The Advent of Dynamic Graphics Statistical Computing

Herbert F. Weisberg, *The Ohio State University*
Charles E. Smith, Jr., *The University of Mississippi*

Easy-to-use microcomputer statistics programs are becoming available, programs that take advantage of the dynamic interaction possibilities of personal computers. This article outlines these dynamic features and describes the performance of several microcomputer statistics programs.

The first generation of microcomputer stat programs was rudimentary; the sole advantage of these programs was that they allowed statistics to be computed on personal computers. Second-generation programs were easier to operate, and, as a result, were more useful for teaching. The development of microcomputer versions of the major mainframe stat programs offered further improvement by providing the full set of statistics available in their mainframe packages, but most were no easier to operate than their mainframe cousins. By contrast, newer dynamic programs provide significant advances that are worth examining.

Our experience with these programs was enriched in the summer of 1992 when the department of political science at the Ohio State University hosted a Faculty Workshop on Undergraduate Instruction in Data Analysis in Political Science sponsored by the National Science Foundation. Under the leadership of Aage Clausen, the participants worked on individual research projects in the afternoon while attending sessions on Microcomputer Applications for Data Analysis Instruction in the morning. The two of us used the morning microcomputer sessions to illustrate how data analysis programs could be used in the undergraduate classroom. Participants were introduced to a series of student exercises requiring basic statistical operations on datasets we assembled. We chose three statistics programs for these exercises: SPSS-PC, plus two less well known programs that we have found easy to use in undergraduate instruction—Mystat on the Mac and MicroCrunch on the PC. Additionally, we demonstrated evaluation copies of several other statistics programs. This process offered us a useful opportunity to compare the ease-of-use of several statistics programs. We soon discovered that the packages varied considerably in their dynamic features.

Dynamic Features in Statistics Programs

The term "dynamic features" refers to a relatively new set of capabilities beginning to emerge in statistics programs for microcomputers. Analysis using these features can be viewed as an extreme form of interactive analysis, variously referred to as "interactive linked-graphics," "bi-directionality," or "two-way analysis." We coin the term "dynamic" as a broad categorization of these features, and outline four program capabilities that come under the heading. The four capabilities are related in that they facilitate the user's ability to conduct several steps of statistical analysis without backtracking. Previous generations of data analysis programs had the user redo the analysis if a slight change was desired (such as adding a regression line to a scatterplot or deleting some cases from an analysis). Dynamic capabilities allow the user to modify results after they are obtained, thus minimizing repetitive efforts.

