Computing Primer for Applied Linear Regression, Third Edition

Using SPSS

Katherine St. Clair & Sanford Weisberg
Department of Mathematics, Colby College
School of Statistics, University of Minnesota

August 3, 2009

Home Website: www.stat.umn.edu/alr

Contents

0 Introduction

This computer primer supplements the book *Applied Linear Regression* (ALR), third edition, by Sanford Weisberg, published by John Wiley & Sons in 2005. It shows you how to do the analyses discussed in ALR using one of several general-purpose programs that are widely available throughout the world. All the programs have capabilities well beyond the uses described here. Different programs are likely to suit different users. We expect to update the primer periodically, so check `www.stat.umn.edu/alr` to see if you have the most recent version. The versions are indicated by the date shown on the cover page of the primer.

Our purpose is largely limited to using the packages with ALR, and we will not attempt to provide a complete introduction to the packages. If you are new to the package you are using you will probably need additional reference material.

There are a number of methods discussed in ALR that are not (as yet) a standard part of statistical analysis, and some methods are not possible without writing your own programs to supplement the package you choose. *The exceptions to this rule are R and S-Plus. For these two packages we have written functions you can easily download and use for nearly everything in the book.*

Here are the programs for which primers are available.

R is a *command line* statistical package, which means that the user types a statement requesting a computation or a graph, and it is executed immediately. You will be able to use a package of functions for R that will let you use all the methods discussed in ALR; we used R when writing the book.

R also has a programming language that allows automating repetitive tasks. R is a favorite program among academic statisticians because it is free, works on Windows, Linux/Unix and Macintosh, and can be used in a great variety of problems. There is also a large literature developing on using R for statistical problems. The main website for R is `www.r-project.org`. From this website you can get to the page for downloading R by clicking on the link for `CRAN`, or, in the US, going to `cran.us.r-project.org`.

Documentation is available for R on-line, from the website, and in several books. We can strongly recommend two books. The book by Fox (2002) provides a fairly gentle introduction to R with emphasis on regression. We will from time to time make use of some of the functions discussed in Fox's book that are not in the base R program. A more comprehensive introduction to R is Venables and Ripley (2002), and we will use the notation VR[3.1], for example, to refer to Section 3.1 of that book. Venables and Ripley has more computerese than does Fox's book, but its coverage is greater and you will be able to use this book for more than linear regression. Other books on R include Verzani (2005), Maindonald and Braun (2002), Venables and Smith (2002), and Dalgaard (2002). We used R Version 2.0.0 on Windows and Linux to write the package. A new version of R is released twice a year, so the version you use will probably be newer. If you have a fast internet connection, downloading and upgrading R is easy, and you should do it regularly.

S-Plus is very similar to R, and most commands that work in R also work in S-Plus. Both are variants of a statistical language called "S" that was written at Bell Laboratories before the breakup of AT&T. Unlike R, S-Plus is a commercial product, which means that it is not free, although there is a free student version available at `elms03.e-academy.com/splus`. The website of the publisher is `www.insightful.com/products/splus`. A library of functions very similar to those for R is also available that will make S-Plus useful for all the methods discussed in ALR.

S-Plus has a well-developed graphical user interface or GUI. Many new users of S-Plus are likely to learn to use this program through the GUI, not through the command-line interface. In this primer, however, we make no use of the GUI.

If you are using S-Plus on a Windows machine, you probably have the manuals that came with the program. If you are using Linux/Unix, you may not have the manuals. In either case the manuals are available online; for Windows see the Help → Online Manuals, and for Linux/Unix use

```
> cd ‘Splus SHOME‘/doc
> ls
```

and see the pdf documents there. Chambers and Hastie (1993) provides the basics of fitting models with S languages like S-Plus and R. For a more general reference, we again recommend Fox (2002) and Venables and Ripley (2002), as we did for R. We used S-Plus Version 6.0 Release 1 for Linux, and S-Plus 6.2 for Windows. Newer versions of both are available.

SAS is the largest and most widely distributed statistical package in both industry and education. SAS also has a GUI. While it is possible to do *some* data analysis using the SAS GUI, the strength of this program is in the ability to write SAS programs, in the editor window, and then submit them for execution, with output returned in an output window. We will therefore view SAS as a *batch* system, and concentrate mostly on writing SAS commands to be executed. The website for SAS is `www.sas.com`.

SAS is very widely documented, including hundreds of books available through amazon.com or from the SAS Institute, and extensive on-line documentation. Muller and Fetterman (2003) is dedicated particularly to regression. We used Version 9.1 for Windows. We find the on-line documentation that accompanies the program to be invaluable, although learning to read and understand SAS documentation isn't easy.

Although SAS is a programming language, adding new functionality can be very awkward and require long, confusing programs. These programs could, however, be turned into SAS *macros* that could be reused over and over, so in principle SAS could be made as useful as R or S-Plus. We have not done this, but would be delighted if readers would take on the challenge of writing macros for methods that are awkward with SAS. Anyone who takes this challenge can send us the results (sandy@stat.umn.edu) for inclusion in later revisions of the primer.

We have, however, prepared *script files* that give the programs that will produce all the output discussed in this primer; you can get the scripts from `www.stat.umn.edu/alr`.

JMP is another product of SAS Institute, and was designed around a clever and useful GUI. A student version of JMP is available. The website is `www.jmp.com`. We used JMP Version 5.1 on Windows.

Documentation for the student version of JMP, called JMP-In, comes with the book written by Sall, Creighton and Lehman (2005), and we will write JMP-START[3] for Chapter 3 of that book, or JMP-START[P360] for page 360. The full version of JMP includes very extensive manuals; the manuals are available on CD only with JMP-In. Fruend, Littell and Creighton (2003) discusses JMP specifically for regression.

JMP has a scripting language that could be used to add functionality to the program. We have little experience using it, and would be happy to hear from readers on their experience using the scripting language to extend JMP to use some of the methods discussed in ALR that are not possible in JMP without scripting.

SPSS evolved from a batch program to have a very extensive graphical user interface. In the primer we use only the GUI for SPSS, which limits the methods that are available. Like SAS, SPSS has many sophisticated tools for data base management. A student version is available. The website for SPSS is `www.spss.com`. SPSS offers hundreds of pages of documentation, including SPSS (2003), with Chapter 26 dedicated to regression models. In mid-2004, amazon.com listed more than two thousand books for which "SPSS" was a keyword. We used SPSS Version 12.0 for Windows. A newer version is available.

This is hardly an exhaustive list of programs that could be used for regression analysis. If your favorite package is missing, please take this as a challenge: try to figure out how to do what is suggested in the text, and write your own primer! Send us a PDF file (sandy@stat.umn.edu) and we will add it to our website, or link to yours.

One program missing from the list of programs for regression analysis is Microsoft's spreadsheet program Excel. While *a few* of the methods described in the book can be computed or graphed in Excel, most would require great endurance and patience on the part of the user. There are many add-on statistics programs for Excel, and one of these may be useful for comprehensive regression analysis; we don't know. If something works for you, please let us know!

