5

Data-Analytic Strategies Using Multiple Regression/Correlation

5.1 RESEARCH QUESTIONS ANSWERED BY CORRELATIONS AND THEIR SQUARES

Until this point we have presented regression/correlation analysis as if the typical investigation proceeded by selecting a single set of IVs and producing a single regression equation that is then used to summarize the findings. Life, however, is seldom so simple for the researcher. The coefficient or set of coefficients that provide the answers depend critically on the questions being asked. There is a wealth of information about the interrelationships among the variables not extractable from a single equation. It is, perhaps, the skill with which other pertinent information can be ferreted out that distinguishes the expert data analyst from the novice. In this chapter we address five major issues of strategy that should be considered in using MRC analysis. The first examines the fit between the research questions and the coefficients that answer them. The second examines some options and considerations for making regression coefficients more substantively interpretable. The third strategic consideration is the use of sequential or hierarchical analysis to wrest the best available answers from the data. The fourth is the employment of sets of independent variables in hierarchical analyses. The final section discusses strategies for controlling and balancing Type I and Type II errors of inference in MRC.

It is often the case that regression coefficients provide the most informative answers to scientific questions. However, there are a number of questions that are best answered by correlation coefficients and their comparisons. Indeed, it is sometimes hard to avoid the suspicion that correlation coefficients and squared correlations of various kinds are not reported or not focused on, even when most relevant, because they are typically so small. There is something rather discouraging about a major effort to study a variable that turns out to account uniquely for 1 or 2 percent of the dependent variable variance. We have tried to indicate that such a small value may still represent a material effect (see Section 2.10.1), but there is no getting around the more customary disparagement of effects of this magnitude.

Different questions are answered by different coefficients and comparisons among them. Standard statistical programs in MRC produce both regression and correlation coefficients for the use of scientists in interpreting their findings. All coefficients, but especially correlation coefficients, need a definable population to which to generalize, of which one has a random, or at least representative or unbiased sample. Without a population framework some coefficients may

be meaningless, and this is especially true of coefficients based (standardized) on the variance of the current sample. Researchers often make the untested assumption that the sample being examined is representative of some larger population. Assumptions about the stability and generalizability of model estimates from the sample at hand to the population of interest are probably more serious sources of bias than many other more familiar or researched statistical problems, such as variable distribution problems.

What kinds of hypotheses and research situations are best examined by comparisons of correlations, partial correlations, and semipartial correlations? The fundamental difference between questions involving correlations and those involving regression coefficients has often been overlooked because they share certain significance tests. Thus, the unique contribution to the prediction of Y, the squared semipartial correlation, shares the t test of statistical significance with the partial regression coefficients, both standardized and in raw units. Nevertheless, a focus on the magnitude and confidence intervals of these different coefficients will be informative to different research questions.

5.1.1 Net Contribution to Prediction

How much does X_i Increment R^2(sr^2)? How much improvement in prediction of Y is associated with the addition of X_i to other predictors? Here our interest is in the value of the squared semipartial correlation. Note that this is fundamentally a prediction question, rather than one of causal inference. For example, how much more predictable is academic success when we add health history to prior achievement and IQ? This is a question of utility rather than of causality. The latter would focus on the regression coefficient.

In addition to applied research that is designed to select variables on the basis of utility, correlation coefficients of various kinds may also answer scientific questions. We have indicated that questions about causal impact are generally best answered with regression coefficients. However, there are a number of scientific questions that are not causal in nature.

5.1.2 Indices of Differential Validity

Is measure A better than measure B? Perhaps some of the most frequently asked questions that are better answered by correlation functions than by regression coefficients involve evaluation of whether some measure or index is a better measure of a common construct than is an alternative, when the two measures are not measured in comparable units. For example, a study attempting to answer the question of whether child or parent is a better informant with regard to parental conflict by correlating scales appropriate to each informant with some criterion is best answered with correlations. Similarly, the question of which of two scales is better with regard to some criterion or outcome is better indexed by a comparison of correlations than by comparison of regression coefficients. For such questions the difference between correlations bounded by the confidence interval on this difference provides the best answer. Although the exact test is a complicated function of the average squared validity and the correlation between the predictors, in general, the standard error of this difference, expressed in Fisher's z', is less than $\sqrt{2/(n-3)}$ when all correlations are positive and the correlation between the predictors is comparable to their average validities (Meng, Rosenthal, & Rubin, 1992).

5.1.3 Comparisons of Predictive Utility

Is X_1 a better predictor than X_2? A question similar to that posed in the previous section may be answered by a comparison of semipartial correlations from a single sample: Which of two predictors is more related to Y net of other influences? For example, is maternal education

or paternal occupational status a better indicator of social class and thus more predictive of academic success, net of IQ? If the predictors were in the same units as, for example, a comparison of the influence of maternal education and paternal education, we might wish to compare Bs to answer the question of whether an additional year of mother's education or of father's education had a higher "payoff" in terms of offspring academic success. However, when the predictors to be compared are in different units, such Bs are not directly comparable. An alternative index would be a comparison of βs; methods for comparing which are described in Appendix 2.

5.1.4 Attribution of a Fraction of the *XY* Relationship to a Third Variable

How much of the XY relationship is accounted for by W? This question may be answered by comparing the zero-order r^2_{XY} with the semipartial $r^2_{YX \cdot W}$. Among other questions, this may represent a test of the mediation of the effect of X by W. This question is not the same as the one answered by a comparison of the zero-order B_{YX} with the partial $B_{YX \cdot W}$. The answer to the first question tells us how much of the prediction of Y by X is accounted for or attributable to W. The second asks whether, for constant values of W—that is, averaged over the values of W—changes in X have the same consequences for Y that they appeared to have when W was ignored. It is quite possible for W to account for a substantial fraction of the XY correlation and still have B_{YX} unchanged when W is partialed. For example, if $r_{XY} = .6$, $r_{YW} = .5$ and $r_{XW} = .83$, the unique contribution of X to the prediction of Y is [by Eq. (3.3.8) squared]

$$sr^2_{YX \cdot W} = \left(\frac{r_{YX} - r_{YW} r_{XW}}{\sqrt{1 - r^2_{XW}}} \right)^2 = \left(\frac{.6 - .5(.83)}{\sqrt{1 - .83^2}} \right)^2 = .11$$

as compared to .36 for the zero-order r^2_{YX}. In contrast,

$$\beta_{YX \cdot W} = \frac{r_{YX} - r_{YW} r_{XW}}{1 - r^2_{XW}} = \frac{.6 - .5(.83)}{1 - .83^2} = .6$$

Since B is equivalent to β with adjustment for standard deviations [Eq. (3.2.5)], it too will remain constant in this situation. Naturally confidence intervals and statistical power for all coefficients will be much affected by the r_{XW}.

5.1.5 Which of Two Variables Accounts for More of the *XY* Relationship?

Olkin and Finn (1995) present an example in which the question of interest is whether change in Y (adolescent drug use) over time ($=Y_2 \cdot Y_1$) is better explained by peer (V) or family (W) variables. Other examples might be whether family member differences are better explained by genetic closeness or period of time they lived together, or whether intergenerational differences on some variable are better explained by socioeconomic status or other differential experiences. Again, the question is posed in terms of variance proportions, and the answer is given by comparing the partial correlation of X with Y_2, partialing Y_1 to the partial correlation of W with Y_2, partialing Y_1.

5.1.6 Are the Various Squared Correlations in One Population Different From Those in Another Given the Same Variables?

This is a sort of popular "bottom line" question, that ignores (for the moment) the theoretically relevant details. Such questions can be asked with regard to zero-order r^2, partial or semipartial r^2, or R^2. Are social values as reflective of mental health in New Zealand as in American adolescents? Can we predict equally well in men and women, old persons and middle aged persons? Note that our overall prediction as assessed by R^2 can be equally good in two populations, although entirely different predictors in the set may be operative, or predictors may have quite different Bs.

5.2 RESEARCH QUESTIONS ANSWERED BY B OR β

5.2.1 Regression Coefficients as Reflections of Causal Effects

The (partial) regression coefficient B, as we have indicated, is often viewed as the premier causal indicator because it informs us about the estimated effect on the dependent variable of a change in the value of a putative cause.[1] Nevertheless, B coefficients as frequently presented have limitations, especially those associated with measurement units that are unfamiliar or that lack intrinsic meaning. These limitations are separable from the problems of equality of scale units, that is, whether the measures being employed can be considered to have the properties of interval scales. (See Cliff, 1982, for a discussion of methods of determining scale qualities and the inseparability of measurement from the scientific enterprise in general.)

As a consequence of a lack of consensus on measures, it is often easier to interpret β than B. β, as we have seen, essentially rescales the effects in terms of the standard deviations of the sample at hand. This is particularly useful when our research question has to do with comparing different variables for their (partialed) effects on Y in a given population represented by this sample. It is also often a necessary convenience when comparing effects of a given (conceptual) X_i on Y across studies, which may differ on the chosen measures of X_i and Y, and may even differ with regard to the population from which the sample was drawn. For example, we may measure depression with different scales in two studies, but wish to determine how similar their relationships are to some common Y. Consequently, it is generally recommended that β be reported, along with its SE in any research report.

5.2.2 Alternative Approaches to Making B_{YX} Substantively Meaningful

Clear interpretation of a (raw unit) B is often absent because the units of our measures are essentially arbitrary or are unfamiliar. Variable scores in the behavioral sciences often consist of simple sums of items, and statistics based on these scores will often be hard to interpret because of reader unfamiliarity with the number of items, the number of response options per item, and the numbers assigned to those response options in a given study. In such cases, even when major concerns about scale quality are absent, it is often difficult to determine anything useful about the meaning of a particular B other than, perhaps, the confidence that it differs

[1] Of course, use of B for these purposes does not, as such, imply that the researcher has provided the necessary theoretical argument that justifies such an interpretation.

from zero in a given direction. However, it is possible to provide data in ways that will enhance the interpretation and utility of the analyses.[2]

Units With Conventionally Shared Meaning

Some units are intrinsically meaningful (e.g., bushels, dollars, inches, years, and degrees, for which our familiarity with the units provides a framework that makes them generally understood). There are a few such cases in the behavioral sciences, in which long familiarity with a scale, and, usually, substantial standardization data has made relatively familiar units out of measures of abstract constructs. In psychology we have become familiar enough with the IQ unit to convey a sense of what are meaningful differences, even though the methods used to assess and calculate IQ vary. A long clinical tradition with the Hamilton Depression Scale has led to some consensus about the meaning and clinical significance of a difference or change of, for example, 5 units. The running example that we have used in this text was selected partly because its variables have such intrinsic meaning—dollars, years, publications, citations, and sex.

The fact that the units are agreed upon does not necessarily mean that a definition of a material difference is clear; rather the meaning is defined in context. Thus, for example, a difference of 4° Fahrenheit may be immaterial when one is cooking, of modest importance when one is deciding whether to go to the beach, and of considerable impact when one is measuring human body temperature (if within a certain range of values in the vicinity of 99°). The difficulty in accomplishing a conversion to centigrade (Celsius) measures of degrees in the United States is testimony to the importance of familiarity in context for ordinary interpretation of temperature units.

When the units (of both dependent and independent variable) are familiar in the context in which they are being used, the B is intrinsically useful. Thus, for example, if we say that the presence of some illness is associated with temperature increases of 5°, we understand its significance (although we would want to know its variability and duration as well as its mean).

Scales With a True Zero: None is None

Perhaps the typical scores that can be said to have true zeros are counts of events or objects. However, when such counts are meant to represent a more abstract construct or lack a familiar context that supplies meaning (e.g., the number of lever pecks a pigeon will complete as a function of the rate of reinforcement), the fact that they are counts is not enough to supply a useful meaning. No doubt experimenters who work in a given specialty area often develop a kind of consensus as to what are "material" effects, just as those working with biological assays do; however, those outside the area are likely to have to take their word for it, having no framework for understanding what is a lot and what is a little.

Under the circumstances in which both independent and dependent variables have true zeros, a useful conversion of B is to a measure of elasticity (E), used by economists to describe the percent change on a variable for which there is an available count (like dollars) but no upper limit (e.g., Goldberger, 1998). Elasticity is the percent change at the mean of Y for each 1% change in X.[3] Thus $B_{YX} = 1.36$ in our running example of faculty salary meant 1.36 publications per year since Ph.D., in relationship to an overall average of 19.93 publications, so there would be a $1.36/19.93 = 7\%$ increase at the mean in publications per year since Ph.D. The average faculty member is 6.79 years post Ph.D., so an increase in one year is a 1/6.79 or

[2] A more detailed presentation of these considerations and options can be found in P. Cohen, J. Cohen, Aiken, and West, 1999.

[3] The coefficient E is defined at the mean of the distribution; the regression is carried out on the raw variables.

14.7% increase. Dividing the 7% by 14.7% gives an elasticity of .476, or nearly a half percent increase in publications per 1% increase in time since Ph.D. (evaluated at the mean). Note that an advantage of this measure is that no upper limit is placed on the magnitude of either measure; an established range is not necessary to produce a meaningful effect size.

Scores for Which Both Zero and a Maximum Are Defined by the Measure

With many novel measures, a convenient unit is the percent of the maximum possible (POMP) score. School grades are often presented in this format, and conventions have developed that tend to lend differences in these units an intrinsic meaning (although again context matters, as one can easily see by comparing the meaning of a score of 80% in high school and in graduate school). The school grades example is also instructive because it makes it clear that the end points of the test do not necessarily define states of zero and complete knowledge of the subject matter. Furthermore, different instructors may devise tests of the same material that are seen as inappropriately easy, difficult, or appropriately discriminating among students.

There is nothing intrinsic about the percent reflecting "correctness," that is, it can equally be the percent of items endorsed in the scored direction regardless of the construct being scaled. Thus, if we have 20 true-false items measuring extroversion, each item endorsed in the extroversion direction would be worth 5 percentage points. Just as in educational tests of subject matter knowledge, a given test may be "hard" or "easy," and may allow for a range of scores that covers the potential range of the construct well or may restrict the range (qualities that may be inferred in part by the distribution of observed scores on the measure). These are qualities that are, in theory, as important as the reliability of the test. The reliability, of course, restricts both the correlation with other measures and the range of the observed scores (as can be seen by the fact that an individual with a true maximum or minimum score would have an observed score closer to the mean due to unreliability).

POMP scores need not be restricted to the number of dichotomous items endorsed in the scored direction. One may extend this procedure to use the full possible range of scores to define the zero and 100% points. For example, for a 10-item scale on which there are four Likert-scaled response options each, scored 0, 1, 2, 3, the potential range is from 0 to 30. A score of 12 would be 12/30 or 40%.

Using Item Response Alternatives as a Scale

An alternative when all item responses are on a Likert scale expressing degree of agreement or frequency is to use the average item score (with items reversed as necessary to conform to the scored direction). Then a unit is the difference between an average score of, for example, "disagree a little" and "agree a little," or "disagree a little" and "disagree a lot." This method is widely used in some research areas but rarely used in others.

When both the IV of interest and Y are treated in this manner, and they have been measured with the same number of response options, the B resulting from this transform will be exactly the same as that resulting from a percentaging of the same scores (POMP scores).

Using the Sample's Standard Deviation as a Scale; z Scores

The most common current method of placing scores in a more familiar unit is to subtract the score from the mean and divide by the standard deviation, thus creating z scores. As we have seen in Chapter 2, this method of scoring X and Y yields a bivariate B_{YX} that is equivalent to r_{XY}. This equivalence also holds between partial B and partial β when both Y and the IV in question have been z scored. Either this method of scoring or the use of partial β have the considerable

advantage of making the units on different variables comparable in one particular, useful sense. They enable us to say, for example, that X_1 had more influence on Y than X_2. However, as noted earlier, standardized measures and statistics are not without potential problems. In addition to potential disadvantages of failure to attend to a measure's units for the advancement of the scientific field, there is the problem that the *sd* unit will typically vary across populations as well as in different samples from a given population. Nevertheless, as things currently stand, β is often the most useful coefficient for answering questions about the influence of one variable on another, with or without other variables in the equation.

5.2.3 Are the Effects of a Set of Independent Variables on Two Different Outcomes in a Sample Different?

There are times when the issue at hand is whether the same IVs have essentially the same influence on two different dependent variables in the same sample. A "net regression" method of testing this issue, both as a single overall question and with regard to individual coefficients is presented in Appendix 2.[4] This method takes advantage of the fact that the predicted value of Y is the sum of the B-weighted IVs. Therefore one can test the difference between these weights, considered simultaneously, by using this $\hat{Y}$ value. Assuming that dependent variables W and Y are in the same units, or creating unit "equivalence" by standardizing them, we begin by estimating either (or each) from the set of IVs. We continue by subtracting $\hat{Y}$ from W and determining the relationship of the (same) IVs to this new $W-\hat{Y}$ variable as a dependent variable. If this equation is statistically significant, we have shown that the IVs are related significantly differently to dependent variable W than they are related to dependent variable Y. In addition, each regression coefficient in this new equation is tested by its *SE*, which indicates whether the influence (weight or B_i) of X_i on W is greater (positive) or less than (negative) its influence on Y. (The symmetrical analysis can be carried out on $Y - \hat{W}$, but the coefficients will simply be reversed in sign.)

