An Overview of the *AXIS* Statistics Package

Robert Stine

Department of Statistics

The Wharton School, University of Pennsylvania

Introduction

*AXIS* is computer software that provides a graphical statistical modeling environment. *AXIS* implements a variety of recent developments in statistics, including scatter plot 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 graphical itself – 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 answers.

*AXIS* is written 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 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 available 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 hardware platform that supports Lisp-Stat. Hence, you can run *AXIS* on Windows PC's, Macintosh's, and various Unix workstations that run Lisp-Stat. The instructions for installing Lisp-Stat on your machine are part of the distribution of that free software package. To install *AXIS* on a machine with Lisp-Stat, copy all of the associated files into a subdirectory of the directory that has the Lisp-Stat application. For example on a PC, if the file WXLS.EXE is located in the XLISP directory, make a subdirectory for the new files with the command mkdir xlisp\axis and then copy the files into the axis subdirectory. Currently the set of files for the *AXIS* interface include the following:

axis.lsp regPatch.lsp  
axisData.lsp compare.lsp  
axisBoot.lsp icons.lsp  
axisIcon.lsp labelPlt.lsp  
axisUtil.lsp axisRegr.lsp  
axisDens.lsp axisRReg.lsp  
axisCmd.lsp axisScat.lsp  
axisSE.lsp

In addition, several data files are also available. Generally these files, such as duncan.dat, are put in a separate subdirectory.

An Illustrative Problem

This example uses a small, well-known sociology dataset – the Duncan dataset containing the income, education and occupational prestige of 45 occupations. The usual interest in this data is the association between the prestige rating of various occupations and the typical income and education of members of those occupations. This data appears in many other papers, such as the Fox (1992) monograph on regression diagnostics. This dataset is also featured in a forthcoming issue of *Sociological Methods and Research* devoted to computing environments for statistics.

Preparing the Data

The first step of our analysis is read the Duncan data into *AXIS*. *AXIS* expects the input data to be in a rather simple, rectangular format in an standard text (ASCII) file. The first line of this file must be a documentation string, delimited by quotation marks. The next line of the file holds the variable names (without quotes or intervening blanks). Each following line contains the data 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 included 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  
 ...

The columns do not need to 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; the occupation title airline\_pilot includes an underline character to avoid an extraneous 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 now launch Lisp-Stat and load the *AXIS* interface. 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 standard **File** menu of the Lisp-Stat application, and select the file axis.lsp. You will be rewarded with a delay while the collection of more than ten *AXIS* files are loaded into Lisp-Stat. Now we are ready to start the data analysis.

As the *AXIS* files are loaded, the software augments the standard 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 list of options. The data menu is pretty short. It only has a single item that reads "Open data file...". The statistics menu has more items. The top of the menu shows several graphing tools, the middle has the analysis routines, and the last is a simulation tool related to bootstrapping.

Reading the Data File

Go back to the **Data** menu and select the item to open a file. This choice will cause the system to present you with a standard file opening 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, you will notice that a new 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 dataset window is open. (This figure is taken from a Macintosh system; that on a Windows PC will look slightly different.)  
   
The rectangular boxes or icons represent the four variables in this dataset: job title, income, education, and prestige. The two ovals represent special features of the dataset and are called the *feature icons*. The filter icon permits subsetting the data, and the other defines case labels for the observations. The icons may be freely moved about the window by *dragging* – clicking on the icon with the mouse cursor and holding down the mouse button while simultaneously moving the cursor.

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 defines the occupations and makes a great labeling choice. To use the values of the TITLE variable as the case labels, double-click on the case label icon. You should now get a dialog box on the screen (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. The icon appearance now changes to have a heavy oval border, indicating that case labels are defined.  
   
We will not be using the title variable further so let's remove it from the display. 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 dataset 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. Just hold down the option key and simultaneously click on an icon – an *option click*. This click causes a short "pop-up" menu to appear at the cursor, offering to print the data (with case labels) or display the data in a histogram or sequence plot. An excerpt of the printed listing of prestige appears below. All of the printed output from AXIS appears in the Lisp-Stat listener window, the main window of the application.

(ACCOUNTANT 82)

(AIRLINE\_PILOT 83)

...

(POLICEMAN 41)

(RESTAURANT\_WAITER 10)

A sequence plot shows the values of the variable plotted on the case number, the position in the original data file duncan.dat. Here is the sequence plot for INCOME, with several cases selected using the mouse. In order to get the case labels to appear, use the **Plot** menu associated with this display. Using case labels makes it very easy to identify the occupations with either very low or very high incomes.

**

The Compare tool from the **Statistics** menu offers the quickest way to browse variables. To use this tool, it helps if we first click on the icons representing the variables of interest. Use a *shift-click* to pick more than one icon (hold down the shift key while selecting the additional icons with the mouse). With INCOME, EDUCATION, and PRESTIGE picked, select the Compare item from the **Statistics** menu. The selected variable names now appear in the resulting dialog box as shown below. Throughout *AXIS*, commands like "Compare" use the variables associated with selected icons. The variable names appear in the command dialog, where they can be edited to include transformations if desired.  
  