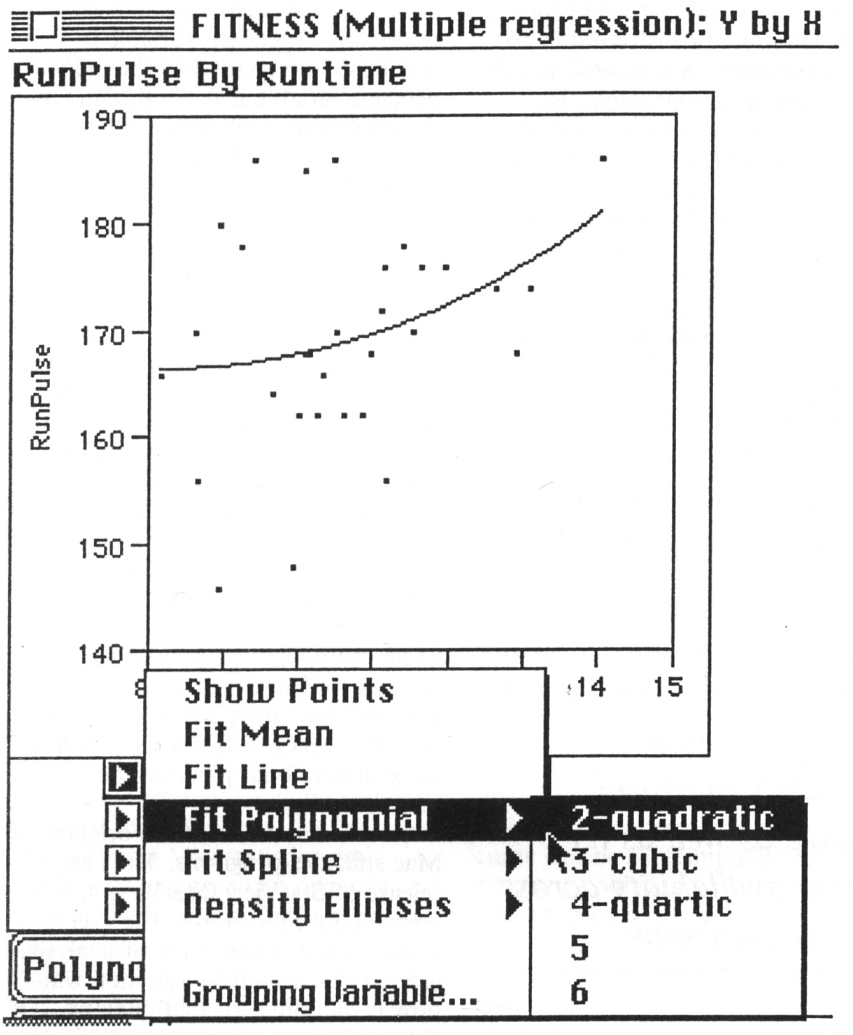
The simplest dynamic capability is *case identification:* the ability to identify data points in graphical displays. For example, the user may be able to highlight the name of a data case (or its value on a particular variable) in a data matrix and identify its corresponding point in a graph. Or, the user may be able to position a "question-mark tool" over a point in a scatterplot, prompting the program to display the name of the corresponding data case on the scatterplot. In both cases, the user is able to examine and integrate both graphic and tabular information rather than focusing exclusively on one or the other.

The second dynamic capability is *graph overlay:* the ability to add further statistical analysis to a graph without executing the commands that generated the original graph. For example, once a scatterplot is on the screen, several programs allow the user to request the addition of a regression line to the graph (see Figure 1 for an example from JMP which permits nonlinear as well as linear curves to be overlaid on the scatterplot). Similarly, the user may be allowed to request a graphic representation of the residuals from a scatterplot. Some programs even include the ability to rotate (or "spin") three-dimensional plots. In each case, the original analysis remains available and serves as the foundation for the overlays, minimizing the need for repetitive runs.

The third dynamic capability is *linked analyses:* the ability to move from statistical results and graphs directly to related analysis and displays without returning to the original command entry mode. This feature is often implemented via the appearance of a new set of menu choices once an analysis is performed. For example, once a correlation matrix is generated, a new set of menu choices may permit the user to select related graphic displays including the plotting of any or all of the bivariate correlations.

Alternatively, linked analysis can be implemented by permitting the user to point (using a mouse) to part of the displayed results and be given a choice of related options. For example, when Data Desk shows a

FIGURE 1.
Adding a Regression Line to a Graph in JMP



bivariate or multivariate analysis, pointing to the name of any variable gives the user a menu of choices which includes producing a histogram of that variable's univariate distribution. Similarly, when Data Desk shows a regression equation, pointing to a particular predictor gives the user a menu of related graphs (including a plot of the regression residuals against the selected predictor). This capability encourages a wide variety of linked analyses. The ideal implementation virtually suggests to the user what additional analysis could complement the current analysis.

The fourth dynamic capability is *recomputation after data modification:* the ability to modify the data and see the effect of the modification on the current results. If a scatterplot, correlation, or regression analysis has already been obtained, for example, the user might delete an observation, change a value on one of the variables in the analysis, or add an observation and then see how that changes the original results. Data Desk permits such changes to the data values while results windows are open on the screen. An exclamation point appears in any window where results have not been updated. Pointing to that symbol gives the user the choice of replacing the original analysis or redoing that analysis in a new window. Redoing the analysis permits the original and new results to be compared.