5.2.4 What Are the Reciprocal Effects of Two Variables on One Another?

As noted earlier, we are usually forced to assume that Y does not cause X when we test our theories with cross-sectional (one-time-point) data. Unfortunately, in the social sciences this is quite often patently unlikely to hold true. For example, we may wish to examine the effect of stressful life events on adaptive function, but we are aware that poor adaptive function is likely to put one at risk of more stressful life events. Or we know that achievement is likely to be hampered by poor student attachment to the school but worry that such attachment may also be lowered by poor achievement.

A classic strategy for estimating such reciprocal effects from data collected at two points in time is called cross-lagged analysis.[5] In these analyses we require each of the two variables, W and Y, to be measured at two points in time. Then W measured at time 1 (W_1) is used to predict Y_2 in an equation that includes Y_1 as an IV, and Y_1 is used to predict W_2 in an equation that includes W_1 as an IV. Other control variables may be included as well, as appropriate to the substantive research issues. The resulting estimates are of the effect of each variable on (regressed) change in the other variable. As will be discussed in Chapter 12, the appropriateness of such estimates are highly dependent on the correct selection of a time between the two measurement occasions.

---

[4] An alternative method employing SEM is described in Chapter 12.

[5] An alternative approach that can sometimes be employed with cross-sectional data is described in Chapter 12.

5.3 HIERARCHICAL ANALYSIS VARIABLES IN MULTIPLE REGRESSION/CORRELATION

A sequential or hierarchical analysis of a set of independent variables may often produce the coefficients necessary to answer the scientific questions at hand.[6] In its simplest form, the k IVs are entered cumulatively in a prespecified sequence and the R^2 and partial regression and correlation coefficients are determined as each IV joins the others. A full hierarchical procedure for k IVs consists of a series of k regression analyses, each with one more variable than its predecessor. (In a subsequent section we see that one may proceed hierarchically with sets of variables rather than single variables.) The choice of a particular cumulative sequence of IVs is made in advance (in contrast with the *stepwise* regression procedures discussed in Section 5.3.3), dictated by the purpose and logic of the research. Some of the basic principles underlying the hierarchical order for entry are causal priority and the removal of confounding or spurious relationships, research relevance, and structural properties of the research factors being studied. As we will see in subsequent chapters, there are also circumstances in which a "tear down" procedure, in which one begins with the full set of variables and removes them selectively if they do not contribute to R^2, may be more in keeping with one's goals than the hierarchical "build up" procedure that we feature here.

5.3.1 Causal Priority and the Removal of Confounding Variables

As seen earlier (Section 3.4), the relationship between any variable and Y may be partly or entirely spurious, that is, due to one or more variables that are a cause of both. Thus, each variable in the investigation should be entered only after other variables that may be a source of spurious relationship have been entered. This leads to an ordering of the variables that reflects their presumed causal priority—no IV entering later should be a presumptive cause of an IV that has been entered earlier.[7]

One advantage of the hierarchical analysis of data is that once the order of the IVs has been specified, a unique partitioning of the total Y variance accounted for by the k IVs, $R^2_{Y.123\ldots k}$, may be made. Indeed, this is the *only* basis on which variance partitioning can proceed with correlated IVs. Because the sr_i^2 at each stage is the increase in R^2 associated with X_i, when all (and only) previously entered variables have been partialed, an *ordered* variance partitioning procedure is made possible by

$$\begin{aligned} R^2_{Y.123\ldots k} &= r^2_{Y1} + r^2_{Y2\cdot 1} + r^2_{Y3\cdot 12} + r^2_{Y4\cdot 123} + \ldots + r^2_{Yk\cdot 123\ldots(k-1)} \\ &= r^2_{Y1} + sr^2_{2\cdot 1} + sr^2_{3\cdot 12} + sr^2_{4\cdot 123} + \ldots + sr^2_{k\cdot 123\ldots(k-1)}. \end{aligned} \tag{5.3.1}$$

Each of the k terms is found from a simultaneous analysis of IVs in the equation at that point in the hierarchy; each gives the increase in Y variance accounted for by the IV entering at that point beyond what has been accounted for by the previously entered IVs. r^2_{Y1} may be thought of as the increment from zero due to the first variable in the hierarchy, an sr^2 with nothing

---

[6] We regret the confusion that sometimes occurs between this older reference to variables entered in a hierarchical sequence with a more recent development of hierarchical linear models (HLM) or hierarchical regression, which refers to a structure of the data in which subject scores are nested within occasions or within some other grouping (e.g., classrooms or families) that tends to prevent independence of observations. The latter procedure is discussed in Chapters 14 and 15.

[7] When a variable X_j that may be an effect of X_i is entered prior to or simultaneously with X_i we have the circumstance referred to by epidemiologists as *overcontrol*, that is, removal from the estimated effect of X_i on Y of some fraction that is mediated or indirect by way of X_j.

partialed. Summing the terms up to a given stage in the hierarchy gives the cumulative R^2 at that stage; for example, $r^2_{Y1} + sr^2_{2.1} + sr^2_{3.12} = R^2_{Y.123}$.

The reader is reminded that the increment attributable to any IV may change considerably if one changes its position in the hierarchy, because this will change what has and what has not been partialed from it. This is indeed why one wishes the IVs to be ordered in terms of the specific questions to be answered by the research, such as causal priority. Otherwise part of the variance in Y due to some cause X_1 is instead attributed to an IV that is another effect of this cause. This stolen (spurious) variance will then mislead the investigator about the relative importance to Y of the cause and its effect.

Of course, it will frequently not be possible to posit a single sequence that is uncontroversially in exact order of causal priority.[8] In such circumstances more than one order may be entertained and the results then considered together. They may not differ with regard to the issue under investigation, but if they do, the resulting ambiguity must be acknowledged.

When the variables can be fully sequenced—that is, when a full causal model can be specified that does not include any effect of Y on IVs or unmeasured common causes, the hierarchical procedure becomes a tool for estimating the effects associated with each cause. Formal causal models use regression coefficients rather than variance proportions to indicate the magnitude of causal effects, as discussed earlier. Because Chapter 12 is devoted to an exposition of the techniques associated with this and other types of causal models, we do not describe them here.

To illustrate a hierarchical analysis, organized in terms of causal priority, we turn again to the academic salary data (from Table 3.5.1). The order of assumed priority is sex (X_3), time (years) since Ph.D. (X_1), number of publications (X_2), and number of citations (X_4) (but note the further discussion of this sequence in Section 12.1). Note that no variable can be affected by one that appears after it; whatever causality occurs among IVs is assumed to be from earlier to later in the sequence. We entered these variables in the specified order and determined the R^2 after each addition. We found $r_{Y3} = .201$, and therefore $R^2_{Y.3}(= r^2_{Y3}) = .040$, that is, 4% of the academic salary variance was accounted for by sex. When time since Ph.D. (X_1) was added to sex, we found that $R^2_{Y.31} = .375$ and we may say that the increment in predicted Y variance of time since Ph.D. over sex, or for time partialing or taking into account the difference in time since Ph.D. between male and female faculty, was $sr^2_{1\cdot3} = R^2_{Y.13} - r^2_{Y3} = .375 - .040 = .335$. Next we added publications (X_2) and found $R^2_{Y.312} = .396$, a very small increment: $sr^2_{2\cdot31} = R^2_{Y.123} - R^2_{Y.31} = .396 - .375 = .021$. Finally, when citations (X_4) was added, we have the R^2 we found in Section 3.5. $R^2_{Y.3124} = .503$, so the increment for X_4 or $sr^2_{4\cdot123} = .503 - .396 = .107$. The final R^2 for the four IVs is necessarily the sum of these increments, by Eq. (3.8.1):

$$.503 = .040 + .335 + .021 + .107.$$

Of course, a different ordering would result in different increments (which would also sum to .503), but to the extent that they violated the direction of causal flow, the findings might be subject to misinterpretation. For example, if entered first, publications would have all of the salary variance associated with its r^2 credited to it, but only on the unlikely premise that time since Ph.D. (or the forces associated with time) did not contribute causally to the number of publications. The causal priority ordering makes it clear that (in these fictitious data) the strong relationship between salary and publications merely reflects the operation of the passage of time.

---

[8]Not infrequently one can identify one or more variables that are thought of as "controls," meaning that although their causal role is not certain, removal of their potential influence will strengthen the inferences that can be made about the role of one or more IVs that will be entered later in the sequence. See *Functional Sets* in Section 5.4.1.

The increments here are sr^2 values, but they are different from those determined previously (Section 3.5.4) and given in Table 3.5.2. For the latter, all the other $k-1(=3)$ IVs were partialed from each, whereas here only those preceding each IV in the hierarchy were partialed. They therefore agree only for the variable entering last. When the significance test of Eq. (3.6.8) is employed for these cumulative sr^2 values (for $n = 62$), it is found that all are significant except the sr^2 for number of publications.[9]

A special case of the hierarchical model is employed in the analysis of change. Under circumstances in which pre- and postscore values are available on some variable and the researcher wishes to determine whether and to what extent treatment or other variables are associated with change, the postscore may be used as the dependent variable, with prescore entered as the first IV in the hierarchy. Unlike the alternative method involving differences (postscores minus prescores), when subsequent IVs are entered into the equation their partial correlations will reflect their relationship with postscores from which prescore influence has been removed. Note that this method effectively removes from consideration any influence that earlier values of the DV may have had on other IVs.[10]

5.3.2 Research Relevance

Not infrequently an investigator gathers data on a number of variables in addition to those IVs that reflect the major goals of the research. Thus, X_1 and X_2 may carry the primary focus of the study but X_3, X_4, and X_5 are also available. The additional IVs may be secondary because they are viewed as having lesser relevance to the dependent variable than do X_1 and X_2, or because hypotheses about their relationships are weak or exploratory. Under these circumstances, X_1 and X_2 may be entered into the equation first (perhaps ordered on the basis of a causal model) and then X_3, X_4, and X_5 may follow, ordered on the basis of their presumed relevance and/or priority. Aside from the clarity in interpretations of the influence of X_1 and X_2 that is likely to result from this approach (because the secondary X_3, X_4, and X_5 variables are not partialed from X_1 and X_2), the statistical power of the test of the major hypothesis is likely to be maximal because the df are not deflated by these less important variables. Under these circumstances, the additional steps answer the question of whether these variables add anything to the prediction of Y.

5.3.3 Examination of Alternative Hierarchical Sequences of Independent Variable Sets

Sometimes the appropriate sequencing of some variable sets is theoretically ambiguous. Although one usually begins with the more distal causes and gradually adds the more proximal causes that may mediate those distal causes, there are times when theory is inadequate to determine such a sequence, when it is likely that there are effects of these IV sets on each other, or when sets are alternative mediators of an unmeasured more distal cause. Such a circumstance might occur, for example, if one set of IVs involved a set of physiological measures and another set consisted of a set of motivational variables, and the study was examining the impact of each on behavior. In such cases the addition of each set to the prediction of Y, over and above the prediction of the other set, would be of interest, and both sequences would usually be reported.

---

[9] An alternative test of significance, in which "Model 2" error is employed, is discussed in Section 5.5.4 in the context of significance tests for sets of IVs entered hierarchically.

[10] This procedure is discussed in greater detail in Chapter 15, where analyses of longitudinal data are more thoroughly reviewed.

5.3.4 Stepwise Regression

There are dangers in letting a computer program sequence the variables for you as happens when one uses the stepwise option, in which variables are selected on the basis of their contribution to R^2. Although stepwise regression has certain surface similarities with hierarchical MRC (and hierarchical MRC and stepwise regression are the same thing when the investigator "forces" the sequencing of the IVs), it is considered separately, primarily because it differs in its underlying philosophy. Stepwise programs are designed to select from a group of IVs the one variable at each stage that has the largest sr^2 and hence makes the largest contribution to R^2. Such programs typically stop admitting IVs into the equation when no IV makes a contribution that is statistically significant at a level specified by the program user.[11] Thus, the stepwise procedure defines an a posteriori order based solely on the relative uniqueness of the variables in the sample at hand.

When an investigator has a large pool of potential IVs and very little theory to guide selection among them, these programs are a sore temptation. If the computer selects the variables, the investigator is relieved of the responsibility of making decisions about their logical or causal priority or relevance before the analysis, although interpretation of the findings may not be made easier. We take a dim view of the routine use of stepwise regression in explanatory research for various reasons (see the following), but mostly because we feel that more orderly advance in the behavioral sciences is likely to occur when researchers armed with theories provide an a priori ordering that reflects causal hypotheses rather than when computers order IVs post and ad hoc for a given sample.

An option that is available on some computer programs allows for the a priori specification ("forcing") of a hierarchy among groups of IVs. An investigator may be clear that some groups of variables are logically, causally, or structurally prior to others, and yet not have a basis for ordering variables within such groups. Under such conditions, variables may be labeled for entering in the equation as one of the first, second, or up to hth group of variables. The sequence of variables within each group is determined by the computer in the usual stepwise manner. This type of analysis is likely to be primarily hierarchical (between classes of IVs) and only incidentally stepwise (within classes), and computer programs so organized may be effectively used to accomplish hierarchical MRC analysis by sets of IVs as described in Section 5.4.4.

Probably the most serious problem in the use of stepwise regression programs arises when a relatively large number of IVs is used. Because the significance tests of each IV's contribution to R^2 and associated confidence intervals proceed in ignorance of the large number of other competing IVs, there can be very serious capitalization on chance and underestimation of confidence intervals. A related problem with the free use of stepwise regression is that in many research problems the ad hoc order produced from a set of IVs in one sample is likely not to be found in other samples from the same population. When among the variables competing for entry at any given step there are trivial differences among their partial relationships with Y, the computer will dutifully choose the largest for addition at that step. In other samples and, more important, in the population, such differences may well be reversed. When the competing IVs are substantially correlated with each other, the problem is likely to be compounded, because the losers in the competition may not make a sufficiently large unique contribution to be entered at any subsequent step before the problem is terminated by the absence of a variable making a statistically significant addition.

---

[11] Some stepwise programs operate backward, that is, by elimination. All k IVs are entered simultaneously and the one making the smallest contribution is dropped. Then the $k - 1$ remaining variables are regressed on Y, and again the one making the smallest contribution is dropped, and so on. The output is given in reverse order of elimination. This order need not agree with that of the forward or accretion method described here.

Although, in general, stepwise programs are designed to approach the maximum R^2 with a minimum number of IVs for the sample at hand, they may not succeed very well in practice. Sometimes, with a large number of IVs, variables that were entered into the equation early no longer have nontrivial relationships after other variables have been added. Some programs provide for the removal of such variables, but others do not. Also, although it is admittedly not a common phenomenon in practice, when there is suppression between two variables, neither may reach the criterion for entrance to the equation, although if both were entered they would make a useful contribution to R^2.

However, our distrust of stepwise regression is not absolute, and decreases to the extent that the following conditions obtain:

1. The research goal is entirely or primarily predictive (technological) and not at all, or only secondarily, explanatory (scientific). The substantive interpretation of stepwise results is made particularly difficult by the problems described earlier.
2. n is very large, and the original k (that is, before stepwise selection) is not too large; a k/n ratio of 1 to at least 40 is prudent.
3. Particularly if the results are to be substantively interpreted, a cross-validation of the stepwise analysis in a new sample should be undertaken and only those conclusions that hold for both samples should be drawn. Alternatively, the original (large) sample may be randomly divided in half and the two half-samples treated in this manner.

5.4 THE ANALYSIS OF SETS OF INDEPENDENT VARIABLES

A set of variables is a group classified as belonging together for some reason. As we will describe them, the grouping of variables into sets may be motivated by structural or formal properties of the variables that the sets include or the sets may have a common functional role in the substantive logic of the research. The basic concepts of proportion of variance accounted for and of correlation (simple, partial, semipartial, multiple) developed in Chapter 3 for single IVs hold as well for sets of IVs. This use of sets as units of analysis in MRC is a powerful tool for the exploitation of data.

5.4.1 Types of Sets

Structural Sets

We use the term *research factor* to identify an influence operating on Y or, more generally, an entity whose relationship to Y is under study. The word *factor* is used here to imply a single construct that may require two or more variables for its complete representation. Such will be the case when the construct consists of multiple independent or overlapping categories, or when more than one variable is required to express the shape of the relationship between a quantitative scale and the dependent variable. Examples of categorical variables are experimental treatment, religion, diagnosis, ethnicity, kinship system, and geographic area. In general it will require $g-1$ variables to represent g groups or categories (see Chapter 8). When they are not mutually exclusive, g variables will generally be required. Thus, in a laboratory experiment in which subjects are randomly assigned to three different experimental groups and two different control groups (hence, $g = 5$), the research factor $\boldsymbol{G}$ of treatment group requires exactly $g-1 = 4$ IVs to fully represent the aspects of $\boldsymbol{G}$, that is, the distinctions among the 5 treatment groups. The several different methods for accomplishing this representation are the subject of Chapter 8, but in each case $g-1$ variables are required to fully represent $\boldsymbol{G}$.

Quantitative scales, such as scores of psychological tests, rating scales, or sociological indices, may also require multiple variables to fully represent them. When one needs to take into account the possibility that a research factor such as age may be related curvilinearly to Y (or to other research factors), other aspects of age must be considered. Age as such represents only one aspect of the A research factor, its linear aspect. Other aspects, which provide for various kinds of nonlinearity in the relationship of A to Y, may be represented by other IVs such as age squared and age cubed. Considerations and methods of representing aspects of scaled research factors are the subject of Chapter 6.

The preceding implies that if we are determining the proportion of variance in a dependent variable Y due to a single research factor, we will (in general) be finding a squared *multiple* correlation, because the latter will require a set of two or more IVs.

