# THE ROLE OF PACKAGE PROGRAMS IN STATISTICAL METHODS COURSES

Thomas A. Bubolz and James E. Gentle
Iowa State University

## I. INTRODUCTION

A substantial proportion of the effort of most academic departments of statistics is expended on the teaching of what are called "statistical methods courses." These courses may be taught at various levels within the university, but most courses deserving of the title are taught at the advanced undergraduate level or at the graduate level. This is because understanding of the content of such courses generally requires previous exposure to investigative thinking and the maturity to appreciate the role of statistical methodology in the scientific method.

The techniques taught in statistical methods courses should be those procedures actually used by data analysts, and problem assignments should simulate practical research situations. The impact of the computer on statistical methods courses has been substantial. This is due to the profound impact of the computer on the statistical methods actually used. Massive amounts of data can be and are being analyzed to an extent that was not feasible before the advent of the computer. Alternate models can be examined and individual cases can be investigated quite readily. Exploratory data analysis is practicable. If the statistical methods courses are to prepare the student for doing data analysis in "the real world", they must provide the student with some hands on computer experience in data analysis. For courses with an applications orientation, computer usage by the student is not merely a superfluous novelty to increase student interest; it is an essential component of the knowledge and skills to be imparted in the course.

Computer usage in statistical methods courses, however, is not an end in itself; therefore the cost of such usage must be reasonable. Computer usage costs that concern us here are of two types: cost of actual usage of computing resources and cost of time spent in instruction. The computer language should be such that instruction in its use has minimal impact on the time devoted to discussion of statistical methods. On the other hand, a software system developed strictly for instructional purposes is of very limited usefulness in statistical methods courses. A primary objective in statistical methods courses is a high rate of transfer of skills developed in the classroom to subsequent work involving data analysis. The students should be taught a system that they are likely to use in their own data analysis.

A major disadvantage of many of the "packaged" statistical software systems is the rigidity they impose on the analyses. In many systems it is inconvenient or impossible to do adequate residual analyses or to fit a linear model using criteria other than ordinary least squares, for example. Yet many of the statistical methods courses introduce such topics; and if these topics are to be explored adequately, the instructor must frequently resort to use of specialized software systems or to use of subprogram packages in an algorithmic language such as FORTRAN. Either alternative requires more class time for instruction in use of the system, or else it imposes stiffer prerequisites on the course.

Additionally, the rigidity of many statistical programs precludes demonstration of the computations for the statistical analyses studied. When everything in the computations is so automatic, memory of the fact that $\hat{\beta} = (X'X)^{-1}X'y$ for example, is not adequately reinforced. Many of the instructors of beginning statistical methods courses, with good reason, like to require students initially to go through the mechanics of the formation of $X'X$, $X'y$, and the sums of squares for analysis of variance of linear models. This accounts for the popularity of the matrix-based languages of some of the statistical packages.

Because of the role of computer analyses in statistical methods courses and the concomitant desirable aspects of a general software system for classroom use, SAS has become the most widely used package in the statistics courses taught at Iowa State University. In the following, we will indicate the nature of the classroom applications of SAS in several of the statistical methods courses. The courses we will consider are:

Statistics 401, 402 - General methods (Snedecor and Cochran level)
Statistics 411 - Experimental design (Cochran and Cox level)
Statistics 407 - Multivariate analysis methods (Morrison level)
Statistics 500, 501 - Intermediate statistical methods (Bancroft and Morrison level)
Statistics 508 - Sociometric statistics
Statistics 538 - Econometric statistics

The courses numbered in the 400's are generally taken by undergraduate statistics majors and graduate students from other departments. Those numbered in the 500's are open to statistics graduate students, who constitute approximately 75-80% of the enrollment in these courses.

## II. OVERVIEW OF STATISTICAL METHODS COURSES

The orientation of the two-course sequence in general methods, Statistical Methods for Research Workers, is toward the application of statistics in other disciplines for the purpose of analyzing data collected from experiments and surveys. The courses are listed at the undergraduate level; however, graduate students not majoring in statistics may take them for credit.

