

RESEARCH FORUM SERIES

Core Curriculum Module E

Working with Large Databases

Part 1: "Preliminary Considerations in Working with Large Data Sets on Microcomputers -- or -- Power Processing with an Attitude"

Part 2: "Sooner Rather Than Later -- or -- Planning for Results through the Application of a priori comparisons"

Mark C. Fulcomer, PhD

Adjunct Associate Professor

Department of Family Medicine

UMDNJ-Robert Wood Johnson Medical School

I. Background

A variety of public health, vital statistics, administrative, and large databases exist that can help address important primary care/health services research issues related to access to care, service utilization, quality, and health outcomes.

Combined with widespread availabilities and low costs of huge data files on easily accessible media such as CDs, the ever-increasing capacities of microcomputers (in terms of processing speeds, hard disk storage, and memory) enable researchers to perform analyses on the desktop that were until recently almost unimaginable even on mainframe computers.

II. Aim

This two-session module will describe and then later demonstrate some important considerations that often arise in utilizing "large" data sources. The first session ("Whys and Wherefores") addresses some early questions confronting the researcher and concludes with some examples of preliminary operations and analyses. A second, follow-up session ("Dancing with Dirty Data") will be scheduled in small groups of up to three fellows to allow some practical "hands-on" work with large files with copies of appropriate software being supplied.

III. Learning Objectives

Part 1: Preliminary Considerations in Working with Large Data Sets on Microcomputers -- or -- Power Processing with an Attitude

- How to plan and carry out initial handling of large data sets (e.g., definitions, documentation, and downloading)

- Methods to assess the "quality" of the information in a file and to ensure appropriate maintenance (e.g., confidentiality, conversion, compression, and backups)
- Steps to prepare already-existing data in a file for subsequent analyses (e.g., variable selection and recoding)
- Combining two or more data sources (e.g., merging, matching, and linking records)

Part 2: Sooner Rather Than Later -- or -- Planning for Results through the Application of *a priori* comparisons

- Some benefits derived from the careful planning of comparisons and trends before data are analyzed
- The simple principles for constructing orthogonal contrasts and polynomials
- How these techniques can enhance the analyses of experimental and non-experimental data alike, particularly by improving linkages between literature reviews and the testing of hypotheses
- The application of these techniques to some "messy" analytic problems, including those involving unequal sample sizes and missing data

IV. Teaching Method(s)

Two 1.5 hour seminars -- didactic lectures, group discussion, and hands-on exercises.

V. Content Outline

Part 1: Preliminary Considerations in Working with Large Data Sets on Microcomputers -- or -- Power Processing with an Attitude

(see attached handout).

Part 2: Sooner Rather Than Later -- or -- Planning for Results through the Application of *a priori* comparisons

(see attached handout).

VI. Application to Actual Research Projects by Fellows

Several NRSA Fellows (Drs. Sajidah Husain, Ambarina Faiz, and Anna Petrova) will share their experiences working with large databases relating to their fellowship research projects.

VII. References/Internet Resources

Part 1

Delwiche, LD and Slaughter, SJ (1995). The Little SAS Book: A Primer. SAS Institute Inc., Cary, NC.

Fulcomer, MC and Kriska, SD (1989). MADMANager Utility Programs Users' Guide: Version B-03. Restat Systems, Inc., Columbus, OH.

Fulcomer, MC, Bastardi, MM, Raza, H, Duffy, M, Dufficy, E, Baron, ML, and Martin, RM (2000). Geocoding of New Jersey Births and Fetal Deaths for 1989-1996. Proceedings of the 1999 National Conference on Health Statistics, National Center for Health Statistics, NCHS Proceedings, CD-ROM No. 1, 9-0595 (7/00) and DHHS Publication No. (PHS) 00-1025.

Horst, P (1963). Matrix Algebra for Social Scientists. Holt, Rinehart and Winston, New York.

Jaro, MA (1989). Advances in record-linkage methodology as applied to matching the 1985 Census in Tampa, Florida. Journal of the American Statistical Association, 84, 414-420.