Functional Sets

Quite apart from structural considerations, IVs are grouped into sets for reasons of their substantive content and the function they play in the logic of the research. Thus, if you are studying the relationship between the psychological variable field dependence (Y) and personality (P) and ability (A) characteristics, P may contain a set of k_P scales from a personality questionnaire and A a set of k_A subtests from an intelligence scale. The question of the relative importance of personality and ability (as represented by these variables) in accounting for Y would be assessed by determining R^2_{YP}, the squared multiple correlation of Y with the k_P IVs of P, and R^2_{YA}, the squared multiple correlation of Y with k_A IVs of A, and then comparing them. Similarly, research that is investigating (among other things) the socioeconomic status (S) of school children might represent S by occupational status of head of household, family income, mother's education, and father's education, a substantive set of four ($=k_S$) IVs. For simplicity, these illustrations have been of sets of single research factors, but a functional set can be made up of research factors that are themselves sets. For example, a demographic set (D) may be made up of structural sets to represent ethnicity, marital status, and age. A group of sets is itself a set and requires no special treatment.

It is often the nature of research that in order to determine the effect of some research factor(s) of interest (a set B), it is necessary to statistically control for (partial out) the Y variance due to causally antecedent variables in the cases under study. A group of variables deemed antecedent either temporally or logically in terms of the purpose of the research could be treated as a functional set for the purpose of partialing out of Y's total variance the portion of the variance due to these antecedent conditions. Thus, in a comparative evaluation of compensatory early education programs (B), with school achievement as Y, the set to be partialed might include such factors as family socioeconomic status, ethnicity, number of older siblings, and pre-experimental reading readiness. This large and diverse group of IVs functions as a single covariate set A in the research described. In research with other goals these IVs might have different functions and be treated separately or in other combinations.

An admonitory word is in order. Because it is possible to do so, the temptation exists to assure coverage of a theoretical construct by measuring it in many ways, with the resulting large number of IVs then constituted as a set. Such practice is to be strongly discouraged, because it tends to result in reduced statistical power and precision for the sets and an increase in spuriously "significant" single-IV results, and generally bespeaks muddy thinking. It is far better to sharply reduce the size of such a set, and by almost any means.[12] One way is

[12]See the discussion of multicollinearity in Section 3.8 and in later chapters, especially Chapter 10.

through a tightened conceptualization of the construct, a priori. In other situations, the large array of measures is understood to cover only a few behavioral dimensions, in which case their reduction to scores on one or more factors by means of factor analysis or latent variable modeling (Chapter 12) is likely to be most salutary for the investigation, with little risk of losing Y-relevant information. Note that such analyses are performed completely independent of the sample values of the r_{Yi} correlations.

It is worth noting here that the organization of IVs into sets of whatever kind bears on the interpretation of MRC results but has no effect on the basic computation. For any Y and k IVs $(X_1, X_2, \ldots, X_k)$ in a given analysis, each X_i's coefficients (sr_i, pr_i, B_i, β_i) and their associated confidence intervals and significance tests are determined as described in Chapter 3.

5.4.2 The Simultaneous and Hierarchical Analyses of Sets

We saw in Chapter 3 that, given k IVs, we can regress Y on all of them simultaneously and obtain $R^2_{Y.12\ldots k}$ as well as partial statistics for each X_i. We will generally write these partial statistics in shorthand notation (i.e., β_i, B_i, sr_i, pr_i), where it is understood that all the other IVs are being partialed. This immediately generalizes to sets of IVs: when sets $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$ are simultaneously regressed on Y, there are a total of $k_U + k_V + k_W = k$ IVs that together determine $R^2_{Y.UVW}$. The partial statistics for each IV in the set $\boldsymbol{U}$ has *all* the remaining $k - 1$ IVs partialed: both those from $\boldsymbol{V}$ and $\boldsymbol{W}$ (numbering k_V and k_W) and also the remaining IVs from its own set. It can be shown that, for example, the adjusted Y means of ANCOVA (analysis of covariance) are functions of the regression coefficients when a covariate set and a set of groups are simultaneously regressed on Y.

In Section 5.3 we saw that each of the k IVs can be entered cumulatively in some specified hierarchy, at each stage of which an R^2 is determined. The R^2 for all k variables can thus be analyzed into cumulative increments in the proportion of Y variance due to the addition of each IV to those higher in the hierarchy. These increments in R^2 were noted to be squared semipartial correlation coefficients, and the formula for the hierarchical procedure for single IVs was given as

$$\textbf{(5.3.1)} \qquad R^2_{Y.12\ldots k} = r^2_{Y1} + r^2_{Y2\cdot 1} + r^2_{Y3\cdot 12} + r^2_{Y4\cdot 123} + \ldots + r_{Yk\cdot 123\ldots(k-1)}$$

The hierarchical procedure is directly generalizable from single IVs to sets of IVs. Replacing k single IVs by h sets of IVs, we can state that these h sets can be entered cumulatively in a specified hierarchical order, and upon the addition of each new set an R^2 is determined. The R^2 for all h sets can thus be analyzed into increments in the proportion of Y variance due to the addition of each new set of IVs to those higher in the hierarchy. These increments in R^2 are, in fact, squared *multiple* semipartial correlation coefficients, and a general hierarchical equation for sets analogous to Eq. (5.3.1) may be written. To avoid awkwardness of notation, we write it for four ($= h$) sets in alphabetical hierarchical order and use the full dot notation; its generalization to any number of sets is intuitively obvious:

$$\textbf{(5.4.1)} \qquad R^2_{Y.TUVW} = R^2_{YT} + R^2_{Y(U\cdot T)} + R^2_{Y(V\cdot TU)} + R^2_{Y(W\cdot TUV)}.$$

We defer a detailed discussion of the multiple semipartial R^2 to the next section. Here it is sufficient to note merely that it is an increment to the proportion of Y variance accounted for by a given set of IVs (of whatever nature) beyond what has already been accounted for by prior sets, that is, sets previously entered in the hierarchy. Further, the amount of the increment in Y variance accounted for by that set cannot be influenced by Y variance associated with subsequent sets; that is, those which are later in the hierarchy.

Consider an investigation of length of hospital stay (Y) of $n = 500$ randomly selected psychiatric admissions to eight mental hospitals in a state system for a given period. Assume that data are gathered and organized to make up the following sets of IVs:

1. Set $\boldsymbol{D}$—Demographic characteristics of patients: age, sex, socioeconomic status, ethnicity. Note that this is a substantive set, and may be thought of as a set of control variables, meaning that they are not themselves of major interest but are there to make sure that effects attributed to later sets are not really due to demographic differences with which they are correlated. Assume $k_D = 9$.
2. Set $\boldsymbol{I}$—Patient illness scores on nine of the scales of the Minnesota Multiphasic Personality Inventory. This set is also substantive, and $k_I = 9$. This set is placed prior to the information on which hospital the patient has been treated in because it is known that patient illness enters into the decision about which hospital will treat them.
3. Set $\boldsymbol{H}$—Hospitals. The hospital to which each patient is admitted is a nominally scaled research factor. With eight hospitals contributing data, we will require a (structural) set of $k_H = 7$ IVs to represent fully the hospital group membership of the patients (see Chapter 8).

Although there are 25 ($k_D + k_I + k_H = k$) IVs, our analysis may proceed in terms of the three ($= h$) sets hierarchically ordered in the assumed causal priority of accounting for variance in length of hospital stay as $\boldsymbol{D,I,H}$.

Suppose that we find that $R^2_{YD} = .20$, indicating that the demographic set, made up of nine IVs accounts for 20% of the Y variance. Note that this ignores any association with illness scores ($\boldsymbol{I}$) or effects of hospital differences ($\boldsymbol{H}$). When we add the IVs of the $\boldsymbol{I}$ set, we find that $R^2_{Y.DI} = .22$; hence, the increment due to $\boldsymbol{I}$ over and above $\boldsymbol{D}$, or with $\boldsymbol{D}$ partialed, is $R^2_{Y(I \cdot D)} = .02$. Thus, an additional 2% of the Y variance is accounted for by illness beyond the demographic set. Finally, the addition of the seven IVs for hospitals (set $\boldsymbol{H}$) produces an $R^2_{Y.DIH} = .33$, an increment over $R^2_{Y.DI}$ of .11, which equals $R^2_{Y(H \cdot DI)}$. Thus, we can say that which hospital patients enter accounts for 11% of the variance in length of stay, after we partial out (or statistically control, or adjust for, or hold constant) the effect of differences in patients' demographic and illness characteristics. We have, in fact, performed by MRC an analysis of covariance (ANCOVA) for the research factor "hospitals," using sets $\boldsymbol{D}$ and $\boldsymbol{I}$ as covariates.[13] Of course, one's substantive interest is likely to focus on the actual B_{Yi} coefficients as each new set is entered. To the extent to which one is interested in the final "adjusted" mean differences in LOS between hospitals, the answer will lie in the final regression coefficients. However, much can also be learned by examination of the extent to which these coefficients differ from those obtained when set $\boldsymbol{H}$ is entered *without* the "covariate" sets $\boldsymbol{D}$ and $\boldsymbol{I}$. For example, it may be that some initially large differences between hospitals were entirely attributable to demographic and symptom differences between patients. This could be concluded when certain B_{Yi} coefficients from the equation with only set $\boldsymbol{H}$ were what could be considered large in the context (and had acceptably narrow confidence limits), but declined substantially in value when the covariate set $\boldsymbol{D}$, or sets $\boldsymbol{D}$ and $\boldsymbol{I}$, were added.

There is much to be said about the hierarchical procedure and, indeed, it is said in the next section and throughout the book. For example, as pointed out in regard to single variables, the increment due to a set may depend critically upon where it appears in the hierarchy; that is, what has been partialed from it, which, in turn, depends on the investigator's theory about the mechanisms that have generated the associations between the variables. As we will see, not all theories permit unambiguous sequencing.

---

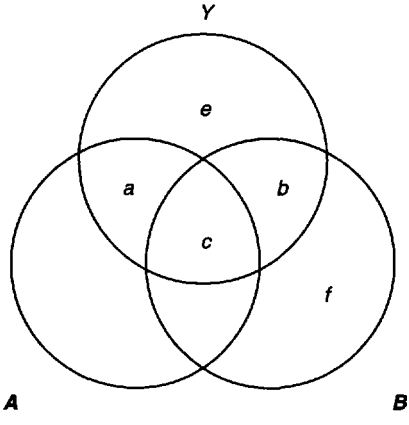
[13] Omitting the significance test; see subsequent section. Also, a valid ANCOVA requires that there be no interaction between $\boldsymbol{H}$ and the aggregate $\boldsymbol{I, D}$ covariate set (see Chapter 9).

5.4.3 Variance Proportions for Sets and the Ballantine Again

We again employ the ballantine to illustrate the structure of relationships of sets of IVs to a dependent variable Y. It was presented in Fig. 3.3.1 for single IVs X_1 and X_2, and we present it as Fig. 5.4.1 here for sets $\boldsymbol{A}$ and $\boldsymbol{B}$. It is changed in no essential regard, and we show how the relationships of sets of IVs to Y, expressed as proportions of Y variance, are directly analogous to similarly expressed relationships of single IVs.

A circle in a ballantine represents the total variance of a variable, and the overlap of two such circles represents shared variance or squared correlation. This seems reasonable enough for single variables, but what does it mean when we attach the set designation $\boldsymbol{A}$ to such a circle? What does the variance of a set of multiple variables mean? Although each of the k_A variables has its own variance, remember that a multiple $R^2_{Y.12\ldots k}$ is in fact a simple r^2 between Y and $\hat{Y}$, the latter optimally estimated from the regression equation of the k_A IVs that make up set $\boldsymbol{A}$ (and similarly for set $\boldsymbol{B}$—i.e., $\hat{Y}_{\boldsymbol{B}}$—or any other set of IVs). Thus, by treating a set in terms of how it bears on Y, we effectively reduce it to a single variable. This lies at the core of the generalizability of the ballantine from single IVs to sets of IVs.

The ballantine in Fig. 5.4.1 presents the general case: $\boldsymbol{A}$ and $\boldsymbol{B}$ share variance with Y, but also with each other.[14] This is, of course, the critical distinction between MRC and the standard



$$R^2_{YA} = a + c; \quad R^2_{YB} = b + c; \quad R^2_{Y.AB} = a + b + c.$$

$$sR^2_A = R^2_{Y(A \cdot B)} = R^2_{Y.AB} - R^2_{Y.B} = a.$$

$$sR^2_B = R^2_{Y(B \cdot A)} = R^2_{Y.AB} - R^2_{Y.A} = b.$$

$$pR^2_A = R^2_{(Y \cdot B)(A \cdot B)} = \frac{R^2_{Y.AB} - R^2_{YB}}{1 - R^2_{YB}} = \frac{a}{a+e}.$$

$$pR^2_B = R^2_{(Y \cdot A)(B \cdot A)} = \frac{R^2_{Y.AB} - R^2_{YA}}{1 - R^2_{YA}} = \frac{b}{b+e}.$$

FIGURE 5.4.1 The ballantine for sets $\boldsymbol{A}$ and $\boldsymbol{B}$.

[14] It can be proved that the correlation between $\boldsymbol{A}$ and $\boldsymbol{B}$, where each is scored by using the equation that predicts Y from the variables within that set, is given by

(5.4.2)
$$r_{\hat{Y}_A \hat{Y}_B} = \frac{\sum \beta_i \beta_j r_{ij}}{R_{YA} R_{YB}}$$

where i indexes an X_i in set $\boldsymbol{A}$, j indexes an X_j in set $\boldsymbol{B}$, and the summation is taken over all i,j pairs (of which there are $k_A k_B$).

orthogonal analysis of variance. In an $A \times B$ factorial design ANOVA, the requirement of proportional cell frequencies makes A and B (specifically $\hat{Y}_A$ and $\hat{Y}_B$) uncorrelated with each other; therefore the A and B circles do not overlap each other, and each accounts for a separate and distinguishable (that is, additive) portion of the Y variance. This is also what makes the computation simpler in ANOVA than in MRC.

The ballantine allows proportions of variance (i.e., squared correlations of various kinds) to be represented as ratios of the corresponding areas of the circle to the whole area of Y, as we saw in Section 3.3. The total variance of Y is taken to equal unity (or 100%) and the Y circle is divided into four distinct areas identified by the letters a, b, c, and e. Because overlap represents shared variance or squared correlation, we can see immediately from Fig. 5.4.1 that set A overlaps Y in areas a and c; hence

$$R^2_{YA} = \frac{YA \text{ overlap}}{sd^2_Y} = \frac{a+c}{1} = a+c. \tag{5.4.3}$$

The c area arises inevitably from the AB overlap, just as it did in the single IV ballantine in Section 3.3, and is conceptually identical with it. It designates the part of the Y circle jointly overlapped by A and B, because

$$R^2_{YB} = \frac{YB \text{ overlap}}{sd^2_Y} = \frac{b+c}{1} = b+c. \tag{5.4.4}$$

Because the c area is part of both A's and B's overlap with Y, for sets, as for single IVs, it is clear that (for the general case, where $\hat{Y}_A$ and $\hat{Y}_B$ are correlated) the proportion of Y variance accounted for by sets A and B together is not simply the sum of their separate contributions, because area c would then be counted twice, but rather

$$R^2_{Y \cdot AB} = \frac{a+b+c}{1} = a+b+c. \tag{5.4.5}$$

Thus, the areas a and b represent the proportions of Y variance uniquely accounted for respectively by set A and set B. By uniquely we mean relative to the other set, thus area b is Y variance not accounted for by set A, but only by set B; the reverse is true for area a.

This idea of unique variance in Y for a set is directly analogous to the unique variance of a single IV discussed in Chapter 3. There we saw that for X_i, the unique variance in Y is the squared semipartial correlation of Y with X_i, which in abbreviated notation we called sr^2_i. It was shown literally to be the r^2 between that part of X_i that could not be estimated from the other IVs and all of Y, the complete cumbersome notation for which is $r^2_{Y(i \cdot 12\ldots(i)\ldots k)}$, the inner parentheses signifying omission. For a set B, we similarly define its unique variance in Y to be the squared *multiple* semipartial correlation of B with the part of Y that is not estimable from A, or $1 - \hat{Y}_A$. Its literal notation would be $R^2_{Y(B \cdot A)}$, or, even more simply, sR^2_B. In the latter notation, Y is understood, as is the other set (or sets) being partialed. (Obviously, all the above holds when A and B are interchanged.)

The Semipartial R^2

The ballantine may again make this visually clear. "That part of B which is not estimable from A" is represented by the part of the B circle not overlapped by the A circle, that is, the combined area made up of b and f. That area overlaps with the (complete) Y circle only in area b, therefore the proportion of the total Y variance accounted for uniquely by set B is

$$sR^2_B = R^2_{Y(B \cdot A)} = \frac{b}{1} = b. \tag{5.4.6}$$

and, by symmetry, the proportion of Y variance accounted for uniquely by set $\boldsymbol{A}$ is

$$sR_A^2 = R^2_{Y(A \cdot B)} = \frac{a}{1} = a. \tag{5.4.7}$$

The ballantine shows how these quantities can be computed. If $R^2_{Y \cdot AB}$ is area $a + b + c$ (Eq. 5.4.5) and $R^2_{Y \cdot A}$ is area $a + c$ (Eq. 5.4.3), then patently

$$b = (a + b + c) - (a + c),$$
$$sR_B^2 = R^2_{Y \cdot AB} - R^2_{YB}. \tag{5.4.8}$$

The sR^2 can readily be found by subtracting from the R^2 for both sets the R^2 for the set to be partialed.