Statistics 401, the first course in the sequence, introduces topics like elementary statistical concepts and models, estimation, tests of significance, linear regression, correlation and one-way analysis of variance. Statistics 402 continues to elaborate and expand the introductory material by including topics such as cross-classification of discrete data, factorial ANOV designs, multiple comparisons, single degree of freedom contrasts, analysis of covariance and multiple regression.

Typically, four or five sections of the courses are offered in the Winter and Spring quarters. Course instructors give fairly uniform treatment so far as the scope of the subject matter in the various sections is presented, although the emphasis given to various topics may differ in order to make the material more meaningful to a particular audience.

The choice of computing tools used to demonstrate techniques and work examples has traditionally been left to the discretion of each instructor. In past years programs like OMNITAB and SPSS have been used in various sections of these courses. Both program languages have disadvantages that have limited their applicability. OMNITAB, by virtue of its worksheet, was not readily adaptable to large scale research problems. SPSS, while providing many desirable facilities, did not have a sufficiently large user community, so that for the majority of students in these courses, the language they used to learn statistics was different from the language used to do research in their own disciplines. Since SAS-76 has been available, it has been the dominant choice of computing tool in these courses.

Statistics 411, Experimental Design for Research Workers, constitutes, for many students in the agricultural and biological sciences, the third course in a sequence encompassing Statistics 401, 402, and 411. Like its predecessors, the course is open to graduate credit for nonmajors in statistics. Its primary orientation is toward an audience whose interest is in applied statistical analysis.

Statistical concepts developed in Statistics 411 build on the material in Statistics 402 but give greater emphasis to varieties of analysis of variance designs. Introduced at this level are topics such as blocking, randomization, replication, split-plots, repeated measures, confounding, fractional factorials, and response surface methodology.

Some instructors in this course continue to rely on SAS-72 although SAS-76 is used quite heavily. Desirable features in the earlier version include the POOL statement and ADJMEANS statement in the ANOVA and REGR procedures.

Statistics 407, Methods in Multivariate Analysis, represents a complementary offering to Statistics 411 in the sequence of methods courses. The scope of the course encompasses methods in which multiple dependent (response) variables are analyzed simultaneously as well as the more familiar multiple independent variable models. Topics covered include MANOVA, Hotelling's $T^2$, principal components analysis, linear discriminant functions, canonical correlation, factor analysis and cluster analysis.

Many of the examples presented in the course have some bearing on agricultural or biological research, which is not unusual since the majority of students taking the course are in these fields. The course strives to convey many of the more important aspects of multivariate analysis without resorting to a highly theoretical approach. Motivation of the topics is sufficiently rigorous to require extensive use of matrix methods, a task for which SAS's matrix procedure is ideally suited.

Like other relatively advanced methods courses, Statistics 407 touches on concepts for which no readily available statistical procedure exists, at least in SAS. Partial correlation is an often used technique that suffers in this respect. In this course, as in the more advanced courses, much stress is placed upon examining foundations of multivariate techniques. Consequently, the interest in SAS procedures is two-fold; some are used in a data-analytic context while others, namely the MATRIX procedure, are used in order to explicate the methodology underlying these procedures.

Statistical methods for graduate students majoring in statistics are taught in the Statistics 500, 501 course sequence. The objective of Statistics 500, Statistical Methods, is to treat, at a more advanced level, some of the topics introduced in Statistics 402 and 411 and to introduce methods that have been found useful in statistical consulting activities.

The orientation of the course is reflected in the assumption that real-world data is neither tidy nor easily manageable, and that the task of analyzing it requires substantially more knowledge than simply specifying a linear model to be fit. The course covers diverse topics such as selection of transformations for inducing normality in data, working around dirty or missing data, data management techniques such as merging, subsetting and concatenating, data condensation and suitable display of raw data for the purpose of seeking interesting or suspicious patterns in the data.

Statistics 501, Intermediate Statistical Methods, introduces topics such as unequal or proportional subclass frequencies and weighted means in ANOV, multiple comparison techniques, continuous multivariate methods and discrete multivariate methods such as log-linear models. In Statistics 501 there is relatively greater emphasis on formal analysis and somewhat less on data manipulation than is found in Statistics 500.

Many of the students enrolled in this course sequence hold research assistantships in which they assist in the analysis of clients' actual data under the direction of a faculty member. It is not entirely coincidental that SAS is the dominant computing tool in these methods courses and the most frequently used package for the analysis of real data. SAS has been found to provide the capability to do "standard" analyses with a minimum of program preparation. Furthermore, "nonstandard" analyses can be put together in the matrix procedure, using the same mathematical principles that were employed in classroom presentations.

