An Overview of the *AXIS* Statistics Package
Version 2.0, July 1995

Robert Stine
stine@wharton.upenn.edu

Department of Statistics
The Wharton School, University of Pennsylvania

Introduction

AXIS provides a graphical statistical modeling environment that runs on most computers. Included are a variety of recent developments in statistics, including scatterplot smoothing, plot linking and brushing, and bootstrap resampling. The focus is the graphical display of data and the manipulation of regression-like models. The user interface is itself graphical – one constructs models by manipulating icons.

This paper introduces *AXIS* with an example that exploits many features of the software. Following some installation instructions, the example shows the essentials of loading data, making plots, and building a regression model. After the example is a list of frequently asked questions (with brief answers) and an overview of the software itself.

AXIS is written entirely in *XLISP* and relies upon the Lisp-Stat software of Tierney (1990). You must have installed Lisp-Stat in order to use *AXIS*. The Lisp-Stat software is available free of charge through e-mail from statlib (send an e-mail message with the single line `send index from xlispstat` to *statlib@lib.stat.cmu.edu*) or via anonymous *ftp* from *umnstat.stat.umn.edu*. The reply message from statlib includes introductory instructions as well as a list of the additional items associated with Lisp-Stat. See the frequently asked questions at the end for further suggestions about keeping up to date.

Installation

AXIS runs on any computer that supports Lisp-Stat. Hence, you can run *AXIS* on a PC with Windows, a Macintosh, and various Unix workstations. The instructions for installing Lisp-Stat on your machine are part of the distribution of that package. To install *AXIS* on a machine with Lisp-Stat, copy all of the *AXIS* files into a directory. For example on a PC, make a subdirectory for the new files with the command `mkdir axis` and then copy the files into this subdirectory. The files for the current version of *AXIS* are:

axis.lsp	axisBoot.lsp	axisCmd.lsp	
axisCorr.lsp	axisData.lsp	axisDens.lsp	
axisIcon.lsp	axisRegr.lsp	axisRobR.lsp	
axisScat.lsp	axisSE.lsp	axisUtil.lsp	controls.lsp
compare.lsp	icons.lsp	labelPlt.lsp	patch.lsp

In addition, several data files are also available. Typically these files, such as `duncan.dat`, are put in a subdirectory, such as `axis\data`. The instructions for starting *AXIS* appear in the illustrative example that follows an overview of the software.

Using AXIS

Before starting the introductory example, its useful to be aware of the general strategy for using this software. Reading a data file produces a window on the screen with rectangular icons (colored boxes) that identify the variables in the data set, one icon for each variable. These icons behave much like file icons defined by Finder on a Macintosh or Windows. In addition, the initial view of the data set contains two oval icons named *Filter* and *Labels*. These icons let you define subsets of the data and label observations. Oval icons always define features of the data set being analyzed; these icons represent functions on the data set whereas the rectangular variable icons denote data.

Double clicking opens a view of the item represented by the icon. For icons denoting variables (rectangular), a double click opens a dialog view that allows you to define various properties of that measurement, such as its name. For feature icons (ovals), a double click opens a dialog that defines a property of the data set, such as the case labels for plots.

The data analysis itself proceeds as in any interactive statistics program. The first step is typically to label the cases. Just "drag and drop" the icon associated with a variable on the oval *Labels* icon to define case labels. Alternatively, double clicking on the *Labels* icon opens a dialog that allows you to enter an expression for the labels variable. Similarly, you can use the *Filter* icon to select a subset of the data. Most of the time, however, one does not start by subsetting the data; this comes later as the analysis becomes more detailed.

With the case labels defined, the analysis turns to graphical displays such as univariate plots and scatterplots. In general, you type less if you will pick the variables before choosing a statistic from the menu. To pick one variables, click on its icon; to pick several, shift click on each. With the variable icons chosen, pick a procedure from the **Statistics** menu. The names of the selected variables are automatically entered in the dialog view, and most times all you need to do to see your results is click the procedure button. For statistics which have two types of factors (such as regression, with a predictor and response), the first icon selected denotes the response variable. Now, on to a small example.

An Illustrative Problem

This example uses a small, well-known sociology data set – the Duncan data set containing the income, education and occupational prestige of 45 occupational categories. The usual feature of interest in this data is the association between the prestige rating of the occupation and the income and education of attained by members. This data appears in many publications, such as the Fox (1992) monograph on regression diagnostics. This data set is also featured in the March, 1995, issue of *Sociological Methods and Research*, an issue devoted to computing environments for statistics. This special issue includes a paper on Lisp-Stat.

Preparing the Data

The first step of our analysis is prepare a data file that holds the Duncan data. *AXIS* expects the input data file to be a text file (ASCII) with a simple, rectangular format. The first line of this file must be a documentation string, delimited by double quotation marks. This string can be as long as you care to enter (spanning several lines if you like) as long as it begins and ends with double quotation marks. The next line of the file holds the variable names (without quotes or spaces within the names); *AXIS* is not case sensitive. Each remaining line of the file contains the values for a single observation on each of the variables listed on the second line of the file. The data file ought to resemble a spreadsheet with variable names as column headings. Spacing is not important and one blank is the same as several blanks. Here are the first few lines of the data file holding the Duncan data (the associated file, `duncan.dat`, is available with the *AXIS* distribution):

```
"Duncan data on occupational status."
TITLE          INCOME  EDUCATION PRESTIGE
accountant            62  86 82
airline_pilot         72  76 83
architect             75  92 90
author                55  90 76
chemist               64  86 90
minister              1  84 87
professor             64  93 93
dentist               80 100 90
...
```

You should notice several features of the data file. The columns need not line up and the variable names are not necessarily aligned over the columns. It is important, though, that each line of the file after the documentation string have the same number of distinct elements (distinct in the sense of being separated by one or more spaces). Data items must be contiguous with no extra spaces. For example, the occupation title `airline_pilot` includes an underline character to avoid an unwanted blank. If a data value is missing, use the symbol `na` to fill in the void. Though the current release of *AXIS* does not handle missing data very well, future versions will (as soon as Lisp-Stat does). The values of variables can be text or numeric. For example, the occupation titles are not numeric, and we will use them in the analysis to label observations.

Launching AXIS

With the data file prepared, we can launch Lisp-Stat and load the *AXIS* interface. *AXIS* requires Lisp-Stat; indeed, one starts *AXIS* once Lisp-Stat has already begun. The manner in which this is done depends upon the system being used. For a Windows PC, get Windows (hopefully at least version 3.1) running, start the Lisp-Stat editor LSPEDIT, and then start Lisp-Stat. On a Macintosh, double-click the Lisp-Stat icon. Now use the **Load** item from the **File** menu of Lisp-Stat, and select the file `axis.lsp`. You will be rewarded with a pause as the collection of 16 program files are loaded into Lisp-Stat. The Lisp-Stat Listener window, which displays the results of commands, should resemble the following.

```
> ; loading "axis.lsp"
; loading "patch.lsp"
; loading "axisUtil.lsp"
; loading "axisCmd.lsp"
; loading "axisBoot.lsp"
; loading "axisData.lsp"
; loading "axisIcon.lsp"
; loading "icons.lsp"
; loading "axisDens.lsp"
; loading "labelPlt.lsp"
; loading "axisScat.lsp"
; loading "axisRobR.lsp"
; loading "axisRegr.lsp"
; loading "axisCorr.lsp"
; loading "axisSE.lsp"
; loading "controls.lsp"
; loading "compare.lsp"
; finished loading "axis.lsp"
```

Now we are ready to start the data analysis.

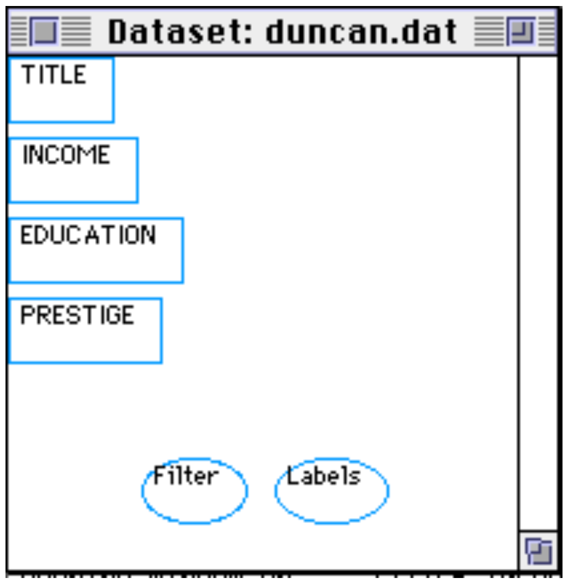
As the *AXIS* files are loaded, the software augments the menu bar of Lisp-Stat. Two additional menu items now appear: **Data** and **Statistics**. Take a look at the options provided by these menus by clicking on the menu name and looking at the pull-down list of options. The **Data** menu is pretty short. It has a two items: Open dataset file... and Write dataset file.... The **Statistics** menu has more items. The menu shows several graphing tools and analysis routines, including a simulation tool related to bootstrapping. The items in the **Statistics** are grayed out and cannot be executed until you create a data set window. We can't use the statistics until the data file has been read.