It is not necessary to provide for the case of more than two sets of IVs in the ballantine,[15] or, indeed, in the preceding equations. Because the result of the aggregation of any number of sets is itself a set, these equations are self-generalizing. Thus, if the unique variance in Y for set $\boldsymbol{B}$ among a group of sets is of interest, we can simply designate the sets other than $\boldsymbol{B}$ collectively as set $\boldsymbol{A}$, and find sR_B^2 from Eq. (5.4.8). This principle is applied successively as each set is added in the hierarchical analysis, each added set being designated $\boldsymbol{B}$ relative to the aggregate of prior sets, designated $\boldsymbol{A}$. We shall see that this framework neatly accommodates both significance testing and power analysis.

We offer one more bit of notation, which, although not strictly necessary, will be found convenient later on in various applications of hierarchical analysis. In the latter, the addition of a new set $\boldsymbol{B}$ (or single IV X_i) results in an increase in R^2 (strictly, a nondecrease). These increases are, of course, the sR_B^2 (or sr_i^2), as already noted. It is a nuisance in presenting such statistics, particularly in tables, to always specify all the prior sets or single IVs that are partialed. Because in hierarchical MRC the hierarchy of sets (or single IVs) is explicit, we will have occasion to identify such sR_B^2 (or sr_i^2) as increments to Y variance at the stage of the hierarchy where $\boldsymbol{B}$ (or X_i) enters.

The Partial R^2

We have already identified the overlap of that part of a set circle that is unique (e.g., areas $b + f$ of set $\boldsymbol{B}$) with the total Y circle as a squared multiple *semi*partial correlation (e.g., $sR_B^2 = R^2_{Y(B \cdot A)}$ = area b). With sets as with single IVs, it is a *semi*partial because we have related the partialed $\boldsymbol{B} \cdot \boldsymbol{A}$ with all of Y. We wrote it as $b/1$ in Eq. (5.4.6) to make it explicit that we were assessing the unique variance b as a proportion of the *total* Y variance of 1. We can however also relate the partialed $\boldsymbol{B} \cdot \boldsymbol{A}$ with the *partialed* Y, that is, we can assess the unique b variance as a proportion not of the total Y variance, but of that part of the Y variance not estimable by set $\boldsymbol{A}$, actually $Y - \hat{Y}_A$. The result is that we have defined the squared multiple partial correlation as

$$pR_B^2 = R^2_{(Y \cdot A)(B \cdot A)} = \frac{b}{b + e} \tag{5.4.9}$$

and symmetrically for set $\boldsymbol{A}$ as

$$pR_A^2 = R^2_{(Y \cdot B)(A \cdot B)} = \frac{a}{a + e}. \tag{5.4.10}$$

---

[15] A fortunate circumstance, because the complete representation of three sets would require a three-dimensional ballantine and, generally, the representation of h sets, an h-dimensional ballantine.

Thus, sR^2 and pR^2 (as sr^2 and pr^2) differ in the base to which they relate the unique variance as a proportion: sR^2 takes as its base the total Y variance whereas pR^2 takes as its base that proportion of the Y variance not accounted for by the other set(s). Inevitably, with its base smaller than (or at most equal to) 1, pR^2 will be larger than (or at least equal to) sR^2 for any given set.

It is easy enough to compute the pR^2. We have seen how, for example, the b area is found by [Eq. (5.4.8)]; the combined areas $b + e$ constitute the Y variance not accounted for by set A, hence $1 - R^2_{Y.A}$. Substituting in Eq. (5.4.9),

$$pR^2_B = \frac{b}{b+e} = \frac{R^2_{Y.AB} - R^2_{YA}}{1 - R^2_{YA}}, \tag{5.4.11}$$

and, symmetrically,

$$pR^2_A = \frac{a}{a+e} = \frac{R^2_{Y.AB} - R^2_{YB}}{1 - R^2_{YB}}. \tag{5.4.12}$$

To illustrate the distinction between sR^2 and pR^2, we refer to the example of the hierarchy of sets of demographics (D), illness (I), and hospitals (H) in relationship to length of hospital stay (Y) of Section 5.3.2. $R^2_{Y.D} = .20$, and when I is added, $R^2_{Y.DI} = .22$. The increment was .02; hence, $sR^2_I = .02$, that is, 2% of the *total* Y variance is uniquely (relative to D) accounted for by I. But if we ask "what proportion of the variance of Y not accounted for by D is uniquely accounted for by I?" our base is not the total Y variance, but only $1 - R^2_{YD} = 1 - .20 = .80$ of it, and the answer is $pR^2_I = .02/.80 = .025$. Letting $D = A$ and $I = B$, we have simply substituted in Eqs. (5.4.7) and (5.4.11).

It was also found that the further addition of H resulted in $R^2_{Y.DIH} = .33$. Thus, H accounted for an additional .11 of the *total* Y variance, hence $sR^2_H = .11$. If we shift our base from total Y variance to Y variance not already accounted for by D and I, the relevant proportion is $.11/(1 - .22) = .141$ (i.e., pR^2_H). Now letting sets $D + I = A$, and $H = B$, we again have simply substituted in Eqs. (5.4.7) and (5.4.11). Any desired combination of sets can be created: If we wished to combine I and H into set B, with $D = A$, we could determine that $sR^2_{IH} = .13$, and $pR^2_{IH} = .13/(1 - .20) = .162$, by the same equations.

It is worth noting that the pR^2 is rather in the spirit of ANCOVA. In ANCOVA, the variance due to the covariates is removed from Y and the effects of research factors are assessed with regard to this adjusted (partialed) Y variance. Thus, in the latter example, D and I may be considered to be covariates whose function is to "equate" the hospitals, so that they may be compared for length of stay, free of any possible hospital differences in the D and I of their patients. In that spirit, we are interested only in the $1 - R^2_{Y.DI}$ portion of the Y variance, and the pR^2_H takes as its base the .78 of the Y variance *not* associated with D and I; hence, $pR^2_H = .11/.78 = .141$ of this adjusted (or partialed, or residual) variance is the quantity of interest.

5.4.4 *B* and β Coefficients for Variables Within Sets

As we have noted, the strongest scientific inferences are likely to come from an examination of raw or standardized regression coefficients. When variables are treated in sets, attention to these coefficients is still indicated. To understand these issues, let us first attend to the influence of other variables in the same set on B and β. We will assume that a functional set is being examined, that is, that there is some theoretical role shared by the variables in the set, such

as control for spurious effects, or representation of a set of related concepts. Effects of other variables in a categorical set are discussed in Chapters 9 and 10.

It is usually the case that members of functional sets are at least somewhat correlated. Thus, the fundamental principal that regression coefficients reflect the influence of a variable net of the influence (controlling for or ceteris paribus) of all other variables in the equation, applies equally to other IVs in the same set and those in other sets. When the correlations among the variables in a set are relatively large, it can happen that no individual variable is significantly related to Y even when the set as a whole accounts for a large and statistically significant proportion of the Y variance. In such a case we will find large *SE*s on at least some Bs and βs, relative to zero-order *SE*s on the bivariate r_{Yi} or B_{yi}.[16] We will discuss this problem at the end of this chapter in a section on balancing Type I and Type II errors (Section 5.7.1). In the present discussion it will suffice for the reader to keep firmly in mind that each of these coefficients is indicating the effect of the individual variables net of others in the set (as well as any variables in other sets that are also in the equation).

As noted earlier, sometimes the causal hierarchical ordering of functional sets cannot be unambiguously asserted. It is not infrequent that variable sets are created to represent domains of influence on Y that may best be thought of as also influencing one another. For example, a previous illustration presented an analysis of the influence of demographic and personality differences of patients in different hospitals on length of hospital stay. Suppose we wanted to add to this analysis a set of variables representing the functional impairment of patients (e.g., competence in activities of daily living, reliability and tenure in occupational settings, ability to relate to and care for others). We may now feel ourselves to be in the all-too-frequent position of ambiguity as to whether the personality (symptom) set or the impairment set should be considered causally prior, and thus added to the equation at an earlier point.

A common practice in such a situation is to examine differences in the estimated regression coefficients (as well as sR^2) when the hierarchical ordering is altered. Again, it is also useful to attend to changes in the *SE*s for these individual partial coefficients, because, as was shown in Eq. (3.6.1), they are enlarged by the multiple correlation of the variable in question with other variables in the equation.

Area c

Finally, returning once more to the ballantine for sets (Fig. 5.4.1), we call the reader's attention to area c, the double overlap of sets $\boldsymbol{A}$ and $\boldsymbol{B}$ in Y. It is conceptually the same as the area c in the ballantine for single IVs (Fig. 3.3.1) and shares its problems. Although in the ballantine it occupies an area, unlike the areas a, b, and e it *cannot* be understood to be a proportion of Y variance, because, unlike these other areas, it may take on a negative value as discussed in Section 3.3. Note that it is never properly interpreted as a proportion of variance, whether or not in any given application it is found to be positive, because we cannot alter the fundamental conception of what a statistic means as a function of its algebraic sign in a particular problem. Because variance is sd^2, a negative quantity leads to sd being an imaginary number, for example, $\sqrt{-.10}$, a circumstance we simply cannot tolerate. Better to let area c stand as a useful metaphor that reflects the fact that $R^2_{Y.\boldsymbol{AB}}$ is not equal in general to $R^2_{Y\boldsymbol{A}} + R^2_{Y\boldsymbol{B}}$, but may be either smaller (positive c) or larger (negative c) than the sum. When area c is negative for sets $\boldsymbol{A}$ and $\boldsymbol{B}$, we have exactly the same relationship of suppression between the two sets as was described for pairs of single IVs in Section 3.4.

---

[16]Note, too, the discussion of this issue in Section 4.5.3.

5.5 SIGNIFICANCE TESTING FOR SETS

We have seen that the addition of a set of variables $\boldsymbol{B}$ to a set $\boldsymbol{A}$ results in an increment in the Y variance accounted for by $R^2_{Y.AB} - R^2_{YA} (= sR^2_B)$, represented by the area b in the ballantine (Fig. 5.4.1). This quantity is properly called an increment because it is not mathematically possible for it to be negative, because $R^2_{Y.AB}$ cannot be smaller than R^2_{YA}.[17]

Our interest, of course, is not in the characteristics of the sample for which these values are determined as such but rather in those of the population from which it comes. Our mechanism of statistical inference posits the null (nil) hypothesis to the effect that in the population, there is literally no increment in Y variance accounted for when $\boldsymbol{B}$ is added to $\boldsymbol{A}$, that is, that $R^2_{Y.AB} - R^2_{YA} = 0$ in the population. When this null hypothesis is rejected, we conclude that set $\boldsymbol{B}$ does account for Y variance beyond that accounted for by set $\boldsymbol{A}$ in the population. This null hypothesis may be tested by means of

$$F = \frac{(R^2_{Y.AB} - R^2_{YA})/k_B}{(1 - R^2_{Y.AB})/(n - k_A - k_B - 1)} \tag{5.5.1}$$

for the source (numerator) $df = k_B$, the error (denominator) $df = n - k_A - k_B - 1$, and referred to the F tables in the appendices (Appendix Tables D.1 and D.2).

This formula is applied repeatedly in varying contexts throughout this book, and its structure is worth some comment. Both the numerator and denominator are proportions of Y variance divided by their respective df; thus both are "normalized" mean squares. The numerator is the normalized mean square for unique $\boldsymbol{B}$ variance (area b of the ballantine) and the denominator is the normalized mean square for a particular estimate of error (i.e., $1 - R^2_{Y.AB}$), that represents Y variation accounted for by neither $\boldsymbol{A}$ nor $\boldsymbol{B}$ (area e of the ballantine). F is the ratio of these mean squares and, when the null hypothesis is true, has an expected value of about one. When F is sufficiently large to meet the significance criterion, as determined by reference to Appendix Tables D.1 and D.2, the null hypothesis is rejected.[18]

For computational purposes Eq. (5.5.1) can be somewhat simplified:

$$F = \frac{R^2_{Y.AB} - R^2_{YA}}{1 - R^2_{Y.AB}} \times \frac{n - k_A - k_B - 1}{k_B} \tag{5.5.2A}$$

(for $df = k_B, n - k_A - k_B - 1$, as before). F may equivalently be determined by means of the regression sums of squares (SS) and mean squares usually provided by the computer output. The numerator for F for the increment equals the difference between the regression SS for the equation including $\boldsymbol{A}$ and $\boldsymbol{B}$ and the regression SS for the equation including only $\boldsymbol{A}$, divided by the df for $\boldsymbol{B}$. The denominator equals the residual MS for the equation with both $\boldsymbol{A}$ and $\boldsymbol{B}$. Thus:

$$F = \frac{(\text{Regression SS}_{AB} - \text{Regression SS}_A)/df_B}{\text{Residual MS}_{AB}} \tag{5.5.2B}$$

When there are additional IV sets to be considered this method is referred to as employing Model 1 error. An alternative strategy is to use the residual MS from the equation that includes all sets, with error df equal to the df for that term. Such a strategy may be considered to employ Model 2 error in the significance tests.

---

[17] This proposition does not hold for R^2 corrected for shrinkage, that is, $\tilde{R}^2_{Y.AB} - \tilde{R}^2_{Y.A}$ may be negative. This will occur whenever the F of Eq. (5.5.1) is less than one.

[18] Readers who know ANOVA will find this all familiar. But the reasoning and structure are not merely analogous but rather mathematically identical, because ANOVA and MRC are applications of the OLS model.

5.5.1 Application in Hierarchical Analysis

To illustrate the application of this formula, let us return to the study presented in the previous sections on the length of stay of 500 hospital admissions, using demographic ($\boldsymbol{D}, k_D = 9$), illness ($\boldsymbol{I}, k_I = 10$), and hospital ($\boldsymbol{H}, k_H = 7$) sets in that hierarchical order. We let $\boldsymbol{A}$ be the set(s) to be partialed and $\boldsymbol{B}$ the set(s) whose unique variance in Y is posited as zero by the null hypothesis. Table 5.5.1 organizes the ingredients of the computation to facilitate the exposition.

The null hypothesis that with $\boldsymbol{D}$ partialed (holding demographic characteristics constant) $\boldsymbol{I}$ accounts for no Y variance in the population is appraised as follows (Table 5.5.1, Example 1). It was given that $R^2_{Y\boldsymbol{D}} = .20$ and $R^2_{Y.\boldsymbol{DI}} = .22$, an increase of .02. To use Eq. (5.5.2), call $\boldsymbol{I}$ set $\boldsymbol{B}$ and $\boldsymbol{D}$ set $\boldsymbol{A}$. For $n = 500, k_B = 10, k_A = 9$, we find

$$F = \frac{.22 - .02}{1 - .22} \times \frac{500 - 9 - 10 - 1}{10} = 1.231,$$

which for df of 10 ($= k_B$) and 480 ($= n - k_A - k_B - 1$) fails to be significant at the $\alpha = .05$ level (the criterion value for $df = 10, 400$ is 1.85, Appendix Table D.2). The increase of .02 of the Y variance accounted for by $\boldsymbol{I}$ over $\boldsymbol{D}$ in the sample is thus consistent with there being no increase in the population.

In Example 2 of Table 5.5.1, we test the null hypothesis that the addition of $\boldsymbol{H}$ (which we will now call set $\boldsymbol{B}$, so $k_B = 7$) to the sets $\boldsymbol{D}$ and $\boldsymbol{I}$ (which together we will call set $\boldsymbol{A}$, so $k_A = 9 + 10 = 19$) results in no increase in Y variance in the population. Because $R^2_{Y.\boldsymbol{DIH}} = .33$ and $R^2_{Y.\boldsymbol{DI}} = .22$ (and hence $sR^2_{\boldsymbol{H}} = .11$), substituting the values for sets $\boldsymbol{A}$ and $\boldsymbol{B}$ as redefined we find

$$F = \frac{.33 - .22}{1 - .33} \times \frac{500 - 19 - 7 - 1}{7} = 11.094,$$

TABLE 5.5.1
Illustrative F Tests Using Model 1 Error

$R^2_{Y\boldsymbol{D}} = .20; R^2_{Y.\boldsymbol{IH}} = .18; R^2_{Y\boldsymbol{I}} = .03; R^2_{Y.\boldsymbol{DI}} = .22; R^2_{Y.\boldsymbol{DH}} = .32; R^2_{Y\boldsymbol{H}} = .17; R^2_{Y.\boldsymbol{DIH}} = .33.$

Example	Set $\boldsymbol{B}$	k_B	Set $\boldsymbol{A}$	k_A	$R^2_{Y.AB}$	$R^2_{Y.A}$	$R^2_{Y(B\cdot A)} = sR^2_B$	Error $1 - R^2_{Y.AB}$	Source df	Error df	F
1	$\boldsymbol{I}$	10	$\boldsymbol{D}$	9	.22	.20	.02	.78	10	480	1.23
2	$\boldsymbol{H}$	7	$\boldsymbol{D,I}$	19	.33	.22	.11	.67	7	473	11.09
3	$\boldsymbol{I,H}$	17	$\boldsymbol{D}$	9	.33	.20	.13	.67	17	473	5.40
4	$\boldsymbol{D}$	9	$\boldsymbol{I,H}$	17	.33	.18	.15	.67	9	473	11.77
5	$\boldsymbol{I}$	10	$\boldsymbol{D,H}$	16	.33	.32	.01	.67	10	473	.71
6	$\boldsymbol{D}$	9	$\boldsymbol{I}$	10	.22	.03	.19	.78	9	480	12.99
7	$\boldsymbol{D}$	9	$\boldsymbol{H}$	7	.32	.17	.15	.68	9	483	11.84
8	$\boldsymbol{H}$	7	$\boldsymbol{D}$	9	.32	.20	.12	.68	7	483	12.18
9	$\boldsymbol{I}$	10	$\boldsymbol{H}$	7	.18	.17	.01	.82	10	482	.59
10	$\boldsymbol{H}$	7	$\boldsymbol{I}$	10	.18	.03	.15	.82	7	482	12.60
11	$\boldsymbol{D}$	9	—	0	.20	0	.20	.80	9	490	13.61
12	$\boldsymbol{I}$	10	—	0	.03	0	.03	.97	10	489	1.51
13	$\boldsymbol{H}$	7	—	0	.17	0	.17	.83	7	492	14.40

$$F = \frac{R^2_{Y.AB} - R^2_{Y.A}}{R^2_{Y.AB}} \times \frac{n - k_A - k_B - 1}{k_B},$$

with source (numerator) $df = k_B$, error I (denominator) $df = n - k_A - k_B - 1$.

which for $df = 7,473$ is highly significant, because the criterion F at $\alpha = .01$ for df 7,400 is 2.69 (Appendix Table D.1).

