline{\beta}^*)'(\underline{a}'(\underline{X}'\underline{X})^{-1}\underline{a})^{-1}(\underline{a}'\underline{b} - \underline{a}'\underline{\beta}^*)}{s^2} \right]^{\frac{1}{2}}$ |
| (4) Set of J independent linear combinations of coefficients: $H_o: \underset{(J\times K)}{\underline{A}}\ \underset{(K\times 1)}{\underline{\beta}} = \underline{A\beta}^*$. | J df F-test: $u = \frac{(\underline{A}\,\underline{b} - \underline{A}\,\underline{\beta}^*)'(\underline{A}(\underline{X}'\underline{X})^{-1}\underline{A}')^{-1}(\underline{A}\,\underline{b} - \underline{A}\,\underline{\beta})/J}{s^2}$ |

The test statistic for each of the above F-tests can be written as:

$$u = \frac{(R^2 - R^2_{Ho})/J}{(1 - R^2)/(N - K)} ,$$

where $R^2$ is the proportion of variance explained by the full, unrestricted model, and $R^2_{Ho}$ is the proportion of variance explained when the model is constrained by the null hypothesis. The statistic u is distributed $F_{J,N-K}$ under the null hypothesis.

## III. Simultaneous Statistical Inference

In most applications of the GLM in survey research more than one of a single type of the above delineated hypotheses is tested or more than one type of hypothesis is tested. Frequently some set of interaction effects are tested jointly and main effects are tested individually. When the effects of a categorical and one or more continuous independent variables are analyzed, often a joint test of the effects of the set of dummy variables representing the categorical variable and one or more individual tests of the coefficients for the continuous variables are performed. In applications of the GLM that involve a single equation, the performance of multiple t-tests on individual slope coefficients is a universal practice. Finally, it is becoming standard practice to do multiple tests on all possible slope coefficients in simple recursive structural equation models.

We have briefly noted above the researcher who performs such multiple hypotheses tests is in a qualitatively different inference situation--that of simultaneous statistical inference--than the researcher who performs only a single hypothesis test. In this section we shall consider both how the single and multiple hypotheses cases differ from the standpoint of inference, and techniques of statistical inference that are appropriate to the mulitple hypotheses test situation. First we shall examine these two issues in general and then we shall consider them as they apply to the standard tests conducted within survey research applications of the GLM.

## A. General Issues in Simultaneous Statistical Inference

An understanding of the basic difference between the single hypothesis and multiple hypotheses cases can best be achieved by first recalling the definition of Type I error in statistical inference. Type I error is the error of falsely rejecting a true null hypothesis. For the researcher who performs only a single null hypothesis test, this definition presents no problem. The probability that he will falsely reject this single null hypothesis is the probability of Type I error in this case. The researcher can straightforwardly proceed by following the suggested standard procedure of specifying a level, $1 - \alpha$, of protection against a Type I error, and then proceed to perform his statistical test accordingly. Now consider what happens if this same researcher sometime in his life performs additional tests of null hypotheses according to the suggested standard procedure. That is, he specifies a level $1 - \alpha$ of desired protection against a Type I error and conducts each of his statistical tests at this level.

If we reflect now on the definition of Type I error we realize that for this researcher the actual level of protection against making a Type I error in the multiple null hypotheses case is less than $1 - \alpha$. Thus, this researcher is overestimating the protection he has against falsely rejecting a true null hypothesis. The problem with employing the conventional procedure for making tests of multiple null hypotheses arises because the probability of making a Type I error in this case is the probability of falsely rejecting any one of the individual null hypotheses--which equals the probability of making a Type I error for the first null hypothesis, or for the second null hypothesis, or for the nth null

hypothesis, or for any combination of the n hypotheses. Except in the case of total dependency among the null hypotheses tested, this probability is grester than the $\alpha$ levèl under which each of the null hypotheses were tested.

Essentially the solution to this problem is provided by the researcher's decision concerning which null hypotheses will be grouped together for the purpose of considering Type I error--generally referred to as the specification of the unit of error rate. Given this decision, the researcher can proceed to do each of the tests of individual null hypotheses in such a fashion that the desired level of protection against making a Type I error for the group of null hypotheses has been provided.

Two extreme groupings of null hypotheses can be identified. First, one could consider as a group all the null hypotheses tests that a researcher or a group of researchers will do in his or their lifetime. By grouping in this manner the researcher would be provided with protection at a specified level against ever falsely rejecting a true null hypothesis. Second, one could consider each individual null hypothesis test as a group for inference purposes. This grouping is generally called a per-comparison unit of error rate and effectively removes one from the simultaneous inference situation. The first extreme grouping essentially has been rejected in discussions of appropriate units of error rate for research, and a general agreement exists that the upper bound for grouping purposes is provided by the group of null hypotheses tested by one researcher in one study. However, there exists no consensus on what is the most appropriate unit of error rate below this upper bound (Ryan, 1959, 1962; Wilson, 1962; Miller, 1966).

The problem addressed by simultaneous statistical inference techniques is that of how to perform tests of individual hypotheses such that one has protection at a specified level against making a Type I error for a group of hypotheses. Numerous techniques of simultaneous statistical inference have been designed to address this problem (Miller, 1966; Kirk, 1968), many of which have been developed for specific types of tests within the GLM framework.[6] However, two techniques, the Bonferroni and Scheffé, are of wide generality.

The Bonferroni technique is based on the Bonferroni inequality, which states that

$$(6) \quad \alpha_G \leq \sum_{i=1}^{N} \alpha_{S_i} ;$$

where $\alpha_G$ equals the significance level for a group of null hypotheses, $\alpha_{S_i}$ equals the significance level for each individual null hypothesis in the group, and N equals the total number of null hypotheses in the group. The Bonferroni technique can be applied to virtually all situations of multiple hypotheses tests where one has prior knowledge of how many tests are to be conducted.

The Scheffé method, in the GLM framework, provides a means of controlling error rate for tests of all possible linear combinations of the least squares estimates of the slope coefficients. The Scheffé technique is based on the common assumptions (distributional and otherwise) of the GLM. Its generality is due to the fact that it allows the researcher to perform an infinite number of tests of linear combinations of $\beta$ coefficients while protecting against a specified value of Type I error for the group.

B. Simultaneous Statistical Inference in GLM Applications in Survey Research

The first question that must be addressed is, "Does one need to be concerned with issues of simultaneous inference?" The implicit answer given to this question in survey research applications of the GLM to date has been, "no." Virtually all analyses of survey data conducted within the GLM framework have implicitly employed a per-comparison error rate. In general, analyses of multiple null hypotheses based on survey data have been performed in the following manner: A value of Type I error is specified and this value is used in tests of each individual null hypothesis. No consideration is given to error rate for any group of hypotheses.

There are compelling reasons for believing that this implicit answer is insufficient. The first reason is that the implicit answer is usually based on a lack of knowledge of the issues in simultaneous inference. Basic textbook treatments of statistical inference, from which most social researchers' knowledge of this subject is obtained, generally ignore simultaneous statistical inference. Consequently, many social researchers are unaware or vaguely aware that a problem may exist in doing multiple tests of null hypotheses.

Beyond this lack of knowledge there are important substantive reasons for considering simultaneous inference. Perhaps the most important of these is the fact that social researchers do not limit their concern to the determination of whether or not a single variable has a statistically significant direct effect on a given dependent variable, but extend their interest to the determination of whether or not a set of independent

variables affects a given dependent variable. Such analyses are frequently done i n the context of a causal model of a process that determines variation in a dependent variable. This is particularly the case in analyses done within the recursive structural equation framework. Here the researcher frequently begins with a specified causal ordering among a set of variables and a set of null hypotheses about the relationships among this set of variables. It can be argued that since the researcher is interested in finding the correct model of a process in a population, the set of multiple null hypotheses used to find this model should be tested simultaneously. That is, the researcher should provide protection against a specified level, $\alpha_G$, of finding an incorrect model of a process; since falsely rejecting any of the multiple null hypotheses is in effect finding an incorrect model of the process.