A final package for regression that we should mention is called Arc. Like R, Arc is free software. It is available from `www.stat.umn.edu/arc`. Like JMP and SPSS it is based around a graphical user interface, so most computations are done via point-and-click. Arc also includes access to a complete computer language, although the language, lisp, is considerably harder to learn than the S or SAS languages. Arc includes all the methods described in the book. The use of Arc is described in Cook and Weisberg (1999), so we will not discuss it further here; see also Weisberg (2005).

0.1 ORGANIZATION OF THIS PRIMER

The primer often refers to specific problems or sections in ALR using notation like ALR[3.2] or ALR[A.5], for a reference to Section 3.2 or Appendix A.5, ALR[P3.1] for Problem 3.1, ALR[F1.1] for Figure 1.1, ALR[E2.6] for an equation and ALR[T2.1] for a table. Reference to, for example, "Figure 7.1," would refer to a figure in this primer, not to ALR. Chapters, sections, and homework problems are numbered in this primer as they are in ALR. Consequently, the section headings in primer refers to the material in ALR, and not necessarily the material in the primer. Many of the sections in this primer don't have any

Table 0.1 The data file `htwt.txt`.

```
Ht Wt
169.6 71.2
166.8 58.2
157.1 56
181.1 64.5
158.4 53
165.6 52.4
166.7 56.8
156.5 49.2
168.1 55.6
165.3 77.8
```

material because that section doesn't introduce any new issues with regard to computing. The index should help you navigate through the primer.

There are four versions of this primer, one for R and S-Plus, and one for each of the other packages. All versions are available for free as PDF files at `www.stat.umn.edu/alr`.

Anything you need to type into the program will always be in `this font`. Output from a program depends on the program, but should be clear from context. We will write File to suggest selecting the menu called "File," and Transform → Recode to suggest selecting an item called "Recode" from a menu called "Transform." You will sometimes need to push a button in a dialog, and we will write "push OK" to mean "click on the button marked 'OK'." For non-English versions of some of the programs, the menus may have different names, and we apologize in advance for any confusion this causes.

0.2 DATA FILES

0.2.1 Documentation

Documentation for nearly all of the data files is contained in ALR; look in the index for the first reference to a data file. Separate documentation can be found in the file `alr3data.pdf` in PDF format at the web site `www.stat.umn.edu/alr`.

The data are available in a *package* for R, in a *library* for S-Plus and for SAS, and as a directory of files in special format for JMP and SPSS. In addition, the files are available as plain text files that can be used with these, or any other, program. Table 0.1 shows a copy of one of the smallest data files called `htwt.txt`, and described in ALR[P3.1]. This file has two variables, named *Ht* and *Wt*, and ten cases, or rows in the data file. The largest file is `wm5.txt` with 62,040 cases and 14 variables. This latter file is so large that it is handled differently from the others; see Section 0.2.4.

A few of the data files have missing values, and these are generally indicated in the file by a place-holder in the place of the missing value. For example, for R and S-Plus, the placeholder is `NA`, while for SAS it is a period "." Different programs handle missing values a little differently; we will discuss this further when we get to the first data set with a missing value in Section 4.5.

0.2.2 Getting the data files for SPSS

Go to the SPSS page at `www.stat.umn.edu/alr`, and follow the directions to download the directory of data files in a special format for use with SPSS. To use a file, you can either double-click on its name, or start SPSS, select File → Open → Data, and and browse to the file name. To data referred to in the text as `heights.txt` will be called `heights.sav`.

0.2.3 Getting the data in text files

You can download the data as a directory of plain text files, or as individual files; see `www.stat.umn.edu/alr/data`. *Missing values on these files are indicated with a `?`. If your program does not use this missing value character, you may need to substitute a different character using an editor.*

0.2.4 An exceptional file

The file `wm5.txt` is not included in any of the compressed files, or in the libraries. This one file is nearly five megabytes long, requiring as much space as all the other files combined. If you need this file, for ALR[P10.12], you can download it separately from `www.stat.umn.edu/alr/data`.

0.3 SCRIPTS

For R, S-Plus, and SAS, we have prepared *script files* that can be used while reading this primer. For R and S-Plus, the scripts will reproduce nearly every computation shown in ALR; indeed, these scripts were used to do the calculations in the first place. For SAS, the scripts correspond to the discussion given in this primer, but will not reproduce everything in ALR. The scripts can be downloaded from `www.stat.umn.edu/alr` for R, S-Plus or SAS.

Although both JMP and SPSS have scripting or programming languages, we have not prepared scripts for these programs. Some of the methods discussed in ALR are not possible in these programs without the use of scripts, and so we encourage readers to write scripts in these languages that implement these ideas. Topics that require scripts include bootstrapping and computer intensive methods, ALR[4.6]; partial one-dimensional models, ALR[6.4], inverse response plots, ALR[7.1, 7.3], multivariate Box-Cox transformations, ALR[7.2], Yeo-Johnson transformations, ALR[7.4], and heteroscedasticity tests, ALR[8.3.2]. There are several other places where usability could be improved with a script.

If you write scripts you would like to share with others, let me know (sandy@stat.umn.edu) and I'll make a link to them or add them to the website.

0.4 THE VERY BASICS

Before you can begin doing any useful computing, you need to be able to read data into the program, and after you are done you need to be able to save and print output and graphs. All the programs are a little different in how they handle input and output, and we give some of the details here.

0.4.1 Reading a data file

Reading data into a program is surprisingly difficult. We have tried to ease this burden for you, at least when using the data files supplied with ALR, by providing the data in a special format for each of the programs. There will come a time when you want to analyze real data, and then you will need to be able to get your data into the program. Here are some hints on how to do it.

SPSS At `www.stat.umn.edu/alr`, you will be able to download all the data files for book (except for `wm5.txt`) in a directory of files in the format preferred by SPSS. These files all end in `.sav`, and are not human readable. To use these files, you simply select File → Open → Data and then browse to the file, or else double-click on the file name.

You can also download and use the *plain text files* that are available on the website. The advantage to the plain text files is that they can be used with many programs besides SPSS[1]. We provide here extensive instructions on how to read a plain text file. We assume the file has a name ending in `.txt`, and looks something like the data in Table 0.1.

Select File → Read Text Data. In the dialog browse to the data file you want to use and press Open. This should open the Text Import Wizard which helps you open the `.txt` in the correct format. When reading an ALR data file follow these six steps:

1. The first screen of the Text Import window shows the first few lines of the file, and asks if you have a predefined format for the file. Unless you have previously saved a format for this particular file, check `No`, and then press NEXT. If you plan on opening the same data file over many

[1] You can use File → Save as to save an SPSS file in many other formats, including plain text.

SPSS sessions, the last step gives you the option of saving the format defined in the following steps.