Systat and NSDstat+ allow the user to select a part of a scatterplot and then rerun the regression using only the cases with values in that part of the scatterplot. MicroCase also permits the user to request outliers to be located in a scatterplot, shows the original correlation between the variables, shows the new correlation if the outlier is deleted, and gives the user a yes/no choice as to dropping the outlier. Unfortunately, MicroCase defines outliers improperly—using values on the dependent variable rather than the size of squared residuals—so this capability is not very useful in practice. However, programs which permit proper identification of outliers (such as Systat and NSDstat+) offer the user a quick, easy, and graphic means of understanding the influence specific data points had on the original results.

Statistics Packages with Dynamic Features

The fact that there are at least four different dynamic features in statistics programs leads to some confusion in distinguishing between stat packages; programs may be advertised as "dynamic" in a very general sense when in fact they have only one dynamic feature, and perhaps not a very important one. Advertised claims that statistics programs are dynamic should therefore be taken with a considerable grain of salt. Some guidance can be had by considering a program's platform heritage as well as its current availability across various platforms. That is, programs which have just been ported over from the mainframe with minimal modifications (such as SPSS/PC and SPSS-Studentware) generally do not provide any dynamic features. Similarly, programs that attempt to be identical across a wide variety of computers (such as Minitab) generally do not provide dynamic features, as the key dynamic features cannot be implemented on all computers.

On the other hand, programs written from the perspective of exploratory data analysis (Data Desk particularly) generally provide a wide

https://doi.org/10.2307/419835 Published online by Cambridge University Press

range of dynamic features. Similarly, programs oriented towards teaching tend to be more dynamic than those oriented towards research. Finally, MS-DOS programs are less likely to contain dynamic features than Macintosh programs, though these capabilities are likely to be available as statistical programs are written for the Windows environment. Indeed, SPSS for Windows exhibits some dynamic features, even though the other personal computer versions of SPSS do not. In general, though, the dynamic features of programs written for Macintosh and MS-DOS environments differ. In the following sections, we review the characteristics of several programs for each environment.

Macintosh Statistics Programs

The graphical user interface on the Mac has been most hospitable to the development of dynamic statistics programs. Data Desk (available in a large-capacity professional edition as well as a more limited capacity student edition) takes full advantage of the interface, and contains the most extensive set of dynamic features of any program we have tested. Authored by Paul Velleman, one of the main advocates of exploratory data analysis, Data Desk effectively captures the essence of the exploratory analysis movement. It provides each of the four dynamic features described above: the ability to identify data points in graphs, the ability to overlay and modify graphs (including spinning three-dimensional plots), the ability to move easily from statistical results to related analyses, and the ability to alter analyses when the data are modified. Each statistical analysis includes "hyperview" menus suggesting possible next steps in the data analysis (such as histograms of the variables or regression diagnostics). Displays generated from the hyperview menu are linked to previous displays and to one another, so selecting a bar in a bar chart highlights the corresponding points in any open scatterplots. The program is an exemplar of how statistical analysis can be made dynamic. While several other programs mentioned in this report have some dynamic elements, Data Desk is thoroughly dynamic in its operation.

Data Desk is best when dealing with correlation and regression for numeric variables or distributions for nominal variables. A wide variety of plotting options are available. By contrast, its crosstabulation features are more minimal. For example, the only statistic provided on crosstabs is a chi-square value. More advanced statistical procedures are also missing from the program. When we demonstrated Data Desk in the data analysis instruction workshop, the Workshop participants were truly excited by its dynamic features, but recognized the effects of these other limitations as well.

Advertised claims that statistics programs are dynamic should therefore be taken with a considerable grain of salt. Some guidance can be had by considering a program's platform heritage as well as its current availability across various platforms.