Statistics 508, Sociometrics, and Statistics 538, Econometrics, are the final courses in the series we have labeled "methods oriented".

These courses are designed for graduate students who need training in state-of-the-art statistical techniques for the social sciences. While some of the concepts overlap topics presented in other methods courses, the presentation of the material is given from a viewpoint familiar to students of sociology and economics.

Statistics 508 examines classical linear regression, hypothesis testing and, finally, the formulation of the general linear model. Special emphasis, befitting the course's title, is given to techniques of model building and path analysis. Also, because of the nature of data used in the social sciences, consideration is given to methods for dealing with measurement error.

Course content in Statistics 538 is oriented toward the types of processes and events that are studied by economists. This course treats topics such as restricted least squares, generalized least squares, autocorrelation, errors in variables and instrumental variables.

The preceding course descriptions illustrate the type of training in applied statistical analysis that is available to undergraduate and graduate students. Tools and techniques acquired in these service courses are diffused throughout the university research community. The widespread use of SAS as a research tool indicates that substantial multiplier effects have been realized in the number of persons formally trained in the use of a tool as a proportion of the total user community.

One factor, common to all the methods courses, is the fact that they all make use of program packages in some degree. The content of these courses has evolved over time, partly in response to changes in user disciplines, and partly as a result of the availability of computing tools that permit students to utilize techniques that previously were too complex to analyze by hand.

## III. APPLICATIONS OF COMPUTING TOOLS

The problems and programs that will be discussed next have, with only a few exceptions, been taken from laboratory exercises given in the respective courses.

Only those exercises in which SAS plays a role have been shown; other examples would only cloud the picture except in the cases where another package is involved with the aid of SAS, as in the SAS BMDP procedure.

The computing situation at Iowa State University is unusual, as university computing facilities go, in that interactive computing is a little used tool. As concerns the major statistical packages, it is unlikely that they will be used in an interactive context in the foreseeable future.

The reason for this state of affairs is that a special high speed batch monitor, installed a few years ago, has throughput capability on the order of one to two thousand jobs per hour. Since SAS and SPSS users may run in this high-speed batch queue, the effective turnaround time of jobs submitted from and retrieved by a remote keyboard terminal is approximately five to thirty seconds. An adept terminal user, exploring a variety of regression models in SAS, can easily generate four to six useful runs per half hour assuming that results of prior runs are used to make decisions about subsequent runs.

### Statistics 401-402

The two course sequence in introductory statistical methods, Statistics 401-402, has the objective of training non-statisticians in the application of statistical data analysis to real-world problems. The first course focuses on data description and display, estimation, hypothesis testing and simple ANOV models, correlation and bivariate regression. The second course covers a more extensive partitioning of sums of squares, analysis of residuals, multiple regression and analysis of covariance.

#### Example 1

This exercise from Statistics 401 was designed to examine whether the rule stating that the variance of means is less than the variance of individual observations actually holds. Students were given summary statistics for a population of 200 body weights and were then asked to compare them to summary statistics for a sample of means, each based on 10 observations from the same population. The program below was used to generate statistics on the means.

```
DATA WT;
SET S401.E2;
RD = UNIFORM(0);
PROC SORT; BY RD;
DATA WT1; SET WT; CT+1;
DATA WT2; SET WT1; GRP = INT ((CT-1)/10);
PROC PRINT;
PROC SPLOT; VAR WT; CLASS GRP;
PROC MEANS; BY GRP; VAR WT;
OUTPUT MEAN = MWT;
PROC MEANS; VAR MWT;
PROC CHART; HBAR WT;
```

#### Example 2

The second example illustrates an assignment from a section of Statistics 402.

The cloud point of liquid is a measure of the degree of crystallization in a stock that can be measured by the refractive index. It has been suggested that the percentage of the chemical "X" in the stock may be a useful predictor of the cloud point.

Students are required to generate an ANOV table using the GLM procedure and create a new dataset which contains residuals and predicted values. Lastly, the residuals are evaluated to determine whether they are normally distributed and to check for possible misspecification of the model.