  
Just press the "Run" button using the mouse to produce comparison boxplots. 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 associated observations being shown.  
   
The Compare dialog, like all *AXIS* command dialogs, does not disappear when the display window opens. We can use this dialog to build either more comparison displays (such as of transformed data) or to modify the associated display using the "Send It" button as shown later. For now, click on the close box to remove this dialog.

This plot window, like all Lisp-Stat plot windows, has its own menu item which we can use to interact with the display. For example, the **Compare** menu for a comparison plot allows us to add kernel densities to the figure, as shown next. Kernel densities are a nice supplement to simple boxplots since. In this case, the kernels reveal that both the prestige and education are bimodal. Boxplots do not reveal this feature.  
   
The comparison plot menu can also produce a scatterplot matrix of the associated variables. Scatterplot matrices are basically visual correlation matrices – except that a scatter plot replaces the single numerical measure of association. The scatterplot matrix produced by this comparison plot appears next. Notice the outlying values.  
   
The associated summary statistics produced by the Print Summary menu item appear in the listener window of Lisp-Stat:

Summary of Comparing 45.

Moments Percentiles

Label Mean S.D. 5% 10% 50% 90% 95%

PRESTIGE 47.7 31.5 7.5 10 41 90 91

EDUCATION 52.6 29.8 18 20 45 91.5 95

INCOME 41.9 24.4 8.5 10.5 42 76 77

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.

For a detailed look at a single variable, use the Density item from the **Statistics** menu. We noted in the Compare summary that the density of EDUCATION was evidently bimodal. To investigate this structure further, pick the EDUCATION icon from the dataset window and then choose the Density menu item. The resulting plot is clearly bimodal and we can see the two clusters of observations.  
   
The Density menu item also allows us to explore the sensitivity of a kernel density estimate to the associated smoothing parameter. Picking the Kernel Slider menu item opens a slider dialog that interactively modifies the level of smoothing. The next plot 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.  
 

Building a Small Model

The obvious use of this small dataset is in a model of the effect of income and education on prestige. The previous scatterplot matrix suggests high correlation between both income and education and prestige. Let's start with a close look at the relationship between income and prestige and use the Scatter Plot command from the **Statistics** menu. This produces the following plot:  
   
If the correlation between the EDUCATION and PRESTIGE were perfect in the form PRESTIGE=INCOME, then the points would all fall on the diagonal line y=x. The "Send It" button allows us to do add this point-of-reference to the plot by exploiting the underlying capabilities of Lisp-Stat. Click on the "Send It" button and type the command "send it :abline 0 1" into the dialog. The dialog should appear as follows. Then click "OK".  
  
   
  
This dialog sends a message to the associated scatter plot telling it (the *it* in the command) to add the line *y = a + b x* with constant term *a* = 0 and slope *b* = 1. Here is the resulting figure. (Interested users can consult Tierney (1990) for other useful messages that modify plots.)  
   
We are done with the command dialog for this window, so you might as well close it.

Let's compare this "perfect" line to the least-squares regression line. At the bottom of the **Scatter** menu associated with this plot is an option Show OLS which adds the least squares regression line to the plot. The slope of the OLS fit is somewhat less than that of the diagonal line added first. How much less? Use the Print lines item to see the fitted equation (as before, this output appears in the listener window):  
 Scatterplot regression lines...

OLS regression: PRESTIGE = 0.284 + 0.902 EDUCATION   
Since we previously defined a label variable for this dataset, the observations in this plot have meaningful labels when selected by using the mouse. A single observation is selected in the plot that follows.  
   
Other menu items add robust regressions and scatter plot smoothers to the display. If the display becomes too cluttered, use the Clear Lines menu item to remove 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

At this point of our analysis, we are going to build and investigate a regression model for prestige. Our preliminary graphical investigation of the Duncan occupational prestige data reveals some interesting features – such as the bimodality and slight outliers – but nothing that would suggest that regression is far off course. Selecting the Regression menu item from the **Statistics** menu with the three icons of interest selected produces the rather imposing dialog  
  
  
Clicking on the run button with the mouse as suggested is this figure builds the regression and prints the following summary of the fit 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 *R2* 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 of the buttons in the regression command dialog do:

**Button Does**    
 *Run* Builds the model. Click on this button first.  
 *Send It* Sends a message to the associated model.  
 *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* Show the partial regression plots, one for each slope.  
 *Partial Resid* Show the partial residual plots, again one for each slope.  
 *Print*  Print the initial summary of the model in the listener window.  
 *Conf Int* Build a confidence interval for the coefficients.  
 *Bootstrap* Bootstrap the fitted model.