It was pointed out in Section 5.4.4 that our appraisal of this .11 increment by $\boldsymbol{H}$ over $\boldsymbol{D}$ and $\boldsymbol{I}$ constitutes the equivalent of an ANCOVA with the 19 IVs of the combined $\boldsymbol{D}$ and $\boldsymbol{I}$ sets as covariates. Indeed, hierarchical MRC may be viewed as equivalent to a series of ANCOVAs, at each stage of which all prior sets are covariates (because they are partialed), whereas the set just entered is the research factor whose effects are under scrutiny. The set just entered may itself be an aggregate of sets. Although it would not likely be of substantive interest in this research, Example 3 of Table 5.5.1 illustrates the F test for the aggregate of $\boldsymbol{I}$ and $\boldsymbol{H}$ (as set $\boldsymbol{B}$) with $\boldsymbol{D}$ (as set $\boldsymbol{A}$) partialed.

5.5.2 Application in Simultaneous Analysis

The F test of Eqs. (5.5.1) and (5.5.2) is also applicable in simultaneous analysis. The latter simply means that, given h sets, we are interested in appraising the variance of one of them with all the remaining $h - 1$ sets partialed. Whereas in the hierarchical model only higher-order (prior) sets are partialed, in the absence of a clear rationale for such a hierarchy it is all other sets that are partialed. For this application of the F test we designate $\boldsymbol{B}$ as the unique source of variance under scrutiny and aggregate the remaining sets that are to be partialed into set $\boldsymbol{A}$.

Let us reconsider the running example of length of stay (Y) as a function of $\boldsymbol{D}$, $\boldsymbol{I}$, and $\boldsymbol{H}$, but now propose that our interest is one of appraising the unique Y variance accounted for by each set. No hierarchy is intended, so by "unique" to a set we mean relative to all (here, both) other sets (i.e., $\boldsymbol{D}$ relative to $\boldsymbol{I}$ and $\boldsymbol{H}$, $\boldsymbol{I}$ relative to $\boldsymbol{D}$ and $\boldsymbol{H}$, and $\boldsymbol{H}$ relative to $\boldsymbol{D}$ and $\boldsymbol{I}$). To proceed, we need some additional R^2 values not previously given in this problem: $R^2_{Y.IH} = .18$ and $R^2_{Y.DH} = .32$.

To determine the unique contribution of $\boldsymbol{D}$ relative to $\boldsymbol{I}$ and $\boldsymbol{H}$, one simply finds $R^2_{Y.DIH} - R^2_{Y.IH} = .33 - .18 = .15 = R^2_{Y(D.IH)}$, the sR^2_D with both $\boldsymbol{I}$ and $\boldsymbol{H}$ partialed. This quantity might be of focal interest to a sociologist in that it represents the proportion of variance in length of stay of patients associated with differences in their demographic ($\boldsymbol{D}$) characteristics, the latter freed of any illness differences ($\boldsymbol{I}$) and differences in admitting hospitals ($\boldsymbol{H}$) associated with $\boldsymbol{D}$. This .15 is a sample quantity, and Example 4 (Table 5.5.1) treats $\boldsymbol{D}$ as set $\boldsymbol{B}$ and aggregates $\boldsymbol{I}$ and $\boldsymbol{H}$ as set $\boldsymbol{A}$ for substitution in Eq. (5.5.2):

$$F = \frac{.33 - .18}{1 - .33} \times \frac{500 - 17 - 9 - 1}{9} = 11.766,$$

which is statistically significant because the criterion F at $\alpha = .01$ for $df = 9{,}400$ is 2.45 (Appendix Table D.1). Note, incidentally, that this example simply reverses the roles of $\boldsymbol{D}$ and $\boldsymbol{I}$, $\boldsymbol{H}$ of Example 3.

The unique variance contribution of $\boldsymbol{I}$ relative to $\boldsymbol{D}$ and $\boldsymbol{H}$ is tested without further elaboration as Example 5. This might be of particular interest to a clinical psychologist or personality measurement specialist interested in controlling demographic variables and systematic differences between hospitals in assessing the relationship of illness to length of stay. The last of this series, the Y variance associated with $\boldsymbol{H} \cdot \boldsymbol{DI}$, has already been presented and discussed as Example 2.

Thus, the investigator's choice of what to partial from what is determined by the logic and purpose of the inquiry. For specificity, assume that the h sets are partitioned into three groups of sets as follows: the groups whose unique source is under scrutiny is, as before, designated set $\boldsymbol{B}$, the covariate group to be partialed from $\boldsymbol{B}$ (again as before) constitutes set $\boldsymbol{A}$, but now the remaining set(s) constitute a group to be ignored, which we designate set $\boldsymbol{C}$. All

we are doing with this scheme is making explicit the obvious fact that not all sets of IVs on which there are data in an investigation need to be active participants in each phase of the analysis. Indeed, the (fully) hierarchical analysis with h sets is simply a predefined sequence of simultaneous analyses in the first of which a prespecified $h - 1$ sets are ignored, in the second a prespecified $h - 2$ sets are ignored, and, generally, in the jth of which a prespecified $h - j$ sets are ignored until finally, in the last of which, none is ignored. The analysis at each stage is simultaneous—all IVs in the equation at that stage are being partialed from each other. Thus, a single simultaneous analysis with all other sets partialed and a strictly hierarchical progression of analyses may be viewed as end points of a continuum of analytic possibilities. A flexible application of MRC permits the selection of some intermediate possibilities when they are dictated by the causal theory and logic of the given research investigation.

5.5.3 Using Computer Output to Determine Statistical Significance

Of course, current data analysts are likely to accept the statistical tests presented by the computer output. In some programs it is possible to specify a set of variables the contribution of which to R^2 is to be evaluated. If such an option is not available, the computer-provided output for various sets and combinations of sets of variables may be employed in Eq. (5.5.2A or B).

In Chapter 3 we saw that the partialed statistics of a single IV, X_i (i.e., sr_i, pr_i, B_i, and β_i) all shared equivalent null (nil) hypotheses and hence the same t test for the same $df = n - k - 1$. Conceptually, this can be explained as due to the fact that when any one of these coefficients equals zero, they all must necessarily equal zero.

For any set B, the same identity in significance tests holds for sR_B^2 and pR_B^2 (hence for sR_B and pR_B). Recall that these are both unique proportions of Y variance, the first to the base unity and the second to the base $1 - R_{Y \cdot A}^2$. In terms of areas of the ballantine for sets (Fig. 5.5.1), $sR_B^2 = b$, and $pR_B^2 = b/(b + e)$. But the null hypothesis posits that area b is zero, hence $pR_B^2 = 0$. Whether one reports sR_B^2 as was done in Table 5.5.1, or divides it by $1 - R_{Y \cdot A}^2$ and reports pR_B^2, or reports both, the F test of Eq. (5.5.2) tests statistical significance of both sR_B^2 and pR_B^2 because the null hypothesis is the same.

One highly desired test is a comparison of the utility of two different sets of variables in predicting the same Y. Because determination of the standard errors of these coefficients and their difference is extremely complicated, involving the covariance among all predictor sets, it is not possible to calculate from the output ordinarily provided to the users of standard statistical programs. Olkin and Finn (1995) provide the test for the special case in which each of the sets consists of a single variable, which is itself complex. It is hoped that a solution to this problem will be found for variable sets and programmed in the next few years.

5.5.4 An Alternative *F* Test: Using Model 2 Error Estimate From the Final Model

An F test is a ratio of two mean square or variance estimates, the numerator associated with a source of Y variance being tested, and the denominator providing a reference amount in the form of an estimate of error or residual variance. In the previous section, identifying $\boldsymbol{A}$ and $\boldsymbol{B}$ as sets or set aggregates, the numerator source was $\boldsymbol{B} \cdot \boldsymbol{A}$, and the denominator contained $1 - R_{Y \cdot \boldsymbol{AB}}^2$ (area e of the ballantine; Fig. 5.4.1) thus treating all Y variance not accounted for by $\boldsymbol{A}$ and $\boldsymbol{B}$ as error in the F test of Eqs. (5.5.1) and (5.5.2). We later introduced the idea of a third set (or set of sets) $\boldsymbol{C}$, whose modest purpose was "to be ignored." Not only was it ignored in that it was not partialed from $\boldsymbol{B}$ in defining $\boldsymbol{B} \cdot \boldsymbol{A}$ as the source for the numerator, but it was

ignored in that whatever Y variance it might uniquely contribute was not included in $R^2_{Y.AB}$ and therefore was part of the error, $1 - R^2_{Y.AB}$.

These two ways of ignoring $\boldsymbol{C}$ are conceptually quite distinct and may be considered independently. We obviously have the option of not partialing whatever we do not wish to partial from $\boldsymbol{B}$. Presumably the source of variance in the numerator is precisely what the theory and logic of the investigation dictates it to be (i.e., $\boldsymbol{B} \cdot \boldsymbol{A}$ and not $\boldsymbol{B} \cdot \boldsymbol{AC}$). We may either choose or not choose to ignore $\boldsymbol{C}$ in defining the error term. The first choice, Model 1 error uses $1 - R^2_{Y.AB}$ in the F test of Eqs. (5.5.1) and (5.5.2) and thus ignores $\boldsymbol{C}$. The alternative, Model 2 error, defines an F ratio for $\boldsymbol{B} \cdot \boldsymbol{A}$ that removes whatever additional unique Y variance can be accounted for by $\boldsymbol{C}$ from the error term, resulting in the following error term and associated *df*, expressed here both in terms of the various R^2 values (proportions of variance) and in terms of the SS and error MS from various equations:

$$\textbf{(5.5.3)} \qquad F = \frac{(R^2_{Y.AB} - R^2_{YA})/k_B}{(1 - R^2_{Y.ABC})/(n - k - 1)} = \frac{(SS_{Y.AB} - SS_{YA})/k_B}{\text{error } MS_{ABC}/(n - k - 1)}$$

where k is the total number of IVs in all sets, that is, $k = k_A + k_B + k_C$ or equivalently,

$$\textbf{(5.5.4)} \qquad F = \frac{(R^2_{Y.AB} - R^2_{YA})}{(1 - R^2_{Y.ABC})} \times \frac{n - k - 1}{k_B},$$

with numerator $df = k_B$, and error $df = n - k - 1$. Note that, as with the F that considers only the sets already entered, this tests both sR^2_B and pR^2_B. The standard F tables (Appendix Tables D.1 and D.2) are used. Of course, although we have discussed Model 2 error in the context of the hierarchical analysis of sets of IVs, any set may consist of a single variable, and the procedure may thus be employed equally appropriately in the case of the determination of statistical significance for a single IV.

Which model to choose? One view notes that because the removal of additional Y variance associated uniquely with $\boldsymbol{C}$ serves to produce a smaller and "purer" error term, one should generally prefer Model 2 error. But although $1 - R^2_{Y.ABC}$ will always be smaller (strictly, not larger) than $1 - R^2_{Y.AB}$ and hence operate so as to increase F, one must pay the price of the reduction of the error *df* by k_C, that is from $n - k_A - k_B - 1$ of Eq. (5.5.2) to $n - k_A - k_B - k_C - 1$ of Eq. (5.5.4), which clearly operates to decrease F. In addition, as error *df* diminish, the criterion F ratio for significance increases and sample estimates become less stable, seriously so when the diminished error *df* are absolutely small. The competing factors of reducing proportion of error variance and reducing error *df*, depending on their magnitudes, may either increase or decrease the F using Model 2 error relative to the F using Model 1 error.

We can illustrate both possibilities with the running example (Table 5.5.1), comparing Model 1 F (Eq. 5.5.2) with Model 2 F (Eq. 5.5.4). If, in testing $\boldsymbol{I} \cdot \boldsymbol{D}$ in Example 1, instead of using Model 1 error, $1 - R^2_{Y.DM} = .78$ with 480 $(= 500 - 9 - 10 - 1)$ *df*, we use Model 2 error, $1 - R^2_{Y.DMH} = .67$ with 473 $(= 500 = 9 - 10 - 7 - 1)$ *df*, F increases to 1.412 from 1.231 (neither significant). On the other hand, shifting to Model 2 error in testing $\boldsymbol{D} \cdot \boldsymbol{H}$ in Example 7 brings F down from 11.838 to 11.766 (both significant at $p < .01$).

In Table 5.5.1 the F ratios of the two models differ little and nowhere lead to different decisions about the null hypothesis. But before one jumps to the conclusion that the choice makes little or no difference in general, certain characteristics of this example should be noted and discussed, particularly the relatively large n, the fact that there are only three sets, and that two of these ($\boldsymbol{D}$ and $\boldsymbol{H}$) account uniquely for relatively large proportions of variance. If n were much smaller, the differences of k_C loss in error *df* in Model 2 could substantially reduce the size and significance of F, particularly in the case where we let $\boldsymbol{I}$ be set $\boldsymbol{C}$: the addition of $\boldsymbol{I}$ to $\boldsymbol{D}$ and $\boldsymbol{H}$ results in only a quite small decrease in error variance, specifically

from $1 - R^2_{Y.DH} = .68$ to $1 - R^2_{Y.DHM} = .67$. If n were 100, the drop in error df from Model 1 to Model 2 would be from 83 to 73. Example 7, which tests $\boldsymbol{D} \cdot \boldsymbol{H}$ would yield a significant Model 1 $F = 2.034$ ($df = 9, 83, p < .05$), but a nonsignificant Model 2 $F = 1.816$ ($df = 9, 73$).

Further, consider the consequence of Model 2 error when the number of sets, and therefore the number of sets in $\boldsymbol{C}$ and, particularly, k_C is large. Many behavioral science investigations can easily involve upward of a dozen sets, so that collectively $\boldsymbol{C}$ may include many IVs and thus df. The optimal strategy in such circumstances may be to order the sets from those judged a priori to be most important to those judged to contribute least, or least confidently judged to account for Y variance, and use Model 1 error. Using the latter successively at each level of the hierarchy, the lower-order sets are ignored and, although their (likely small) contribution to reducing the proportion of error variance is given up, their large contribution to the error df is retained.

On the other hand, if sets are few and powerful in accounting uniquely for Y variance, Model 1 error will contain important sources of variance due to the ignored $\boldsymbol{C}$, and may well sharply negatively bias (reduce) F at a relatively small gain in error df. No simple advice can be offered on the choice between error models in hierarchical analysis of MRC. In general, large n, few sets, small k, and sets whose sR^2 are large move us toward a preference for Model 2 error. One is understandably uneasy with the prospect of not removing from the Model 1 error the variability due to a set suspected a priori of having a large sR^2 (e.g., Examples 9 through 13) during the planning of an investigation, which ideally is when it should be made. Unfortunately, because most computer programs do not offer Model 2 error as an option, the data analyst who relies completely on the program-produced tests of statistical significance will necessarily be using Model 1 error.

5.6 POWER ANALYSIS FOR SETS

In Chapters 2 and 3 we focused on the precision of estimates and the statistical power for detecting differences from various null hypotheses for the relationship, zero-order or partial, between a single IV and Y. In determining the power against the null hypothesis of no contribution to the Y variance in the population for a set of variables, we will again generally use a computer program to determine:

1. The power of the F test of significance for partialed sets, given the sample size (n), k, the significance criterion (α), and the effect size (ES), an alternative to the null (nil) hypothetical value for the population. This ES is a ratio of two variances, that due to the predictor(s) being considered (sR^2_B), and the error variance ($1 - R^2_Y$).
2. The necessary sample size (n^*) for the significance test of a set involving k variables, given the desired power, α, and the alternate hypothetical value for the contribution to R^2, relative to the null hypothesis of no population effect.

Assume that an investigation is being planned in which at some point the proportion of Y variance accounted for by a set $\boldsymbol{B}$, over and above that accounted for by a set $\boldsymbol{A}$, will be determined. We have seen that this critically important quantity is $R^2_{Y.AB} - R^2_{YA}$ and has variously and equivalently been identified as the increment due to $\boldsymbol{B}$, the squared multiple semipartial correlations for $\boldsymbol{B}$ (sR^2_B or $R^2_{Y(B \cdot A)}$) and as area b in the ballantine for sets (Fig. 5.4.1). This sample quantity will then be tested for significance, that is, the status of the null hypothesis that its value in the population is zero will be determined by means of an F test.