2. The files for ALR are formatted as space separated columns with each variable named at the top of its column. On the second screen, make sure `Delimited` is checked as the variable arrangement and `Yes` is checked under variable name inclusion, and press NEXT.

3. On the third screen, since SPSS was already told that the variable names are included at the top of each column the default line number for the first case of data should be 2. If it is not, make that change. The default values for the next two questions should be correct so simply check that `Each line represents a case` and `All of the cases` are chosen, and then press NEXT.

4. On the fourth screen, the delimiter used to separate the columns is a space so make sure SPSS has chosen this option. There should not be any text qualifiers so `None` should be checked for this question. Click NEXT.

5. The fifth screen gives you the option of editing the name of each variable, and setting or changing its *type*. SPSS has several types of variables, but the usual type we will use is *numeric*. Other types include *string* for text variables, *date* for dates, and so on. To check the specifications for each variable click anywhere on its column in the Data Preview section of this screen. Most default specifications should be correct. Some of the data files have an extra blank after the last variable on the line, and this causes SPSS to add an additional variable that is all blanks. While harmless, you might find this extra variable unesthetic, and you can delete it now by clicking on it and selecting `Do Not Import` from the data format list; you can delete it later as well by selecting the variable from the spreadsheet and then Edit → Cut. Once the variables are satisfactory press NEXT.

6. On the final screen you have the option of saving the format entered for this data file. By saving this format you can save time when reading the same data file again in a different session by selecting its predefined format in step 1. Alternatively, you can also save the data file in the SPSS `.sav` format which can be opened without any of the formatting needed for a `.tex` file. Pressing FINISH completes the formatting steps of the text wizard.

After completion of this (seemingly endless) list of steps, the data will appear in the `Data Editor` window. The editor offers two views of the data: the *data view*, which is much like a spreadsheet, and the *variable view*, which lists variables and their properties. Because SPSS is a general purpose program, each of the variables in a data set can have many *attributes*, including its

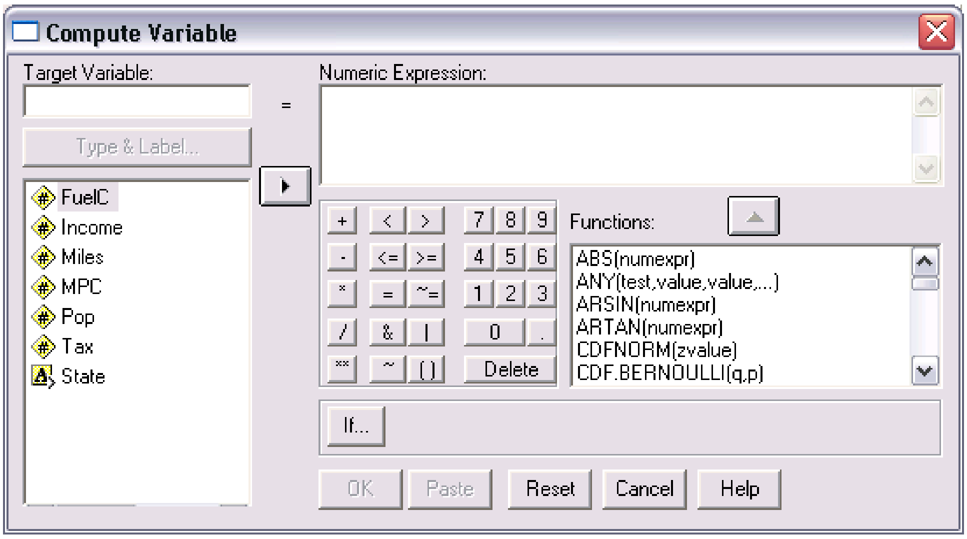


Fig. 0.1 SPSS transformation dialog.

type, as we have already seen, and its *measure*, allowing you to specify if the variable is *scale*, or continuous, *nominal*, meaning an indicator for categories, or *ordinal*, meaning ordered categories. SPSS will guess the right measure, but it will sometimes guess wrong. For example with `forbes.txt`, all variables are set to nominal by default, but the correct measure to plot or analyze the data would be the scale measure.

You can transform variables by selecting Transform → Compute and entering the appropriate formula in the expression editor. Figure 0.1 shows the dialog used for defining new variables when the data file `fuel2001.txt` is open. The target variable will be assigned the expression value. The name of this variable can be a new variable name or an existing variable name which will have the effect of overwriting the current values with the transformed values. Examples of transformations will be given in Chapter 1.

The generality of SPSS can cause new users lots of frustration, particularly if the defaults selected by the program for types and measures are not appropriate for the data. Taking a little time at the beginning of an analysis to be sure that the program has correctly read and defined your data can save you lots of grief later.

0.4.2 Saving text output and graphs

All the programs have many ways of saving text output and graphs. We will make no attempt to be comprehensive here.

SPSS Once you run a procedure in SPSS the results are displayed in a `Viewer` window, which is composed of an outline pane on the left and a content pane on the right. The content pane contains the results from the procedure and the outline pane allows you to choose which tables or graphs you want to see by opening or closing the small book icon next to each result with a double click. Many results are presented in a *pivot table* which can be manipulated in a variety of ways to create a data summary to your own liking. Chapter 11 of SPSS (2003) is a good reference on the many ways to edit these tables. If you would prefer text output over a pivot table you can use a `Draft Viewer` window instead of the standard `Viewer` window by selecting File → New → Draft Output before running the desired procedure. When two or more output viewers are present the output to any analysis will be directed to the *designated* viewer. A viewer is designated if the status bar at the bottom of the window shows a red "!". To change the output designation press the red ! button on the toolbar of the desired viewer.

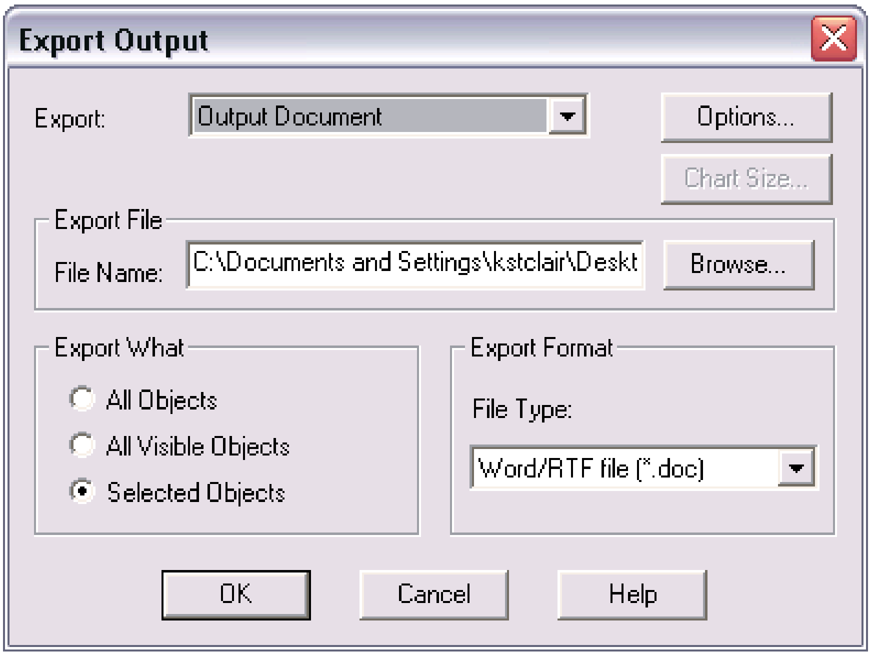
The tables and graphs in a `Draft Viewer` window can be exported or saved as a `.txt` or `.rtf` file. You have the choice of exporting all output or only the selected graphs and tables. Select File → Export for your export options or choose File → Save As to save all the output. You can also copy selected graphs and tables and then paste them directly into a word processing document.