JMP (from the SAS Institute) and JMP-IN (the student version of JMP) are intellectually intriguing programs. The SAS Institute has included six basic analysis options in the programs: distribution of Y's, fit Y by X, fit Y by X's, nonlinear regression, spin, and Y by Y's. The user does not command the program to perform a type of analysis (such as regression). Instead the user chooses the desired analysis option after specifying which are the X and Y variables, and then the program automatically chooses the analysis technique appropriate for the levels of measurement of the specified variables. For example, if the user chooses the "fit Y by X" option, the program performs regression analysis if both variables are interval level, one-way analysis of variance if the dependent variable (Y) is interval while the independent variable (X) is ordinal or nominal, logistic regression if the dependent variable is ordinal or nominal while the independent variable is interval, and contingency table analysis if neither variable is interval. The approach is interesting, but analysts may find it limiting, and teachers may wish the program allowed students to decide which statistical strategies are appropriate for various levels of measurement.

However, JMP and JMP-IN have several attractive dynamic features. First, clicking on a point in a scatterplot shows the corresponding case number; points can also be "brushed" in the scatterplot to identify the corresponding rows in the data matrix. Second, related statistics can be requested when an analysis window is open, such as adding a linear regression line or a quadratic curve to fit the points in a scatterplot, or computing a normality test on a frequency distribution. Third, three-dimensional plots can be rotated freely.

StatView was one of the original Mac statistics programs. The new release of StatView (StatView 4.0) is remarkably easy to use. The user clicks on the desired type of analysis, after which a variable selection window is modified for that analysis. When regression is chosen, the variable selection window allows the user to select a variable and then designate (by clicking one box or another) that variable to be a Y variable, an X variable, or a "select by" (control) variable. However, it is dynamic only in the sense that it recalculates all open analysis results when cases are modified (such as when their rows are deleted from the data matrix). StatView Student does have a few dynamic features. Graphic overlays are implemented to the extent of permitting the user to add confidence bands to the scatterplot for a regression or to add error bars to the scatterplot for a correlation. Also, StatView Student recomputes open statistical windows when cases are deleted from the data matrix: double-clicking on the row of a case in the data matrix results in the recomputation of any visible statistical windows with that corresponding case deleted.

https://doi.org/10.2307/419835 Published online by Cambridge University Press

Systat, developed by statistician Leland Wilkinson, is always rated as one of the top microcomputer statistics programs. Its student version, Mystat, is an effective, inexpensive teaching tool that both of us have used in the undergraduate classroom. However, the dynamic features of these programs are more limited than those of Data Desk and JMP. Systat and Mystat for the Mac provide the ability to identify data points in graphical displays. Systat has the further capability of limiting subsequent analysis to a subset of cases that has been selected from a graph. It can also rotate three-dimensional plots.

We have not examined the Mac version of **Statistica,** but the program is designed to be interactive. Statistica/Mac puts all of its output into "scrollsheets" which can be manipulated directly. For example, pointing to a variable in a descriptive statistics table leads to the presentation of its histogram with a normal curve drawn over the histogram. Given our experience with its PC cousin, we would, however, expect its number of dynamic features to be much less than that of Data Desk.

While most Mac statistics programs have at least some dynamic features, it is notable that two programs related to mainframe packages are not at all dynamic: **SPSS** for the Mac and ***Minitab.*** SPSS for the Mac does use a command generator to help build commands, and Minitab does use a pull-down menu procedure to help build commands. However, both programs execute the commands in batch fashion. In SPSS, for example, all the output goes to a single output window, and there is no possibility of making part of that output (such as a chart) active and subject to further manipulation.

All in all, Data Desk and JMP are the most dynamic of the Macintosh statistical programs reviewed here. Indeed, Data Desk is so complete in its dynamic features that it virtually defines what dynamic statistical analysis is all about. Political scientists who enjoy doing statistical research on computers should take a look at Data Desk to see what this mode of analysis can involve. We hope that Data Desk is soon ported over to the Windows environment on MS-DOS computers, but meanwhile some MS-DOS programs are promising in their incorporation of dynamic elements.

MS-DOS Statistics Programs

Dynamic features are also available in several MS-DOS programs. One of the most interesting new statistical programs for MS-DOS is **NSDstat+,** from the Norwegian Social Science Data Archives. A simpler version of this program was originally developed for high schools in Norway, which correctly suggests how easy it is to use. The "plus" version incorporates a wider range of statistical features. It has been adopted in other Scandinavian nations and in Germany, but it is just beginning to be available in English-speaking markets. Its ability to import SPSS portable files is attractive, as is its use of modern VGA graphics and generation of high quality plots. Its mapping facility is also particularly well-executed.