5.6.1 Determining n^* for the F Test of sR_B^2 with Model 1 or Model 2 Error

As was the case for determining n^* for an F test on $R^2_{Y.12...k}$ (Section 3.7), the procedure for determining n^* for an F test on $sR_B^2 = R^2_{Y.AB} - R^2_{YA}$ proceeds with the following steps:

1. Set the significance criterion to be used, α.
2. Set desired power for the F test.
3. Identify the number of IVs to be included in Set A, in Set B, and, if Model 2 error is to be used, in Set C.
4. Identify the alternate hypothetical ES in the population for which n^* is to be determined, that is, the population sR_B^2.
5. Identify the anticipated error variance, that is, $(1 - R^2_{Y.AB})$ for Model 1 or $(1 - R^2_{Y.ABC})$ for Model 2 error.

If a computer power analysis program is used to determine n^*, these values are entered into the program and the necessary n^* is read out. If the computation is done by hand, the next step is to look up the value of L for the given k_B (row) and desired power (column) in a table for the selected α (Appendix Table E.1 or E.2). One then determines the ES index f^2, which is the ratio of the variances determined in steps 4 and 5. In determining n^* to test $R^2_{Y.AB} - R^2_{YA}$ using Model 1 error,

$$f^2 = \frac{R^2_{Y.AB} - R^2_{YA}}{1 - R^2_{Y.AB}}, \tag{5.6.1}$$

or, using Model 2 error,

$$f^2 = \frac{R^2_{Y.AB} - R^2_{Y.A}}{1 - R^2_{Y.ABC}}. \tag{5.6.2}$$

We remind the reader that these R^2s are alternate hypothetical values referring to the population, *not* sample values. When the same ratio for *sample* values is combined with the *df*, the formulas are equivalent to those for F [Eqs. (5.5.1) and (5.5.3)]. This occurs after there *is* a sample, whereas in the planning taking place *before* the investigation the formulation is "*if* f^2 is thus and such in the population, given α and the desired probability of rejecting the null, what n^* do I need?" To estimate this one draws on past experience in the research area, theory, intuition, or conventional values to answer the questions "What additional proportion of Y variance do I expect B to account for beyond A? (the numerator), and "What proportion of Y variance will no be accounted for by A or B, or not by A or B or C"? (the denominators for Model 1 and Model 2 error, respectively). The values from these steps are then substituted in

$$n^* = \frac{L}{f^2} + k_A + k_B + 1 \tag{5.6.3}$$

for Model 1 error, or

$$n^* = \frac{L}{f^2} + k_A + k_B + k_C + 1 \tag{5.6.4}$$

for Model 2 error. The result is the number of cases necessary to have the specified probability of rejecting the null hypothesis (power) at the α level of significance when f^2 in the population is as posited.

For illustration, we return to the running example, where length of stay of psychiatric admissions (Y) is studied as a function of sets of variables representing their demographic characteristics ($\boldsymbol{D}$), their illness scores ($\boldsymbol{I}$), and the hospitals where they were admitted ($\boldsymbol{H}$) as described originally in Section 5.4.4. To this point this example has been discussed after the fact—results from a sample of 500 cases were presented and used to illustrate significance testing. Now we shift our perspective backward in time to illustrate the power analysis associated with the planning of this research.

In planning this investigation we know that we will eventually be testing the null hypothesis (among others) that $\boldsymbol{I}$ will account for no variance in Y beyond what is accounted for by $\boldsymbol{D}$. Thus, $\boldsymbol{I}$ is the set $\boldsymbol{B}$ and $\boldsymbol{D}$ is the set that is partialed from $\boldsymbol{B}$, set $\mathbf{A}$, and this null hypothesis is that $R^2_{Y.DI} - R^2_{YD} = R^2_{Y.AB} - R^2_{YA} = 0$, to be tested with Model 1 error, $1 - R^2_{Y.DI} = 1 - R^2_{Y.AB}$ (that is, the test eventually performed as Example 1, Table 5.5.1). Assume that we intend to use as significance criterion $\alpha = .05$ (step 1) and that we wish the probability of rejecting this hypothesis (the power of the test) to be .90 (step 2). There are 9 variables in set $\boldsymbol{D}$ and 10 IVs in set $\boldsymbol{I}$, so $k_{\boldsymbol{B}} = 10$ (step 3). We estimate the actual population value for $sR^2_{\boldsymbol{I}} = R^2_{Y.DI} - R^2_{YD} = .03$ and $R^2_{Y.DI} = .18$ (and hence, necessarily, $R^2_{YD} = .15$). If determining the necessary n^* by computer program, these values are entered and $n^* = 580$ is read out. For hand calculation from Eq. (5.6.1),

$$f^2 = \frac{.03}{1 - .18} = \frac{.03}{.82} = .0366$$

(step 4). Looking up the value for L in Appendix Table E.2 for $\alpha = .05$, in row $k_{\boldsymbol{B}} = 10$, column power = .90, we find $L = 20.53$, and, solving Eq. (5.6.1) for

$$n^* = \frac{20.53}{.0366} + 9 + 10 + 1 = 581,$$

approximately the same value provided by the program.[19] Thus, if the unique Y variance of $\boldsymbol{I} \cdot \boldsymbol{D}$ in the population ($sR^2_{\boldsymbol{I}}$) is .03, and Model 1 error is $1 - .18 = .82$, then in order to have a .90 probability of rejecting the null hypothesis at $\alpha = .05$, the sample should contain 581 cases. As was the case for single variables, lowering α, the desired power, or the number of variables will reduce the estimated number of cases required, whereas lowering the estimate of the population effect size, or the proportion of Y variance accounted for by other sets will increase the number of cases required.

What if we had decided to use Model 2 error? In this case we add an estimate of the (net) effect of differences among the hospitals on Y, length of stay, that is, estimate $R^2_{Y.DIH}$. Suppose we posit this value to be .25 so that Model 2 error is $1 - .25 = .75$. Therefore,

$$f^2 = \frac{.03}{1 - .25} = \frac{.03}{.75} = .04$$

from Eq. (5.6.2), which, of course, cannot be smaller than the f^2 for Model 1, which was .0366. Again, we either enter these values in the computer program or go on to look up L in the relevant Appendix Table E. We find $L = 20.53$, as it was for the Model 1 calculations. Solving Eq. (5.6.4), we find

$$n^* = \frac{20.53}{.04} + 9 + 10 + 7 + 1 = 540$$

[19] Hand calculation inevitably involves the use of tabled approximations, whereas the computer provides a more nearly exact value. Of course the degree of precision is adequate in either, as one can see by the crude approximation of the population ES that is necessary for these estimates.

for Model 2, compared with 581 for Model 1. In this case we found that n^* was smaller for Model 2 than for Model 1 for the same specifications. However, this case should not be overgeneralized, as we have already argued. The relative size of the n^* of the two models depends on how much reduction in the (alternate hypothetical) proportion of error variance occurs relative to the cost of df due to the addition of the k_C IVs of set $\boldsymbol{C}$. Model 2 will require smaller n^* than Model 1 when sR^2_C is large relative to k_C, but larger n^* than Model 1 when sR^2_C is small relative to k_C. If, for example, we had posited $R^2_{Y.DIH}$ to be .19, so that the Model 2 $f^2 = .03/(1 - .19) = .0370$. Solving Eq. (5.6.4), we find

$$n^* = \frac{20.53}{.037} + 9 + 10 + 7 + 1 = 582$$

as compared to 581 for Model 1. (The exact values provided by the computer program actually finds n^* to be equivalent in the two cases.) It is interesting to note that even when we posited that the differences between the eight hospitals uniquely accounted for only 1% of the Y variance, the Model 2 n^* was very close to the Model 1 n^*.

5.6.2 Estimating the Population sR^2 Values

The key decision required in the power analysis necessary for research planning in MRC, and generally the most difficult, is estimating the population ESs. One obviously cannot *know* the various population R^2 values, or the research would be unnecessary. Nevertheless, unless some estimates are made in advance, there is no rational basis for planning. Furthermore, unless they bear some reasonable resemblance to the true state of affairs in the population, sample sizes will be too large or (more often) too small, or, when sample sizes are not under the control of the researcher, the power of the research will be under- or (more often) overestimated.

The best way to proceed is to muster all one's resources of empirical knowledge, both hard and soft, about the substantive field of study and apply them, together with some insight into how magnitudes of phenomena are translated into proportions of variance, in order to make the estimates of the population R^2 values that are required. Some guidance may be obtained from a handbook of power analysis (J. Cohen, 1988), which proposed operational definitions or conventions that link qualitative adjectives to amounts of correlation broadly appropriate to the behavioral sciences. Translated into proportion of variance terms (r^2 or sr^2), these are "small," .01; "medium," .09; and "large," .25. The rationale for these quantities and cautions about their use are given by J. Cohen (1962, 1988).

One may think of f^2 as the approximate percentage of the Y variance *not* accounted for by the other variables (in the error term) that *is* accounted for by the set ($\boldsymbol{B}$) under consideration. With some hesitation, we offer the following as a frame of reference: "small" = .02, "medium" = .15, and "large" = .35. Our hesitation arises from the following considerations. First, there is the general consideration of obvious diversity of the areas of study covered by the rubric "behavioral and social sciences." For example, what is large for a personality psychologist may well be small for a sociologist. The conventional values offered can only strike a rough average. Second, because we are required to estimate two or three distinct quantities (proportions of Y variance), their confection into a single quantity offers opportunities for judgment to go astray. Thus, what might be thought of as a medium-sized expected sR^2_B (numerator) may well result in either a large or quite modest variance ratio, depending on whether the expected contributions to R^2 of sets $\boldsymbol{A}$ and $\boldsymbol{C}$ are small or large. Furthermore, 15% may be appropriately thought of as a "medium" ES in the context of 5 or 10 IVs in a set but seems too small when $k = 15$ or more, indicating that, on the average, these variables account for, at most (.15/15 =) .01 of the Y variance. Nevertheless, conventions have their uses, and the ones modestly offered here

should serve to give the reader *some* sense of the ES to attach to these verbal formulations, particularly when it is hard to cope with estimating the population values themselves. The latter is, as we have said, the preferred route to determining power and sample size. For further discussion of this issue, see J. Cohen (1988, Chapters 8 and 9).

5.6.3 Setting Power for n^*

In the form of power analysis discussed thus far, we find the necessary sample size n^* for a given desired power (given also α and f^2). What power do we desire? If we follow our natural inclinations and set it quite large (say at .95 or .99), we quickly discover that except for very large f^2, n^* gets to be very large, often beyond our resources. (For example, in the first example of Section 5.6.1, the test of I, for $\alpha = .05$ and power $= .99, n^*$ works out to be 905, about double what is required at power $= .80$) If we set power at a low value (say at .50 or .60), n^* is relatively small (for this example, at power $= .50, n^* = 271$), but we are not likely to be content to have only a 50-50 chance of rejecting the null hypothesis (when it is as false as we expect it to be).

The decision as to what power to set is a complex one. It depends upon the result of weighing the costs of failing to reject the null hypothesis (Type II error in statistical inference) against the costs of gathering and processing research data. The latter are usually not hard to estimate objectively, whereas the former include the costs of such imponderables as failing to advance knowledge, losing face, and editorial rejections, and of such painful ponderables as not getting continued research support from funding agencies. This weighing of costs is obviously unique to each investigation or even to each null hypothesis to be tested. This having been carefully done, the investigator can then formulate the power value desired.

Although there will be exceptions in special circumstances, most investigators choose some value in the .70 to .90 range. A value in the lower part of this range may seem reasonable when the dollar cost per case is large or when the more intangible cost of a Type II error in inference is not great (i.e., when rejecting the null hypothesis in question is of relatively small importance). Conversely, a value at or near the upper end of this range would be chosen when the additional cost of collecting and processing cases is not large, or when the hypothesis is an important one.

It has been proposed, in the absence of some preference to the contrary, that power be set at .80 (J. Cohen, 1965, 1988). This value falls in the middle of the .70 to .90 range and is a reasonable one to use as a convention when such is needed.

5.6.4 Reconciling Different n^*s

When more than one hypothesis is to be tested in a given investigation, the application of the methods described earlier will result in multiple n^*s. Because a single investigation will have a single n, these different n^*s will require reconciliation.

For concreteness, assume plans to test three null hypotheses (H_i) whose specifications have resulted in $n_1^* = 100, n_2^* = 300$, and $n_3^* = 400$. If we decide to use $n = 400$ in the study, we will meet the specifications of H_3 and have much more power than specified for H_1 and more for H_2. This if fine if, in assessing our resources and weighing them against the importance of H_3, we deem it worthwhile. Alternatively, if we proceed with $n = 100$ we will meet the specification of H_1 but fall short of the power desires for H_2 and H_3. Finally, if we strike an average of these n^*s and proceed with $n = 267$, we shall have more power than specified for H_1, slightly less for H_2, and much less for H_3.

There is of course no way to have a single n that will simultaneously meet the n^* specifications of multiple hypotheses. No problem arises when resources are sufficient to proceed

with the largest n^*; obviously there is no harm in exceeding the desired power for the other hypotheses and improving the precision of those estimates. But such is not the usual case, and difficult choices may be posed for the investigator. Some help is afforded if one can determine exactly how much power drops from the desired value when n is to be less than n^* for some given hypothesis. Stated more generally, it is useful to be able to estimate the power of a test given some specified n, the inverse of the problem of determining n^* given some specified desired power. The next section is devoted to the solution of this problem.

5.6.5 Power as a Function of n

Thus far, we have been pursuing that particular form of statistical power analysis wherein n^* is determined for a specified desired power value (for given α and f^2). Although this is probably the most frequently useful form of power analysis, we have just seen the utility of inverting n and power, that is, determining the power that would result for some specified n (for given α and f^2). The latter is not only useful in the reconciliation of different n_i^*, but in other circumstances, such as when the n available for study is fixed or when a power analysis is done on a hypothesis post hoc as in a power survey (J. Cohen, 1962). To find power as a function of n, we enter the computer program with n, α, and f^2, and read out the power.

If this calculation is to be done by hand, one needs to use the L tables (J. Cohen, 1988) backward. Enter the table for the significance criterion α to be used in the row for k_B, and read across to find where the obtained L^* falls. Then read off at the column heading the power values that bracket it.[20]

To illustrate: In Section 5.6.1 we considered a test of $R^2_{Y.DI} - R^2_{YD}$ using Model 1 error at $\alpha = .05$, where $k_I = k_B = 10, k_D = k_A = 9$, and $f^2 = .0366$. Instead of positing desired power (e.g., .80) and determining $n^*(= 581)$, let us instead assume that (for whatever reason) we will be using $n = 350$ cases. Enter these values into the computer program to determine the power or alternatively, use hand calculation to find

$$L^* = f^2(n - k - 1), \tag{5.6.5}$$

where k is the number of variables contributing to the R^2 in the denominator of f^2 (the error term), whether one is using Model 1 or Model 2 error. In our illustration $L^* = .0366\ (350 - 9 - 10 - 1) = 12.07$.

Recall that L is itself a function of k_B, α, and power. To find power one simply uses the L tables backward. Enter Appendix E Table E.1 or E.2 for the significance criterion α to be used in the row for k_B, and read across to find where the obtained L^* falls. Then read off at the column heading the power values that bracket it. We find that for this example, $L^* = 12.07$ falls between $L = 11.15$ at power $= .60$ and $L = 13.40$ at power $= .70$. Thus, with $n = 350$ for these specifications, power is between .60 and .70. (Linear interpolation gives us an approximate value of .64, which agrees closely with the computer program value of .65). Power may similarly be found for the other hypotheses to be tested in this data set, with a specified n of 350.

A major advantage of computer programs such as those cited here is the possibility of plotting the power as a function of n and in general obtaining a clearer picture of the consequences for each of the parameters in the equation as a function of changes in other parameters.

---

[20] It is for such applications that the tables provide for low power values (.10 to .60). When a specified n results in low power, it is useful to have some idea of what the power actually is.

5.6.6 Tactics of Power Analysis

We noted earlier that power analysis concerns relationships among four parameters: power, n, α, and ES (indexed by f^2 in these applications). Mathematically, any one of these parameters is determined by the other three. We have considered the cases where n and power are each functions of three others. It may also be useful to exploit the other two possibilities. For example, if one specifies desired power, n, and α for a hypothesis, the computer program will provide the detectable ES, that is, the population f^2 one can expect to detect using this α, with probability given by the specified power desired in a sample of n cases. One can also determine what α one should use, given the ES, desired power, and a sample of size n.

It must be understood that these mathematical relationships among the four parameters should serve as tools in the service of the behavioral scientist turned applied statistician, not as formalisms for their own sake. We have in the interest of expository simplicity implicitly assumed that when we seek to determine n^*, there is only one possible α, one possible value for desired power, and one possible ES. Similarly, in seeking to determine power, we have largely operated as if only one value each for α, ES, and n is to be considered. But the realities of research planning often are such that more than one value for one of these parameters can and indeed must be entertained. Thus, if one finds that for a hypothesis for which $\alpha = .01$, power $= .80$, and $f^2 = .04$, the resulting n^* is 600, and this number far exceeds our resources, it is sensible to see what n^* results when we change α to .05. If that is also too large, we can invert the problem, specify the largest n we can manage, and see what power results for this n at $\alpha = .05$. If this is too low, we might examine our conscience and see if it is reasonable to entertain the possibility that f^2 is larger, perhaps .05 instead of .04. If so, what does that do for power at that given n? At the end of the line of such reasoning, the investigator either has found a combination of parameters that makes sense in the substantive context or has decided to abandon the research, at least as originally planned. Many examples of such reasoning among a priori alternatives are given in J. Cohen (1988).