Reading the Data File

Go back to the **Data** menu and select the item Open dataset file... This choice will cause the system to present you with a standard open-file dialog (standard for whatever system you are using). Locate the file `duncan.dat` on your system and request that the file be opened. Lots of things start to happen.

First, a lot of text appears in the Listener window. Don't worry if you have trouble reading it as it goes by – the Listener window is scrollable and saves recent lines that have rolled off the top. As the text flows by, a window with several *icons* appears on the screen. This window should resemble the following figure. The item **Icons** also appears on the menu bar whenever this data set window is open. (This figure is taken from a Macintosh system; the window on a Windows PC will look slightly different.)

Figure 1. Initial view of the Duncan data showing 4 variable icons and the two feature icons that control subsetting and case labels.



The rectangular boxes or icons represent the four variables in this data set: job title, income, education, and prestige. The two ovals represent special properties of the data set and are called feature icons. Any of these icons may be freely moved about the window by *dragging* – clicking on the icon using the mouse cursor and holding down the mouse button while simultaneously moving the mouse. Basically, this view of the data set resembles the file view presented by the Macintosh Finder or Windows.

Some Graphical Preliminaries

It is useful in interactive data analysis to have case labels that identify observations associated with the points that appear in plots. In this example, the TITLE variable identifies the occupations and makes a great labeling choice. To use the values of the TITLE variable as case labels, drag the TITLE icon over the *Labels* icon and release the mouse (drag and drop). Make sure that the center of the icon (as denoted by the small crossing lines which are shown as you drag the icon) is within the feature icon before releasing the mouse button; otherwise, you have simply moved the icon. Harmless, but not what we need to do. After dropping the titles over the labels icon, the icon appearance changes as shown in Figure 2 to have a heavy oval border, indicating that case labels are defined.

Figure 2. The border of the Labels feature icon becomes heavy once the associated labels variable is defined.

Alternatively, if you have trouble with drag and drop or need more flexibility in defining the labels, there is a second way to define the contents of a feature icon. A double-click on the case label icon opens a dialog window (double-clicking takes some practice in Lisp-Stat, so be patient, particularly with Windows). Position the mouse cursor in the text field of the dialog and type `title` in the box, and then click on the OK button. Remember, Lisp-Stat and *AXIS* are not case sensitive, so you can mix upper and lower case characters in names.

We will not be using the title variable further so let's remove it from the view of the Duncan data. Click on the TITLE icon with the mouse (which makes the icon filled in or highlighted) and select Cut from the **Icon** menu. After a little visual feedback, this icon disappears. Only the icon disappears – the TITLE variable has not been altered. Icons act as visual reminders of the contents of a data set and *cannot* delete variables. Now let's look at some plots of this data.

The remaining variable icons (INCOME, EDUCATION, and PRESTIGE) provide quick tools for browsing the associated data. Hold down the option key and simultaneously click on an icon – an *option click*. This action produces a "pop-up" menu to appear, offering to

- print the data for that variable with case labels,
- describe it with some summary statistics,
- draw its histogram with a superimposed kernel density, or
- plot the data in the order of the cases in the data file (a sequence plot).

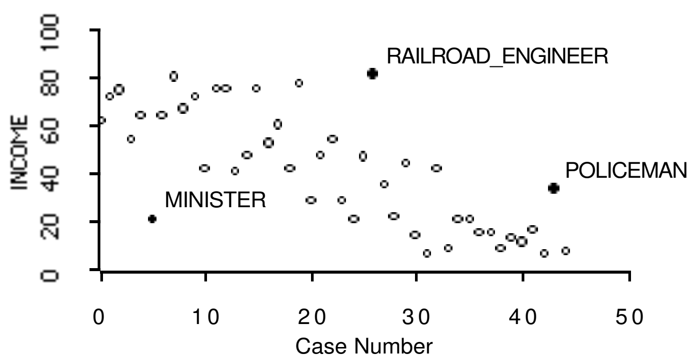
An excerpt of the printed listing of INCOME appears below; the command prints 45 lines, one for each occupation, on the screen the in Listener window. All of the printed output from *AXIS* appears in the Lisp-Stat Listener window, the main window of the application.

```
(ACCOUNTANT                     62)
(AIRLINE_PILOT                  72)
        . . .  41 more occupations  . . .
(POLICEMAN                      34)
```

```
(RESTAURANT_WAITER                8)
```

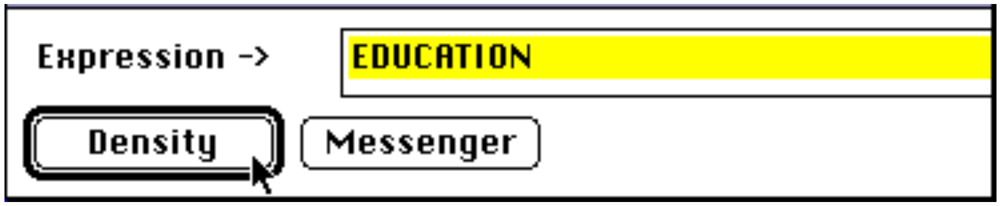
As shown in the partial listing of the cases, each occupation is labeled by its title. Had we not defined the label variable by filling that feature icon, cases numbers would have been used instead. Figure 3 shows the sequence plot for INCOME, with several cases highlighted. A sequence plot shows the values of the variable plotted on the case number, the position in the original data file `duncan.dat`. In order to highlight observations (here, minister, policeman, and railroad engineer), use the mouse to position the cursor over the point denoting an occupation. Then click the mouse. Use a shift click to select several points. The presence of case labels makes it very easy to identify the occupations with either very low or very high incomes.

Figure 3. The sequence plot for INCOME shows the values of income for each occupation, plotted in the order in which the observations appear in the data file.



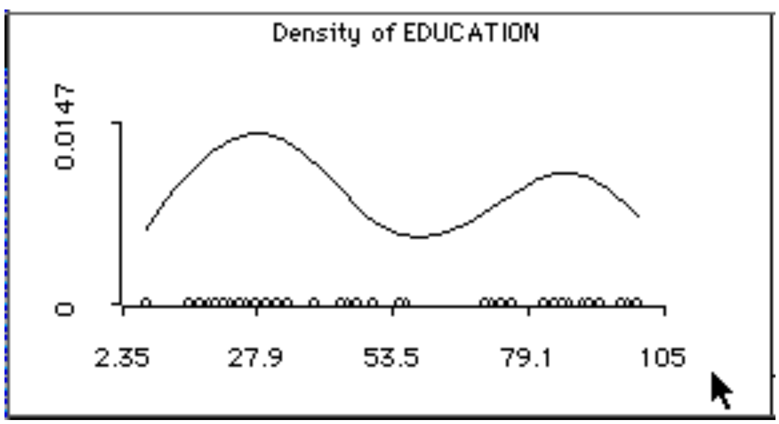
Before investigating relationships among these variables, it is useful to examine each one more closely. The Univariate item on the **Statistics** menu from offers a variety of univariate summary plots, including histograms, kernel density plots, and normal quantile plots. For this example, select the EDUCATION icon (this will save having to type the name of the variable in the coming dialog window). Next, select the Univariate menu item. The dialog in Figure 4 appears on the screen.

Figure 4. Choosing the Univariate menu item a variable selected opens a dialog whose text field is initialized to the name of the selected variable. In this case, the EDUCATION variable is entered.



The dialog opens with the names of the selected variables appearing in a text-edit field. This dialog item allows you to edit the name of the variable or define a transformation, such as one to take logs of the data. Clicking the mouse on the Density button (as suggested by the position of the cursor in Figure 4) opens a kernel density plot of the data defined by the contents of the text-edit field. In this case, the EDUCATION data. (The Messenger button is described later. Basically, it sends "messages" to the plot that the command builds.) Figure 5 shows the resulting density plot – it seems quite bimodal. The points along the base of the figure denote the observations.

Figure 5. The univariate command in Figure 4 produces a kernel density plot of the EDUCATION data for the 45 occupations. The points along the base of the plot indicate the education values for each occupation

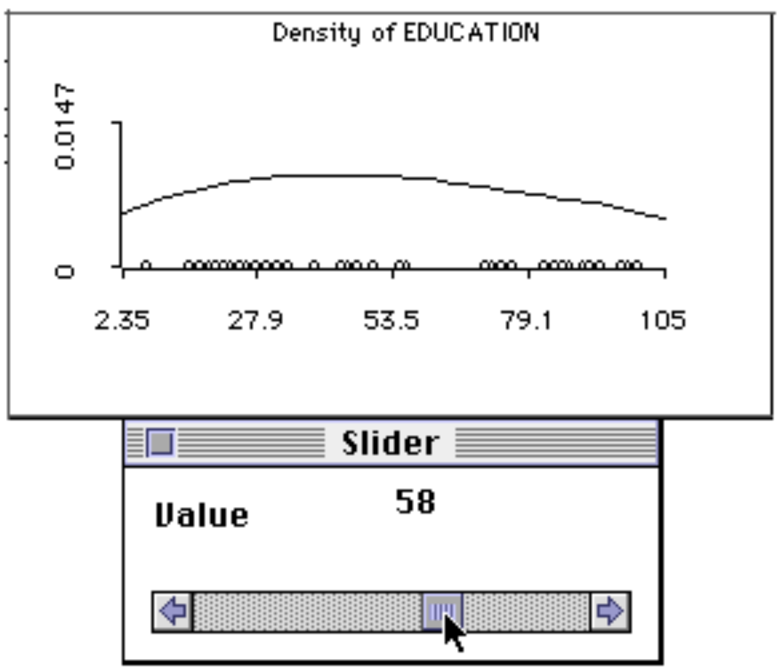


The dialog for the univariate command remains on the screen after the plot appears, just in case you want to change the variable (such as define a transformation) or entered the wrong name. To make the dialog disappear, click on the window's close box (whose location varies from system to system).