NSDstat+ also implements several dynamic features. First, graphic overlays are implemented in several procedures. For example, a normal curve can be superimposed over a frequency polygon and a box-and-whisker plot can be added to the same graph. Similarly, a regression line can be added to a scatterplot. Linked analysis is also encouraged. Most statistical procedures generate new menus, suggesting and allowing more analysis. When a correlation matrix is obtained, for example, the user can then choose a graphical display which gives all the corresponding bivariate plots (plus the corresponding univariate histograms if they are next requested). Also, it is possible to subset cases from a graph. When a scatterplot is obtained, the user can extract parts of the scatterplot to keep (and recompute the bivariate regression for those cases that are kept). Participants in the undergraduate data analysis workshop were particularly impressed by the dynamic features of NSDstat+, and they appreciated its more standard capabilities as well.

CSS:Statistica is a high-end statistics program for MS-DOS computers that contains an unusually wide variety of statistical procedures. The program is quite expensive, and there is no inexpensive student version. Its graphics are impressive looking, but slow and awkward to produce and control. CSS:Statistica has some dynamic features. Three- and four-dimensional plots can be freely rotated. Also, after obtaining a correlation matrix, the user can highlight a particular correlation cell and then ask for the corresponding bivariate scatterplot. The user can also have the program highlight the significant correlations in a matrix. Linked analysis is implemented through new menu choices after each statistical computation. However, regardless of the implications of their advertisements, the dynamic features of CSS:Statistica are much more limited than for the most dynamic programs.

MicroCase is being marketed aggressively as a stand-alone statistics program and also bundled along with American government and statistics texts. MicroCase has limited dynamic features. Its best is the ability to add regression lines and show residuals on a scatterplot. As mentioned above, it can also identify outliers in that plot and show the effect of their removal, but its definition of outliers is not conventional, which renders this capability misleading. Its mapping routine permits a user to see the data value corresponding to a geographic region (such as a state's Democratic vote proportion in a recent election). However, this is implemented rather clumsily; rather than permitting the user to point directly to the state on the map, users must repeatedly press the arrow keys to direct the program to give the data value for each state in turn. Crosstabulations can be explored by going back and forth from frequencies to percentages, expected frequencies, and association statistics, though this does mean that it is impossible to see cell frequencies and percentages simultaneously. The CGA graphics in MicroCase result in less professional graphs than those in NSDstat+ and Statistica.

CHIPendale has some dynamic features, but is limited to crosstabulation analysis. Originated by sociological researcher James Davis, this program is a useful means of exploring causal order and seeing

https://doi.org/10.2307/419835 Published online by Cambridge University Press

how controls affect relationships in tables. It is student-oriented, and Jere Bruner is developing a textbook that uses CHIP with political science applications. The data entry is unusual in that CHIP does not read raw data; instead the instructor must prepare large multi-way crosstabs (such as all the combinations of cell frequencies for five variables) and input them into the program. Afterwards, the student can easily generate any crosstab based on a subset of those variables. This program is best seen as more specialized than the other programs mentioned in this article, though it may be attractive for some teaching situations.

Some other MS-DOS statistical programs with which we are familiar do not have dynamic features. In particular, **MicroCrunch** (which we consider one of the best statistics packages for students on MS-DOS machines) lacks dynamic features. **Mystat** on the PC also does not provide such features (and is much less user friendly than its Mac cousin). **Minitab** is nondynamic since it seeks to be identical in capabilities across a wide variety of computers. Neither **SPSS-PC** nor SPSS-Studentware offer anything in the way of dynamics, but the preliminary glance we have given to SPSS for Windows suggests that its pull-down menu procedure may provide some dynamic features.