Searle, SR (1966). Matrix Algebra for the Biological Sciences (Including Applications in Statistics). John Wiley and Sons, Inc., New York.

Part 2

Cohen, J and Cohen, P (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. (Second Edition). Lawrence Erlbaum Associates, Hillsdale, NJ.

Fisher, RA and Yates, F (1953). Statistical Tables for Biological, Agricultural, and Medical Research. (Fourth edition). Oliver and Boyd, Edinburgh.

Hays, WL (1994). Statistics. (Fifth edition). Harcourt Brace College Publishers, Fort Worth.

Kerlinger, FN and Pedhazur, EJ (1973). Multiple Regression in Behavioral Research. Holt, Rinehart, and Winston, New York.

Winer, BJ (1971). Statistical Principles in Experimental Design. (Second edition). McGraw-Hill Book Company, New York.

RESEARCH FORUM SERIES

Core Curriculum Module E

Working with Large Databases

Handout

Part 1: "Preliminary Considerations in Working with Large Data Sets on Microcomputers -- or -- Power Processing with an Attitude"

Mark C. Fulcomer, PhD

Adjunct Associate Professor

Department of Family Medicine

UMDNJ-Robert Wood Johnson Medical School

BACKGROUND

A. What is a "large" data set (or file)?

The answer to this question depends on one's perspectives and purposes, so an idea of "size" is needed.

1. For example, one might be interested in addressing some research questions that refer to a population-based data file such as vital records for births (110,000 or so records each year in New Jersey) or deaths (about 70,000 or so records each year). Another somewhat familiar example is the UB-92 hospital discharge file with nearly 1.4 million records in this state each data year. In both types of examples, the number of "rows" (i.e., records) in the resulting data matrix are quite big so we could talk of a "tall" matrix (a term originally employed by Paul Horst).

2. In other instances, there may be a large number of variables found in each of the records in a data file, reflecting considerable breadth and complexity in the content of the data files. We might refer to the resulting data matrices as "wide". Again, vital records for births and hospital discharge records are familiar examples, with 370 and 950 columns in each of the original records, respectively. With 2800+ columns, records from New Jersey's recently-developed Electronic Birth Certificate (or EBC) system are even larger.

3. Then, there are data matrices that could be considered both "tall" and "wide" - that is, they contain a large number of both records/rows and variables/columns. An example of a file with 971,830 records is found in an earlier paper (Fulcomer et al., 2000)

4. Quite apart from sheer dimensions, other aspects might lead one to identify a data matrix as "large". For example, a data file of hospital discharge records may involve considerable "recoding" of diagnostic values before analyses can be done, resulting in considerable up-front programming effort. In other instances such as the EBC, data collection itself might be complex.

Using the numbers of records (N) and columns (p) in a data file, its "size" (S) can be approximated with the formula

$$S = N*(p + 2)$$

where +2 in the term (p + 2) reflects the addition of two characters to delimit (or separate) each record. For example, the data file mentioned previously required 507,295,260 bytes - 971,830 times 522 (520 + 2) characters. Of course, this

formula is less helpful when one is viewing largeness from the fourth perspective, say more in terms of "importance".

B. How have changes in microcomputer storage capacities, memory, and processing speeds changed affected the ability to work with "large" files (in a "size" sense only)?

1. Typical hardware configurations have expanded dramatically since 1982.
2. Clearly, most of the examples of "large" data files could not be approached on microcomputers at all prior to 1989 or so.
3. Furthermore, it wasn't until the mid-1990's that storage capacities increased sufficiently so that these data sets, including all of examples cited, could be processed effectively on 80486-class systems, even though at much slower speeds than are now possible.
4. Perhaps even more remarkable, CD-ROMs and other portable mass storage devices now permit the easy transfer of large data files (e.g., up to a maximum of 650 Mb on a single CD platter) at exceptionally low costs and with surprising speeds.

C. Although advances in microcomputing certainly facilitate the handling of large data files, the researcher still needs monitor the process closely.

