Open Access Repository
www.ssoar.info

Historical Software Issue 11: Biomedical Computer Programs/ BMDP

Thaller, Manfred

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Thaller, M. (1984). Historical Software Issue 11: Biomedical Computer Programs/ BMDP. *Historical Social Research*, *9*(1), 78-85. https://doi.org/10.12759/hsr.9.1984.1.78-85

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-52149

ISSUE 11: BIOMEDICAL COMPUTER PROGRAMS / BMDP

GENERAL CONSIDERATIONS

While not necessarily the first statistical package designed, BMDP has been one of the very first big ones packages and is currently the one that has been in constant use for the longest time (since the late fifties, the manual of the forerunner BMD having been published first in 1961).

Such a lengthy history (almost unique in the quickly developing and changing field of software) makes it clear, that there have to be important arguments for the product. On the other hand one thinks immediately of FORTRAN, which, as a programming language of course, has an even longer tradition, continuously ignoring new trends and developments as long as possible.

We called SAS not so much a package, but rather a core of statistical programs, embedded into a consistent software environment. This, at the same time, is exactly what BMDP is not. It is a statistical package, held together relatively loosely by a common file concept and certain common characteristics of the command language of several independent executable programs.

Unlike the more recent systems, like SPSS or SAS, BMDP does not consist of one program, which is put to different uses by selecting the appropriate command, but of 42 independent programs, each of which is dedicated to one rather restricted purpose, as e.g: "Histograms and Univariate Plots" (the program called BMDP5D), "Nonlinear Regression" (BMDP3D) or "Univariate and Bivariate Spectral Analysis" (BMDP1T). This design has the advantage, that, writing those programs, their authors were relatively free from all considerations about how to fit them into the system without making it so large as to be unmanageable when loaded into core as a whole. This is not the only reason for it, but has probably contributed to BMDP being certainly the most powerful general purpose system for statistical analysis which is around. Indeed it seems to be rather typical, that one of the BMDP components, BMDP4V ("General Univariate and Multivariate Analysis of Variance and Covariance, Including Repeated Measures"), was originally developed as a special purpose program for this particular type of analysis and only afterwards integrated into the system as a whole. One might describe BMDP in general as a loose bundle of extremely powerful special purpose programs for statistical analysis.

Beyond the high statistical quality and astonishing flexibility of the different programs, this extremely modular character allowed BMDP to become the first system to go completely Micro: as we already mentioned in Software Section 7, BMDP offers a micro computer called StatCat which can be bought together with an arbitrary selection out of all BMDP programs available, each of them being executable independently. While exact quotations of European prices could not be gotten, and the modules are sold piecemeal, a sensibly equipped configuration is available for US $ 11.000 or less. So

Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, D-3400 Göttingen

BMDP is currently the only big statistical system that can be bought by individual institutes together with their own hardware. Having honored the advantages of this approach, one has also to point at its less bright side. BMDP as a system consists of a set of common input conventions and of a very restricted set of data definition and modification commands applicable (almost) uniformly within the command language of all the 42 modules. Even the concept of a system file (called BMDP Save File) came rather late in the development: only relatively recently it became a recommended practise to create such a save file instead of reentering all the necessary commands for the definition and transformation of the data together with them for every step in the analysis. So the features for describing and modifying (or transforming) the data were developed, when it was assumed that they would just be used for a single step in the analysis - and are extremely limited therefore. Or, to paraphrase the words of a recent review(1): BMDP is a statistical package which to some degree still clinges to the principle of a subroutine library, providing excellent statistical algorithms, but expecting the user to take care of all non-trivial I/O problems himself. This has a happy side effect by the way: BMDP offers a lot of well defined interfaces to user written FORTRAN programs, not only for data transformations (Manual 57 - 58) but also in a number of routines for the provision of special statistical functions wanted by a user, but not incorporated into BMDP (e.g. Manual 295 - 296 on how to include your own non-linear regression function into the general non-linear regression program). Even worse: the documentation, while pretty good for the statistical procedures is often obscure for data definition and transformation. Just one example: BMDP recognized originally only Fixed Formats. How to define FORTRAN formats for that purpose is described in the manual(2) on pp. 21 - 23 under the heading "Specifying the Data Format" as part of chapter 3 "Using BMDP Programs". A "Summary of differences Between IBM FORTRAN Input and the (sc. Fixed Format) BMDP Data Reader" on the other hand appears on pp. 46/47 as "Additional Options and Features for Formatted Data Reading" within chapter 5 "Describing Data and Variables with BMDP instructions". This chapitself mentions Fixed Formats just briefly and advises people not familiar with FORTRAN to go back to chapter 3 (see above). (Quite well designed) Free Format input facilities are, on the other hand, described in detail in chapter 5 (pp. 44-46) and mentioned only briefly in chapter 3. Confused? So was I.