Data consist of 19 observations in which measurements were taken on the percentage of the chemical "X" and the cloud point.

```
DATA CLOUD_1;
INPUT NUMBER X Y;
LABEL X = PERCENTAGE OF CHEMICAL X
      Y = CLOUD POINT;
CARDS;
PROC PRINT;
  TITLE CLOUD POINT DATA;
```

```
PROC PLOT; PLOT X * Y
PROC SORT; BY X;
PROC MEANS; VAR Y; BY X;
PROC GLM; MODEL Y=X/P CLM;
  OUTPUT OUT = CLOUD_2 PREDICTED = YHAT
        RESIDUAL = RESID;
PROC PRINT;
PROC PLOT; PLOT YHAT * RESID;
PROC RANK NORMAL = BLOM OUT = CLOUD_3;
  VAR RESID; RANKS RANKITS;
PROC PRINT;
PROC PLOT; PLOT RANKITS * RESID;
```

Example 3

In this example the matrix procedure is used to obtain an ANOV table and carry out the least squares analysis for a simple linear regression problem. Students are required to find a) the least squares regression line, b) vectors of residuals and predicted values, c) the ANOV table, d) residual plots, e) confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$, and f) confidence limits for y at X = 70.

```
PROC MATRIX;
YT = 83 70 80 67 58 55 30;
XT = 5 20 45 55 65 80 95;
Y = YT'; X1 = XT'; N = NROW (Y);
XO = J(N, 1);
X = XO||X1;
D = INV(X' * X);
BHAT = D *X' * Y;
YHAT = X * BHAT;
RESID = Y - YHAT;
SSTOT = Y' * Y;
YMEAN = (XO' * Y) #/N;
CORRECTN = YMEAN * YMEAN * N;
SSCTOT = SSTOT - CORRECTN;
SSREG = YHAT' * YHAT - CORRECTN;
SSRESID = RESID' * RESID;
MSRESID = SSRESID #/(N-2);
PRINT BHAT SSREG SSRESID SSCTOT CORRECTN MSRESID;
COVBHAT = MSRESID * D;
C = 1/70;
MEANPRED = C' * BHAT;
VARPRED = MSERROR * C' * D * C;
PRINT BHAT COVBHAT MEANPRED VARPRED;
```

## Statistics 411

Statistics 411, Experimental Design for Research Workers, provides students with an understanding of the basic principles of experimental design and a catalog of useful designs, as well as their properties and methods of analysis. The course is taken primarily by graduate students from departments other than statistics. The most commonly used text is Cochran and Cox, Experimental Design. Some of the assignments given in this course require construction of designs with certain confounding patterns or satisfying other criteria for the design. There may be no data to analyze. Other assignments that do require data analysis are performed with SAS. PROC PLAN is usually discussed and illustrated with handouts of listings. PROC ANOVA, GLM, and PLOT are the most commonly used procedures for the assignments. Techniques for analysis of simple designs with missing data are discussed and the equivalence of analysis of covariance and least squares fitting for missing data is shown.

Example 1

In an experiment using randomized complete blocks to ascertain the effect of various treatments for corn root borer, the datum for the second treatment in the third block was lost. The students were asked to use analysis of covariance to account for this, using the data already stored as a SAS dataset.

```
DATA NEW;
SET S.CORN; X = 0; IF BLK = 3 AND TRT = 2 THEN
        X = 1;
PROC ANOVA DATA = NEW;
CLASSES BLK TRT;
MODEL DAM = BLK TRT X;
```

Example 2

The students are asked to imagine that they are the consulting statistician on a project given to the experimental laboratory of a large industrial firm for determining the effect of temperature and two catalysts on the yield of a certain chemical reaction. Along with the handout for the assignment the students are given a listing of data from a $2^3$ explanatory experiment in the form:

| TEMP (DEGREE C) | CATALYST A (GRAMS) | CATALYST B (GRAMS) | YIELD (GRAMS) |
|---|---|---|---|
| 245 | 105 | 140 | 1484 |
| 245 | 105 | 140 | 1455 |
| 245 | 105 | 175 | 2700 |
| . | . | . | . |

There are sixteen observations in this data set. The yield is the amount of product per fixed amount (5 kg) of reactants after equilibrium is reached.