1. In some cases, popular software packages (e.g., most spreadsheet programs) may not be able to manipulate files with many records (rows) or variables (columns), even if there is sufficient storage space available.
2. Also, several statistical packages may require that a large amount of temporary storage space be available (often an amount at least equal to the size of the original file) when data files are initially downloaded.
3. Moreover, some files may be in formats that are difficult to transfer over different computer environments.
4. Finally, given their size/complexity, it is sometimes difficult to "browse" or "inspect" a large data file and this problem doesn't appear to have improved much in the Windows operating systems either.

Thus, while current-day microcomputer capacities allow the productive handling of large files in theory, many researchers may feel grave trepidations about how to proceed in actual practice. Our most important purpose with this topic is to help you remove the obstacles of "fear" from any possible work with "large" data files by "de-mystifying" these preliminary considerations into a series of relatively straight-forward steps.

INITIAL DOWNLOADING OF LARGE DATA FILES (OR FIRST ITTY-BITTY STEPS)

- A. Checking the size of the file.
- B. Inspection of the initial file.

C. **Mystery Step #1:** Copying/Converting to ASCII Characters.

D. **Mystery Step #2:** Bordering.

E. **Mystery Step #3:** Counting.

OOPS. DID WE SAY DOCUMENTATION?

Good, solid documentation is an absolute "must" for large data sets. Typically, this is in the form of a "codebook", listing the columnar layout (i.e., beginning and ending positions for each field), values coded, and other details. Without documentation, it is difficult to decipher the contents of data files.

A. Key concepts to look for in good documentation:

1. Treatment of confidentiality.
2. Unique Identifiers.
3. Columnar locations.
4. Missing data values.
5. Fields with special coding issues (e.g., diagnoses).
6. Special data organization issues (e.g., concatenation).

B. **Mystery Step #4:** Frequencies ("marginals").

C. **Mystery Step #5:** Reformatting.

D. **Mystery Step #6:** Converting/Recoding.

File Maintenance

A. Secure Storage.

B. **Mystery Step #7:** Compression.

C. **Mystery Step #8:** Backups, Backups, Backups.

Follow-up session topic: Parsimony (or "less is more")

A. **Mystery Step #9:** Forming special subsets of records and variables

Follow-up session topic: Merging, Matching, and Linking Records (or "outcomes anyone?")

A. *Mystery Step #10*: Probabilistic record linkage (AUTOMATCH)

RESEARCH FORUM SERIES

Core Curriculum Module E

Working with Large Databases

Handout

Part 2: "Sooner Rather Than Later -- or -- Planning for Results through the Application of a priori comparisons"

Mark C. Fulcomer, PhD

Adjunct Associate Professor

Department of Family Medicine

UMDNJ-Robert Wood Johnson Medical School

INTRODUCTION

For several decades, almost all detailed treatments of analysis of variance (ANOVA) have suggested the use of planned, orthogonal comparisons in place of the usual calculations for data from experiments (e.g., Winer, 1971; Hays, 1994). Similarly, through the application of multiple regression (MR) techniques, a priori tests have more recently been encouraged for non-experimental data as well (e.g., Cohen and Cohen, 1983). Unfortunately, the advice to incorporate planned comparisons in analyses is not always followed in actual practice.

Orthogonal contrasts are sets of uncorrelated, weighted combinations of means which allow the testing of one research hypothesis for each degree of freedom in the systematic portion of a design. The sums of squares for the contrasts constructed from a research factor are unique and additive (i.e., they add to the usual sums of squares for the entire research factor in ANOVA or MR calculations). Furthermore, such contrasts have the advantage of allowing the most powerful, directional tests of hypotheses to be conducted while, at the same time, eliminating the need for somewhat cumbersome post-hoc tests. In addition to assisting the researcher to better focus analytic efforts on the hypotheses of interest, planned comparisons can also be applied to some more difficult issues such as unequal cell sizes and missing data.