There are multiple ways to save output from a `Viewer` window as outlined below.

Copy/Paste Graphs and pivot tables can be copied by right clicking (in Windows) and selecting Copy. They can then be pasted into a word processing document or Excel spreadsheet.

Export You can export selected results by choosing File → Export. The export dialog is shown in Figure 0.2. You first choose what to export by selecting one of the following Export options: `Output Document` will export tables, text, and charts (graphs), `Output Document (No Charts)` will export only tables and text, or `Charts Only` will export only charts. In the Export File section you specify the destination file name. Next choose what to export: all objects produced as output of your procedures, all object visible (open book) in the content pane, or only the objects selected in the content pane. Finally you choose the format of the exported output. When exporting only charts you have eight formats to choose from, three of which are `.eps`, `.jpg`, and `.bmp`. For the other two types of output there are four format options to choose from: `.htm`, `.txt`, `.xls`, and `.doc`. Any of these formats can be edited further by selecting OPTIONS after picking the file type. More details on exportation can be found in Chapter 9 of SPSS (2003).

SPSS File The entire `Viewer` window can be saved as a `.spo` file by selecting File → Save. This type of file can be opened again as a `Viewer` window in SPSS.



Export: Output Document

File Name: C:\Documents and Settings\kstclair\Deskt

○ All Objects

○ All Visible Objects

● Selected Objects

File Type: Word/RTF file (*.doc)

Fig. 0.2 SPSS Export dialog for `Viewer` window output.

0.4.3 Normal, F, t and χ^2 tables

ALR does not include tables for looking up critical values and significance levels for standard distributions like the t, F and χ^2. Although these values can be computed with any of the programs we discuss in the primers, doing so is easy only with R and S-Plus. Also, the computation is fairly easy with Microsoft Excel. Table 0.2 shows the functions you need using Excel.

SPSS SPSS does include functions for computing both significance levels and critical values, as defined in Table 0.3. To use one of the functions, you must first have an active data set, and then select Transform → Compute. In the resulting dialog, you must select a name for the Target Variable, and then in the “Numeric Expression” area you can type the expression based on Table 0.3 that does the calculation you want. After selecting OK, the result of the calculation will be added to the data set, and repeated once for each observation.

Table 0.2 Functions for computing p-values and critical values using Microsoft Excel. The definitions for these functions are not consistent, sometimes corresponding to two-tailed tests, sometimes giving upper tails, and sometimes lower tails. Read the definitions carefully. The algorithms used to compute probability functions in Excel are of dubious quality, but for the purpose of determining p-values or critical values, they should be adequate; see Knüsel (2005) for more discussion.

Function	What it does
`normsinv(p)`	Returns a value q such that the area to the left of q for a standard normal random variable is p.
`normsdist(q)`	The area to the left of q. For example, `normsdist(1.96)` equals 0.975 to three decimals.
`tinv(p,df)`	Returns a value q such that the area to the left of $-\|q\|$ *and* the area to the right of $+\|q\|$ for a $t(\mathrm{df})$ distribution equals q. This gives the critical value for a two-tailed test.
`tdist(q,df,tails)`	Returns p, the area to the left of q for a $t(df)$ distribution if $tails = 1$, and returns the sum of the areas to the left of $-\|q\|$ and to the right of $+\|q\|$ if $tails = 2$, corresponding to a two-tailed test.
`finv(p,df1,df2)`	Returns a value q such that the area to the *right* of q on a $F(\mathrm{df}_1, \mathrm{df}_2)$ distribution is p. For example, `finv(.05,3,20)` returns the 95% point of the $F(3, 20)$ distribution.
`fdist(q,df1,df2)`	Returns p, the area to the *right* of q on a $F(\mathrm{df}_1, \mathrm{df}_2)$ distribution.
`chiinv(p,df)`	Returns a value q such that the area to the *right* of q on a $\chi^2(\mathrm{df})$ distribution is p.
`chidist(q,df)`	Returns p, the area to the *right* of q on a $\chi^2(\mathrm{df})$ distribution.

0.5 ABBREVIATIONS TO REMEMBER

ALR refers to the textbook, Weisberg (2005). VR refers to Venables and Ripley (2002), our primary reference for R and S-Plus. JMP-START refers to Sall, Creighton and Lehman (2005), the primary reference for JMP. Information typed by the user looks like `this`. References to menu items looks like File or Transform → Recode. The name of a BUTTON to push in a dialog uses this font.

Table 0.3 Functions for computing p-values and critical values using SPSS. These functions may have additional arguments useful for other purposes.

Function	What it does
`CDF.NORM(p)`	Returns a value q such that the area to the left of q for a standard normal random variable is p.
`IDF.NORM(q)`	Returns a value p such that the area to the left of q on a standard normal is p.
`CDF.T(p,df)`	Returns a value q such that the area to the left of q on a t(df) distribution equals q.
`IDF.T(q,df)`	Returns p, the area to the left of q for a $t(df)$ distribution
`CDF.F(p,df1,df2)`	Returns a value q such that the area to the left of q on a $F(\mathrm{df}_1, \mathrm{df}_2)$ distribution is p. For example, `qf(.95,3,20)` returns the 95% points of the $F(3, 20)$ distribution.
`IDF.F(q,df1,df2)`	Returns p, the area to the left of q on a $F(\mathrm{df}_1, \mathrm{df}_2)$ distribution.
`CDF.CHISQ(p,df)`	Returns a value q such that the area to the left of q on a χ^2(df) distribution is p.
`IDF.CHISQ(q,df)`	Returns p, the area to the left of q on a χ^2(df) distribution.

0.6 COPYRIGHT AND PRINTING THIS PRIMER

1
Scatterplots and Regression

1.1 SCATTERPLOTS

A principal tool in regression analysis is the two-dimensional scatterplot. All statistical packages can draw these plots. We concentrate mostly on the basics of drawing the plot. Most programs have options for modifying the appearance of the plot. For these, you should consult documentation for the program you are using.