Another reason for being concerned with units of error rate other than the per-comparison unit is the scientific dictum of conservatism and parsimony. It is generally thought that the acceptance of a false null hypothesis is more desirable scientifically than the rejection of a true null hypothesis. Such a principle, it is proposed, keeps the scientific literature from becoming unduly confused by false research findings and keeps scientific theories from becoming overly complex. Since the employment of a unit for error rate other than the per-comparison unit makes it more difficult to capitalize on chance in conducting tests of multiple null hypotheses, the scientific dictums of conservatism and parsimony argue for the use of simultaneous inference techniques.

A final reason is specific to the practice of "data snooping" or "data dredging." Both terms are used in reference to the practice of doing some previously unspecified number of tests within a body of data in an attempt to discover relationships among the set of variables analyzed. Such a practice is undertaken either because the researcher has no prior hypotheses about the relationships among a set of variables or wishes to supplement an analysis of prior hypotheses. In this situation it is argued that since one approaches an analysis with an unspecified number of null hypotheses to be tested--the number tested could be one, several, or all possible tests--the scientifically honest procedure is to use a simultaneous inference technique that provides protection against a level of Type I error for all possible tests.

Given that one has concluded that it is desirable to employ simultaneous statistical inference techniques, the next question that must be addressed is "What unit of error rate should be employed?" The most straightforward answer that can be given to this question is simply that there are no hard and fast rules. The unit of error rate used is dependent upon the researcher's judgement of what unit best suits the research proposes. We can make suggestions, however, about what seems to be appropriate units for certain applications of the common hypotheses tests delineated in Table 1. We will consider three such applications: (1) the prediction situation, (2) the use of "theory trimming" in simple recursive structural equation models, and (3) the use of various hypothesis tests in post hoc analyses of linear models.

Consider first the situation in which the researcher is simply interested in determining which variables among a set of independent variables have significant, direct effects on a dependent variable. The intent here is usually that of discovering what variables are important determinants of variation in some dependent variable.. The usual procedure in this case is the performance of an individual test of the hypothesis that $\beta_i$ equals zero for each independent variable. We suggest that all of the individual hypothesis tests of the $\beta$'s be grouped together for purposes of considering error rate. Such a grouping seems appropriate since the focus of this type of research is on the correct prediction of values of a given dependent variable. By grouping in this fashion, the researcher is protected at the level $1 - \alpha_G$ against making a Type I error in predicting values of a given dependent variable.

Secondly, consider the "theory trimming" strategy (Heise, 1969) often employed in the analysis of simple recursive structural equation models. A common procedure in social research is the specification of a recursive causal ordering among a set of variables and the employment of multiple t-tests of individual $\beta$ coefficients (or their standardized counterparts) to determine which effects among those possible in a recursive causal ordering are significant.[7] The intent here is usually that of determining the most plausible model of some process in a population. We propose that all of the null hypotheses tested in the "theory trimming" process be considered as a group for error rate purposes. Because in this case, the researcher is interested in finding the correct model of

a process in a population, grouping in this fashion is appropriate. Since falsely rejecting any one of the null hypotheses about the individual $\beta$ coefficients means that the researcher has found an incorrect model of the process, protection should be provided against falsely rejecting any one of the null hypotheses. By treating all of the null hypotheses as a unit for purposes of considering Type I error the researcher does so at the level $1 - \alpha_G$ .

Finally, consider the use of the common hypotheses tests in the post hoc case. Frequently the results of one's analysis do not conform to the original expectations. The attainment of unexpected results may at least partially be attributed to initial assumptions not holding. For example, one may have assumed that the relationships among a set of variables are linear and additive when in fact they are nonlinear or not additive. Many of these assumptions are testable with the data at hand and in such a situation the researcher may wish to perform a number of hypothesis tests to determine if the unexpected results are attributable to the failure of meeting one's assumptions.

Additionally, the results of one's analysis may suggest further tests that may be interesting to the researcher or the researcher may simply wish to snoop around in the data in the hope of discovering an interesting result. In all these post hoc analyses the researcher is "data dredging." For the reasons of scientific honesty elaborated above we suggest that the researcher employ as the unit of error rate all possible tests of the $\beta$ coefficients and employ the Scheffé technique.

After one has determined the unit of error rate to be employed, a simultaneous inference technique must then be chosen. The Bonferroni and Scheffé techniques can both be applied to all of the common hypotheses tests performed on slope coefficients listed in Table 1.[8] The two techniques do differ, however, in the advantages each presents in specific situations. Two criteria are of importance in weighing the relative advantages of each technique: (1) the ability the techniques present to detect specific alternative hypotheses (i.e. statistical power), and (2) their applicability to a priori versus post hoc statistical tests. We shall first consider the mechanics of applying each technique and then weighing their relative advantages in terms of these two criteria.

C. The Bonferroni Technique

To provide protection at level $1 - \alpha_G$ for a group of null hypotheses via the Bonferroni technique one first determines the total number of individual null hypotheses to be tested, m. Then, one divides $\alpha_G$ by m and tests each individual null hypothesis with a significance level equal to $\alpha_G/m$. For example, if one wishes to test whether a subset of coefficients are jointly equal to zero and to test whether four additional coefficients are individually equal to zero, with a group probability error rate of .05.one would simply conduct the tests corresponding to hypotheses (1) and (2) in Table 1, each with $\alpha_{S_i}$ equal to .01.

D. The Scheffé Technique

The Scheffé technique is applied to the various F-tests specified in Table 1. It requires that one first perform a test of the joint null hypothesis that all of the coefficients from which subsequent tests of the nature of those in Table 1 will be conducted, with $1 - \alpha_G$ equal to the desired level of protection against a Type I error. If this test is nonsignificant one stops here and performs no further tests, since tests of any linear combination of these $\beta$ coefficients (this includes as well hypotheses (1) and (2) in Table 1) will prove nonsignificant. If, on the other hand, one can reject this joint null hypothesis, then one carries out any and all of the tests in Table 1 by using as the critical value of the test statistic for all individual null hypotheses tests, the quantity

$$JF^{\alpha_G}(J, N - k) ,$$

where J equals the degrees of freedom from the joint null hypothesis test that all coefficients equal to zero. For example, if the researcher wishes to conduct individual tests of three different linear combinations of four $\beta$ coefficients (tests of the form of hypothesis (3) in Table 1) one would use the specified test in Table 1 with the critical value of the test statistic equal to

$$4F^{\alpha_G}(4, N - K) .$$

E. Bonferroni and Scheffé Procedures: Some Comparisons

Note first that the Bonferroni and Scheffé procedures are conservative. In general the actual value of the group error rate will be less than that desired. Hence, one will have greater protection against a Type I error than initially specified. The Bonferroni procedure, as we had mentioned, is based on the Bonferroni inequality and the fact that it produces only approximations to the actual group error rate that can be readily seen. The Scheffé technique provides an exact value of the group error rate for all possible linear combinations. However, it is only a finite subset of these linear combinations that is ever tested, and consequently it, like the Bonferroni technique, is conservative.

The fact that the Scheffé procedure provides an error rate for all possible linear combinations, while the Bonferroni procedure is based on a finite number of tests provides some insight into the ability of each to allow the rejection of individual null hypotheses. Intuitively, it appears that the Scheffé technique will be less powerful than the Bonferroni technique because the former is based on an infinite number of tests while the latter is not. In fact the Scheffé technique will always be less powerful for the rejection of individual null hypotheses when m, the actual number of tests made, is less than or equal to J, the degrees of freedom for the numerator of the F statistic. On the other hand, when m is considerably bigger than J, the inexactitude of the Bonferroni procedure is such that the Scheffé procedure provides greater power for the

rejection of individual null hypotheses (Miller, 1966, pp. 62-63; Dunn, 1959). Rather than rely on the somewhat sparse literature comparing the power of simultaneous inference techniques, the researcher can apply both the Bonferroni and Scheffé techniques (or others) and use the one that provides greatest power. Doing so is fully permissible in the a priori case since the choice of technique is independent of the data collected.