The students are told that after analyzing the preliminary data in any way they wish, they may request the lab to conduct additional experiments at whatever levels of temperature and catalyst they specify. The objective, of course, is to ascertain the optimal combination for maximum yield. The experiments are requested by creating a SAS dataset with variable names TEMP, CATA, and CATB, with the variables given values for the levels desired; that is, the student's job may be as follows:

```
//      JOB
//   EXEC SAS
//SYSIN   DD      *
DATA;
INPUT TEMP CATA CATB;
CARDS;
245    110    140
245    110    175
       etc.
$COPY DD1=EXP,DISP=SHR
/*
```

The effect of the $COPY card is to concatenate the card image file in EXP to the SYSIN file. The file EXP is as follows:

```
OPTIONS NOSOURCE;
DATA; SET  _LAST_;
YIELD = 350 * TEMP + 420 * CATB + TEMP * CATB -
        TEMP * TEMP - 2 * CATB * CATB - 78155 +
        12 * NORMAL (0);
```

```
PROC PRINT;
TITLE DATA FROM THE EXPERIMENT YOU REQUESTED;
```

The students do not get to see these SAS statements.

Execution of the job shown above provides the student with a listing of simulated data from a quadratic response surface involving temperature and catalyst B. Catalyst A does not affect the yield.

The students are told that the laboratory experiments are relatively expensive to conduct, so they should not request more than 32 observations in any one experiment and not request more than three experiments.

After obtaining their data they are to analyze it however they wish, probably with PROC GLM.

The report for this exercise is to consist of the following parts:

A. Preliminary analysis of the data from the exploratory experiment.
B. Plan for the next experiment: description, models, rationale, etc.
C. Data listing for the experiment in B.
D. Analysis of the experiment in B.
E. F. G. Repeat B, C, D if appropriate.
H. I. J. Repeat B, C, D if appropriate.
K. Discussion and conclusions.

Part K should include the final fitted model and a discussion of the optimal experimental settings with consideration being given to the cost of the catalyst and the value of the product.

Statistics 407

The objective of Statistics 407, Methods of Multivariate Analysis, is to introduce, at the advanced undergraduate level, some widely used techniques for multivariate data analysis. The course builds upon material presented in Statistics 402 and emphasizes application rather than theory. Statistics 407 focuses on topics such as multivariate analysis of variance, Hotelling's $T^2$, principal components analysis, canonical correlation and discriminant analysis.

For computer calculations, SAS is the only program language that is used in the course. The explication of multivariate techniques is handled best by the matrix procedure. Built-in procedures such as CLUSTER and FACTOR are used to demonstrate the applications that may be undertaken once the mathematical details of these methods have been investigated using matrix algebra. Furthermore there are many problems encountered in multivariate analysis that are "nonstandard" and cannot be handled adequately by any built-in procedures. Students in this course come to view the matrix procedure as an extremely useful adjunct to other SAS procedures rather than as a mere pedagogical device.

Example 1

Students are asked to analyze the relationship between reproductive growth and vegetative growth in apple trees. Data were gathered for a 5-year period. Results from previous experiments indicate that reproductive growth is small when vegetative growth is large and vice versa. The output from SAS is used to interpret the data in light of the assumptions.

Students must obtain a) plots of all pairs of variables, b) the correlation matrix, c) the partial correlation between growth and girth controlling for crop, d) the multiple correlation between reproductive growth and variables representing vegetative growth and e) the variance covariance matrix of all variables.

```
DATA GROWTH;
INPUT YEAR 2 Y1 4-7 Y2 9-12 Y3 14-17;
LABELS Y1 = CROP
       Y2 = GROWTH
       Y3 = GIRTH;
Y1=LOG10(Y1); Y2=LOG10(Y2); Y3=LOG10(Y3);
CARDS;
PROC PLOT; PLOT Y1*Y2 = YEAR Y1*Y3 = YEAR
         Y2*Y3 = YEAR;
PROC CORR; VAR Y1 - Y3;
PROC GLM; MODEL Y1 = Y2 Y3;
```

Example 2

In this problem students are asked to cluster a set of 10 objects such as a dictionary, newspaper, SAS manual and a deck of playing cards. They begin by taking 8 to 11 measurements, of attributes they believe are relevant, on each of the 10 objects. The measurements may be discrete or continuous. For example, the object may be used for work or leisure and it may contain greater or lesser amounts of information.