a. STATISTICAL METHODS AVAILABLE

Each of the programs of BMDP implements a certain statistical method: as many methods overlap, very many of the more simple computations can be taken care of by different programs in a slightly different way. There is e.g. no command to compute correlation coefficients: but 17 of the 42 programs can be used to compute a variety of them. So one better selects as example for the control language a typical application of one of the method dedicated programs. Quite representative is the setup for a multiple linear regression:

```
/ PROBLEM  PROVIDE AN EXAMPLE FOR THE COMMAND LANGUAGE
/ INPUT    UNIT IS 9.
           CODE IS MYDATA.
/ REGRESS  DEPENDENT IS FOOD.
           INDEPENDENT ARE AGE, INCOME, NOFCHILD.
/ END
```

This sequence of commands, when interpreted by BMDP1R could be used to check, if the proportion of expenditure for food in a family budget can be explained by the cumulative effect of the age of the head of household, his income and the number of children in the family.

A few formal comments: the format of the command language is free, so the "columns" recognizable in the example above have just been used to improve readability. The / PROBLEM, / INPUT and / END constructs (called "paragraphs" within BMDP) are common for all BMDP programs, specifying the title to be printed on top of the page, where the data are to be gotten from and the end of the directives needed to execute a complete step of processing (called "problem" within BMDP) respectively.

The / REGRESS paragraph is specific in its syntax to the multiple linear regression program (BMDP1R). It consists in our example of two "sentences" (starting with a "BMDP word" - DEPENDENT, INDEPENDENT - and ending with a period), which assign variables to two key words, with a meaning quite obvious. While in our example only these two sentences of the / REGRESS paragraph can influence the intended computations, in more complex programs as, e.g., BMDP4F (for the tabulation of data and the log-linear analysis of those tables) three or more different paragraphs with 50 or more types of sentences can be used to define with astonishing flexibility how the computations shall be undertaken.

A very nice feature of BMDP is the possibility to use in many programs / PRINT and / PLOT paragraphs, which allow the printing and/or (line printer) plotting of selected information about the computations not ordinarily displayed. If we add to our example above

```
/ PRINT CORRELATION.
/ PLOT  RESIDUALS.
```

we would get, in addition to the standard results of our regression command, a display of the correlation matrix and two plots of (a) the residuals and (b) the squared residuals against the predicted values.

Another general feature which is good to have is the possibility to use the / GROUP paragraph in some, though not all, of the programs to split the data into a number of subsets which shall be analysed separately. Unfortunately this feature can be used at the same time for purposes of case selection and value labelling as well, while in some of the programs special provisions for grouping the cases into subsets exist - so what is in principle, as we stated already, a very good feature, will become a source of obscurity particularly for the beginner.

Which statistical features exist? BMDP provides of course quite a few possibilities for descriptive statistics (Manual 73 - 91), including a variety of line printer histograms, scatter plots and so on (Manual 123 - 142). Of a more principal interest are a number of programs which are dedicated to the analysis of missing values and their estimation (Manual 207 - 234). Historical data being what they are, the possibility to estimate all missing values by a regression function of all values available, just entering a single line of command, is very valuable indeed. Unfortunately this services are provided mainly for continuous data on interval scales. Even less fortunate: most of the other BMDP programs are very inflexible when it comes to the handling of missing data. This seems to be no drawback in the biological sciences, but it reduces certainly the value of BMDP for historical social research.