Now use the mouse to click on the density plot itself. This click on the density plot window brings this window "to the front" and adds a new item to the menu bar – the **Density** menu item. Most of the plots in *AXIS* offer their own menu of commands for modifying the display; the special window-specific menu only appears for the front-most window. If you do not see the sought menu item, make sure that you have selected the plot by clicking on it once with the mouse button. In this case, we can also show a boxplot or a normal curve. The default display options, to show the data and the kernel estimate, are checked in the pull-down **Density** menu.

The **Density** menu item also allows us to explore the sensitivity of a kernel density estimate to the associated smoothing parameter. The Kernel Slider menu item opens a slider dialog that interactively modifies the level of smoothing. Figure 6 shows what happens when we increase the level of smoothing using the slider. Too much smoothing conceals the bimodality and gives an estimate that resembles a normal distribution. You can compare this density to the standard normal approximation using the Show Normal menu option. A second slider lets you vary the power in a power transformation of the data. The transformation slider is useful for determining if a simple monotone transformation of the data can induce symmetry.

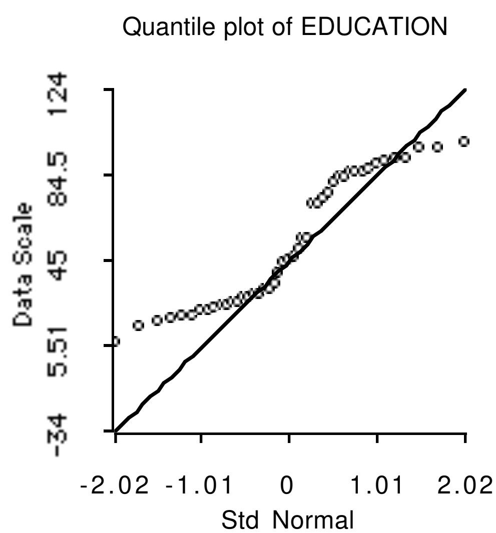
Figure 6. The density menu item Kernel Slider opens a slider that controls the amount of smoothing in the kernel density plot. Moving the slider to the right as in this example increases the amount of smoothing in the kernel density.



The **Density** menu is also the gateway to two other useful univariate plots: the normal quantile plot (abbreviated as QQ-plot) and the symmetry plot. A normal quantile plot shows the ordered items from the data plotted on the quantiles of a standard normal distribution. These quantiles indicate how the data ought to be distributed if indeed the data are a sample from a normal population. If the data are a sample from a normal population, the values in the QQ-plot lie near a 45 degree line. Systematic deviations from this diagonal indicate that the data are not a normal sample. In Figure 7, the S shape in the QQ-plot of EDUCATION shows a strong deviation from the values expected under normality.

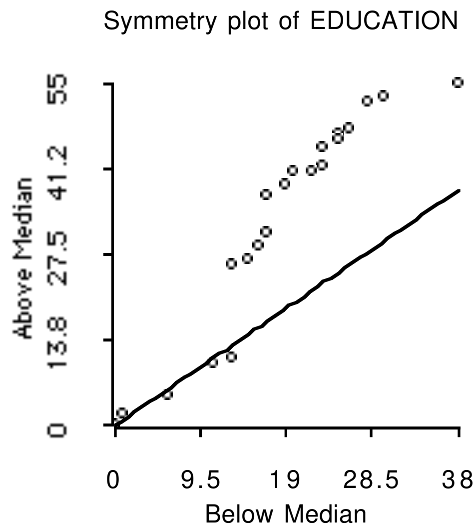
A QQ-plot focuses on deviations from normality. A symmetry plot emphasizes deviations from symmetry, such as skewness. Data, such as EDUCATION, might be non-normal, but still symmetric about its median. Figure 8 shows the symmetry plot for EDUCATION. Again, consider the deviations from reference line. In a symmetry plot, deviations from the line indicate a lack of symmetry.

Figure 7. An item on the **Density** menu builds a QQ plot of the data shown in a kernel density plot. Here the data is EDUCATION. Deviations from the shown diagonal line indicate that the data are not from a normal population.



The coordinates that make up a symmetry plot are deviations from the median of the data. Points on the x-axis are the ordered absolute distances of observations *below* the median where as the y-axis are the ordered absolute deviations above the median. The plot shows pairs formed by pairing the first above with the first below, the second above with the second below, and so forth. Thus, symmetry plots only have half as many points as there are observations.

Figure 8. Deviations from the line drawn in a symmetry plot indicate a lack of symmetry in the data. An item on the **Density** menu builds the symmetry plot.

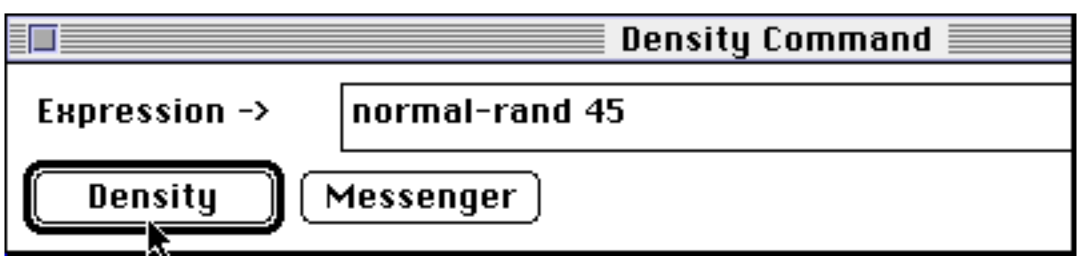


These diagnostic plots can be confusing since so much depends upon a visual impression, leaving the question: What's a big deviation? Simulated samples help answer this question. The following univariate plots are based on a simulated sample of 45 observations from a normal population. To get these, just go back to the density dialog and enter the expression

```
normal-rand 45
```

so that the dialog is as in Figure 9. This command, typed in the input of the univariate dialog, generates a sample of 45 values from a standard normal distribution.

Figure 9. One can enter an expression in the density dialog to show plots associated with samples from a normal distribution. This dialog generates plots a sample of size 45 from a standard normal.



First, consider the density plot itself. What should this look like for a sample from a normal population? Click on the density button in the command dialog several times. Each click generates a new, independent sample from the same normal population. Figure 10 shows density plots for two normal samples. As the plot on the left shows, one has to be careful about interpreting bumps in a density plot. Two bumps are evident, yet we know that this data are from a single normal population.

Figure 10. Density plots for two samples of size 45 from a standard normal population illustrate the difficulty of detecting normality in a density plot. The density plots are quite different even though both samples come from the same population.

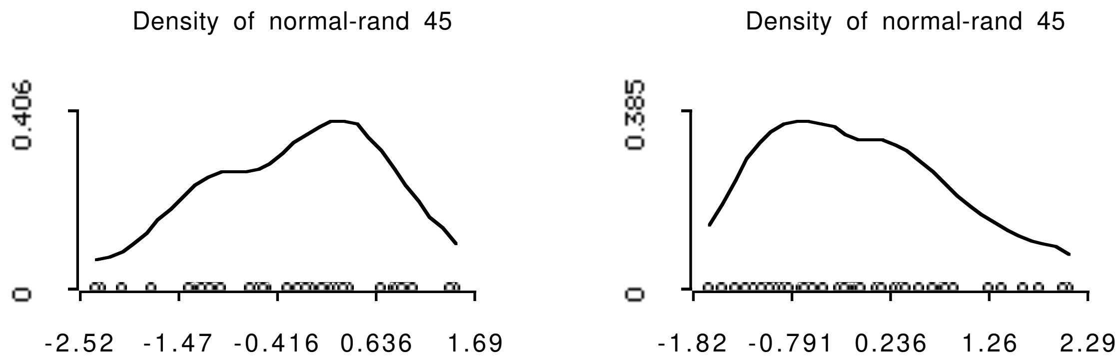
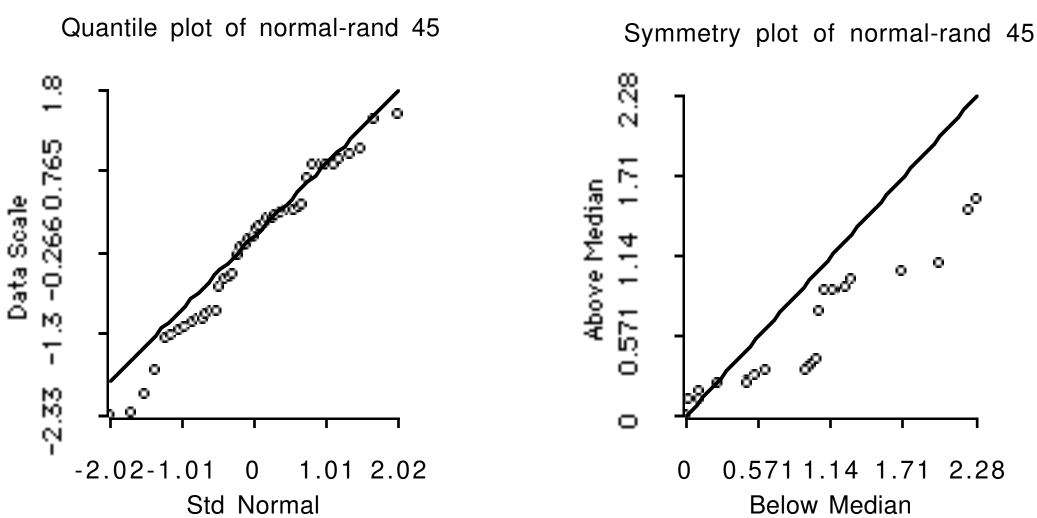