These tools make it easy and convenient to investigate the validity of the fitted model. The strategy is basically to move from left to right across the middle line 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. The plot produced by this button enjoys all of the features of the scatter plot used earlier. For example, we can add the a scatter plot smoother with the Show Smooth button of the **Scatter** menu. Systematic deviations of the smoother from the horizontal line at zero indicate problems. In this case, the smoother is 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 suggest.  
 

The Diagnostics button produces some tabular results as well as two more plots. The tabular results 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

The variance inflation factors (VIF) do not indicate a problem with multicollinearity, and the Durbin-Watson is not terribly meaningful unless we can decern some meaning to the ordering of the observations in the original data file. The other two plots, though, are rather interesting and show some features not noticed in prior figures. These plots are *linked* to each other so that selecting points in one of the plots highlights them in both of the plots at once. (You can link other plots to these as well using the Link View item which appears on each of the graphics menus.)  
   
These plots reveal three distinct outlying occupations: minister, railroad engineer and conductor. 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*. Here an occupation is termed influential if fitted regression coefficients would change if we deleted this observation from the analysis. Linking these plots to a scatter plot of income on education shows why the three occupations are influential. 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.  
 

Partial regression plots given by the Partial Regr button of the regression command dialog make the effects of these outliers even more clear. Again, the same three occupations are very distinct.  


The effect of these three occupations on the estimated regression coefficient of the education variable appear rather innocent. (We will check this later on.) The effect of these three on the income coefficient is very dramatic as the fitted line associated with the regression 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 – its the same as the line from the multiple regression.) Use the Remove Selection item from the **Scatter** menu. The Show OLS option reveals the line fitted to the remaining occupations. This line omitting the three outlying occupations is the steeper fit shown in the following plot.  
 

To see just how these three observations affect the regression model, let's refit the model omitting them. This problem illustrates many of the other features of *AXIS*. The first component of this task is to build a variable that clearly delineates these three in the data, and the leverage plot suggests that this variable is a good candidate to identify the three. To use the regression leverages as a variable in our analysis, we have to first add them as a variable to the data set. This is easily done with the Send It button of the regression command dialog. Push this button and enter (def leverage (send it :leverages)). This command defines a new variable named leverages which is built from the leverages of this model; an icon for the variable appears in the data set icon window. Now use the filter feature icon of the data set window. Double-click this icon and enter the expression (< leverage 0.15), that is use only the observations which have leverage less than 1.5 (the syntax is Lisp prefix notation). Building a new regression model gives the printed summary shown next:

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 *R2* 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)

Some other questions for further analysis include: Where did the two groups that we originally noticed go, and what happened to the reporters?

Bootstrap Methods

*AXIS* supports bootstrapping in several domains. First, you can use the bootstrap to compare the performance of estimators via the Simulate item on the **Statistics** menu. Second, you can use the bootstrap to measure the uncertainty in a scatter plot smooth. And last, you can use the bootstrap with regression.

The Simulate item 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 means for estimating the center of a normally distributed population. We can mathematically show that the average is best, and simulation confirms this fact. Here is the simulation dialog for simulating the mean, *trimmed mean*, and median for using 100 samples from a normal distribution.  
  
  
Clicking on the Run button builds a *new* dataset, 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.  
   
   
Now we can explore the simulated data using the same tools. The comparison tool works nicely here, particularly with the vertical option checked so that the boxplots appear on the same scales as shown in the following plot with kernels added. The simulated sampling distribution of the mean is more compact than those of the other two – but only somewhat tighter.  
  
 

The same procedure produces a bootstrap comparison. 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 of the sampling rule item which becomes, for example:  


Bootstrapping a regression operates in a similar fashion. Use the "Bootstrap" button of a regression command dialog to built 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 datasets, say of at most several hundred observations and 5-10 variables. Though social science datasets 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 dataset.   
 That said, the only thing that limits the size of the dataset 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 dataset, the more memory you will need. I generally suggest running with a 3MB segment on a Macintosh. For Windows, you probably need a machine configured with 8MB of memory.

*What if I can't figure something out? Who can I call?*  
Send an e-mail message describing your problem to *bob@hilbert.wharton.upenn.edu* and I will try to 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 hilbert.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: **bob@hilbert.wharton.upenn.edu**  
To get the files you want, change directories (careful, *ftp* is case sensitive):  
 > **cd pub/xlispstat/AXIS**  
 > **dir**  
 SYSTEM  
 DATASETS  
 DOC  
The latest program files are in the SYSTEM subdirectory. To get all of these, change to that directory and use the mget command:  
 > **cd SYSTEM**  
 > **mget \***  
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 directories on hilbert offer sample datasets and some documentation (namely this document). The documentation comes in two forms, one for MS Word on the Mac, one for Word for Windows. You can also get the latest version of the XLISP software itself from hilbert. The directory pub/xlispstat/windows has the latest windows version; the Mac version lives in pub/xlispstat/mac.

References

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

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