BACKGROUND

Before going into any detail about planned, orthogonal contrasts and polynomials, this presentation will begin with a simple example of an ANOVA design with two research factors considered to be fixed effects. In the case of fixed-effects ANOVA models, the usual omnibus F-tests can be replaced by a set of planned, a priori

comparisons, one for each degree of freedom. Then, the computational equivalence of ANOVA and MR will be described, a correspondence that enables the application of orthogonal comparisons to be carried into non-experimental situations as well. Such comparisons are essentially simple t-tests between means from two independent samples. Because independent t-tests are, in turn, equivalent to point-biserial correlation coefficients (or "r point-biserial") between a "true" dichotomy and a continuous variance, orthogonal comparisons are quite convenient to interpret in terms of the proportion of variance accounted for (i.e., the coefficient of determination).

A. SIMPLE ANOVA WITH TWO FIXED RESEARCH FACTORS

Suppose an experimental design with two research factors, A and B, has been employed to investigate the impact of the combined levels of these two fixed effects in explaining some continuous outcome measure. For example, in explaining how quickly rats traverse a maze (i.e., measured speed would be the outcome measure), factor A (rows) might be the type of reinforcement and factor B (columns) might be the length of deprivation and test animals would be randomly assigned to the A x B combinations. Suppose that factor A has "p" levels, factor B has "q" levels, and that "n" test animals are allocated to each of the pq combinations listed below.

FACTOR B (COLUMNS) b1 b2 . . . bq

----- ----- ----- -----	a1	. . .	----- ----- ----- -----	a2	. .
.	FACTOR A	----- ----- ----- -----	(ROWS)	
----- ----- ----- -----	ap	. . .	----- ----- ----- -----		

Ideally, the outcome measure should be "normally" distributed, although this is not necessary. That is, strictly speaking, it is the "residuals" that are assumed to be normally distributed, a very robust assumption given the "central limit theorem".

The ANOVA results for such a design are traditionally reported in a summary table like that below, with columns describing the source of the variation, the sums of squares, the degrees of freedom, the mean squares (i.e., sum of squares divided by the corresponding degrees of freedom), and a F-ratio (if a test of significance is appropriate).

ANALYSIS OF VARIANCE SUMMARY TABLE

DEGREES OF SOURCE	SUM OF SQUARES	FREEDOM	MEAN SQUARES	F-RATIO-----
-----	-----	-----	-----	-----
FACTOR A	SSa	p-1	SSa/(p-1)	MSa/MSeROWS = reinforcement
FACTOR B	SSb	q-1	SSb/(q-1)	MSb/MSeCOLUMNS = deprivation
INTERACTION	SSab	(p-1)(q-1)	SSab/(p-1)(q-1)	MSab/MSe
(A x B)ERROR	SSE	pq(n-1)	SSE/pq(n-1)	-- (WITHIN CELLS)-----
-----	-----	TOTAL	SSt	pqn - 1 SSt/(pqn-1) --

B. HYPOTHESIS TESTING IN ANOVA

When used in this omnibus fashion, a significant F-ratio reflects difference among "some" of a group of means. Sometimes, the occurrence of a "significant finding" is followed by subsequent tests (e.g., those proposed by Scheffe, Tukey, and Newman-Keuls) to unravel the nature of the observed differences. However, such post hoc tests present important, difficult choices to the investigator, particularly to account for considerations such as type I and type II errors when making multiple comparisons among the means since the differences have already been observed. In a sense, the imposition of some of the cumbersome constraints on ex post facto tests is quite similar to need to control for cheating on an examination.

Of course, it is hardly surprising that Sir Ronald Fisher and others working on the early development of analysis of variance techniques had long advocated the use of a priori (or planned) orthogonal comparisons to circumvent the problems associated with multiple post hoc comparisons. For example, the Fisher-Yates weights for orthogonal polynomials were probably first published before their appearance in the original 1938 publication by those same two authors (Fisher and Yates, 1953).

In the decades following the appearance of these tables in Fisher and Yates, the advice to utilize orthogonal contrasts has been offered in most textbook treatments of ANOVA, including more recent treatments that have drawn heavily on the equivalence uses of MR for non-experimental data (e.g., Kerlinger and Pedhazur, 1973; Cohen and Cohen, 1983). Clearly, for industrial and agricultural applications that rely heavily on sophisticated experimental designs to optimize cost-effectiveness, this advice has been heeded with tremendous benefits. However, for other more applied applications in the medical and social sciences, it is still rather curious that orthogonal contrasts, or at least some set of planned (not necessarily uncorrelated) comparisons, are not routinely applied to fixed-effects designs. Most importantly, the definition of fixed effects implies that the entire population of levels of the research factors have been included in the design. As a consequence, only hypotheses about the levels actually included in the design are allowed. This is in sharp contrast to designs with random effects which allow inferences to levels not included in a particular experiment because of the process of drawing the random sample of levels that are incorporated by the investigator.