One of the highlights of BMDP is the program called BMDP4F dedicated to

"Two-Way and Multiway Frequency Tables -- Measures of Association and the Log-Linear Model" (Manual 143 - 206). This program allows the analysis of tables from the display of simple two dimensional cross-tabulations right thru to log-linear analysis (supporting two methods for the analysis of structurally empty cells). And, what is most important, all this with a command language that is consistent and easier to learn than the one of dedicated special programs like ECTA. Very good supported are of course the more traditional methods like analysis of variance (Manual 93 - 122, 345 - 436) and regression (Manual 235 - 288). This reviewer is afraid however, that data as typical for historical social research, will only rarely have the quality to allow fully to take advantage of what is better here than in other packages. Nonparametric methods are unfortunately supported relatively poor (Manual 437 - 446). Other extensions of the traditional regression type of analysis, like nonlinear regression, are undoubtedly very interesting in themselves (Manual 289 - 344) and BMDP has certainly advantages over other packages in this field, but here again we have some doubts, if our data will fulfill the requirements very often.

Two of the fields of statistical development singled out in our introductory paper as particularly useful for historical social research, Cluster Analysis and Time Series Analysis, are supported by BMDP. The routines for Cluster Analysis seem to be rather restricted to this reviewer, providing almost none but Euclidean measures of distance and supporting only very few approaches (leaving out such important ones, as, e.g., Wards Method). Time series analysis is supported very well in contrast (Manual 595 - 660). Given the importance ascribed to this method by some of our readers we would like to point out, that Spectral analysis is supported very well, including precautions for missing data. This reviewer has to admit, however, that he is still a bit uneasy about historical data fulfilling the requirements for Spectral analysis - so he thinks it particularly helpful, that the Box-Jenkins Time Series Analysis, looking at least to him more transparent, is supported as well. Laudable: the general introduction into time series analysis as a statistical method is more detailed as are most of the other general introductions (Manual 595 - 603).

Finally BMDP provides (besides Survival Analysis: Manual 555 - 594) a whole array for the usual multivariate approaches (Manual 479 - 554). Factor Analysis, Canonical Correlation, Partial Correlation and Discriminant Analysis are supported, with more flexible options as in most comparable programs. What we would like to hint at: BMDP supports Boolean Factor Analysis (Manual 538 - 546), that is, the factor analysis of variables that can be coded with yes and no (i.e, with 1 and 0). Included are tools to recode efficiently variables on nominal scale into such variables. This approach could make factor analysis applicable to quite a few fields, where the traditional approach has to be looked at distrustfully, due to the (lacking) quality of the data.

To sum it up: BMDP offers an excellent statistical armory. Most of the tools provided are aimed at high quality data as the are available within the natural sciences. Within the wealth offered, there is a large number of methods, however, which are applicable to historical material better than traditional methods have been.

b. LEARNING AND TEACHING

One of the ways of selecting a case for processing in BMDP consists of assigning a value to a variable called USE. This value can consist of a logical expression. If USE becomes greater than 0, the case is used for analysis and is included into the BMDP Save File, if one happens to be created. If USE equals 0, the case is not used in the analysis, but saved in the BMDP Save file. If USE becomes negative, the case is neither used for the analysis, nor saved. So much the basic description on "Selecting Subpopulations and Deleting Cases" on p. 55 of the Manual. If you read additionally on p. 57 the remarks on "FORTRAN transformations using the BIMEDT procedure", you are informed furthermore that, if USE equals -100 BMDP assumes an EOF after this case in the input data; this makes now understandable an unexplained example to the same effect on p. 55. If you read furthermore p. 63 in the chapter on "Multipass Transformations" you get even acquainted with the fact, that, if USE is negative, the case is included, as described above, but the sequence numbers of the cases are not changed. When this reviewer had to teach EDP to historians which were not acquainted with it, he had to spend so much time on explaining how to handle much simpler selection mechanisms, that he hopes for sure, he will never have to explain data handling in BMDP to a real novice with computers.

BMDP offers excellent statistical tools; it assumes however, that the user has a sound statistical background already. The data handling capabilities are much to poor to allow using BMDP to simulate the basic statistical computations. The user has an enormous influence how the statistical computations are undertaken but he has to know very well how those computations work or he is threatened by wrong results caused by a misunderstood parameter. An excellent tool for producing statistical results, a dangerous one for the beginner.

c. ENHANCED DATA HANDLING CAPABILITIES

With a few exceptions BMDP does not provide very much in this area. As we mentioned, the concept of a BMDP Save File is not very central do the system. The basic logic runs like this:

Every BMDP program has an / INPUT paragraph. It describes where the data to be processed come from. This can be any input medium of the host computer (if a unit or channel number can be assigned to it by the language of the operating system). The data can be entered as input data in Fixed Field or a number of Free Field formats or as a BMDP Save File created in a previous job. The only difference between raw data and a BMDP Save File is, that in the first case the / INPUT paragraph has to contain the number of variables used and an input format (a FORTRAN Format or a keyword, selecting the appropriate variety of Free Format), while in the second case this does not apply.