SPSS There are two types of scatterplots available in SPSS: the standard plot and the interactive plot. An interactive plot allows for some modification after it has been created, such as adding additional variables to a plot. Once the data file has been changed, however, the plot will become detached and you cannot use any newly made variables in the plot. We generally prefer the presentation and resolution of the interactive plots over that of the standard plots, but standard plots have some built-in options not available in the interactive plots. For instance, standard plots have a large selection of lines which can be inserted into them, such as loess curves or quadratic or cubic regression lines, while interactive plots have a smaller selection of such lines. All scatterplot instructions below will create interactive plots, except for ALR[F1.10] which fits a loess curve. SPSS refers to any type of graph which has been produced as a *chart*.

After the data file `heights.txt` has been read into SPSS, we can create ALR[F1.1] by selecting Graphs → Interactive → Scatterplot from the `Data Editor` window. In the dialog popup, click and drag the variable *Dheight* to

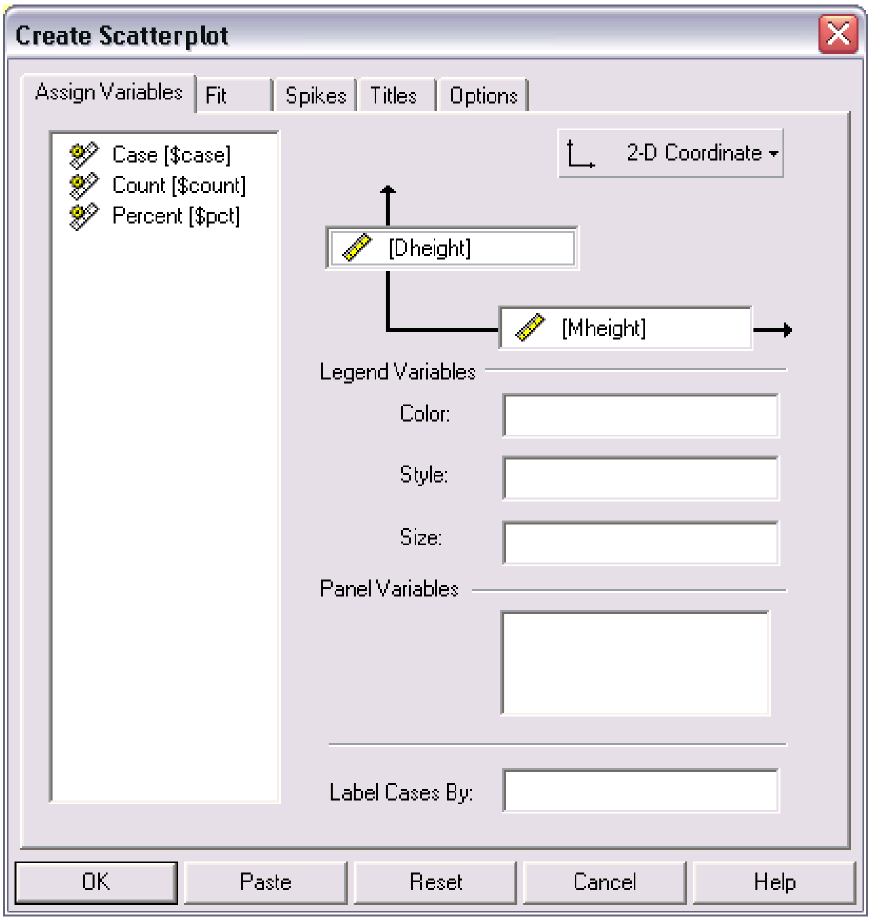
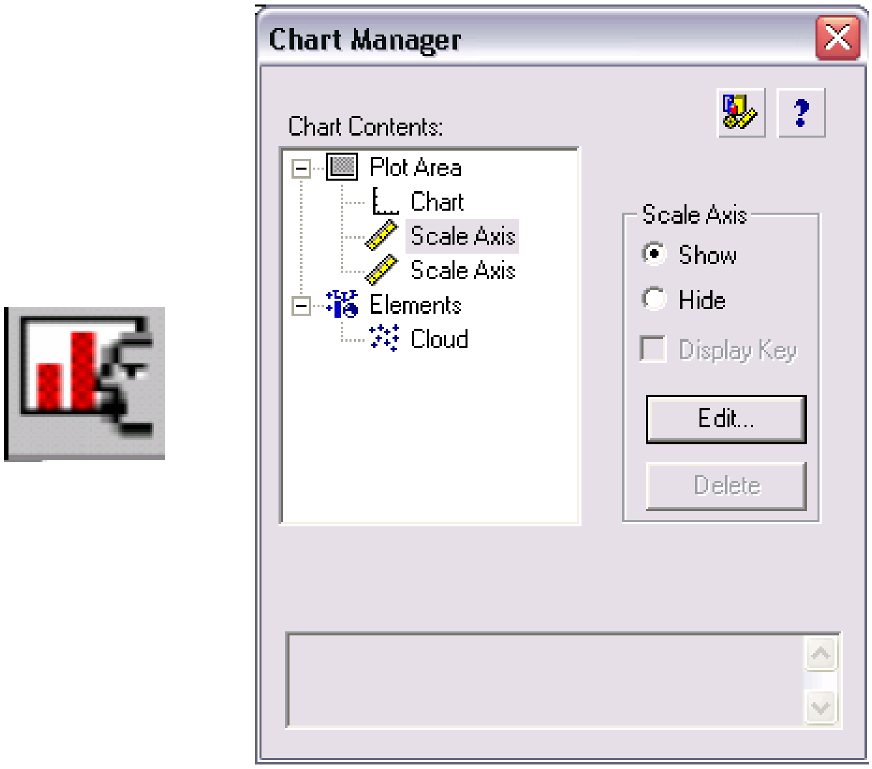


Fig. 1.1 Interactive plot dialog for the data `Heights.txt`.

the vertical axis and click and drag the variable *Mheight* to the horizontal axis. This dialog should now look like Figure 1.1. You can select the Titles tab to give a title, subtitle, or caption to the scatterplot, then press OK. The scatterplot will be displayed in the designated `Viewer` window. To edit this plot double click anywhere on the plot or right click and select SPSS Interactive Graph Object → Edit. This will create moveable toolbars which can be used to modify the plot.