The Scheffé procedure presents an advantage over the Bonferroni technique in post hoc tests. For reasons of scientific honesty elaborated earlier, the Scheffé technique is more suitable for searching one's data in an attempt to discover the nature of the relationships among a set of variables.

We turn now to an examination of issues concerning the power of statistical tests of GLM hypotheses. Before doing so, an important implication of the conservative nature of the Bonferroni and Scheffé procedures for the estimation of the power of GLM hypothesis tests must be noted. As we shall discuss below, the smaller the $\alpha$ level for a hypothesis test, the less power of that test (holding other factors constant). As a consequence, the ease of applicability of the two procedures is purchased at the cost of an overestimate of the power of a test (where the degree of overestimate depends on the size of the discrepency between the conservative $\alpha$ and the true $\alpha$). This fact, together with the conditional nature of the power calculations noted above, makes it imperative that we stress that the power calculations to be presented below should be taken as absolute minimum Type II error rates.[9]

## IV Power

Once appropriate statistical tests and Type I error rate have been selected, to calculate power, the probability of rejecting a false null hypothesis for any test, we must determine the probability that the test statistic for the hypothesis exceeds the critical value when a given alternative hypothesis is true. The GLM test statistic u is distributed as a noncentral F when an alternative hypothesis is true, with J and N - K degrees of freedom and noncentrality parameter $\delta^2$. The distribution has the property that the probability of the statistic u exceeding a given critical value (and consequently power) increases monotonically with $\delta^2$. The noncentrality parameter is a function of (among other things) the degree to which the null hypothesis is false. For the most general GLM test, the test of a set of J linear combinations of coefficients is

$$(7) \quad H_0: \ \underline{A\beta} = \underline{A\beta}^* \quad \text{and,}$$

the noncentrality parameter, $\delta^2$ is

$$(8) \quad \delta^2 = \frac{(\underline{A\beta} - \underline{A\beta}^*)' \ (\underline{A}(\underline{X}'\underline{X})^{-1} \underline{A}')^{-1} \ (\underline{A\beta} - \underline{A\beta}^*)}{\sigma^2_{y.x}} .$$

Figure 1 presents a plot of power as a function of $\delta^2$ for various combinations of Type I error rates and numerator degrees of freedom, and arbitrarily large denominator degrees of freedom.[10] It can be seen that for a given $\alpha$ and $\delta^2$, power decreases with numerator degrees of freedom J, and that for given J, power is monotonically related to the

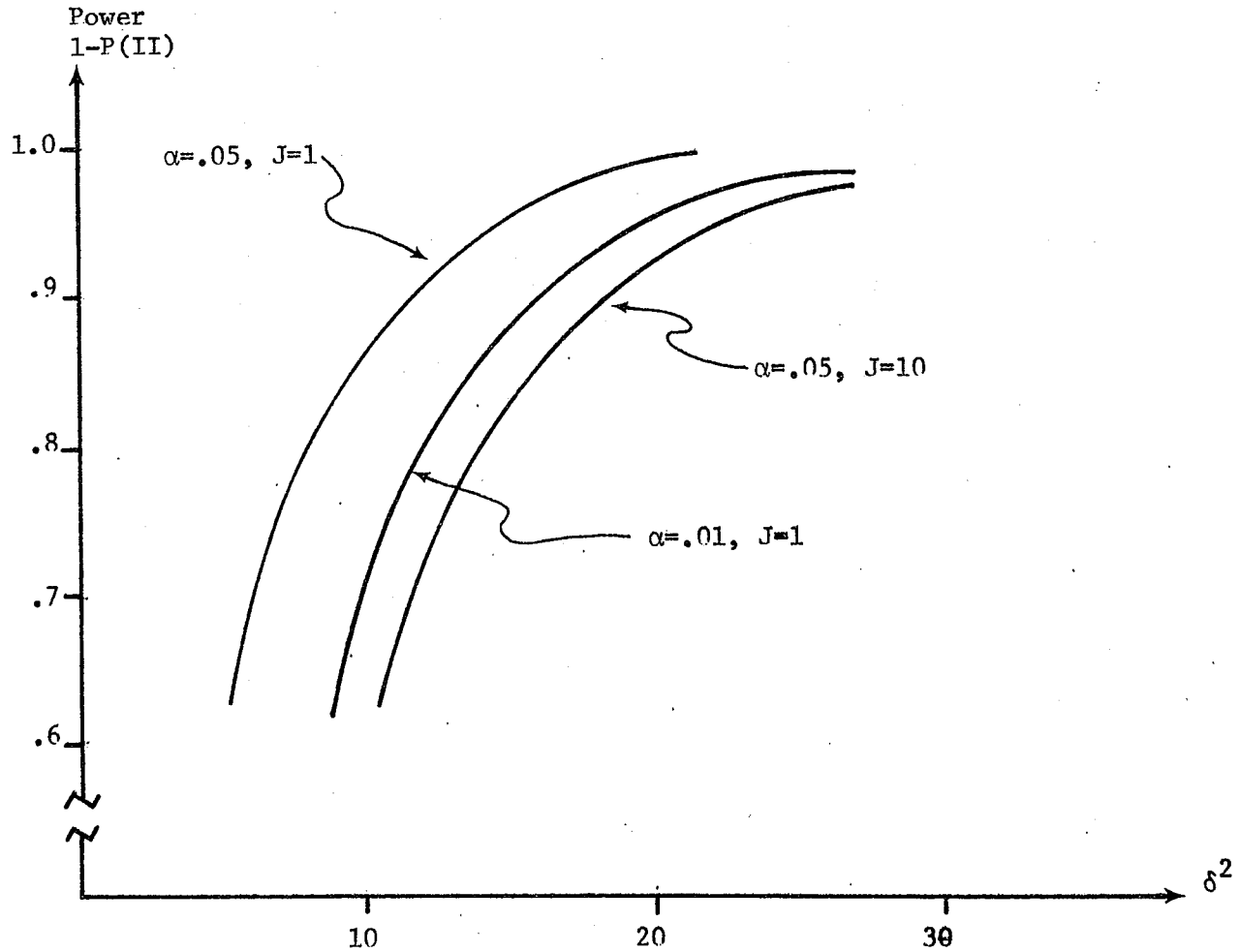


Figure 1: Power as a function of non-centrality parameter

probability of a Type I error. All three of these relationships are important when we consider the implications of simultaneous inference for power.

In Table 2 we present alternative expressions for the noncentrality parameters for the test of an individual coefficient and the joint test of J ($J \leq K - 1$) coefficients. The noncentrality parameters are presented as functions of the original GLM parameters and standardized parameters.[11] Looking first at the test of the $k^{th}$ individual coefficient we see immediately that $\delta^2$ (and therefore power) increases with the degree $\beta_k$ departs from its hypothesized value $\beta_k^*$. Noting that $\underline{x}'\ \underline{M}^*\ \underline{x}$ is just the sum of squared residuals for the regression of the $k^{th}$ independent variable on the remaining K - 2 independent variables, we conclude also that power increases with the orthogonality of the $k^{th}$ independent variable to the others. We see this again in the $(1 - R^2_{x_k \cdot X^*})$ term in the standardized expression, and note also that, of course, power increases with sample size. From the standardized expression we also see that power increases with the proportion of variance explained (i.e. as $\sigma^2_{y.x}/\sigma^2_y$ decreases). None of these results should be surprising. The ability to reject a false null hypothesis increases with the degree to which it is false, the degree to which the effect being tested is non-redundant with the effects of other parameters, the amount of data available, and the overall power of the linear model.