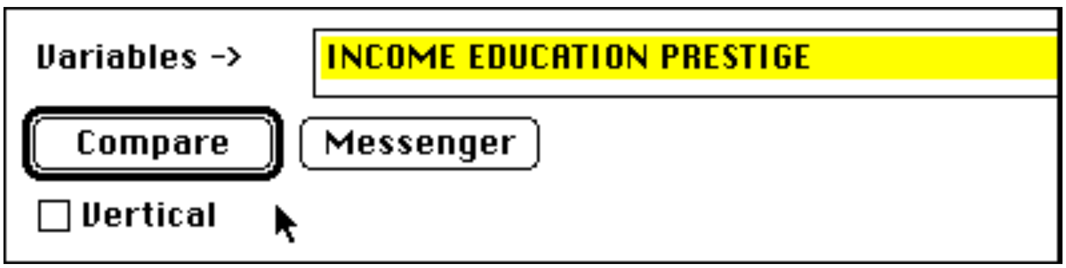
Figure 11 shows the QQ-plot and symmetry plot for the sample on the left of Figure 10. The QQ-plot indicates that the data are rather close to normal. Oddly, the symmetry plot would appear to suggest that the data come from a non-symmetric population. One has to be careful interpreting a symmetry plot since the points in the upper right corner of the plot are the extremes of the data and are quite unstable.

Figure 11. The QQ plot and symmetry plot for the normal sample on the left of Figure 10 indicate that the data seem to be normal, but not symmetric! Clearly, the symmetry plot requires a large sample size before concluding that the population is not symmetric.



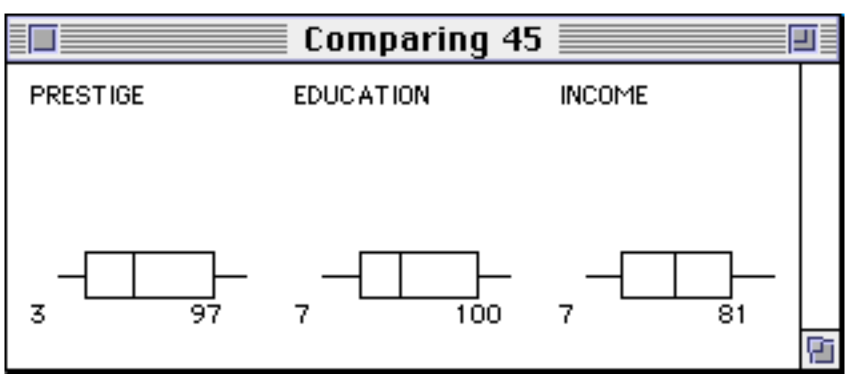
The Compare tool from the **Statistics** menu offers a quick way to browse and compare several variables at one time. To use this tool, it helps if we first click on the icons representing the variables of interest. Use a *shift click* to select the INCOME, EDUCATION, and PRESTIGE icons (hold down the shift key while selecting the additional icons with the mouse). With these chosen, select the Compare item from the **Statistics** menu. The names of the selected variables appear in the resulting dialog as shown in Figure 12. Throughout *AXIS*, commands use the names of selected variables to initialize dialogs. Since the variable names appear in the command dialog, they can be edited to include transformations if desired.

Figure 12. The Compare tool builds plots that compare several variables. Here is the dialog for comparing the income, education, and prestige of the 45 occupations in the Duncan data.



Pressing the Compare button using the mouse produces side-by-side comparison boxplots as shown in Figure 13. By default, the Compare tool shows the boxplots with respect to different scales. Clicking on the vertical box in the Compare dialog shows the boxplots on a common scale. The title of the window gives the number of observations being shown.

Figure 13. The Compare tool generates side-by-side boxplots of the variables included in the command dialog.

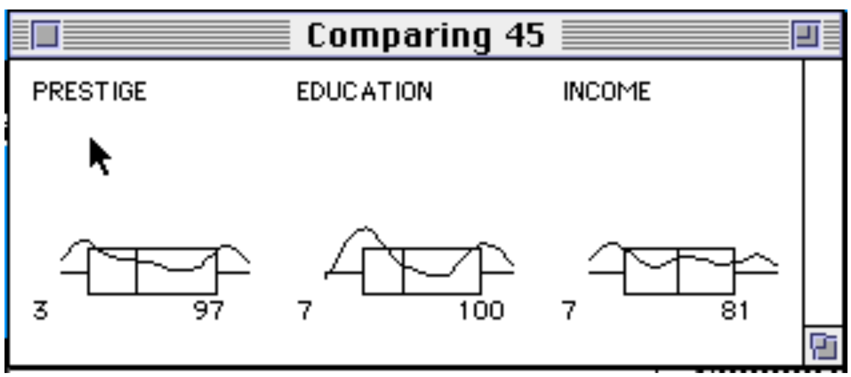


The Compare dialog, like the Univariate dialog, does not disappear when the display window opens. We can edit this dialog to build more comparison displays (such as of transformed data) or to modify the most recent display using the Messenger button (as shown later). For now, click on the close box to remove this dialog.

This Compare window, like all Lisp-Stat plot windows, has its own menu item which we can use to modify the display. For example, the **Compare** menu associated with the plot in Figure 13 allows us to add kernel densities so that the figure appears as in Figure 14. Kernel densities are a nice supplement to simple boxplots since they show more information, such as bimodality. In this case, the kernels reveal that both the prestige and education are bimodal. Boxplots do not reveal this feature.

The **Compare** menu can also produce a scatterplot matrix of the associated variables. Scatterplot matrices are basically visual correlation matrices – only that a scatter plot replaces the single numerical measure of association. The scatterplot matrix produced by this comparison plot appears in Figure 15. Notice the outlying occupations, particularly in the plot at the upper left corner which shows PRESTIGE on INCOME.

Figure 14. An item on the **Compare** menu adds kernel densities to each of the boxplots shown in Figure 13.



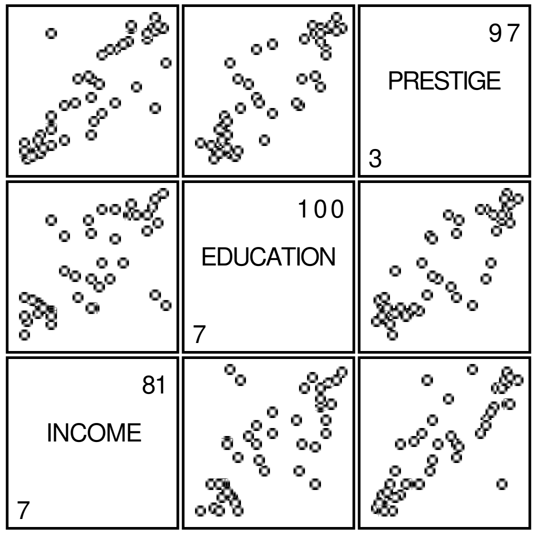
Finally, the Print Summary item of the **Compare** menu prints a table of several summary statistics for each variable shown in the plot. The output for this comparison window is:

Summary of Comparing 45.

	Moments		Percentiles				
Label	Mean	S.D.	5%	10%	50%	90%	95%
INCOME	41.9	24.4	8.5	10.5	42	76	77
EDUCATION	52.6	29.8	18	20	45	91.5	95
PRESTIGE	47.7	31.5	7.5	10	41	90	91

In addition to the usual mean and standard deviation, this summary includes several percentiles of the data. These percentiles are very useful in bootstrap resampling.

Figure 15. The **Compare** menu is also able to build a scatterplot matrix of the data. Note the outlier in the frame showing PRESTIGE ON INCOME in the upper left corner of the display.



Building a Small Model

The obvious use of the Duncan occupational status data set is to estimate a model of the association of income and education with prestige. The previous scatterplot matrix suggests high correlation between both income and education and prestige, but with a few outlying occupations. Let's start with a close look at the relationship between prestige and education using the Scatter Plot command from the **Statistics** menu. To generate the plot, go to the icon window and click first on the PRESTIGE icon, and then the shift click to add the EDUCATION icon. Choosing the Scatter Plot item generates the dialog shown in Figure 16. Notice that by choosing the icons in this order, the dialog arrives configured so that no typing is needed! Clicking on the Scatter Plot button generates the plot in Figure 17. This dialog box goes away automatically after the plot is shown.