Presumably, a researcher employing a fixed effects models has presented some a priori rationale for incorporating the set into a design. Usually, such rationale would appear in the literature review and methods sections along with some implicit comparisons suggested by the results of earlier research. In addition, as will be demonstrated, the construction and use of orthogonal contrasts is surprisingly straight-forward. Because an orthogonal contrast is essentially a simple independent t-test between two means, a significant result is very easy to interpret. [Note that in such cases the square of the t-test result is equal to the corresponding F-ratio.] Furthermore, because point-biserial correlation coefficients, used to relate "true" dichotomous independent variables to continuous outcome measures, are equivalent to independent t-tests, the proportion of variance allocated to a contrast conveys its explanatory "importance". Finally, the uncorrelated nature of the orthogonal contrasts reduces the problems of computational accuracy, still a somewhat-overlooked issue on microcomputers.

The analogy between MR and ANOVA rests on the notion of the fixed effects. In MR the effects may overlap (or relate) somewhat, whereas in ANOVA the components of an experimental design are generally constructed to be orthogonal from the outset. Clearly, as computers became more powerful and more widely available over the last three decades or so, MR procedures have become increasingly popular, especially since simple orthogonal ANOVA designs can be handled as special cases. Nonetheless, for more intricate designs, say those involving complex interactions, ANOVA approaches still have much to offer.

SIMPLE PRINCIPLES FOR CONSTRUCTING ORTHOGONAL CONTRASTS

The use of orthogonal contrasts requires the imposition of two constraints: (1) that the weights sum to zero (i.e., that the overall mean of each contrast be equal to zero); and (2) that the sum of the cross-products of the weights from different comparisons be equal to zero (i.e., that the complete set of contrasts be mutually uncorrelated).

A. ADJUSTING THE CONTRAST WEIGHTS FOR UNEQUAL SAMPLE SIZES

B. ORTHOGONAL POLYNOMIALS FOR QUANTITATIVE VARIABLES

While most fixed effects refer to "qualitative" independent variables, variations on the same set of "tricks" can be applied to equally-spaced quantitative variables using the weights for orthogonal polynomials, originally developed by Fisher and Yates. With a similar approach for such quantitative variables it is possible to apply trend component weights (e.g., for linear, quadratic, cubic) for each degree of freedom in the design. Provided that the forms of the independent-dependent variable relationships are well-known, MR approaches to quantitative variables provide greater flexibility in that it is no longer required that the fixed-effect values be equally-spaced.

C. HANDLING TWO OR MORE RESEARCH FACTORS

D. CODING OTHER INDEPENDENT VARIABLES

TABLE

FACTOR B (COLUMNS)

	b1	b2	...	bq
a1			...	
a2			...	
FACTOR A				

```

(ROWS) | . | . | ... | . |
        | . | . | ... | . |
        |-----|-----|-----|-----|
ap |      |      | ...  |      |
        |-----|-----|-----|-----|

```

DEGREES OF

SOURCE SUM OF SQUARES FREEDOM MEAN SQUARES F-RATIO

FACTOR A	SSa	p-1	SSa/(p-1)	MSa/MSe
ROWS				
reinforcement				
FACTOR B	SSb	q-1	SSb/(q-1)	MSb/MSe
COLUMNS				
deprivation				
INTERACTION	SSab	(p-1)(q-1)	SSab/(p-1)(q-1)	MSab/MSe(A x B)
ERROR	SSE	pq(n-1)	SSE/pq(n-1)	--
(WITHIN				
CELLS)				

TOTAL	SSt	pqn - 1	SSt/(pqn-1)	--
-------	-----	---------	-------------	----