With the multiple hypotheses that generally characterize MRC analysis, the need for exploring such alternatives among combinations of parameters is likely to increase. If H_1 requires $n^* = 300$ for desired power of .80, and 300 cases give power of .50 for H_2 and .60 for H_3, etc., only a consideration of alternate parameters for one or more of these hypotheses may result in a research plan that is worth undertaking.

To conclude this section with an optimistic note, we should point out that we do not always work in an economy of scarcity. It sometimes occurs that an initial set of specifications results in n^* much smaller than our resources permit. Then we may find that when the parameters are made quite conservative (for example, $\alpha = .01$, desired power $= .95, f^2$ at the lower end of our range of reasonable expectation), we still find n^* smaller than our resources permit. We might then use the power analysis to avoid "overkill," and perhaps use our additional resources for obtaining better data, for testing additional hypotheses, or even for additional investigations.

5.7 STATISTICAL INFERENCE STRATEGY IN MULTIPLE REGRESSION/CORRELATION

5.7.1 Controlling and Balancing Type I and Type II Errors in Inference

In the preceding sections we have set forth in some detail the methods of hypothesis testing and power analysis for sets. Testing for significance is the procedure of applying criteria designed to control at some rate α, the making of a Type I error in inference, that is, the error of rejecting

true null hypotheses or, less formally, finding things that are not there. Power analysis focuses on the other side of the coin of statistical inference, and seeks to control the making of a Type II error, the error of failing to reject false null hypotheses and failing to find things that *are* there. Of course current thinking (e.g., Harlow, Mulaik, & Steiger, 1997) notes that it is likely very rare that an effect in a population will be *precisely* zero, and that failing to find an effect to be significantly different from zero should *never* be so interpreted. However, the making of provisional judgments about the presence in the population of an effect of practical or material magnitude is, in many cases, aided by the use of tests of statistical significance (Abelson, 1995, 1997). Thus, one fundamental demand of an effective strategy of statistical inference is the balancing of Type I and Type II errors in a manner consistent with the substantive issues of the research. In practice, this takes the form of seeking to maintain a reasonably low rate of Type I errors while not allowing the rate of Type II errors to become unduly large or, equivalently, maintaining good power for realistic alternatives to the null hypothesis.

For any discrete null hypothesis, given the usual statistical assumptions and the requisite specification, the procedures for significance testing and power analysis are relatively simple, as we have seen. When one must deal with multiple hypotheses, however, statistical inference becomes exceedingly complex. One dimension of this complexity has to do with whether the Type I error rate is calculated per hypothesis, per group of related hypotheses ("experiment-wise"), or for even larger units ("investigation-wise"). Another is whether α is held constant over the multiple hypotheses or is varied. Still another is whether the hypotheses are planned in advance or stated after the data have been examined (post hoc) the latter being sometimes referred to as "data snooping." And there are yet others. Each of the possible combinations of these alternatives has one or more specific procedures for testing the multiple hypotheses, and each procedure has its own set of implications to the statistical power of the tests it performs.

An example may help clarify the preceding. Assume an investigator is concerned with hypotheses about the means of a dependent variable across levels of a research factor, G, made up of 6 ($= g$) groups. Any of the following kinds of multiple hypotheses may be of interest, and each has its own procedure(s):

1. ***All simple comparisons between means.*** There are $g(g-1)/2 = 15$ different pairs of means and 15 simple comparisons and their null hypotheses. Assume each is t tested at $\alpha = .05$; thus the Type I error rate *per hypothesis* is controlled at .05. But if, in fact, the population means are all equal, it is intuitively evident that the probability that *at least one* comparison will be "significant" (i.e., the experiment-wise error rate) is greater than .05. The actual rate for $g = 6$ is approximately .40.[21] Thus, the separate α's escalate in this case to .40. This error rate may well be unacceptable to the investigator, and almost certainly so to scientific peers. But each t test at $\alpha = .05$ will be relatively powerful.

There is a large collection of statistical methods designed to cope with the problem of making all simple comparisons among g means. These vary in their definition of the problem, particularly in their conceptualization of Type I error, and they therefore vary in power and in their results. For example, the Tukey HSD test (Winer, 1971, pp. 197–198) controls the experiment-wise error rate at α. The Newman-Keuls test and the Duncan test both approach Type I error via "protection levels" that are functions of α, but the per-hypothesis Type I error risks for the former are constant and for the latter vary systematically (Winer, 1971, pp. 196–198). Bonferroni tests employ the principle of dividing an overall α into as many (usually equal) parts as there are hypotheses, and then setting the per-hypothesis significance criterion accordingly;

[21] The calculation requires special tables and the result depends somewhat on sample size. Some other experimentwise error rates for these conditions are (approximately) for $g = 10$, .60 and for $g = 20$, .90. Even for $g = 3$ it is .13. Only for $g = 2$ is it .05.

thus, for $\alpha = .05$, each of the 15 comparisons would be tested with significance criterion set at $\alpha = .05/15 = .0033$ (R. G. Miller, 1966, pp. 67–70). The preceding tests of all pairs of means are the most frequently employed, and by no means exhaustive (Games, 1971).

One of the oldest and simplest procedures for all pairs of g means is Fisher's "protected t" (or *LSD*) test (Carmer & Swanson, 1973). First, an ordinary (ANOVA) overall F test is performed on the set of g means ($df = g - 1, n - g$). If F is not significant, no pair-wise comparisons are made. Only if F is significant at the α criterion level are the means compared; this being done by an ordinary t test. The t tests are protected from large experiment-wise Type I error by the requirement that the preliminary F test must meet the α criterion. As we will see, this procedure is readily adapted for general use in MRC analysis.

Note that each of these tests approaches the control of Type I errors differently, and that therefore each carries different implications to the rate of Type II errors and hence to the test's power.

2. *Some simple comparisons between means.* With g means, only differences between some pairs may be of interest. A frequent instance of this case occurs when $g - 1$ of the groups are to be compared with a single control or reference group, which thus calls for $g - 1$ hypotheses that are simple comparisons. In this special case the Dunnett test, whose α is controlled experiment-wise, applies (Winer, 1971, pp. 201–204). For the more general case where not all pair-wise hypotheses are to be tested, protected t and Bonferroni tests (and others) may be used. Again these different tests, with their different strategies of Type I error control have different power characteristics.

3. *Orthogonal comparisons.* With g groups, it is possible to test up to $g - 1$ null hypotheses on comparisons (linear contrasts) that are orthogonal (i.e., independent of each other). These may be simple or complex. A complex comparison is one that involves more than two means, for example, M_1 versus the mean of M_3, M_4, M_5, or the mean of M_1 and M_2 versus the mean of M_3 and M_5. These two complex "mean of means" comparisons are, however, not orthogonal. (The criterion for orthogonality of contrasts and some examples are given in Chapter 8.) When the maximum possible number of orthogonal contrasts, $g - 1$, are each tested at α, the experiment-wise Type I error rate is larger, specifically, it is approximately $1 - (1 - \alpha)^{g-1} = .226$. It is common practice, however, not to reduce the per-contrast rate α below its customary value in order to reduce the experiment-wise rate when orthogonal contrasts are used (Games, 1971).

Planned (a priori) orthogonal comparisons are generally considered the most elegant multiple comparison procedure and have good power characteristics, but alas, they can only infrequently be employed in behavioral science investigations because the questions to be put to the data are simply not usually independent (e.g., those described in paragraphs 1 and 2 previously discussed and in the next paragraph).

4. *Nonorthogonal, many, and post hoc comparisons.* Although only $g - 1$ orthogonal contrasts are mathematically possible, the total number of different mean of means contrasts is large, and the total number of different contrasts of all kinds is infinite for $g > 2$. An investigator may wish to make more than $g - 1$ (and therefore necessarily nonorthogonal) comparisons, or may wish to make comparisons that were not contemplated in advance of data collection, but rather suggested post hoc by the sample means found in the research. Such "data snooping" is an important part of the research process, but unless Type I error is controlled in accordance with this practice, the experiment-wise rate of spuriously "significant" t values on comparisons becomes unacceptably high. The Scheffé test (Edwards, 1972; Games, 1971; R. G. Miller, 1966) is designed for these circumstances. It permits *all possible* comparisons, orthogonal or nonorthogonal, planned or post hoc, to be made subject to a controlled experiment-wise Type I error rate. Because it is so permissive, however, in most applications it results in very conservative tests, i.e., in tests of relatively low power (Games, 1971).

The reasons for presenting this brief survey are twofold. The first is to alert the reader to the fact that for specific and well defined circumstances of hypothesis formulation and Type I error definition, there exist specific statistical test procedures. But even for the simple case of a single nominally scaled research factor G made up of g groups, the basis for choice among the alternatives is complex. Indeed, an entire book addressed to mathematical statisticians has been written in this area (R. G. Miller, 1966).

The second reason for presenting this survey is to emphasize the fact that given the variety of approaches to the conception of Type I errors, there are differential consequences to the rate of Type II errors and thus to the statistical power of the tests. Conventional statistical inference is effectively employed only to the extent that Type I and Type II error risks are appropriately balanced. The investigator can neither afford to make spurious positive claims (Type I) nor fail to find important relationships (Type II). Since, all other things equal, these two types of errors are inversely related, some balance is needed. Yet the complexity that we encountered earlier when confronted only with the special case of a single nominal scale makes it clear that any effort to treat this problem in comprehensive detail is far outside the bounds of practicality and not in keeping with this book's purpose and data-analytic philosophy, nor with the needs of its intended audience.

What is required instead are some general principles and simple methods that, over the wide range of circumstances in research in the behavioral and social sciences, will serve to provide a practical basis for keeping both types of errors acceptably low and in reasonable balance. The major elements of this approach include parsimony in the number of variables employed, the use of a hierarchical strategy, and the adaptation of the Fisher protected t test to MRC.

5.7.2 Less Is More

A frequent dilemma of the investigator in behavioral science arises in regard to the number of variables she will employ in a given investigation. On the one hand is the need to make sure that the substantive issues are well covered and nothing is overlooked, and on the other is the need to keep in bounds the cost in time, money, and increased complexity that is incurred with an increase in variables. Unfortunately, the dilemma very often is resolved in favor of having more variables to assure coverage.

In addition to the time and money costs of more variables (which are frequently negligible, hence easily incurred), there are more important costs to the validity of statistical inference that are very often overlooked. The more variables, dependent or independent, there are in an investigation, the more hypotheses are tested (either formally or implicitly). The more hypotheses are tested, the greater the probability of occurrence of spurious significance (investigation-wise Type I error). Thus, with 5 dependent and 12 independent variables analyzed by 5 MRC analyses (one per dependent variable), there are a total of 60 potential t tests on null hypotheses for partial coefficients alone. At $\alpha = .05$ per hypothesis, if all these null hypotheses were true, the probability that one or more ts would be found "significant" approaches unity. Even at $\alpha = .01$ per hypothesis, the investigation-wise rate would be in the vicinity of .50.[22] It is rare in research reports to find their results appraised from this perspective, and many investigations are not reported in sufficient detail to make it possible for a reader to do so—variables that "don't work" may never surface in the final report of a research.

One might think that profligacy in the number of variables would at least increase the probability of finding true effects when they are present in the population, even at the risk of

[22]Because the 60 tests are not independent, exact investigation-wise error rates cannot be given. If they were independent, the two investigation-wise Type I error rates would be $(1 - .95)^{60} = .954$ (for $\alpha = .05$) and $(1 - .99)^{60} = .453$ (for $\alpha = .01$).

finding spurious ones. But another consideration arises. In each MRC, the greater the number of IVs ($= k$), the lower the power of the test on each IV (or set of IVs). We have seen this in several ways. First, for any given n, the error $df = n - k - 1$ and are thus diminished as k increases. Second, a glance at the L tables (Appendix Tables E.1 and E.2) quickly reveals that all other things being equal, as k_B increases, L increases, and therefore power decreases for any given f^2, α, and n. Also, it is likely that as k increases, the R_is among the IVs increase, which in turn increases the standard errors of partial coefficients (e.g., SE_{B_i}s) and reduces the t_is and hence the power. Thus, having more variables when fewer are possible increases the risks of both finding things that are not so and failing to find things that are. These are serious costs, indeed.

Note that a large n does not solve the difficulties in inference that accompany large numbers of variables. True, the error df will be large, which, taken by itself, increases power. The investigation-wise Type I error rate depends, however, on the number of hypotheses and not on n. And even potentially high power conferred by large n may be dissipated by large k, and by the large R_is (low tolerances) that large k may produce.

Within the goals of a research study, the investigator usually has considerable leeway in the number of variables to include, and too frequently the choice is made for more rather than fewer. The probability of this increases with the "softness" of the research area and the degree to which the investigation is exploratory in character, but no area is immune. When a theoretical construct is to be represented in data, a large number of variables may be used to represent it in the interest of "thoroughness" and "just to make sure" that the construct is covered. It is almost always the case, however, that the large number is unnecessary. It may be that a few (or even one) of the variables are really central to the construct and the remainder peripheral and largely redundant. The latter are better excluded. Or, the variables may all be about equally related to the construct and define a common factor in the factor-analytic sense, in which case they should be combined into an index, factor score, or sum (or treated in a latent variable model, see Chapter 12). The latter not only will represent the construct with greater reliability and validity, but will do so with a single variable (recall in this connection the lessons of unit weighting in Section 3.8.3). Perhaps more than one common factor is required, but this is still far more effective than a large number of single variables designed to cover (actually smother) the construct. These remarks obtain for constructs in both dependent and independent variables.

Other problems in research inference are attendant upon using many variables in the representation of a construct. When used as successive dependent variables, they frequently lead to inconsistent results that, as they stand, are difficult to interpret. When used as a set of IVs, the partialing process highlights their uniqueness, tends to dissipate whatever common factor they share, may produce paradoxical suppression effects, and is thus also likely to create severe difficulties in interpretation.

5.7.3 Least Is Last

The hierarchical model with Model 1 error can be an important element in an effective strategy of inference. We have already commented briefly on its use when IVs may be classified into levels of research relevance (Section 5.3.2). This type of application is appropriate in investigations that are designed to test a small number of central hypotheses but may have data on some additional research factors that are of exploratory interest, and also in studies that are largely or wholly exploratory in character. In such circumstances, the IVs can be grouped into two or more classes and the classes ordered with regard to their status in centrality or relevance. Each class is made up of one or more research factors, which are generally sets of IVs. Thus, for example, the first group of IVs may represent the research factors whose effects the research was designed to appraise, the second some research factors of distinctly secondary

interest, and the third those of the "I wonder if" or "just in case" variety. Depending on the investigator's interest and the internal structure of the research, the levels of the hierarchy may simply be the priority classes or one or more of these may also be internally ordered by research factors or single IVs.

The use of the hierarchical model, particularly when used with Model 1 error at each priority class level, then prevents variables of lower priority, which are likely to account uniquely for little Y variance, from reducing the power of the tests on those of higher priority by stealing some of their variance, increasing the standard errors of their partial coefficients, and reducing the df for error. In using this stratagem, it is also a good idea to lend less credence to significant results for research factors of low priority, particularly so when many IVs are involved, because the investigation-wise Type I error rate over their IVs is likely to be large. We thus avoid diluting the significance of the high priority research factors. This is in keeping with the sound research philosophy that holds that what is properly obtained from exploratory research are not conclusions, but hypotheses to be tested in subsequent investigations.

When hierarchical MRC is used for relevance ordering, it is recommended that Model 1 error be used at each level of relevance, that is, the first class ($\boldsymbol{U}$) made up of the centrally relevant research factors uses $1 - R^2_{Y\boldsymbol{U}}$ (with $df = n - k_{\boldsymbol{U}} - 1$) as the error term for its F and t tests, the second class ($\boldsymbol{V}$) made up of more peripheral research factors used $1 - R^2_{Y.\boldsymbol{UV}}$ (with $df = n - k_{\boldsymbol{U}} - k_{\boldsymbol{V}} - 1$), and so on. This tends to make it probable that the tests at each level have minimum error variance per df and thus maximal power. Of course, it is always possible that a test using Model 1 error is negatively biased by an important source of variance remaining in its error term, but the declining gradient of unique relevance in this type of application makes this rather unlikely. It is, in fact, in analyses of this kind that Model 1 error has its major justification and use.

We summarize this principle, then, as "least is last"—when research factors can be ordered as to their centrality, those of least relevance are appraised last in the hierarchy and their results taken as indicative rather than conclusive.

5.7.4 Adaptation of Fisher's Protected t Test

The preceding sections of the chapter have been devoted to the use of sets of IVs as units of analysis in MRC. We have seen how Y variance associated with a set or partialed set can be determined, tested for significance, and power analyzed. The chapters that follow show how research factors can be represented as sets of IVs. It should thus not come as a surprise that in formulating a general strategy of statistical inference in MRC, we accord the set a central role.