If the association between the EDUCATION and PRESTIGE were exactly linear with PRESTIGE=INCOME, then all of the points would all fall on the diagonal line x=y. The Messenger item in the **Scatter** menu associated with this plot allows us to add this reference line to the plot. (Remember to bring this plot to the front by clicking on this window to obtain the associated **Scatter** menu.) Clicking on the Messenger menu item opens a dialog; type the command ":abline 0 1" (with the leading colon) into the dialog as in Figure 18 and then click "Send It". This dialog sends a message to the associated scatter plot telling it to add the line $y = a + b\,x$ with constant term $a = 0$ and slope $b = 1$. The resulting figure appears in Figure 19. The Messenger exploits the underlying object-oriented capabilities of *Lisp-Stat.* A list of the possible messages (with documentation shown when available) appears as you move the slider below the message input box. Interested users can consult Tierney (1990) for details of this and other useful messages that modify plots.

Figure 16. If we first select the PRESTIGE and then the add the EDUCATION icon with a shift-click, the Scatter Plot menu item opens a dialog configured with the variables as we need them.

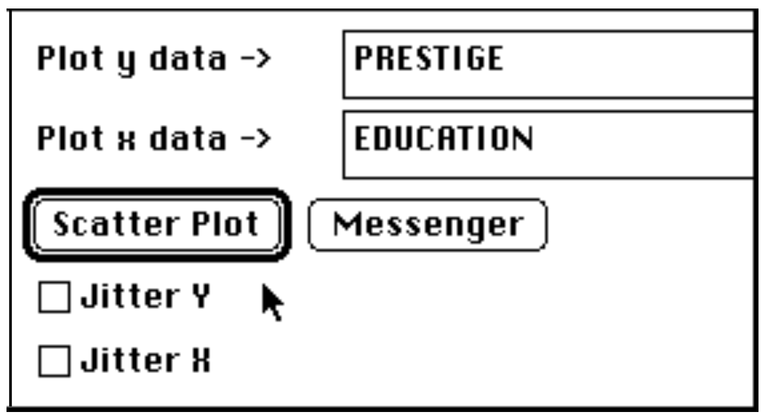


Figure 17. Clicking on the Scatter Plot button in the dialog of Figure 16 generates the scatterplot of prestige on education.

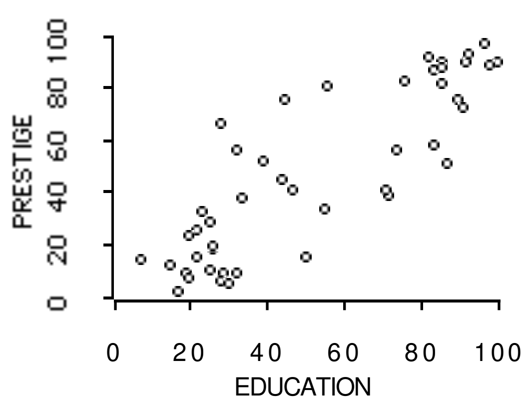


Figure 18. The Messenger item of the **Scatter** menu opens a dialog that lets us send messages to the plot that cause the plot to change. The message shown here adds a line with the form y = 0 + 1 x to the plot – a diagonal line in this case. The modified plot appears in Figure 19.

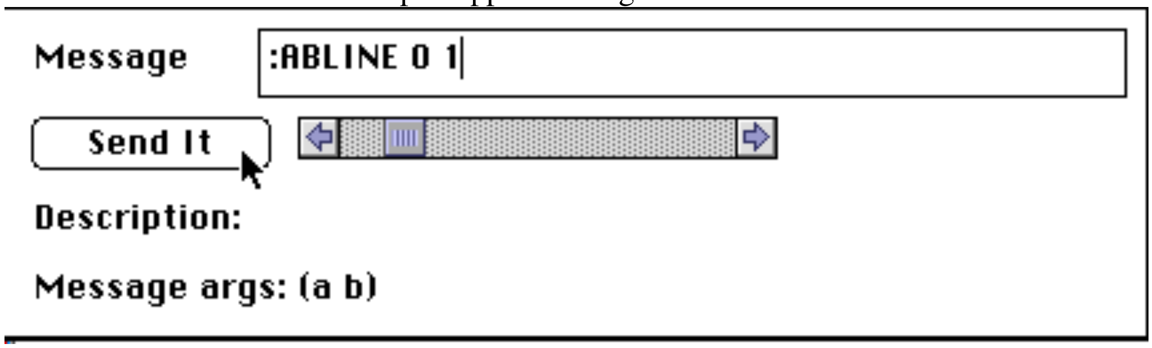
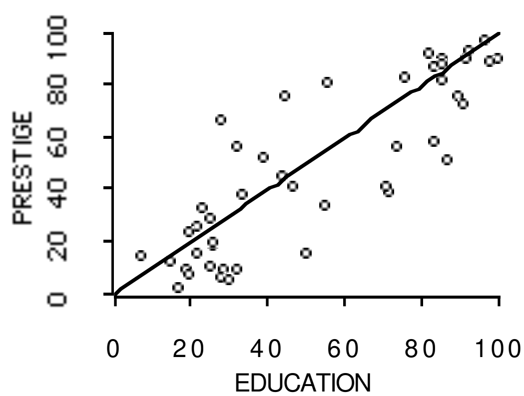
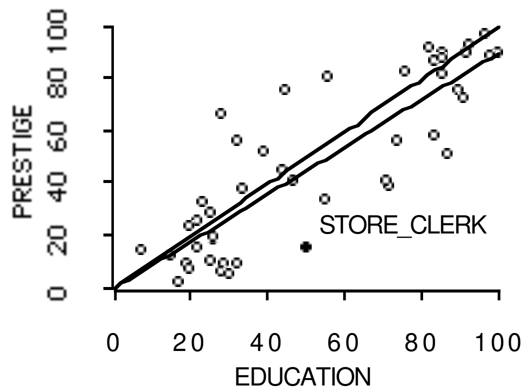


Figure 19. The messenger dialog has added a diagonal line to the plot of prestige on education initially shown in Figure 17.



Let's compare this diagonal line to the least-squares regression line. Near the bottom of the **Scatter** menu is the item Show OLS which adds the least squares regression line to the plot. The revised plot appears in Figure 20.

Figure 20. The Show OLS item from the **Scatter** menu adds the least squares regression line to scatterplot of prestige on education. The slope of the OLS line is less than one. Observations are highlighted by clicking on the points using the mouse.



The slope of the OLS fit is somewhat less than that of the diagonal line added first. How much less? The Print lines item (also from the **Scatter** menu) prints the equation of the fitted OLS line in the Listener window:

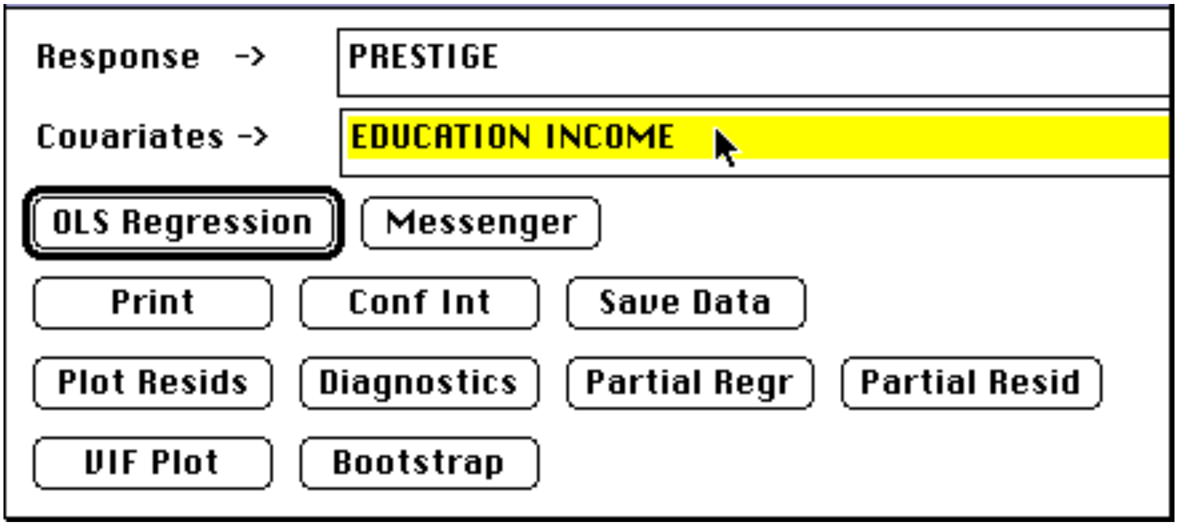
```
Scatterplot regression lines...
OLS  regression: PRESTIGE  = 0.284 + 0.902 EDUCATION
```

Since we previously defined case labels for this data set, the observations in this plot have meaningful labels when selected by the mouse. A single observation is selected in the scatterplot shown in Figure 20. Other menu items add robust regressions and scatter plot smoothers to the display. If the display becomes too cluttered, the Clear Lines menu item removes the lines that have been added to the original plot. This scatter plot only enables us to display and print regression lines. To find the statistical properties of the associated regression estimates, we need to use the more capable regression tool.