The chart manager can be used to change the components which make up the plot. We will use it to change the axes of the plot because, as discussed in ALR[1.1], we would like to draw this scatter plot so that the horizontal and vertical axes are the same. To access the chart manager either right click on the region outside the plot and select Chart Manager or click the chart manager



● Show

○ Hide

□ Display Key

Fig. 1.2 The Chart Manager icon and dialog.

tool shown in Figure 1.2. This figure also contains the chart manager dialog that appears after the icon is clicked and outlines the chart contents which can be modified. From this outline click on the first `Scale Axis` option then select EDIT. This will allow the horizontal axis (*Mheight*) to be manipulated in a variety of ways. Under the scale tab we can change the default settings by unchecking the auto box behind each scale option and entering the desired value. In this manner, set the minimum to 55 and the maximum to 75. By then selecting the button APPLY, you can see the results of the change without closing the editing window. Next, set the tick interval to 5 and the number of ticks to 5 and click OK. This should produce an axis identical to that in figure ALR[F1.1]. Back at the chart manager dialog select the second `Scale Axis` option and repeat the previous steps to edit the vertical axis. You can change the dimensions of the overall scatterplot by choosing and editing the `Chart` option in the chart manager window. Finally, you can change the plotting symbols by selecting the `Cloud` option. After pressing EDIT, selecting the symbols tab from the popup window allows you to change the plotted

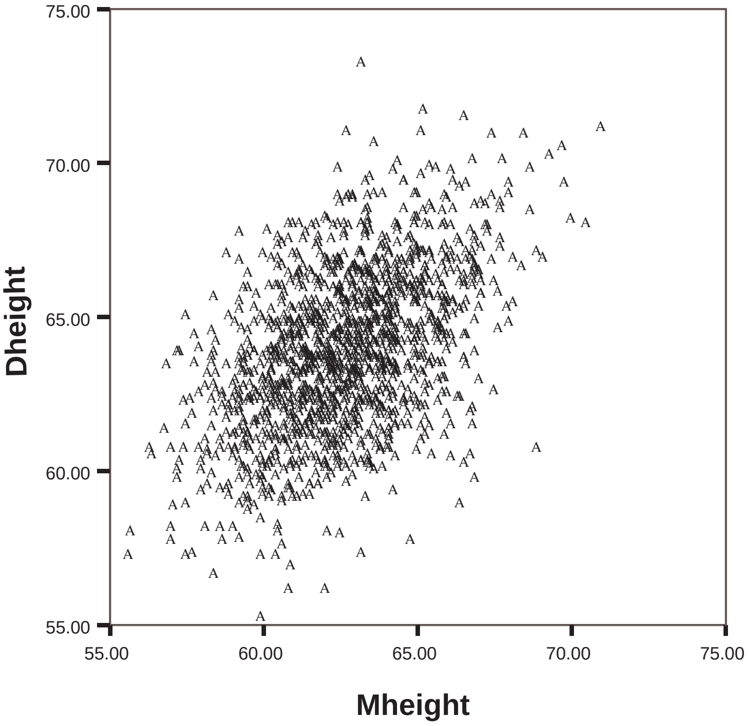


Fig. 1.3 The SPSS version of ALR[F1.1] drawn as an interactive plot. The plotting symbol used in SPSS was an open circle which changed when exporting the plot as a `.eps` file.

points to the size, style, and color you want. The SPSS version of ALR[F1.1] is shown in Figure 1.3.

On all interactive plots you can identify any point by its case number. When the plot is interactive, change from the "arrow tool" cursor to the "point id tool" cursor located on the interactive plot toolbar. If you click on a point in the data cloud, its case number will be displayed next to it. To remove the case number from the plot, click a second time on the point.

ALR[F1.2] can be obtained by selecting the cases to plot and then following the steps above to draw ALR[F1.1]. To make the selection choose Data → Select Cases from any SPSS window. With the popup window we can specify which cases we want to select and by doing so we can filter the unselected cases from any analysis or graphs. The cases we want to select

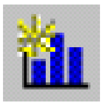


Fig. 1.4 Insert Element tool.

are any which satisfy $57.5 < \textit{Mheight} \leq 58.5$, *or* $62.5 < \textit{Mheight} \leq 63.5$ *or* $67.5 < \textit{Mheight} \leq 68.5$. To specify these cases select *Mheight* in the popup window and check the option `If condition is satisfied`, then press the newly activated IF button. The dialog box for this choice allows you to enter a conditional expression which will select the cases evaluated as true. We type the conditions into the text box using the logical "and" symbol `&` and "or" symbol `|` when needed. To select the cases for ALR[F1.2] the following conditions are entered into the conditional text area:

```
((57.5 < Mheight) & (Mheight <= 58.5)) |
((62.5 < Mheight) & (Mheight <= 63.5)) |
((67.5 < Mheight) & (Mheight <= 68.5))
```

Press continue then, if needed, check `Filtered` as the action to take with the unselected variables and press OK. This should add a filter variable to the data table and put a line through the case numbers of unselected cases. Follow the steps used above to create ALR[F1.1] and the resulting scatterplot should filter out the unselected cases and produce ALR[F1.2]. If you wish to remove the filter (and obtain ALR[F1.1]) activate the graph by double clicking and then select Edit → Assign Variables from the `Viewer` menu. Choose the tab Cases from the popup window and click and drag the filter conditions to the variable list. This will automatically remove the filter from the plot.

ALR[F1.3] uses the `forbes.txt` data file so, as mentioned in Section 0.4.1, change the variable measure to scale if it something different. SPSS can not work with two open data sets, so to read this new file you can either close the previous data file in the `Data Editor` or start a new SPSS session. To draw ALR[F1.3A] follow the standard steps for creating an interactive scatterplot. The regression line can be added by selecting Insert → Fit Line → Regression or clicking the insert element tool shown in Figure 1.4 and choosing Regression Fit. This will add the OLS regression line to the plot. It also adds the equation for the fitted line next to the plotted line. To remove this label select it with a click, then right click and choose Hide Label from the popup.

To obtain ALR[F1.3B] you will need to analyze the data using linear regression, and then save the residuals, which you will plot against *Temp*. To start the analysis select Analyze → Regression → Linear from any window. In the popup dialog enter *Pressure* as the *dependent variable*, which is ALR is called the *response* and *Temp* as the *independent variable*, called in ALR either a predictor or a term, depending on context. Any statistic that can be obtained from this linear regression can be saved by selecting the button SAVE.

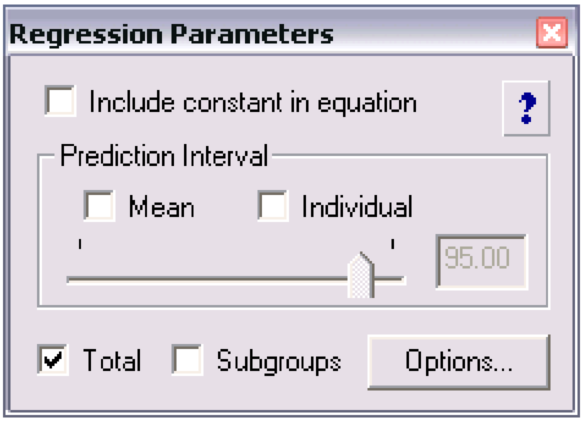
This opens a window from which we can save the residual values by checking `Unstandardized` from the residuals category then pressing continue. Click OK back in the linear regression window and SPSS will then calculate the regression. In the designated `Viewer` window the output for the model fit, coefficients, and analysis of variance will be displayed. To obtain ALR[F1.3B] we must construct a scatterplot by selecting Graphs → Interactive → Scatterplots, then place *Temp* on the horizontal axis and *Unstandardized Residual* on the vertical axis. We can automatically add the mean line (as opposed to inserting it afterwards) by selecting the tab Fit and choosing `Mean` as the method and then pressing OK.