To analyze the data, the measurements are standardized and a measure of association, the product moment correlation, is calculated between all pairs of objects. The thrust of the exercise is focused more on the problems of measurement rather than the actual computations required.

```
PROC MATRIX; FETCH X;
NOBS = NROW(X); NVAR = NCOL(X);
MEAN = J(1, NOBS)*X#/NOBS;
X = X - J(NOBS, 1)*MEAN;
SS = J(1, NOBS)*(X#X);
STD = SQRT (SS'#/(NOBS-1));
X = X*DIAG (1#/STD);
PRINT X;
X = X'; NOBS = NROW(X); NVAR = NCOL(X);
SUM = J (1, NOBS) * X; XPX = X' * X;
XPX = XPX - (SUM' * SUM)#/NOBS;
S = 1#/SQRT (DIAG(XPX));
CORR = S*XPX*S;
PRINT CORR;
```

Statistics 500-501

Statistics 500 is a graduate level course that focuses on the application of least squares using classifactory and functional models. Considerable importance is placed on the ability to manipulate data. Transformations, subsetting, concatenating and merging are treated extensively. A major tenet of the course is that the solution of a problem is greatly simplified once the data have been suitably organized.

Statistics 501, Intermediate Methods, continues the development of ANOV and ANCOV models for 2 to N-way comparisons with unequal subclass N's. Polynomial regression, orthogonal polynomials, MANOVA and Hotelling's $T^2$ are introduced as multivariate methods for continuous data. Log-linear models highlight discrete multivariate methods.

The program languages used in this two course sequence are SAS and BMDP. In Statistics 500 the objective is to impart data analytic skills, and SAS provides the necessary tools without requiring course content to be diluted by the inclusion of extraneous topics such as intermediate JCL (Job Control Language) and IBM's dataset utilities.

Procedures in the BMDP package are accessed via the BMDP procedure in SAS. This means a considerable savings in class time is realized since only very small parts of a distinctively different procedural language need to be taught.

Example 1

This exercise was designed to develop a student's skill in formulating an appropriate model. Emphasis is placed on developing informal and intuitive perspectives. Answers to the problem are necessarily ambiguous and require judgement rather than blind acceptance or rejection.

In the model under consideration, minimum temperature of a given city in January is assumed to be a function of only latitude and longitude though other measurements are available. Data were collected at 56 different sites. The objective of the exercise is to illustrate procedures that may be used to probe for specification error.

```
DATA WEATHER;
INPUT TMIN LAT LONG ALT PRECIP;
CARDS;
PROC GLM; MODEL TMIN = LAT LONG;
  OUT = NEW PREDICTED = TMIN_HAT RESIDUAL =
        TMIN_RES;
PROC PLOT; PLOT TMIN_RES * TMIN_HAT
           TMIN_RES * LAT
  TMIN_RES * LONG TMIN_RES * ALT
  TMIN_RES * PRECIP;
PROC RANK                 OUT = ORDERED;
          VAR = TMIN_RES;
  RANKS RANKF;
DATA NPLOT; SET ORDERED;
FNTMIN = PROBIT ((2 * RANKF - 1)/112);
PROC PLOT; PLOT FNTMIN * TMIN_RES;
  TITLE FULL NORMAL PLOT;
```

Example 2

This exercise was developed to illustrate strategies that may be employed to build linear, least squares models, assuming all pertinent variables are accessible. As in the previous example, a model to predict the minimum temperature in January is sought. Students are asked to compare and contrast 4 and 5-variable models generated by the forward, backward and stepwise selection methods, obtain the best 3-variable model generated by the Max $R^2$ and Min $R^2$ methods, and use the RSQUARE procedure to obtain all possible regressions. Students are to explain why different models are selected by ostensibly similar procedures.

```
DATA WEATHER;
INPUT TMIN LAT LONG ALT PRECIP HIGH SMAX SMIN;
CARDS;
PROC STEPWISE;
  MODEL TMIN = LAT LONG ALT PRECIP HIGH SMAX
               SMIN/ STEPWISE FORWARD
               BACKWARD SLE = .25;
PROC STEPWISE;
  MODEL TMIN = LAT LONG ALT PRECIP HIGH SMAX
               SMIN/ MAXR MINR SLE = .25
               STOP = 3;
PROC RSQUARE;
  MODEL TMIN = LAT LONG ALT PRECIP HIGH SMAX
               SMIN/ START = 3 STOP = 5;
```