When data are first entered, the variables do not have names, but indices. Every number is a legal variable index in a BMDP user program: i.e. a number is usually interpreted as signifying a variable. Blanks are treated as missing. All numbers read are considered to be legal. This behavior can be changed with the / VARIABLE paragraph.

The user has at any stage the possibility to assign NAMES to a variable, which consist of 8 characters or less. If they start with a digit or contain blanks or certain other special signs, the have to be surrounded by apostrophes. Names can be assigned in sequential order, as in

```
NAMES ARE AGE, INCOME, JOBCODE, EXPEND, FOOD, CLOTHING,
          NCHILD.
```

or, by using a variable index, specifically to a subset of the variables, as, e.g.

NAME(2) IS INCOME.

These indices can always be used instead of the variable names: the following two / REGRESS paragraphs (our initial example) would therefore be equivalent:

```
/ REGRESS DEPENDENT = INCOME.
/ REGRESS DEPENDENT = 2.
```

Variable naming being a rather new element of BMDP, there are still a number of commands, where one has to use the indices and must not use the variable names - e.g. when assigning MISSING values or defining which values of a variable are the maximal or minimal ones legal. (Looking at recent improvements to the command language, one can assume this restriction to be removed ere long.) When a case is encountered which has a variable which is missing or above or below the assigned minimum or maximum, it is internally recoded, so the original numeric value of this variable gets lost.

Very limited possibilities for recoding and assigning labels to values exist. In most cases, however, one will have to use explicit IF commands, if a more complicated recoding of a variable shall be done. Modifications of existing variables and creations of new ones have to be described in the / TRANSFORM paragraph. If new variables are to be created in the / TRANSFORM paragraph, their number has to be declared in advance in the / VARIABLE paragraph (!).

An existing variable can be modified, or a new one created, by an unconditional value assignment (e.g. AGEATMAR = (DAYS(MARIAGDT) - DAYS(BIRTH))/365) or by a conditional one (e.g. IF (AGEATMAR GT 70) ASTONISH=1). The logical operators and built-in-functions are roughly equivalent to the ones provided by SPSS (pre-X)), plus efficient handling of calendar dates after January 1, 1960. Two valuable extensions exist: 26 built-in-functions allow the definition of statistical measures within a case. 2 functions exist for the interpolation of missing data within one case. Lagging is cumbersome.

For case selection the error prone mechanism already described exists; additionally one can explicitly exclude a set of cases of which one knows the sequence numbers and furthermore some other commands have an implicit case selecting effect.

A facility for "generating a number of similar instructions" (i.e. for doing what can be done with DO REPEAT/END REPEAT in SPSS) exists. It is, however, certainly inferior to the one just mentioned. This inferiority in comparisons to the equivalent within SPSSś data modification language holds true for all features described that far. But BMDP has (besides the additional built-in-functions mentioned already) three features that are definitely beyond SPSS and are very helpful:

a. If a computational expression contains a value that is missing, the computed variable gets a missing value assigned according to very plausible rules.
b. BMDP contains a (not overly flexible) Macro facility, i.e., a project can define a central library of complex transformations needed frequently, the individual user including those she or he need without reformulating them.
c. Most important (and even better than the data handling facilities of SAS,

which are in all other respects incomparably superior to the ones of BMDP) the package contains a special routine for "multipass transformations". This is the BMDP wording for the case, that one wants to compute a variable which is a function not only of the values contained in the current case, but of the parameters of the distribution of a number of variables. For this purpose the BMDP1S program (Manual 59 - 64) contains a very flexible mechanism.

As already mentioned, BMDP provides an interface to an user specific FORTRAN program for complex data transformations which can not be handled within the data modification language of BMDP. While of course very useful and lending considerable flexibility to the program, that must not obscure the fact, that one will have to go this way all too often, if a more sophisticated data modification has to be taken care of.