In summary, MS-DOS statistical programs do not incorporate dynamic fetaures as extensively as do some Macintosh programs. Of the programs we have managed to review, NSDstat+ is the most promising MS-DOS program for this type of analysis. Some of the other MS-DOS statistics programs promise more dynamic operation than they deliver.

Conclusions

Dynamic capabilities are, by their nature, most likely to be included in programs that are written for microcomputers and that are not similar to larger statistical programs used on mainframe computers. Thus, we are left with the paradox that the dynamic environment on microcomputers is ideal for teaching statistics though the current programs do not generalize to mainframe computers. For now, we can only hope that students learning data analysis on dynamic microcomputer programs will be able to transfer what they learn to other data analysis programs in later life.

Eventually, dynamic features may modify how we perform our own data analysis as well as how we teach statistics. Instructors teaching data analysis to undergraduates would enjoy trying some of the programs that incorporate dynamic features. So far, the program that stands out most in terms of dynamic capabilities is Data Desk on the Mac. Other programs to consider in this category are JMP (also on the Mac), and, as it becomes available in the United States, NSDstat+ for the MS-DOS platform. These programs are not related to mainframe computer programs, but they are easy to use and easy to learn.

Availability of Programs Mentioned in This Article

CHIPendale 3: ZetaData, 25 Haskins Rd., Hanover, NH 03755. Phone: (603) 643-6103 ($500 for department site-license).

CSS:Statistica (MS-DOS) and Statistica/Mac: StatSoft, 2325 E. 13th St., Tulsa, OK 74104. Phone: (918) 583-4149 ($795 for PC version and $495 for Mac version; smaller subsets of both programs are available for $295).

Data Desk 4.0 (Mac only): Data Description, P.O. Box 4555, Ithaca, NY 14852. Phone: (607) 257-1000 ($595; academic price $357; between $17 and $29 for *Learning Data Analysis with Data Desk* by Paul Velleman, which includes Data Desk Student through W.H. Freeman & Co., 41 Madison Ave., New York, NY 10010).

JMP and JMP-IN (Mac only): SAS Institute Inc., SAS Circle, Box 8000, Cary, NC 27512. Phone: (919) 677-8000 ($40 for JMP-IN).

MicroCase (MS-DOS only): MicroCase Corp., P.O. Box 2180, West Lafayette, IN 47906. Phone: (317) 497-9999 ($395 for full version; student versions are available with an American government book and with a statistics book).

Minitab (Mac and MS-DOS): Minitab, Inc., 3081 Enterprise Dr., State College, PA 16801. Phone: (814) 238-3280. A student version is available.

MicroCrunch (MS-DOS only): SofTex Micro Systems, 7915 Glenbrae, Houston, TX 77061 ($129 for a single copy; $300 for a department site license; $500 for a university site license; plus $4 for shipping/handling).

Mystat (Mac and MS-DOS): Course Technologies. Phone: (800) 648-7450 ($14).

NSDstat+ (MS-DOS only): Norwegian Social Science Data Services (NSD), Hans Holmboesgate 22, N-5007 Bergen, Norway. Phone: 011-47-5-21 21 17; Fax: 011-47-5-96 06 60; E-mail: nsd@cc.uib.no (U.S. pricing not yet determined).

SPSS Studentware (MS-DOS only): SPSS, Inc., Chicago, IL. Phone: (800) 543-9263 ($40).

StatView 4.0 and StatView Student (Mac only): Abacus Concepts, 1984 Bonita Ave., Berkeley, CA 94704. Phone: (800) 666-STAT ($595 for StatView; $99 for StatView Student).

Systat (Mac and MS-DOS): SYSTAT, Inc., 1800 Sherman Ave., Evanston, IL 60201. Phone: (800) 428-STAT.

About the Authors

Herbert F. Weisberg is professor of political science at the Ohio State University. His most recent books are *Classics in Voting Behavior* and *Controversies in Voting Behavior* (both coedited with Richard Niemi).

Charles E. Smith is currently an advanced graduate student in political science at the Ohio State University and, beginning August 1993, assistant professor of political science at the University of Mississippi. His areas of specialty include voting behavior and research methods.