In Section 5.7.1, in our brief review of alternative schemes of testing multiple hypotheses for the special case where the research factor is a nominal scale $\boldsymbol{G}$, we noted that among the methods available for the comparison of pairs of groups' means was a method attributed to R. A. Fisher: The usual ANOVA overall F test over the set of g means is first performed, and if it proves to be significant at the α level specified, the investigator may go on to test any or all pairs at the same α level, using the ordinary t test for this purpose, and interpret results in the usual way. If F fails to be significant, no t tests are performed—all g population means are taken to be potentially equivalent, based on the evidence, so that no difference between means (or any other linear contrast function of them) can be asserted, whatever value it may yield. This two-stage procedure combines the good power characteristics of the individual t tests at a conventional level of α with the protection against large experiment-wise Type I error afforded by the requirement that the overall F also meet the α significance criterion. For example, in Section 5.7.1, we saw that when all 15 pair-wise comparisons among 6 means are performed by t tests using $\alpha = .05$ per comparison, the probability that one or more will be found "significant" when all 6 population means are equal (i.e., the experiment-wise Type I

error rate) is about .40. But if these tests are performed only if the overall F meets the .05 criterion, the latter prevents us from comparing the sample means 95% of the time when the overall null hypothesis is true. Thus, the t tests are protected from the mounting up of small per-comparison α to large experiment-wise error rates.

The virtues of simplicity and practicality of the protected t test procedure are evident. What is surprising is how effective it is in keeping Type I errors low while affording good power. In an extensive investigation of 10 pair-wise procedures for means compared empirically over a wide variety of conditions, it was unexcelled in its general performance characteristics (Carmer & Swanson, 1973).

To adapt and generalize the protected t test to the MRC system, we use the framework of sets as developed in this chapter and used throughout the book. We discussed in Section 5.4.1 the principle that information on research factors of all kinds can be organized into sets of IVs for structural and functional reasons and in the ensuing sections illustrated how these sets may then serve as the primary units of MRC analysis. Now, the protected t test described previously covers only one type of set—a research factor defined by a nominal scale (i.e., a collection of g groups). We generalize the protected t procedure, applying it to the functional sets that organize an MRC analysis, whatever their nature. Specifically,

1. The MRC analysis proceeds by sets, using whatever analytic structure (hierarchical, simultaneous, or intermediate) is appropriate.
2. The contribution to Y variance of each set (or partialed set) is tested for significance at the α level by the appropriate standard F test of Eqs. (5.6.2), (5.6.6), or their variants.
3. If the F for a given set is significant, the individual IVs (aspects) that make it up are each tested for significance at α by means of a standard t test (or its equivalent $t^2 = F$ for numerator $df = 1$). It is the partial contribution of each X_i that is t tested, and any of the equivalent tests for its sr_i, pr_i, or B_i may be used [Eq. (3.6.8)]. All standard MRC computer programs provide this significance test, usually for B_i.
4. If the setwise F is not significant, no tests on the set's constituent IVs are permitted. (The computer program will do them automatically, but the t_is, no matter how large, are ignored.) Overriding this rule removes the protection against large setwise Type I error rates, which is the whole point of the procedure.

This procedure is effective in statistical inference in MRC for several reasons. Because the number of sets is typically small, the investigation-wise Type I error rate does not mount up to anywhere nearly as large a value over the tests for sets as it would over the tests for the frequently large total number of IVs. Then, the tests of single IVs are protected against inflated setwise Type I error rates by the requirement that their set's F meet the α significance criterion. Further, with Type I errors under control, both the F and t tests are relatively powerful (for any given n and f^2). Thus, both types of errors in inference are kept relatively low and in good balance.

To illustrate this procedure, we return to the running example of this chapter: length of hospital stay (Y) for $n = 500$ psychiatric patients was regressed hierarchically on three sets of IVs, demographic ($\boldsymbol{D}, k_{\boldsymbol{D}} = 9$), illness ($\boldsymbol{I}, k_{\boldsymbol{I}} = 10$), and a nominal scale of 8) hospitals ($\boldsymbol{H}, k_{\boldsymbol{H}} = 7$), in that order. Using F tests with Model 1 error and $\alpha = .05$ as the criterion for significance, it was found that set $\boldsymbol{D}$ was significant. Note that the primary focus on sets helps control the investigation-wise Type I error risk. Even for $\alpha = .05$. The latter is in the vicinity of .14; the more conservative $\alpha = .01$ for significance per set would put the investigation-wise Type I error rate in the vicinity of .03.[23]

[23] Again, because the tests are not independent, exact rates cannot be determined. The rates given are "ballpark" estimates computed on the assumption of independence, that is, $1 - .95^3$ and $1 - .99^3$, respectively.

Because set $\boldsymbol{D}$ was found to be significant, one may perform a t test (at $\alpha = .05$) on each of the nine IVs that represent unique aspects of patient demography.[24] Because these t tests are protected by the significance of F, the mounting up of set-wise Type I error is prevented. Without this protection, the set-wise error rate for nine t tests, each at $\alpha = .05$, would be in the vicinity of $(1 - .95)^9 = .37$.

Set $\boldsymbol{I}$'s increment (over $\boldsymbol{D}$) to R^2 was found to be nonsignificant, so no t tests on the unique (within $\boldsymbol{I}$ and $\boldsymbol{D}$) contributions of its 10 IVs are admissible. It would come as no surprise if the computer output showed that 1 of these 10 t values exceeded the nominal $\alpha = .05$ level. With 10 tests each at $\alpha = .05$, the set-wise Type I error rate would be large. Although the tests are not independent, the "ballpark" estimate of the error rate computed with an assumption of independence is $(1 - .95)^{10} = .40$. In the protected t strategy, the failure of F for the set to be significant is treated so as to mean that all IVs in the set have zero population partial coefficients, a conclusion that cannot be reversed by their individual ts.

Finally, the increment of set $\boldsymbol{H}$ (over $\boldsymbol{D}$ and $\boldsymbol{I}$) to R^2 was found to be significant, and its constituent $k_{\boldsymbol{H}} = g - 1 = 7$ IVs were t tested, and the significant ones interpreted. These seven aspects of hospital-group membership may include simple (pair-wise) or complex (involving more than two hospitals) comparisons among the eight hospital Y means, depending on which of several different methods of representing group membership was employed. (The latter is the subject matter of Chapter 8). The method used in this example was presumably chosen so that the seven partialed IVs would represent those comparisons (aspects) of central interest to the investigator and their protected ts test these aspects. Thus far, we have proceeded as with any other set. However, because $\boldsymbol{H}$ is a nominal scale and is thus made up of g groups, we admit under the protection of the F test any comparisons of interest in addition to the 7 $(= g - 1)$ carried by the IVs. Thus, in full compliance with both the letter and spirit of Fisher's original protected t test, one could t test any of the $(8 \times 7)/2 = 28$ pair-wise simple comparisons (not already tested) that may be of substantive interest.[25]

We reiterate the generality of our adaptation of the protected t test to MRC. Whether one is dealing with one set or several, whether they are related to Y hierarchically or simultaneously, whether error Model 1 or 2 is used, and whatever the substantive nature of the set(s), the same procedure is used: The first order of inference is with regard to the set(s), and only when a set's significance is established by its F test are its contents further scrutinized for significance.

In using the protected t procedure, it may happen that after a set's F is found to be significant, none of its IVs yields a significant t. This is apparently an inconsistency, because the significant F's message is that at least one of the IVs has a nonzero population partial coefficient, yet each t finds its null hypothesis tenable. A technically correct interpretation is that collectively (set-wise) there is sufficient evidence that there is something there, but individually, not enough evidence to identify what it is. A risky but not unreasonable resolution of this dilemma is to tentatively interpret as significant any IV whose t is almost large enough to meet the significance criterion; whatever is lost by the inflation of the Type I error is likely to be compensated by the reduction of Type II error and the resolution of the apparent inconsistency. It is also very prudent to examine the tolerance for each variable $(= 1 - R^2_{i.123\ldots(i)\ldots k})$ to try to identify high levels of correlations that may have greatly increased the standard errors for some variables that have apparent high levels of influence as suggested by β. Fortunately, the occurrence of this anomaly is rare, and virtually nonexistent when error df are not very small.

---

[24] A refinement of this procedure would be to test the sR^2s for subsets (for example, the nominal scale for ethnicity) by F and then perform t tests on the subset's IVs only if F is significant. This gives added protection to the t tests, which is probably a good idea when k for the set is large, as it is here.

[25] In compliance at least with the spirit of Fisher's procedure, one could also test any complex comparisons of interest, but there would usually be few, if any, that had not been included as IVs.

Another difficulty that may arise with the protected t test is best described by a hypothetical example. Assume, for a set made up of many IVs, that one or two of them have large population partial coefficients and that the remainder all have population partial coefficients equal to zero. Now when we draw a random sample of reasonable size from this population, we will likely find that the F for the set is statistically significant. This result is quite valid, because this F tests the composite hypothesis that all the IVs in the set have zero population coefficient, and we have posited that this is not true. But using the protected t test, the significance of F confers upon us the right to t test *all* the IVs, including those for which the null hypothesis *is* approximately true. For that large group of IVs, the subset-wise Type I error rate will obviously mount up and be high. Of course, we do not know the state of affairs in the population when we analyze the sample. We cannot distinguish between an IV whose t is large because its null hypothesis is false from one (of many t values) that is large because of chance. Obviously, in circumstances such as these our t tests are not as protected as we should like.

Fortunately, a means of coping with this problem is available to us: We invoke the principle of Section 5.7.2—"less is more." By having few rather than many IVs in a set, a significant F protects fewer ts for IVs whose null hypotheses may be true and inhibits the mounting up of Type I error. Moreover, if the investigator's substantive knowledge is used to carefully select or construct these few IVs, fewer still of the ts are likely to be testing true null hypotheses. In this connection, we must acknowledge the possibility that sets $\boldsymbol{D}$ and $\boldsymbol{I}$ in our running example are larger than they need have been; the former may be benefited from reduction by a priori selection and the latter by either selection or factor-analytic reduction.

5.7.5 Statistical Inference and the Stage of Scientific Investigations

Some of the problems of statistical inference can be seen in better perspective when the stage or level of information already available about the phenomena under investigation is taken into account. In early studies, tests of statistical inference will be useful in establishing the direction of effect and in aiding decisions about the probable approximate magnitude of effects in certain populations. These estimates will then be useful in planning future studies. As we will see in subsequent chapters, such significance tests on individual variables or variable sets can be very useful in aiding decisions about whether certain covariate sets will be needed in current or future studies, whether nonlinear functions of variables contribute materially to the prediction of Y, and whether important interactive effects among IVs are present. All such significance tests have to be treated as providing very tentative answers to our questions, that is, answers badly in need of confirmation by replication. (However, the unwary researcher is warned about the problems associated with modest power in "replicating" studies).

Having established the presence and direction of effects, latter stage investigations may be devoted to improving the precision of estimated effects and determining the limits of generalizations across populations and changes in methods, measures, or controlled variables. When this sequence is followed it is likely that problems associated with Type I errors will fade. Problems associated with inadequate statistical power should also be minimized once it is determined what population ES is likely, providing that such estimates can be and are used in the planning of subsequent studies.

5.8 SUMMARY

This chapter introduces five of the major strategic considerations for the employment of MRC analyses to answer research questions. Sections 5.1 and 5.2 discuss considerations in selecting

the coefficients that will best answer particular research questions. Section 5.2 also discusses methods of making regression coefficients maximally informative. Sections 5.3 and 5.4 present the hierarchical analyses of individual IVs and of sets of IVs as a strategic aid to answering research questions. Section 5.4 presents the utility of considering IVs in sets, and multiple partial correlation coefficients (the ballantine again). Sections 5.5 and 5.6 present significance testing and power analysis for sets of variables. The final section considers strategies for controlling and balancing Type I and Type II errors in making inferences from findings.

There are a number of different correlation and regression coefficients produced by MRC analyses, and each of them is optimal for particular kinds of research questions. The squared semipartial correlation tells us how much of the variance in Y is uniquely attributable to that IV, a figure that is particularly useful in situations in which the predictive utility or a comparison of predictive utility is at issue. A comparison of a zero-order r^2_{YX} with $sr^2_{X \cdot W}$ tells us how much the prediction of Y by X is attributable to W. In contrast, a comparison of $B_{Y \cdot X}$ with $B_{YX \cdot W}$ tells us whether averaged over the values of W, changes in X had the same consequences for Y that they had when W was ignored. Some investigations may target the question of differential partialed correlations of two different predictors with Y. A series of questions about differences between populations in squared zero-order, partial, or semipartial correlations of variance proportions may be appropriate targets of research studies (Section 5.1).

In order for zero-order and partialed B to provide useful answers to research questions they must reflect meaningful units of Y and the IV in question. When meaningful units are used, B is the premier causal indicator. Meaningful units may come from long familiarity and established conventions. When units represent counts in both Y and X, an alternative to B is elasticity (E), the percent change on Y per percent change in X, measured at the mean. When measures are novel, there are several options to be preferred to the simple sum that is frequently used. These include POMP scores (percent of the maximum possible score) and item averages (when all items have the same response options). Often the sample's *sd* may be used as a scale; if both X and Y are so standardized zero-order and partial $B = \beta$ (Section 5.2).

The choice of the analytic model will determine the amount of information extracted from a data set. Hierarchical analysis allows appropriate consideration of causal priorities and removal of confounding variables. It may also be used to reflect the research relevance or structural properties of variables. An alternative strategy, "step-wise" MRC, in which IVs are entered in a sequence determined by the size of their increment to R^2 is also discussed. Use of this strategy is generally discouraged because of the necessarily post hoc nature of interpretation of the findings and the substantial probability of capitalizing on chance (Section 5.3).

Sets of IVs may be treated as units or fundamental entities in the analysis of data. From this perspective, the single IVs and single group of k IVs treated in Chapter 3 are special cases. Two types of sets are described: sets that come about because of the *structure* of the research factors they represent (for example, religion is represented as a set of IVs because it is a categorical variable), and sets that have a specific *function* in the logic of the research, such as a set of control variables that must be adjusted for, or a set of variables that collectively represent "demographic characteristics"). Groups of sets are also described, and such set aggregates are treated simply as sets (Section 5.4.1).

The simultaneous and hierarchical procedures of MRC are shown to apply to sets as units of analysis, and it is shown that Y variance associated with set $\boldsymbol{A}$ may be partialed from that associated with set $\boldsymbol{B}$, just as with single IVs. For h sets, the simultaneous procedure appraises Y variance for a given set with the remaining $h - 1$ sets partialed. The hierarchical procedure orders the sets into an a priori hierarchy and proceeds sequentially: For each set in hierarchical order of succession, all higher level sets (and no lower level sets) are partialed. The chief quantities of interest are the increments of Y variance accounted for by each set uniquely, relative to sets of higher order of priority (Section 5.4.2).

The ballantine for sets is presented as a device for visualizing proportions of Y variance associated with sets ($\boldsymbol{A}$,$\boldsymbol{B}$) and with partialed sets, analogously with those for single IVs. The increments referred to above are squared multiple semipartial correlations and represent the proportion of total Y variance associated with $\boldsymbol{B} \cdot \boldsymbol{A}$. Similarly, we define the squared multiple partial correlation of $\boldsymbol{B}$ as the proportion of Y variance not accounted for by $\boldsymbol{A}$ that is associated with $\boldsymbol{B} \cdot \boldsymbol{A}$. These two statistics are compared and exemplified. As with single IVs, the troublesome area of overlap of sets $\boldsymbol{A}$ and $\boldsymbol{B}$ with Y, area c, cannot be interpreted as a proportion of variance because it may be negative, in which case we have an instance of suppression between sets (Section 5.4.3).

A general test of statistical significance for the Y variance due to a partialed set $\boldsymbol{B} \cdot \boldsymbol{A}$ is presented. Two error models for this test are described. Model 1 error is $1 - R^2_{Y.\boldsymbol{AB}}$, with sets other than $\boldsymbol{A}$ or $\boldsymbol{B}$ (collectively, set $\boldsymbol{C}$) ignored. Model 2 error is $1 - R^2_{Y.\boldsymbol{ABC}}$, so the Y variance unique to $\boldsymbol{C}$ (together with its *df*) is additionally excluded from error in the significance test of $\boldsymbol{B} \cdot \boldsymbol{A}$s Y variance. The applicability of the two error models to the hierarchical and simultaneous procedures of MRC is described and exemplified (Section 5.5).

Methods of statistical power analysis for partialed sets, necessary for research planning and assessment, together with the use of a computer program or hand calculations are presented. The determination of n^*, the necessary sample size to attain a desired degree of power to reject the null hypothesis at the α level of significance for a given population effect size (ES), is given for both error models, as well as the means of estimating power that results from a specified n, α, and ES. The section concludes with a discussion of the tactics of power analysis in research planning, particularly those of specifying alternative combinations of parameters and studying their implications (Section 5.6).

Finally, the issue of a general strategy of statistical inference in MRC is addressed. One element of such a strategy involves minimizing the number of IVs used in representing research factor constructs by judicious selection and the use of composites. Another exploits the hierarchical procedure of MRC and the use of Model 1 error in significance testing. A generalization of the protected t test is offered as a simple but effective means of coping with the multiplicity of null hypotheses in MRC. This procedure prevents the rapid inflation of set-wise and investigation-wise Type I error that would occur if the individual ts were not so protected and at the same time enjoys the good power characteristics of the t test. When an investigation is placed in a sequence of scientific research, problems associated with both Type I and Type II errors should be minimized. Early analyses and replications may establish the presence, direction, and estimated magnitude of effects. Subsequent research can be planned with realistic levels of statistical power. This work can then improve the precision of estimates and their variability across changes in methods, measures, and populations (Section 5.7).