The $\delta^2$ parameter for the joint test of J ($J \leq K - 1$) coefficients can be interpreted as a multivariate extension of the single coefficient

Table 2. Noncentrality Parameters for some GLM Tests

| | Unstandardized | Standardized |
|---|---|---|
| 1. The test of an individual coefficient: $H_o: \beta_k = \beta_k^*$ | $\delta^2 = \frac{(\beta_k - \beta_k^*)^2 \underline{x}'_k \underline{M}^* \underline{x}_k}{\sigma^2_{y.X}}$ | $\delta^2 = \frac{n(\tilde{\beta}_k - \tilde{\beta}_k^*)^2 (1 - R^2_{x_k.X^*})}{\sigma^2_{y.X}/\sigma^2_y}$ |
| 2. The joint test on the last J coefficients: $H_o: \underline{\beta}_{(2)} = \underline{\beta}^*_{(2)}$ | $\delta^2 = \frac{(\beta_{-(2)} - \beta^*_{-(2)})' \underline{X}_2 \underline{M}_1 \underline{X}_2 (\beta_{-(2)} - \beta^*_{-(2)})}{\sigma^2_{y.X}}$ | $\delta^2 = \frac{n(\tilde{\beta}_{-(2)} - \tilde{\beta}^*_{-(2)})' R_{-22.1} (\tilde{\beta}_{-(2)} - \tilde{\beta}_{-(2)})}{\sigma^2_{y.X}/\sigma^2_y}$ |

Notation and Definitions:

1. $\underline{\beta}$ is the K - 1 element vector of standardized coefficients, $\tilde{\beta}_k = \frac{s_x}{s_y}\beta_k$ for k = 2, . . ., K.

2. The vector $\underline{x}_k$ is the $k^{th}$ column of $\underline{X}$. The matrix $\underline{X}^*$ is the N x (K - 1) data matrix $\underline{X}$ with $k^{th}$ column omitted.

3. The coefficient vector $\underline{\beta}'$ may be partitioned as $(\underline{\beta}'_{(1)}\ \underline{\beta}'_{(2)})$, where $\underline{\beta}_{(2)}$ is the vector of the last J coefficients. The matrix $\underline{X}$ is similarly partitioned by columns as $(\underline{X}_1\ \underline{X}_2)$.

4. $\underline{M}^*$ is the idempotent matrix $(\underline{I} - \underline{X}^*(\underline{X}^{*\prime}\underline{X}^*)^{-1}\underline{X}^{*\prime})$. $M_1$ is the idempotent matrix $(I - X_1(X_1'X_1)^{-1}X_1')$. For the properties of such idempotent matrices see Theil (1971: 113-114).

5. $R^2_{x_k.X^*}$ is the squared multiple correlation coefficient of $\underline{x}_k$ on $\underline{X}^*$. $\underline{R}_{22.1}$ is the matrix of partial correlations among the J variables in $\underline{X}_2$ with the variables in $\underline{X}_1$ partialled out. The diagonal elements of this matrix are the proportions of variance in the J variables orthogonal to the K - J - 1 variables in $\underline{X}_1$.

case. Both $\underline{X}_2' \underline{M}_1 \underline{X}_2$ and $\underline{R}_{22.1}$ are measures of the degree to which the covariation among the J variables with coefficients being tested is orthogonal to the covariation among the remaining K - J - 1 independent variables, and $\underline{B}_2 - \underline{B}_2^*$ is the vector discrepancy between the true values of the J coefficients and their hypothesized values. Indeed for all GLM tests we can conceptualize the noncentrality parameter as a scaler measure of the degree to which the null hypothesis is false, weighted by the amount of independent information available.

Given a substantively meaningful alternative hypothesis one would wish to be able to detect, the configuration of independent variables in the model, and the completeness of the model as measured by the proportion of variance explained, it is a trivial matter to program a computer to compute $\delta^2$ as expressed in equation (8) for any general linear hypothesis and any specified alternative. Thus given $\delta^2$ and n, one has enough information to determine the power of the test from Pearson and Hartley charts (Scheffé, 1959: 438-445; Kirk, 1968: 520-547). We now present examples of calculations of power as functions of n and the degree to which the null hypothesis is false.

## A. Determining the Power of GLM Tests: Some Examples

Is it indeed the case that survey researchers are typically confronted with testing situations where they have "too much" power, i.e. is it usually true that trivial departures form the null hypothesis result in statistically significant tests? It is impossible to answer this question with any accuracy without doing power calculations . Our calculations

presented below show that having too much power is by no means generally the case. Furthermore, we argue that if a researcher is to report the results of statistical tests, it is always imperative that the magnitude of the effects, trivial or nontrivial, which are likely to be detected also be presented.

Consider the following linear model:

$$(9) \quad y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i ,$$

where:

$y$ = Income,

$X_2$ = Education,

$X_3$ = Occupation,

$X_4$ = Parental income,

$X_5$ = Father's occupation,

$X_6$ = Father's income.

When models of the socioeconomic achievement process such as equation (9) are estimated, the researcher is usually interested in hypotheses about all five coefficients (excluding the intercept), $\beta_2, \ldots, \beta_6$. To maintain an overall protection of $1 - \alpha_g = .95$ against Type I error we can conduct simple t-tests on the five coefficients at the .01 level (Bonferroni), or compare the usual 1 degree of freedom F-tests (the square of the t-test) to the $5F^{.05}_{5,N-K}$ critical value (Scheffé).

Let us consider the power of the test of the education coefficient, $H_0 : \beta_2 = 0$, when the true net return to a year of education is \$150, $H_1 : \beta_2 = 150$. Given this null hypothesis and a meaningful alternative

hypothesis, how large of a sample size is required to have a reasonable likelihood of detecting an education effect of $150 a year in samples drawn from the United States labor force? We know from the expression for the noncentrality parameter in Table 2 that the power of the test also depends on the variation in and covariation among the independent variables, and the proportion of variation explained in the model. Assuming that the five social variables explain about 15 percent of the income variation in the United States labor force, and using published correlations and variances of the five variables from a sample of Wisconsin high school graduates (Sewell and Hauser, 1972), we can calculate the noncentrality parameter as a function of sample size. From Pearson-Hartley charts we can then plot power as a function of sample size.[12]

In Figure 2 we present the plot of power as a function of sample size for a Type I error rate of .05 for: (1) the Bonferroni test, (2) the Scheffé projection, and (3) a simple t-test not controlling for overall Type I error rate (the noncentrality parameter is identical for all three tests). We see, as noted above, that the Bonferroni test is slightly more powerful than the Scheffé projection, and that compared to the simple t-test one must pay a price in terms of power in order to control for overall error rate. Thus, if one is going to test the net education effect with a Bonferroni test, a sample size of at least 2200 observations is required to achieve a power of .90.

Does the above result imply that national samples of more than 10,000 observations, tests on coefficients of the income determination model will have "more than enough power"? This is only true for the