Fitting a Least Squares Regression

The preliminary graphical investigation of the Duncan occupational prestige data reveals some interesting features – such as the apparent bimodality and several outliers – but nothing that would suggest that OLS regression is going to be misleading. To prepare for building the regression, return to the icon view of the data set and first select PRESTIGE, then shift click to add EDUCATION and INCOME. Selecting the Regression menu item from the **Statistics** menu with these three icons selected produces the rather imposing dialog shown in Figure 21. As with the scatterplot, clicking first on the icon of the response, then using shift-clicks to add the covariates leads to a properly initialized dialog. (One can, of course, just type the names into the dialog).

Figure 21. Selecting the PRESTIGE icon first and using shift-clicks to add the covariates leads to the shown dialog for building a regression model.



Clicking on the OLS Regression button with the mouse fits the regression and prints the following summary in the Listener window (perhaps with differing numbers of decimals in the output):

```
Building regression model for PRESTIGE...
Least Squares Estimates:

    Variable    Estimate    Std.Err.    t-Ratio
    Constant     -6.065      4.271       -1.4
    EDUCATION     0.546      0.098        5.6
    INCOME        0.598      0.119        5

R Squared:                  0.828
Sigma hat:                 13.369
Number of cases:               45
Degrees of freedom:            42
```

The R^2 statistic is rather large and both of the t-statistics for the individual coefficients are significant at any reasonable level. Before we can conclude that this is model is satisfactory, though, we need to explore its structure more carefully with regression diagnostics.

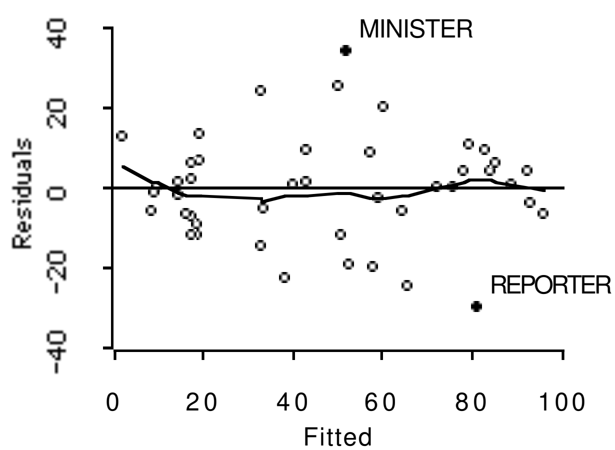
The additional buttons in the regression dialog produce a variety of useful diagnostic plots and statistics. Here is a brief summary of what each button in the regression command dialog does:

Button	**Does**
OLS	Fits the model. Click on this button first.
Messenger	Sends a message to an estimated model.
Print	Prints summary of the model in the Listener window.
Conf Int	Builds a confidence interval for each coefficient.
Save Data	Saves features of the model (*eg,* residuals) in the data set.

Plot Resids Plots the least squares residuals on the fitted values.
Diagnostics Plots the studentized residuals on the leverages *and* plots Cook's D statistic on the case numbers. The variance inflation factors and Durbin-Watson statistic appear in the listener.
Partial Regr Shows partial regression plot for each slope (added-variable plots).
Partial Resid Shows partial residual plots, one for each slope (component+residual plots).
Bootstrap Bootstrap the fitted model.

These buttons make it convenient to diagnose the validity of the fitted model. The strategy is basically to move from left to right across the middle row of buttons. In this way, the first plot to consider is the familiar plot of residuals on fitted values produced by the Plot Resids button. This plot appears in Figure 22. The menu of the plot produced by this button has all of the items of the scatter plot used earlier. For example, we can add the a scatter plot smoother by choosing the Show Smooth item of the **Scatter** menu. Systematic deviations of the smoother from the horizontal line at zero indicate problems. In this case, the smooth fit stays rather close to the zero line. No problems here, though ministers seems to be an overrated occupation with prestige unaccounted for by this model (not so surprising). Reporters are less prestigious than their education, income, and Dan Rather would have us believe (at least at the time of this data).

Figure 22. The scatterplot smoother does not indicate a problem in the plot of residuals on the fitted values from the model built by the dialog of Figure 21 (Prestige on Income and Education).



The Diagnostics button of the regression dialog prints several summary diagnostics and plots the leverage and influence of the fitted model. The printed results include variance inflation factors (actually, the square root of the VIF's), the Durbin-Watson statistic, and the runs test. The values of these numerical diagnostics printed in the Listener window for this model are:

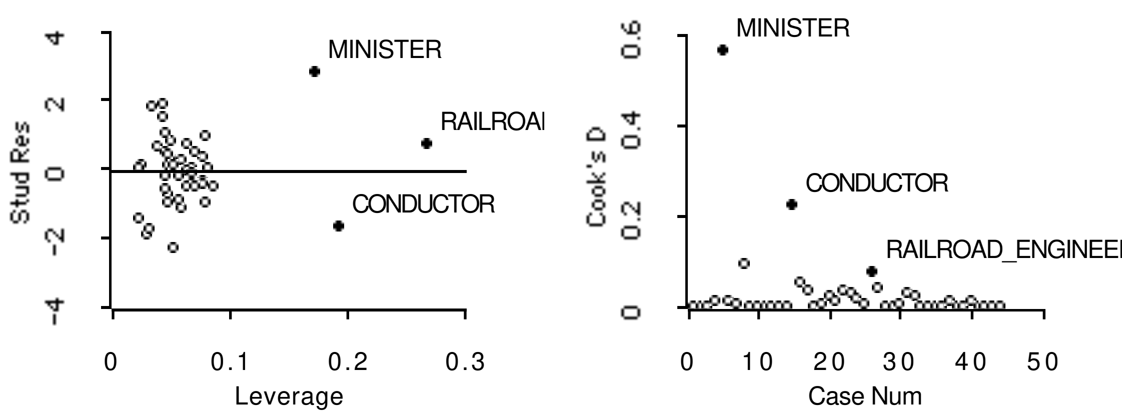
```
Variables and Square Roots of VIF's.
EDUCATION --> 1.45083
```

```
INCOME    --> 1.45083
Durbin-Watson statistic = 1.45833 (rho^ =  0.271)
22 runs with z = 0.237571 (p= 0.812)
```

The variance inflation factors do not indicate a problem with multicollinearity, and the Durbin-Watson is not terribly meaningful unless we can deduce some meaning to the ordering of the observations in the original data file.

The two leverage/influence plots, which are also part of the output generated by the Diagnostics button, are rather interesting and show some features not evident in prior figures. Both of these appear in Figure 23. The diagnostic plots in Figure 23 are *linked* to each other; selecting points in one of the plots highlights the observations in both of the plots at once. These plots reveal three distinct outlying occupations: minister, railroad engineer and conductor. The points for these occupations are highlighted in Figure 23. The observations for these three are *influential* not because of large deviations on the prestige scale, but rather because they are unusual in the space of income and education – these occupations are highly *leveraged.* An occupation is termed influential if the fitted regression coefficients would change by much if we deleted this observation from the analysis.

Figure 23. The Diagnostics button of the regression dialog produces plots of the studentized residuals on the leverage (left) and Cook's D statistic for influence on the case number (right). The highlighted observations in these linked plots are leveraged and influential.



Linking these plots to a scatter plot of income on education shows why the three occupations are influential. You can link other plots to these using the Link View item which appears in many of the graphics menus. In this example, linking the diagnostic plots to a scatter of income on education (Figure 24) shows why these observations are so highly leveraged. The railroad occupations have rather high income for their level of education (railroad unions were very effective), whereas ministers have small incomes given the level of education.

The remaining buttons Partial Regr and Partial Regr of the regression dialog generate multiple two-dimensions views of the data used to fit the regression. Partial regression plots obtained via the Partial Regr button make the effects of these outliers even more clear. In Figure 25, the same three occupations are very distinct. The effect of these three occupations on the estimated regression coefficient of the education variable appears rather slight. (We will check this later.) In contrast, the effect of these three points on the income coefficient is large; the fitted line misses the pattern in the bulk of the data. The slope for income would be much larger without the combined effect of these three outliers. To see just how different, we can interactively remove these points from the scatter plot and ask that it show the OLS line fitted to the remaining data. (Try fitting the OLS line without changing the data – it is the same as the line from the multiple regression.) Use the Remove Selection item from the **Scatter** menu. With these three hidden, the Show OLS item estimates the line fitted to the remaining occupations. This line omitting the three outlying occupations is the steeper fit in Figure 26.

Figure 24. Linking the diagnostics plots to a scatter of income on education shows that the highlighted points in Figure 23 are leveraged because income and education of the associated occupations do not obey the diagonal pattern of the other observations.

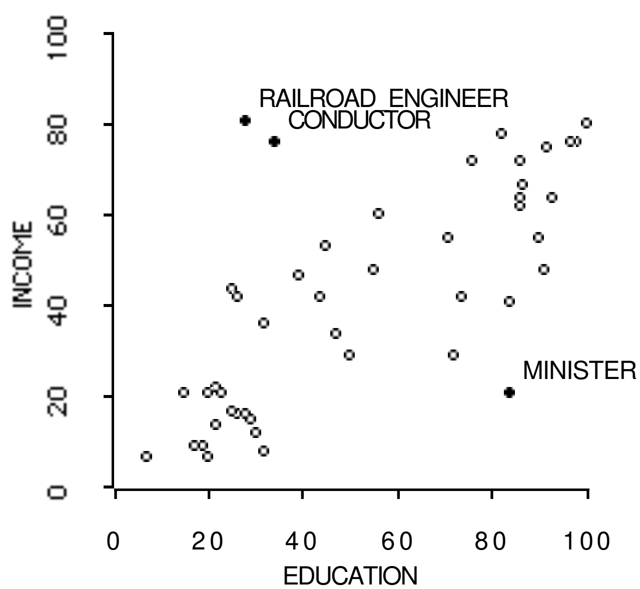


Figure 25. Partial regression plots for the model of prestige on education and income show the effect the of the leveraged occupations on the regression fit. These three highlighted occupations attenuate the slope of income and slightly accentuate the slope of education.