The data modification facilities of BMDP can be compared to the possibilities provided by current SPSS; they have to be rated poor in the context of a software section dedicated to recent advances in the statistical packages beyond the earlier state of the art. Indeed, one would strongly advice the user with data requiring elaborate transformations, to use BMDP only for its statistical power and muster the services of another package for the data transformations. In this context it is reassuring, that fully developed interfaces are said to exist towards SAS and P-STAT, while SIR can at least write BMDP Save Files. If, for a given application, the additional statistical power of BMDP balances the problems inherent in the concurrent use of more than one package, the user has to decide(3).

d. MORE COMPLEX FILE HANDLING

BMDP has interfaces to FORTRAN, SAS, P-STAT and SIR. That is about the best what can be said about its capabilities to handle non-rectangular files. Even the approach described by this reviewer in issue 2 of this section (HSR 20/ September 1981) will create considerable difficulties with BMDP.

The package has some file handling capabilities which are useful, when administering the data of projects which use many different files or of such, which have to keep track of a large number of generations of a particular set of files which have undergone successive costly transformations. The BMDP Save File (Manual 65 - 71) is actually a library, containing within one physical file of the host computer a potentially large number of saved data sets, together with a large number of correlation and other matrices generated successively and stored for further analysis.

The manual mentions the merging of different data sets as one of the reasons, why one may have to write his or her own FORTRAN data transformation routine. No specific features are available for this purpose.

e. STRING HANDLING IN STATISTICAL PACKAGES

BMDP has not even rudimentary string handling capabilities; Report Generation is not supported either.

f. IMPROVED DISPLAY FACILITIES

BMDP does not support any of the modern plotting devices, nor does it support a specific interface to one of the specialized packages. Our judgement can not be as simple in this case, however, as it was with regard to string handling: a surprisingly high proportion of the BMDP programs recognizes a / PLOT paragraph, which allows the specification of lineprinter plots which visualize distributions analysed, residuals remaining or other properties of the data. The commands available for this purpose are rather flexible and allow, e.g., the printing of histograms side-by-side to make comparisons easier.

EDA is basically unsupported. There exists a possibility however, to produce stem-and-leaf displays (Manual 84/85). What this reviewer thinks more important than that is, that in quite a few places of the manual, not overly explicit statistically otherwise, it is hinted at the fact, that certain procedures have a rather exploratory character to begin with and should better not be interpreted as explanatory statistics. (E.g. Manual 208 or 602.)

CONCLUSIONS

BMDP is a statistical package. It provides excellent statistical instruments for a user with a good statistical background. One will have a hard time to use it for teaching the complete newcomer to statistics and/or EDP how to take advantage of its undisputed merits. The data handling capabilities are insufficient for any, but the most simple applications. Modern features, like refined file handling or improved methods of display are either missing or available only with considerable restrictions.

The package will be used advantageously in two set-ups: either in research programs, where all statistical and/or computer work is delegated to specialists. (With all the dangers inherent in such an approach.) Or in places, where another statistical package exists and is used for all simple computations and for all data preparation, using BMDP finally for the purpose it has been written for: the statistical analysis of data made machine readable with an explicit statistical method already in mind.

NOTES

1) Helmut Wilke: Datenmanagement mit Statistischen Programmpaketen - Möglichkeiten und Grenzen, in: Helmut Wilke et al. (Edd.): Statistik-Software in der Sozialforschung, Berlin: QUORUM, 1983, 211 - 252, here: 235.
2) This review is based upon: W.J. Dixon et al. (Edd.): BMDP Statistical Software. 1983 Printing with Additions, Berkely etc.: University of California Press, 1983. German Readers are strongly advised to consult: G. Bollinger et al.: BMDP. Statistikprogramme für die Bio-, Human- und Sozialwissenschaften, Stuttgart and New York: 1983. While necessarily selective in the description of the available statistical features, this is a much better introduction into the logic of filehandling than the system manual itself. The BMDP newsletter, called "BMDP communications", contains quite important contributions which clarify parts of the manual, describe additional features or report tricks. The substantial articles of former newsletters are therefore integrated into the current issue of the manual (pp. 698 - 707 in the 1983 edition). If it does not do so that far: convince your computing centre that the "communications" should be made accessible to the user community instead of being locked away somewhere.
3) To remain consistent within our reviews: we consider it a great argument for a system like SAS or P-STAT to be able to make accessible the statistical power of BMDP on top of what else those packages provide. We are considerably less convinced by a system as such, which has to be put on top of the data handling capabilities of another one.