ALR[1.4] uses a base 10 log transformation of *Pressure* as the dependent variable, then draws the scatterplots following the same steps used to get ALR[1.3]. This transformed variable is *Lpres* in the Forbes data file. If this variable were not provided, we could transform *Pressure* as follows. Select Transform → Compute to obtain the dialog shown in Figure 0.1. Enter the name *logPressure* as the target variable, then select LG10 as the function and use the arrow button to move it to the expression text box. Select *Pressure* as the argument for the log function and press OK. This will add the new variable to the data table.

To plot ALR[1.5] read the `wblake.txt` file into SPSS and change all variables to the *measure* type scale in the variable view tab of the `Data Editor`. As done above, follow the commands for making an interactive scatterplot and insert the regression line. Double click on the plot to activate it, then press the insert element icon and select Dot-line. This will draw a line connecting the mean length of each age. To change the line type select the line and right click, then choose Dots and Lines and select the style of line you want.

If *Age* is a nominal type variable SPSS will not insert a regression line. If you have changed the variable type in the `Data Editor` but the dialog to create the scatterplot still shows it as nominal (i.e. there isn't a little ruler by it), then click reset. Press OK in the popup and SPSS will read in the data values again, this time with the correct measure type.

To draw figure ALR[F1.7] read `turkey.txt` in SPSS and check that the variables *Gain* and *A* are scale and *S* is ordinal or nominal. If the latter is scale you cannot use it to determine the plotting symbols, but will be prompted to change the type when drawing the scatterplot. Follow the usual steps to create the interactive scatterplot. Enter *A* on the horizontal axis and *Gain* on the vertical. Then drag the variable *S* to either the `Color` or `Style` option under legend variables. Both options will create a plot with each type of *S* drawn in either a different color or different symbol, but not both. To change the type of symbol or color used, click on the chart manager icon and edit the `Color Legend`.



Regression Parameters

□ Include constant in equation

Prediction Interval

□ Mean □ Individual

95.00

☑ Total □ Subgroups

Fig. 1.5 SPSS regression parameters dialog for modifying an OLS line added to an interactive scatterplot.

1.2 MEAN FUNCTIONS

SPSS ALR[F1.8] cannot be duplicated in SPSS because the dashed line, the regression line for E(*Dheight* | *Mheight*) = *Mheight*, cannot be added to the scatterplot formed in ALR[F1.1]. We can add a regression line which constrains the intercept to zero, but we cannot force it's slope to equal to one. To insert the regression line with no intercept we first add the standard OLS line to ALR[F1.1]. Select the line with a mouse click, then right click on the highlighted line and choose Regression Parameters. Uncheck the box for the option `Include constant in equation` as shown in Figure 1.5.

1.3 VARIANCE FUNCTIONS

1.4 SUMMARY GRAPH

1.5 TOOLS FOR LOOKING AT SCATTERPLOTS

SPSS ALR[F1.10] adds a loess smooth to the scatterplot of the `heights.txt` data file. We cannot insert this line into the interactive plot created earlier for ALR[F1.1] so we must redraw ALR[F1.1] using a standard static scatterplot. To do this, select Graphs → Scatter then choose the Simple plotting option in the popup dialog and press DEFINE. To enter the axes select a variable and click the arrow button next to the appropriate axis, then press OK. A scatterplot similar to ALR[F1.1] will appear in the `Viewer` window and we can modify this graph with a `Chart Editor` which appears after double clicking on the scatterplot. With this editor we can make the same changes to the horizontal axis which were made for ALR[F1.1] by selecting Edit → Select X Axis or by clicking the large X on the toolbar. Choose the scale tab to edit the range and increments plotted on the horizontal axis. Repeat these steps for the Y axis to modify the vertical axis.

To add any line to the plot we must first select the point cloud by clicking somewhere on it. Next, choose Chart → Add Chart Element → Fit Line at Total to add the regression line. In the dialog window select the Fit Line tab and check the `Linear` fit option, then press APPLY and close the dialog. To add the loess curve, highlight the point cloud and follow the same menu options used to fit the regression line, but in the Fit Line tab of the dialog check the `Loess` fit option and press APPLY. To change the style of the line select the Lines tab and apply the line style desired and close the window. ALR[F1.10] should now appear in the editor window and to apply these changes to the scatterplot in the `Viewer` window simply close the editor.

1.6 SCATTERPLOT MATRICES

SPSS To draw ALR[F1.11] we must first transform the variables in the data file `fuel2001.txt`. This is done as demonstrated earlier when drawing ALR[F1.4]. To transform the four variables follow Transform → Compute and enter the appropriate function expression for each:

```
Dlic = 1000*Drivers/Pop
Fuel= 1000*FuelC/Pop
Income = Income/1000
logMiles = LG10(Miles)/LG10(2)
```

By naming a transformed variable the same name as a current variable you will be asked if you want to change the existing variable, to which you press OK. includes log functions only for natural logs and for base 10 logs. To get logs to the base two that will match the text, use the fact that $\log_b x = \log_{10} x / \log_{10} b$, then the transformation above is equal to the base 2 log of *Miles*. Because if this extra step in computing the logs, we suspect that the use of base two logs will be relatively unusual with .

There is no interactive scatterplot matrix option, so to draw ALR[F1.11] we must use a standard graphics scatterplot. Select Graphs → Scatter and in the popup dialog choose the Matrix option and press DEFINE. Then using the arrow button next to the `Matrix Variables` box, enter the variables *Tax*, *Dlic*, *Income*, *logMiles*, and *Fuel* and press OK. You can change the axes from the Edit menu, but any range or increment changes you make will be applied to *all* plots in the matrix. Ticks marks and their values can be added by checking the option `Display ticks` given in the Ticks and Grids tab, then checking `Display labels` from the Axis Labels tab and pressing APPLY.

Problems

1.1. Boxplots would be useful in a problem like this because they display level (median) and variability (distance between the quartiles) simultaneously.

SPSS To examine a variable at different levels of a second variable select Analyze → Descriptive Statistics → Explore. Enter *Length* as the dependent variable (the response) and *Age* as the factor (a term). You can choose the statistics and plots you would like to see, but the default settings are usually adequate.

To plot standard deviation versus *Age*, choose Graphs → Interactive → Line and place *Age* on the horizontal axis and *Length* on the vertical axis. A box should then appear at the bottom of the dialog window from which you choose `Standard Deviations` as the value which the dots and lines will represent. Press OK and the standard deviation plot will be drawn.

1.2.