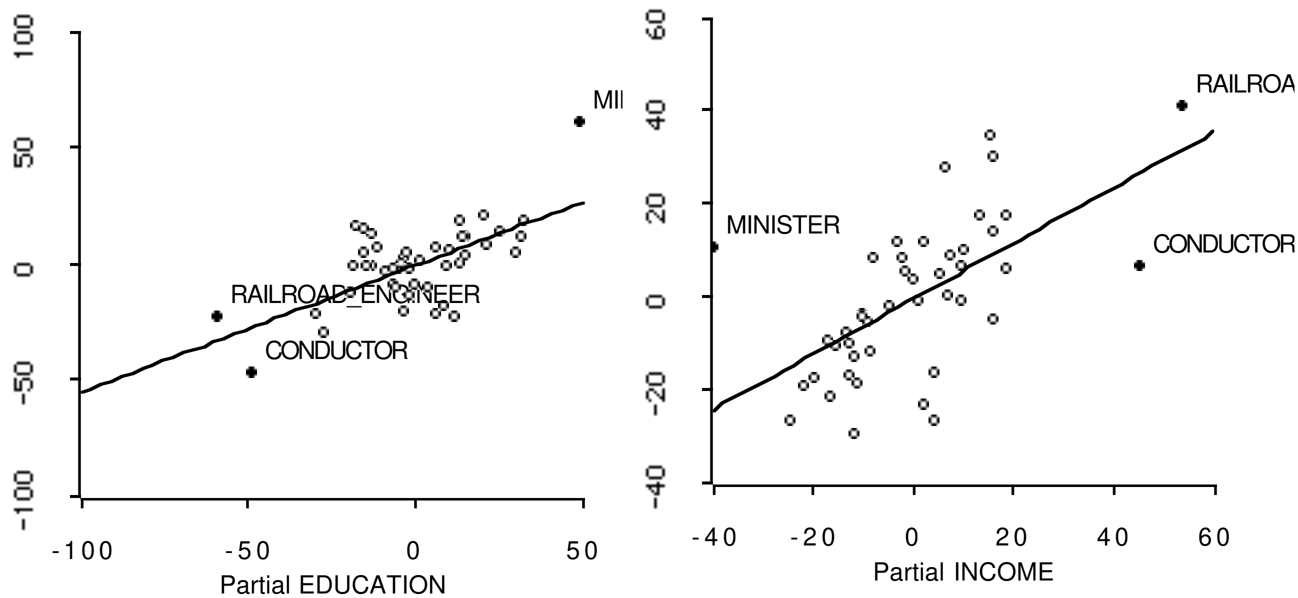
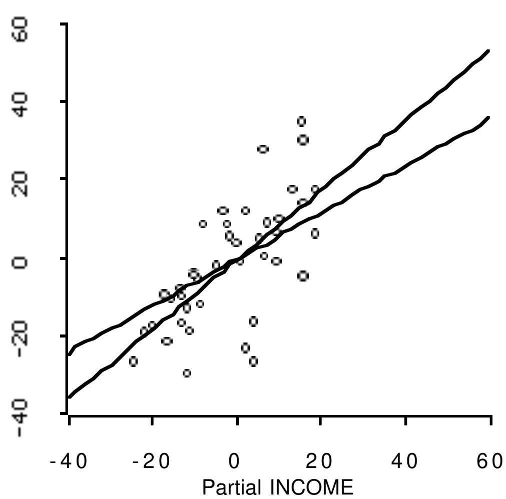
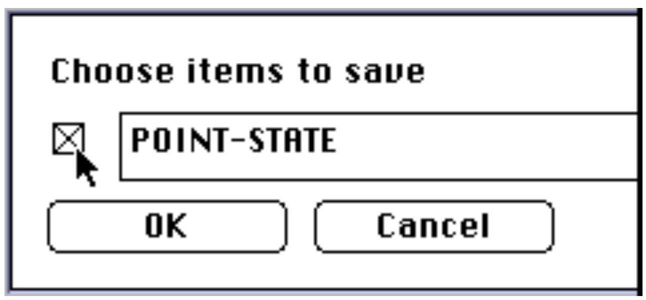


Figure 26. The Remove Selection item from the **Scatter** menu "hides" the highlighted three occupations so that the Show OLS item can reveal what would happen to the OLS slope without these three points. The resulting slope is much steeper.



To see how these three observations affect the rest of the regression model, we need to refit the model omitting them. The easiest way to identify these observations is in a plot (rather than with some sort of "if" statement.) The Save Data item from the **Scatter** menu makes it easy to build a variable from a plot. First, go back the plot of the studentized residuals on the leverage in which these three stand out. Choosing the Save Data item from the menu produces the dialog shown in Figure 27. Click on the check-box to the left of the text field (which can be used to name the variable) and then click on the OK button. This sequence saves a variable in the data set whose values are either "normal" or "selected" (print the values of the new variable as we did at the start using the popup dialog accessed by the option click). By default, the name of the new variable icon is POINT-STATE – you could have changed the name in the dialog of Figure 27.

Figure 27. The Save Data item from the **Scatter** menu presents a dialog that allows you to save the highlight status of points in the plot as a variable in the data set. The resulting variable (named Point-State by default) can then define a variable useful for subsetting the data.

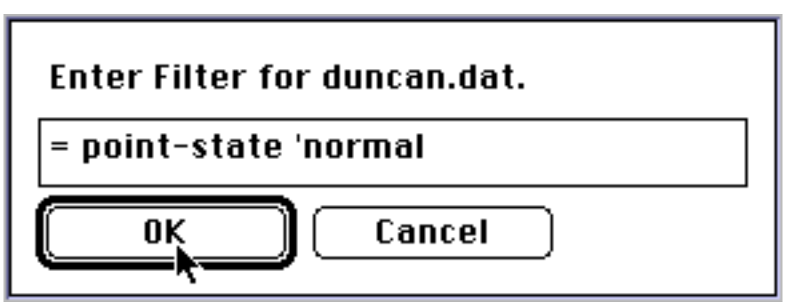


To subset the data using this new variable, double-click on the Filter feature icon in the data set view and enter the command

```
= point-state 'normal
```

as shown in Figure 28. Be sure to type just a single quote before the word "normal". The filter icon becomes highlighted (wide border). Statistics in the data set will now be computed from just those occupations which were not selected (i.e., displayed normally) in Figure 23.

Figure 28. Double-click on the Filter icon in the data set view to define a selection filter for the data.



To see the results of omitting these three cases, return to the **Statistics** menu and build a new regression model. The fitted summary of the revised model fit to just the 42 remaining cases is:

```
Building regression model for PRESTIGE...
Least Squares Estimates:
    Variable    Estimate    Std.Err.    t-Ratio
    Constant      -6.32     3.67        -1.7
    EDUCATION      0.28     0.12         2.3
    INCOME         0.93     0.15         6.1
R Squared:                  0.876
Sigma hat:                 11.49
Number of cases:           42
Degrees of freedom:        39
```

The model fits much better and the income slope is much larger than before, changing from .598 to .876. The overall fit suggested by the R^2 is also better. But is this model really that much better, or does it have problems as well? Why does the coefficient of EDUCATION become so much smaller (.546 down to .28)? Also: Where did the two groups that we originally noticed go, and what happened to the reporters?

Bootstrap Methods

AXIS supports bootstrapping in several domains. For example, you can use the bootstrap to compare the performance of estimators via the Simulate item on the **Statistics** menu. Alternatively, you can use the bootstrap to measure the uncertainty of a complicated estimator such as a scatter plot smooth. Finally, you can use the bootstrap to build confidence intervals.

The Bootstrap item on the **Statistics** menu constructs small simulation analyses of the behavior of statistics under various sampling models. As an example, consider the assertion that the sample average is the best estimator of the center of a normal population. Mathematics show that the average is best, and simulation confirms this fact. The dialog for building the simulation of the mean, trimmed mean, and median for 100 samples from a normal distribution appears in Figure 29. Clicking on the Bootstrap button builds a *new* data set, which appears in the next window. The "# Trials" feature icon permits one to add more trials to the experiment – a useful feature when you are not sure how long it will take to generate the simulated results and would rather start with just a few trials.

Although this new data set has been simulated, we can explore it using the same tools as we used in the analysis of the Duncan data set. The **Compare** tool works nicely here, particularly with the vertical option checked so that the boxplots appear on common scales as shown in Figure 31 with added kernels. The simulated sampling distribution of the mean is more compact than those of the other two – but only somewhat tighter.

Figure 29. The Bootstrap item on the **Statistics** offers a dialog that can be used to construct simulations that compare several estimators. This dialog compares three estimates of location for 100 samples from a standard normal population (mean zero, variance one).