Example 3

The third example illustrates an application of the BMDP procedure. The problem is based on data which measures the concentration of toxic metals in various soil types sampled at various distances from coal fired electric power plants. Measures of soil types are nested within distance. After categorizing, the concentration is identified in terms of its toxicity on laboratory animals.

```
DATA METALS;
INPUT CONC SOIL DISTANCE;
CARDS;
PROC BMDP PROG = BMDP3F UNIT = 3;
PARMCARDS;
/PROBLEM TITLE IS 'TOXIC METAL ANALYSIS'.
/INPUT    UNIT = 3.
/TABLE    INDICES ARE CONC, SOIL, DISTANCE.
          SYMBOLS ARE C, S, D.
          ALL.
/CATEGORY CODES (1) are 1, 2, 3.
          NAMES (1) ARE 'LO-TOXIC', 'MED-TOXIC',
                        'HI-TOXIC'.
          NAMES (2) ARE 'PERMEABLE',
                        'SEMI-PERMEABLE',
                        'IMPERMEABLE'.
          NAMES (3) ARE '1_KM', '10_KM',
                        '50_KM'.
/FIT      MODEL IS CONC, SOIL, DISTANCE.
          LAMBDA.
          BETA.
/END
/FINISH
```

Statistics 508-538

These courses focus on the application of linear and nonlinear models to applications such as path analysis and demand forecasting. Statistics 508, Sociometrics, is concerned with topics such as selection of "best" predictor variables, estimation, dummy variables, measurement error, index construction, composite measures and causal inference. Statistics 538, Econometrics, treats topics such as generalized linear regression models, instrumental variables, multicollinearity, nonlinear estimation and errors in variables.

The content of these courses varies from time to time, primarily to keep abreast of state-of-the-art developments in the respective disciplines. Consequently, the computing tools are more specialized and include programs like

SPSS, LISREL and SAS, which are used in Sociometrics, and SUPER CARP and SAS, used in Econometrics. SAS applications in these courses include the use of GLM and REGR in Sociometrics and AUTOREG, GLM, SYSREG and NLIN in Econometrics.

Example 1

In this example Statistics 508 students are asked to find the total effect, direct effect and indirect effects for a path model in which

$$X_3 = f(X_1\ X_2)$$
$$X_4 = f(X_1\ X_2\ X_3)$$

```
PROC SAS72;
PARMCARDS4;
PROC REGR SIMPLE CORR;
  MODEL X3 = X1 X2;
  MODEL X4 = X1 X2 X3;
```

Example 2

The parameters of a model for a production process are to be estimated. The model allows for two methods of production and relates output to input linearly over certain regions and quadratically over other regions but continuously over all regions.

```
DATA PROD;
INPUT M X Y;
Z1 = 0; Z2 = 0; XX = X * X;
IF X > 13 THEN Z1 = (X - 13)**2;
IF X > 50 THEN Z2 = (X - 50)**2;
CARDS;
PROC GLM;
CLASSES M;
MODEL   Y = M X XX Z1 Z2;
```

## IV. CONCLUSIONS

Packaged statistical software has come to play a major role in statistical method courses. The packages have allowed consideration of more interesting statistical topics and enabled students to perform more complex analysis by relieving them of the computational burdens. When the details of the computations are deemed worthy of consideration the students can be asked to make use of the algorithmic flexibility of SAS, generally through PROC MATRIX, to follow the sequence of computations.

In addition to the role of package programs as aids in teaching statistical methods, use of the packages is a statistical method; and, thus, familiarity with packages such as SAS has become part of the content of statistical methods courses.

## V. ACKNOWLEDGEMENTS

We would like to thank our colleagues at Iowa State University for sharing with us, and allowing us to report, their classroom techniques and examples. In particular we would like to acknowledge the contributions of Drs. T. Bailey, D. Cox, W. Fuller, P. Hinz, K. Koehler and R. Warren.