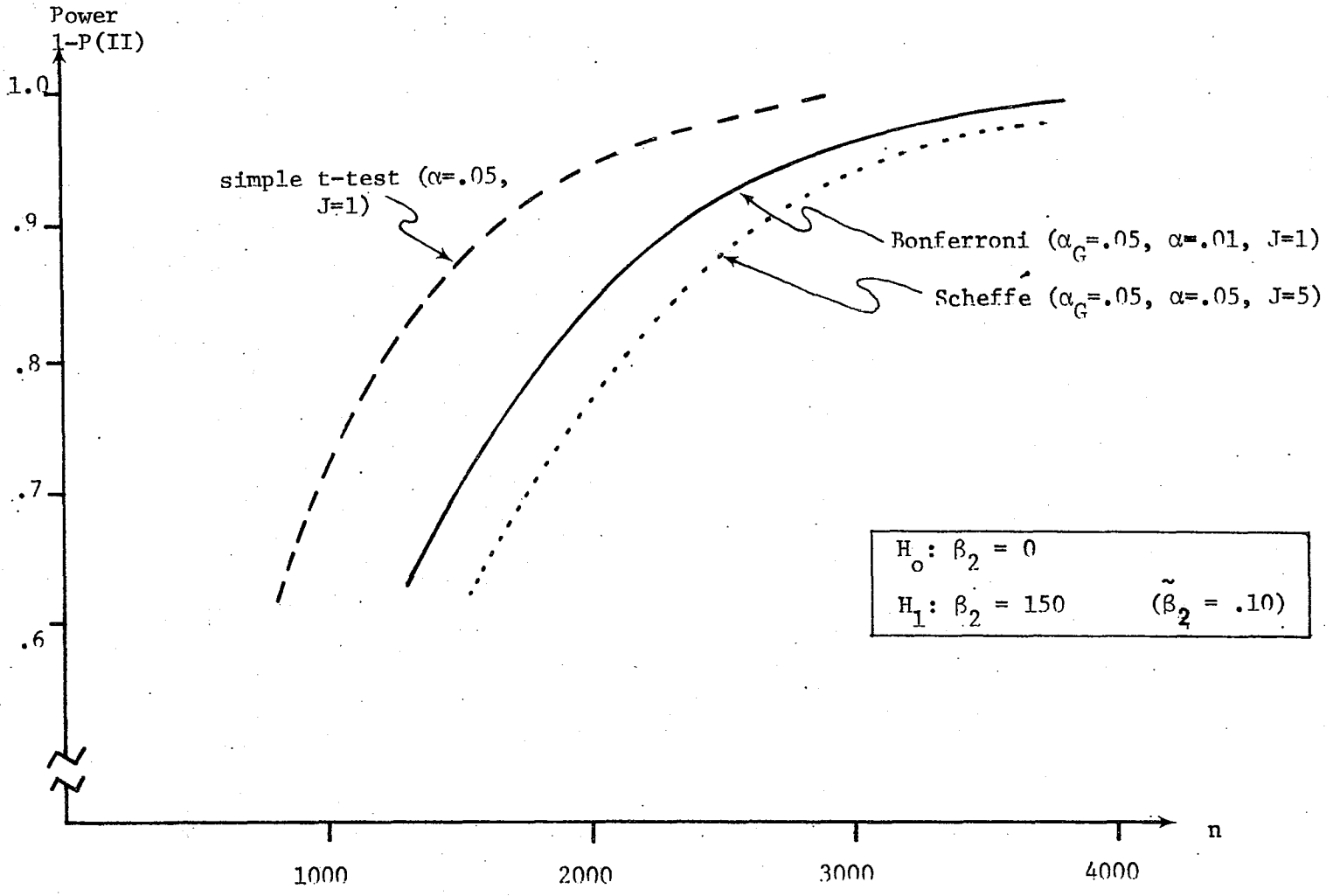


Figure 2: Income equation--Power as a function of sample size for the test on an individual coefficient

specific null and alternative hypotheses specified above. Consider a different hypothesis on the same education coefficient. Suppose that from a census of the population we know that the net effect of a year of education in 1960 was \$150. In 1975 we are to collect a sample in order to detect changes in $\beta_2$ through $\beta_5$ and we want to be able to detect a change in $\beta_2$ of \$30 a year in either direction. In this case we are testing a nonzero null hypothesis, $H_0 : \beta_2 = \beta_2^* = 150$, against a nondirectional alternative: $H_1 : \left| \beta_2 - \beta_2^* \right| = 30$. Using the same information as in the previous example, we have determined power as a function of sample size for this test and present the plot in Figure 3. In order to detect a change of \$30 a year with a Bonferroni test with a power of .90, a sample of about 60,000 observations would be required.[13] The power of <u>joint</u> tests on coefficients is generally greater for a given sample size. Figures 4 and 5 present the power of joint tests on $\beta_2$, $\beta_3$ and $\beta_4$ (where we have assumed that no hypotheses concerning $\beta_5$ and $\beta_6$ are to be treated). Figure 4 presents the power of the test of the joint null hypothesis that $\beta_2$, $\beta_3$ and $\beta_4$ are all zero versus the alternative that they each have standardized effects of .10. Figure 5 presents the power to detect a joint standardized change of .02 in each coefficient. Once again, the sample size needed to detect the joint change with a power of .90 is relatively large, nearly 9,000 for $\alpha = .01$.

We conclude from the above examples that it is by no means guaranteed that tests based on large sample surveys have "more than enough" power. While it is true that as our theories become more powerful and our models become more precise representatives of empirical processes, the increase

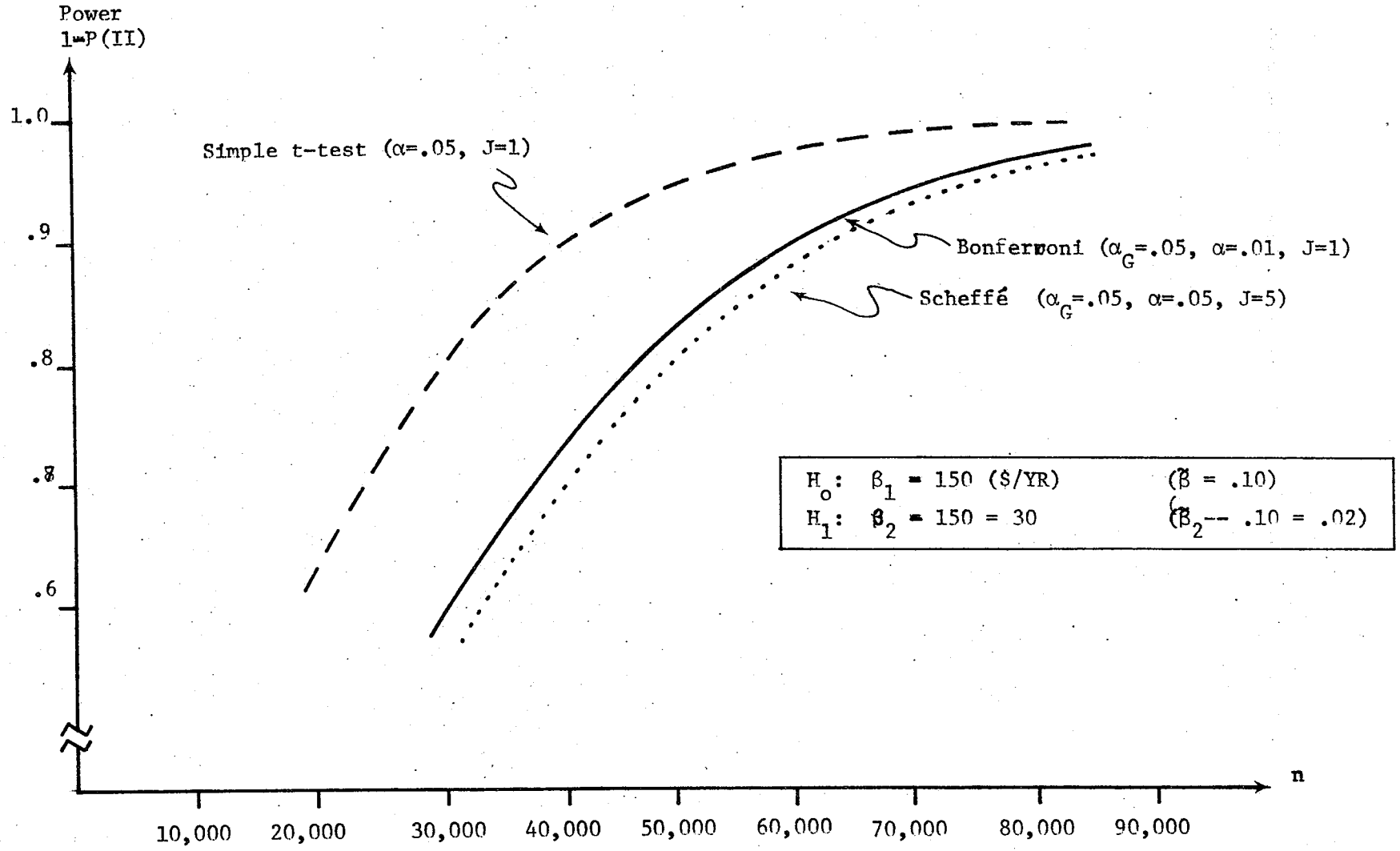


Figure 3: Income equation--Power as a function of sample size for the test on an individual coefficient