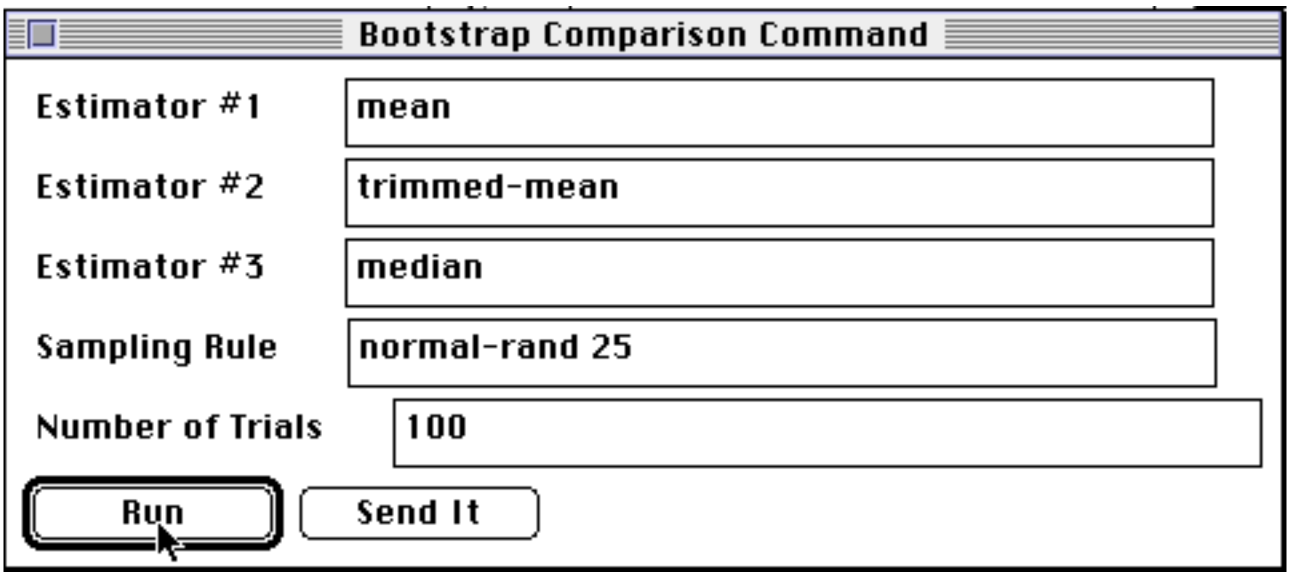


Figure 30. The data set built by the simulation has a variable corresponding to each of the three estimators in the dialog of Figure 29. The "# Trials" icon allows you to augment the simulation with more trials.

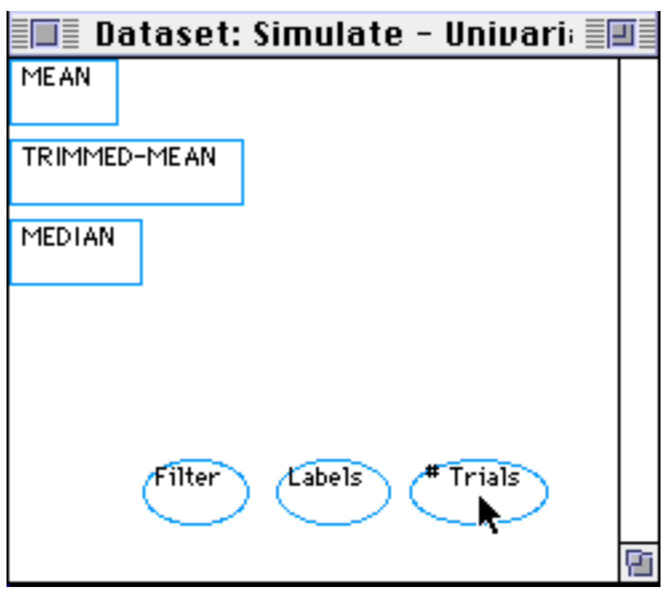
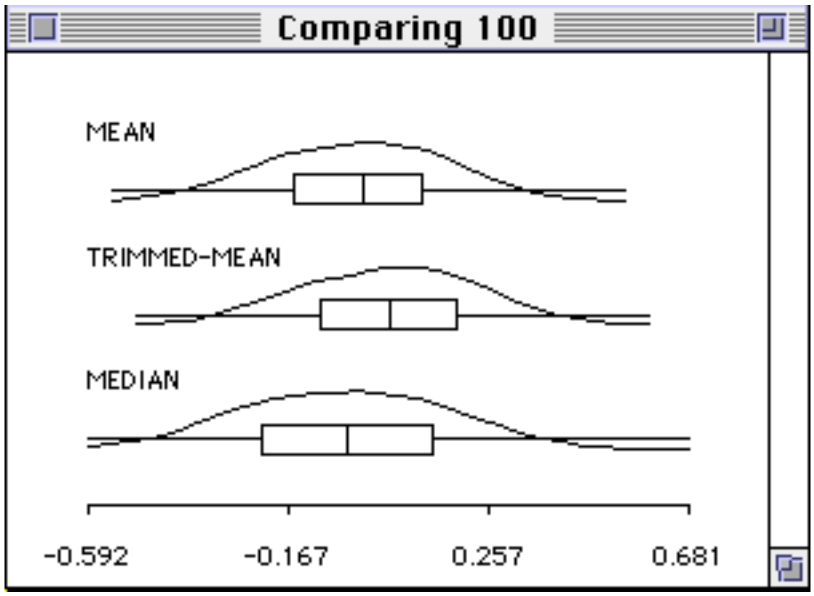


Figure 31. Comparison boxplots of the 100 simulated estimates of location computed in three ways confirm that the mean is best (its box is more narrow than those of the others), but also shows that the other two are quite close.



The same procedure produces a bootstrap simulation that compares these three. Rather than sample from a hypothetical distribution such as the normal, draw samples with replacement from a variable. The simulation dialog is the same with the exception that the sampling rule item becomes, for example,

```
resample prestige
```

as in Figure 32.

Figure 32. Changing the sampling rule in the dialog of Figure 29 as shown generates a bootstrap simulation of the same three location estimators.

Bootstrapping a regression operates in a similar fashion. Use the "Bootstrap" button of a regression command dialog to build the bootstrap data set. The bootstrap data set has one variable for each of the coefficients (including the constant) in the regression model.

Frequently Asked Questions

How many cases can AXIS accommodate?

I generally use *AXIS* with small data sets, say of at most several hundred observations and 10-20 variables. The faster the machine, the more data you can use. Lisp-Stat uses whatever memory you can give it. Though social science data sets often exceed these dimensions, restricting attention to homogeneous subsets of the data often gives a better impression of what is going on and yields a small data set.

The only limit to the size of the data set that you can use in *Lisp-Stat* and *AXIS* is the amount of available memory in the computer that you are using. The bigger the data set, the more memory you will need. I generally suggest running with a 4MB segment on a Macintosh. For Windows, you probably need a machine configured with at least 8MB of memory.

Why is the statistics menu grayed out even after I have opened a data set?

The software sometimes gets confused and believes that you have not opened a data set even though you have. You can clear this up by clicking once on the data set window (the one with the variable icons), then clicking on something else (like the background screen), and then clicking on the data set window again. The menu items ought to be accessible now.

I can't find the menu that allows me to change the plot. Where is it?

Window-specific menus only appear for the front most window, the window most recently activated by a mouse click. If you do not see the desired menu, click on the window in question and it should appear.

What if I can't figure something out? Who can I call?

Send an e-mail message describing your problem to *stine@wharton.upenn.edu* and I will answer your question if I can. Please be patient, though.

How do I get the most recent copy of AXIS?

Macintosh users should use the free *Fetch* program to get these files. For PC users, you can get the latest version of the *AXIS* files by using the Internet file transfer program know as *ftp*. This program allows you to log into my workstation and get the latest version of the software. The use of *ftp* varies from system to system, but here are the basic steps. First start *ftp* with the address of my system (the things that you need to type are in boldface):

`C:>` **`ftp compstat.wharton.upenn.edu`**

If this does not work, you will have to get in touch with your system administrator. Assuming it does, *ftp* will prompt you for your login name. You probably don't have one on my system, so use the login name anonymous and give your e-mail address as the password:

Enter login name:	**anonymous**
Password:	**your email address, to be polite**

To get the files you want, change directories (careful, *ftp* is case sensitive):

```
>       cd pub/software/lispstat/axis
>       dir
        A list of many files!
```

The latest program files have the .lsp suffix. To get all of these, use the mget command:

```
>       mget *.lsp
```

and follow the prompts. All of the files will appear on your machine in the directory or folder from which you started *ftp*. Assuming that's all you need, end the session with the quit command.

```
>       quit
```

Other files and my WWW page offer some documentation (namely this document and some others). The documentation comes as a postscript file that you can print. You can get the latest version of the XLISP software itself from umnstat.stat.umn.edu.

References

Fox, J. (1992). *Regression Diagnostics*. Sage, Newbury Park, CA.

Tierney, L. (1990). *Lisp-Stat.* Wiley, New York.

——— (1995). Data analysis in Lisp-Stat. . *Sociological Methods and Research*, 23.