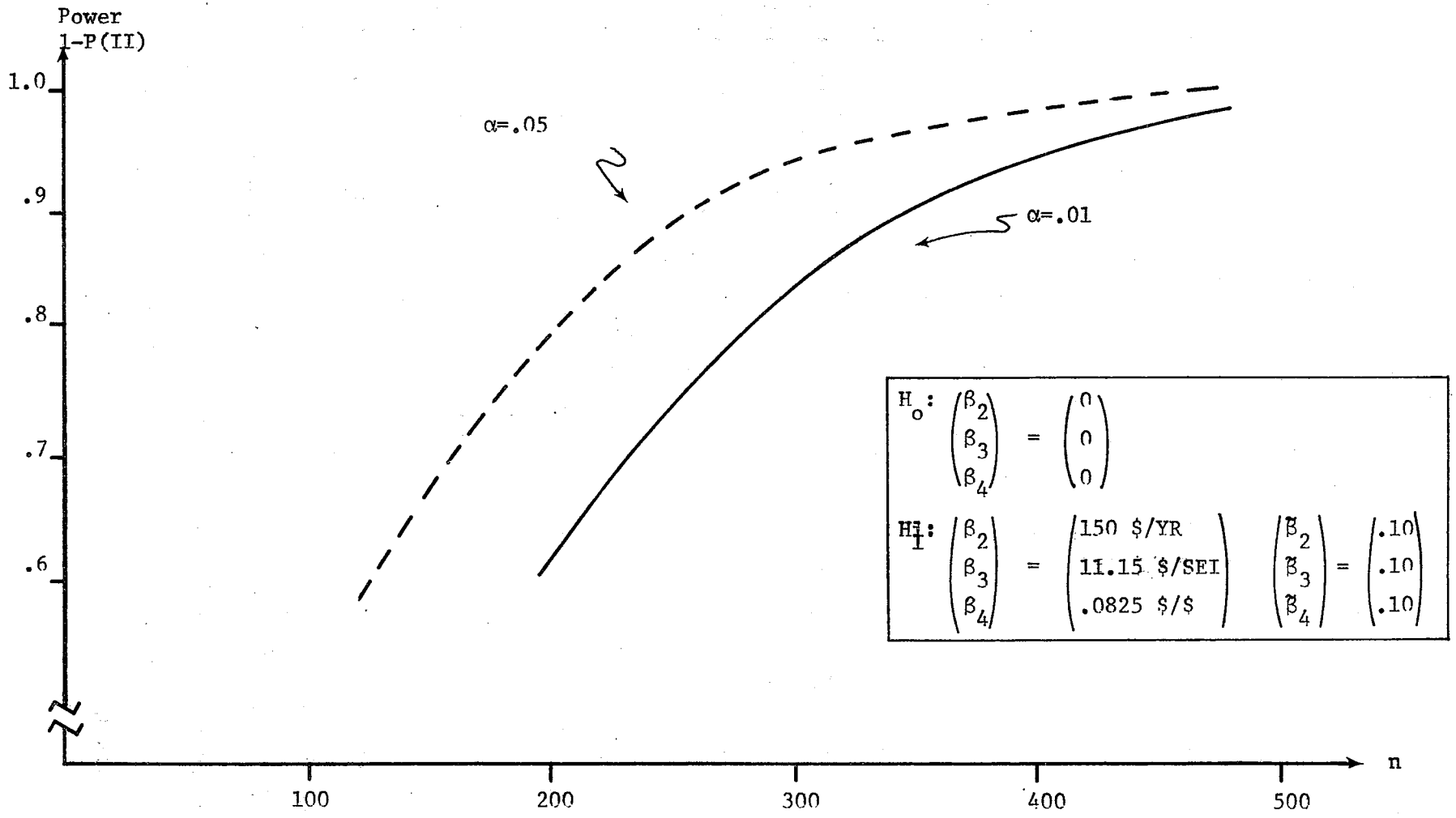


Figure 4: Income equation--Power as a function of sample size for a joint F-test on three coefficients

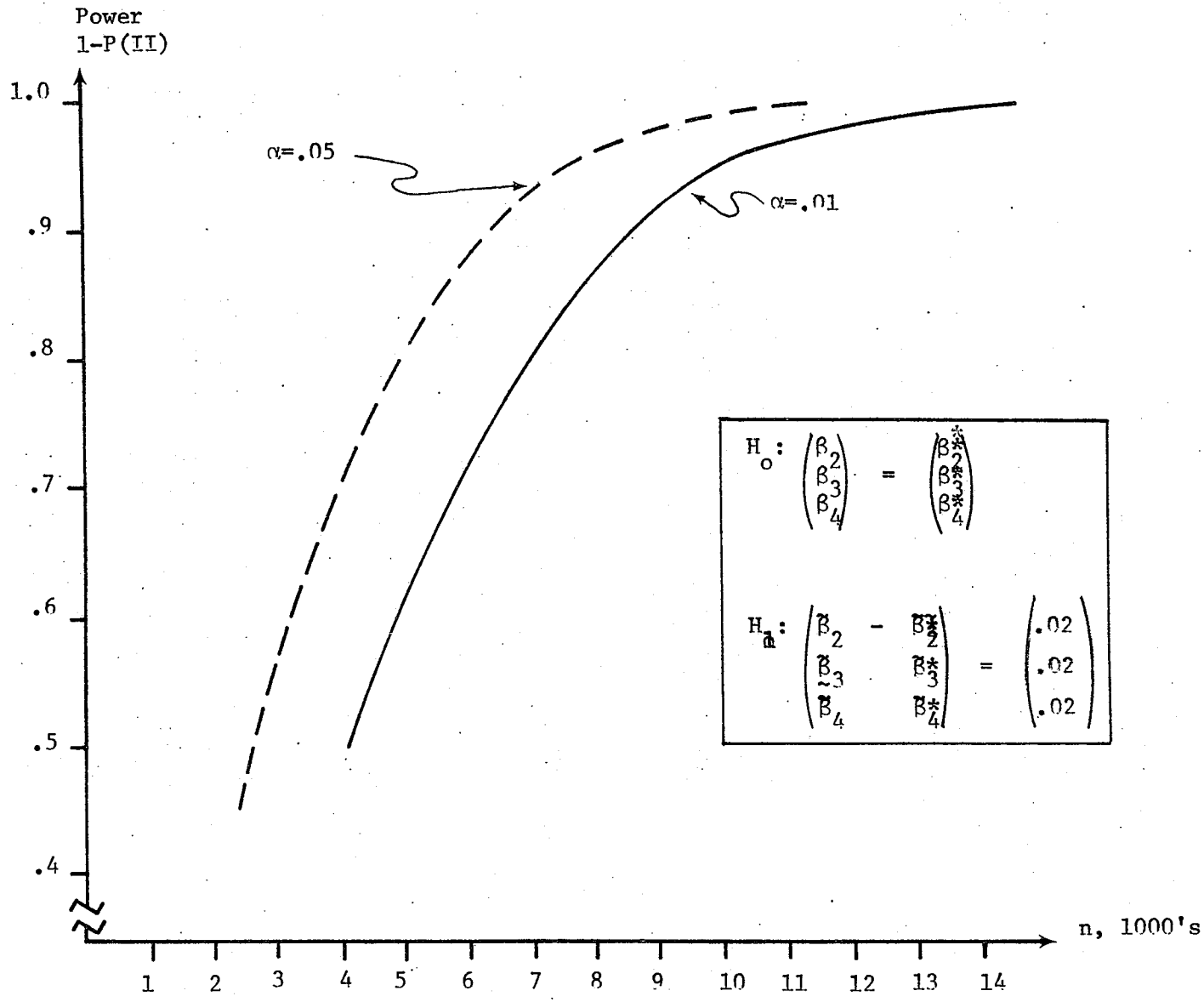


Figure 5: Income equation--Power as a function of sample size for a joint F-test on three coefficients

in the proportion of variance we can explain will unilaterally increase the power of our statistical tests, we will also become interested in detecting increasingly smaller effects. Indeed, in situations where we are interested in detecting change through replications of surveys, it is likely that we will wish to detect relatively small effects with little or no increase in the proportion of variance explained in the replication.

Analysts of survey data are most often confronted with a situation where the data have already been collected. Sample size and the configuration of independent variables are given, and the researcher wishes to test hypotheses of the parameters of a model on the given data set. In such a situation the relevant power calculation is power as a function of the degree to which the null hypothesis is false. From equation (8) and Table 2 we see that the only additional information needed to compute $\delta^2$ is a value for $\sigma^2_{y.x}$ or $\sigma^2_{y.x}/\sigma^2_y$. Should the researcher find that one or more tests are not powerful enough to detect substantively meaningful effects, two actions are possible. The researcher can increase $\alpha_G$, lowering the protection against making a Type I error. If this is unacceptable, the researcher must simply conclude that the data are inadequate for testing those particular hypotheses.

In Figure 6 we present power as a function of the degree to which the null hypothesis is false for our example of the Bonferroni test of the education coefficient in the income model. For fixed sample sizes from 250 to 10,000 observations, the power of the test of the hypothesis, $H_0 : \beta_2 = \beta_2^*$ is presented as a function of the absolute magnitude of the standardized measure of the degree to which the null hypothesis is false, $\left|\tilde{\beta}_2 - \tilde{\beta}_2^*\right|$. With a sample size of 250, we see that the standardized effect

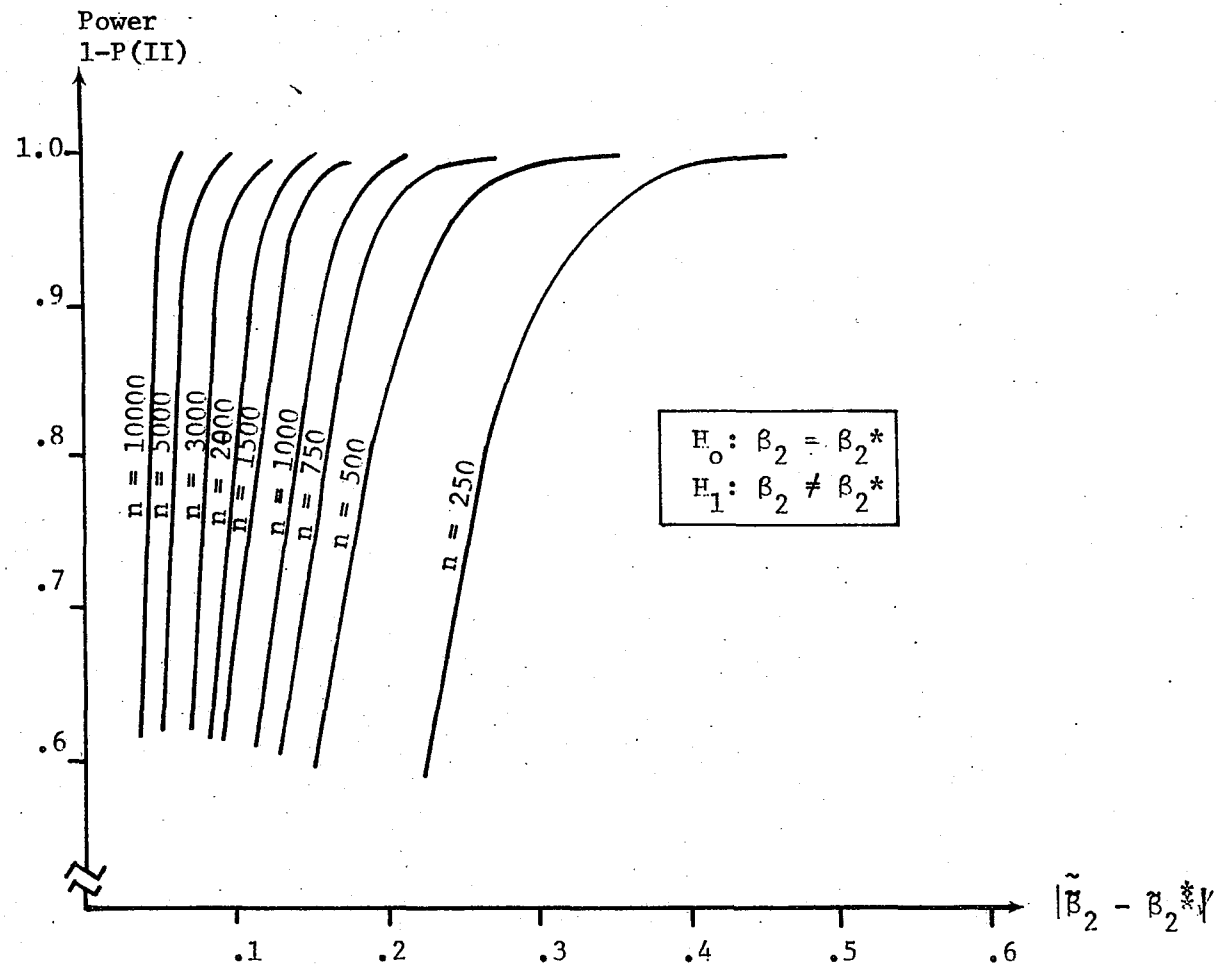


Figure 6: Income equation--Power of the Bonferroni test on an individual coefficient as a function of standardized effect to be detected. $\alpha_G = .05$, $\alpha = .01$.

would need to be as large as .30 to be detected with any regularity (power of .90), while for a sample of 10,000 observations, an effect as small as .07 can be detected with near certainty. To detect an effect of .15, a researcher with a sample of 250 observations would have to conclude that the data are inadequate, while a researcher with a sample of 10,000 would be in danger of finding "trivial" effects statistically significant and would perhaps decide to increase protection against Type I error substantially.

A single short Fortran computer program based on equation (8) has allowed us to compute all of the power calculations presented in this section. The logistics of these calculations are simple[14] and could be routinely incorporated into regression or GLM computer packages. If for no other reason than to force investigators to decide what effects in the population they would find substantively important, power considerations should be incorporated into our GLM hypothesis testing procedures. It is our hope that by incorporating power and simultaneous inference considerations into our hypothesis testing procedures we can narrow the gap between "statistical significance" and "substantive significance."

## V. Conclusion

Classical hypothesis testing as presented in most texts in a two-step procedure. One chooses a level of Type I error, $\alpha$, and compares the test statistic to a critical value based on that $\alpha$. After reviewing the neglected issues of simultaneous inference and power, we find

classical hypothesis testing inadequate for the purposes of social research. The intelligent use of statistical inference demands control over the overall level of Type I error and knowledge of the magnitude of effects one is likely to detect. Our examination of specific techniques for dealing with the power and simultaneous inference problems have led us to conclude that these techniques can be routinely incorporated into our procedures for the statistical analysis of survey data. Therefore, we suggest the following procedures precede the testing of GLM hypotheses:

1. Specify the hypotheses to be tested in terms of the parameters of the linear model.

2. Choose an acceptable Type I error rate, $\alpha_G$, for the group of hypotheses.

3. Select the appropriate test statistics, Bonferroni or Scheffé, which provide protection against Type I error at the $1 - \alpha_G$ level.

4. Compute the noncentrality parameter, $\delta^2$, and power of the tests as a function of the magnitude of the effects to be detected.

The above steps will provide information such that meaningful decisions can be made about the hypotheses being tested. This information may be used in survey design for the rational choice of sample size, or for assessing the adequacy of available data for hypothesis testing.

In recent years the use of statistical inference as a criteria in scientific decision-making has been the subject of increasing criticism (Morrison and Henkel, 1970). Part of this criticism has been directed at the failure of standard procedures of statistical inference to provide the kind of information required for meaningful scientific decision-making.

In spite of this criticism standard procedures of statistical inference continue to be employed. The continued use of the standard procedures can, at least in part, be attributed to the lack of a viable alternative when survey samples are analyzed. The use of our alternative procedure in our estimation, can contribute to a more informed use of statistical inference in scientific decision-making. It does so by requiring that the researcher give more attention to the goals of his research in the use of statistical inference as a scientific decision-making aid. The procedure we suggest requires that the researcher consider the purpose of the research in the selection of a meaningful unit of error rate. It also requires that the researcher give attention to the size of effect believed to be substantively significant in judging the adequacy of a given sample for decision-making purposes.

NOTES

[1]The choice of sample size in a survey design is always subject to cost constraints, and it is simply impossible to take into consideration the impact of sample size on all hypotheses which will be subsequently tested on the data. With multivariate sampling, it is impossible to fix a priori the variation in each independent variable and the covariation among the independent variables. Thus, even those analysts fortunate enough to be involved in survey design have only limited control over the power of their subsequent tests.

[2]In models in which an intercept is specified, the first column of $\underline{X}$ is a vector of ones, (1 1. . . 1)′; and the first element of the $\underline{\beta}$ vector is the intercept parameter. All models considered in this paper will have intercepts specified. Thus K - 1 rather than K is the number of independent variables.

[3]This requirement is stronger than that of uncorrelatedness of $\underline{\varepsilon}$ and $\underline{X}$; it is a weaker assumption than statistical independence of $\underline{\varepsilon}$ and $\underline{X}$.

[4]The null distributions of GLM test statistics do not depend upon the configuration of the $\underline{X}$ matrix, and consequently the conditional and unconditional Type I error rates are equivalent. The nonnull distributions of the GLM test statistics do depend upon the $\underline{X}$ matrix configuration (see the expression for the noncentrality parameter presented below).

By ignoring sampling variability in the $\underline{X}$ matrix , a source of variability in the nonnull distribution of the test statistics is being ignored. Consequently, the unconditional probability of Type II error is underestimated. Unfortunately, the unconditional nonnull test statistic distribution theory is quite complex, and incorporating it into our presentation would take us out of the context of the classical general linear model.

[5]The matrix expression for the t-test of hypothesis (3) is merely $\underline{a}' \underline{b} - \underline{a}' \underline{\beta}^*$ divided by the standard error of the linear combination $\underline{a}' \underline{b}$, where the standard error is $s^2 (\underline{a}'(\underline{X}'\underline{X})^{-1}\underline{a})$.

[6]Miller (1966) provides an exhaustive treatment of the statistical bases and applications of the many techniques of simultaneous statistical inference. A less exhaustive and more applications-oriented review is provided by Kirk (1968).

[7]The case of "theory trimming" described here is qualitatively different from the situation of testing an *a priori* hypothesized model. A number of procedures have been proposed for testing the fit for an *a priori* model where certain structural coefficients are hypothesized to be equal to zero (Land, 1973; McPherson and Huang, 1974; Specht, 1975). McPherson and Huang (1974) present an equation-by-equation scheme for testing the fit of an hypothesized recursive structural equation model that explicitly incorporates simultaneous inference considerations. If a single test of the global fit of an hypothesized model is performed, then one is effectively removed from the simultaneous inference case. If a comparison of the fit of several models is performed (cf. Specht,

1975), or if an attempt is made to diagnose what specific structural parameters are responsible for the failure of an hypothesized model to hold, then considerations raised in our discussion of simultaneous inference issues again become relevant.

[8]There are, of course, other techniques that are applicable to the common hypotheses tests on slope coefficients. For example, Williams (1972) discusses the application of Tukey's technique for making pairwise multiple comparisons of means within the regression framework. We restrict our attention to the Bonferroni and Scheffé techniques because of their wide generality and ease of application.

[9]Other factors such as departures from random sampling and measurement error affect both Type I and Type II error rates. Again, we have slighted important issues in order to remain within the context of the classical general linear model as it is most often applied in research by sociologists. Our point is not primarily that approximate error rate calculations are better than none. A more fundamental point is that the conceptualization of the appropriate unit of error rate and of meaningful effects to be detected will enhance our understanding of what our statistical analyses of survey data can and cannot tell us.

[10]Power tables approach an asymtope at about $N - K = 100$ denominator degrees of freedom. Since we are concerned with survey samples with generally many more than 100 observations, our calculations are based on tabled power for "infinite" denominator degrees of freedom.

[11]It must be assumed here that the hypotheses being tested are with respect to the unstandardized parameters and that the standardized parameters are merely an arbitrary rescaling of their unstandardized counterparts. The GLM distribution theory does not apply to the direct estimation of standardized parameters. The distributions of the standardized estimates and test statistics can become quite complex. The application of such distributions to direct statistical inference with respect to standardized parameters is virtually nonexistent in the social survey literature.

[12]An intermediate step is required to use these charts. They are presented in terms of a parameter $\Phi$ where $\Phi = \sqrt{\delta^2 / (J + 1)}$, where J is the numerator degrees of freedom of the test. For Scheffé projections, J is the numerator degrees of freedom from the preliminary joint test.

[13]Note also that our value for $\beta_2$ in 1960 was assumed to be based on a census and therefore not subject to sampling variability. If this were not the case, the power curves would be still lower.

[14]While the calculations are simple, the intermediate step of calculating the $\Phi$ parameter and finding power in the Pearson-Hartley charts can be annoying. It would be much more convenient if charts were available for power as a function of $\delta^2$ directly (as in Figure 1) for a number of $\alpha$ levels and numerator degrees of freedom. The Pearson-Hartley charts are further limited in that they have been tabulated only for $\alpha$ = .01, .05 and .10.

REFERENCES

Blau, Peter M., and Duncan, Otis Dudley

The American Occupational Structure. New York: John Wiley and Sons, Inc., 1967.

Dunn, Olive J.

"Confidence Intervals for Means of Dependent, Normally Distributed Variables." Journal of the American Statistical Association 54 (September 1959): 613-621.

Graybill, Franklin A.

An Introduction to Linear Statistical Models. New York: McGraw-Hill Book Company, Inc., 1961.

Heise, David R.

"Problems in Path Analysis and Causal Inference." in Sociological Methodology: 1969, edited by Edgar F. Borgatta, pp. 38-72. San Francisco: Jossey-Bass, 1969.

Kirk, Roger E.

Experimental Design Procedures for the Behavioral Sciences. Belmont, California: Brooks/Cole Publishing Company, 1968.

Land, Kenneth C.

"Identification, Parameter Estimation, and Hypothesis Testing in Recursive Sociological Models." in Structural Equation Models in the Social Science, edited by A.S. Goldberger and O.D. Duncan, pp. 19-49. New York: Seminar Press, 1973.

McPherson, J. Miller, and Huang, Cliff J.

"Hypothesis Testing in Path Models." Social Science Research 3 (June 1974): 127-139.

Miller, Rupert G., Jr.

Simultaneous Statistical Inference. New York: McGraw-Hill Book Company, 1966.

Morrison, Denton E., and Henkel, Ramon E.

The Significance Test Controversy. Chicago: Aldine Publishing Company, 1970.

Ryan, T.A.

"Multiple Comparisons in Psychological Research." Psychological Bulletin 56 (January 1959): 26-47.

Ryan, T.A.

"The Experiment as the Unit for Computing Rates of Error." Psychological Bulletin 59 (March 1962): 301-305.

Sampson, Allan R.

"A Tale of Two Regressions." Journal of the American Statistical Association 60 (September 1974): 682-689.

Scheffé, Henry.

The Analysis of Variance. New York: John Wiley and Sons, Inc., 1959.

Sewell, William H., and Hauser, Robert M.

"Causes and Consequences of Higher Education: Models of the Status Attainment Process." American Journal of Agricultural Economics 54 (December 1972): 851-861.

Specht, David A.

"On the Evaluation of Causal Models." Social Science Research 4 (June 1975): 113-133.

Theil, Henri.

Principles of Econometrics. New York: John Wiley and Sons, Inc., 1971.

Williams, John D.

"Multiple Comparisons in a Regression Approach." Psychological Reports 30 (March 1972): 639-647.

Wilson, Warner.

"A Note on the Inconsistency Inherent in the Necessity to Perform Multiple Comparisons." Psychological Bulletin 59 (March 1962): 296-300.