SPSS To resize an interactive scatterplot edit the Scale option in the Chart Manager but make sure to uncheck `Maintain aspect ratio` in order to change the width but not the height. To resize a standard scatterplot simply click once on the plot and resize it using the mouse.

1.3.

SPSS Details on transforming to $\log_2$ are given in Section 1.6 when explaining ALR[F1.11]. For drawing graphs, the base of logarithms is irrelevant.

2

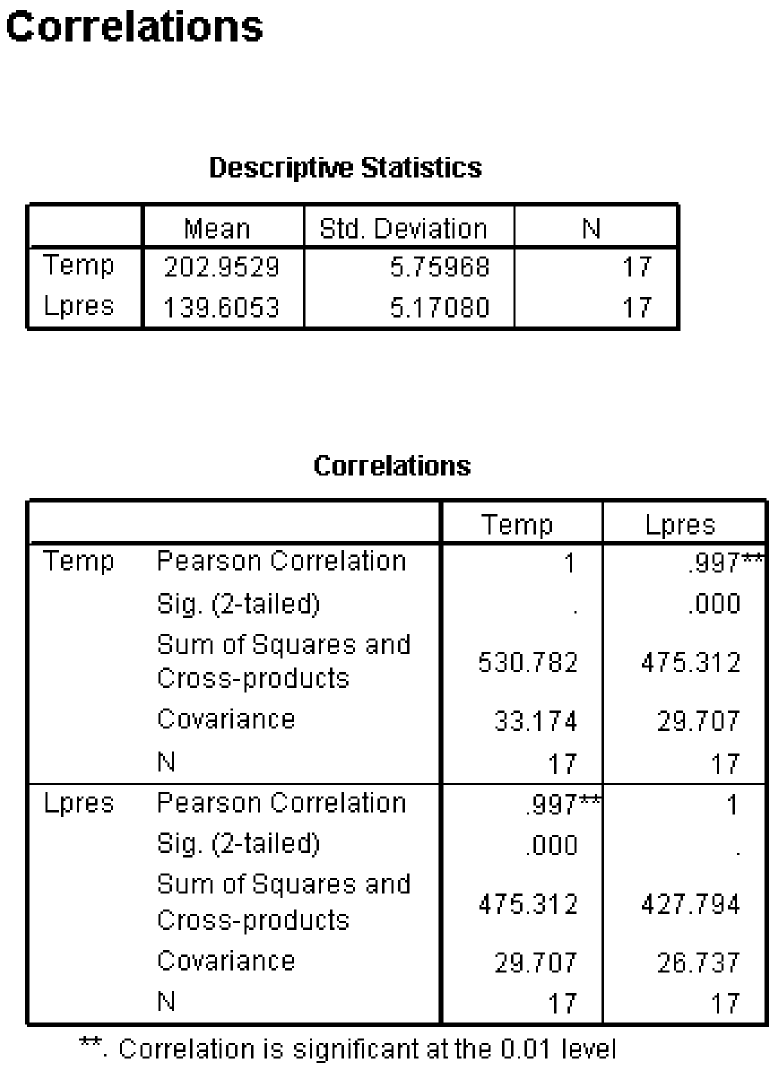
Simple Linear Regression

2.1 ORDINARY LEAST SQUARES ESTIMATION

All the computations for simple regression depend on only a few summary statistics; the formulas are given in the text, and in this section we show how to do the computations step–by-step. All computer packages will do these computations automatically, as we show in Section 2.6.

2.2 LEAST SQUARES CRITERION

SPSS The means and sum of squares used in computing the least squares estimators for simple linear regression can be quickly calculated in SPSS. After the `forbes.txt` data is entered, select Analyze → Correlate → Bivariate and place *Temp* and *Lpres* in the variables box. Next, click the OPTIONS button and check both Statistics options `Means and SD` and `Cross-product deviations and covariances`. Press Continue, then OK, and the results will be displayed in the `Viewer` window as shown in Figure 2.1. SXX is given in the top left square of the *Sum of Squares and Cross-products* row, SYY is given in the bottom right square, and SXY is given in the other two squares. The means are given as part of the descriptive statistics output, so to calculate the least squares estimates use a calculator and the formulas in ALR[2.2].



Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
Temp	202.9529	5.75968	17
Lpres	139.6053	5.17080	17

Correlations

		Temp	Lpres
Temp	Pearson Correlation	1	.997**
	Sig. (2-tailed)	.	.000
	Sum of Squares and Cross-products	530.782	475.312
	Covariance	33.174	29.707
	N	17	17
Lpres	Pearson Correlation	.997**	1
	Sig. (2-tailed)	.000	.
	Sum of Squares and Cross-products	475.312	427.794
	Covariance	29.707	26.737
	N	17	17

**. Correlation is significant at the 0.01 level

Fig. 2.1 SPSS bivariate correlation output for the Forbes data.

2.3 ESTIMATING σ^2

2.4 PROPERTIES OF LEAST SQUARES ESTIMATES

2.5 ESTIMATED VARIANCES

The estimated variances of coefficient estimates are computed using the summary statistics we have already obtained. These will also be computed automatically linear regression fitting methods, as shown in the next section.

2.6 COMPARING MODELS: THE ANALYSIS OF VARIANCE

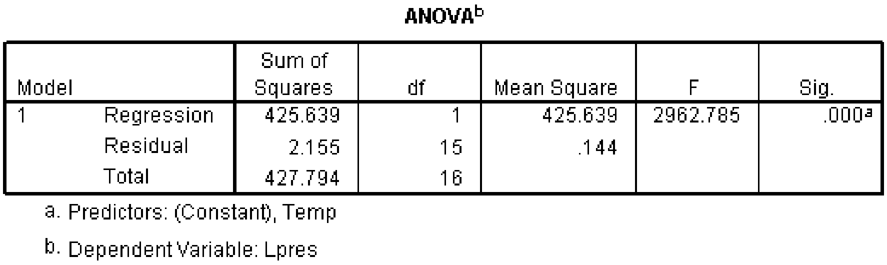
Computing the analysis of variance and F test by hand requires only the value of RSS and of $SSreg = SYY - RSS$. We can then follow the outline given in ALR[2.6].

SPSS In Section 1.1 we first showed how to fit linear regression in SPSS when drawing the residual plot for the Forbes data. The standard analysis is SPSS returns coefficient estimates, estimates and their standard errors, R^2 and $\hat{\sigma}^2$. If you want to see other statistics or plots you must specify them *before* running the analysis. Some plots will require that you save quantities like residuals and fitted values, and then plot them outside the regression command.

Load the Forbes data into SPSS, and select Analyze → Regression → Linear. Place the *response Lpres* in the Dependent box and *predicor Temp* in the Independent box. The Method popup menu located in the dialog near the Independent variable box gives you the option of changing the way predictors enter into the analysis, but for simple regression use the standard method `Enter`. In SPSS the default mean function for linear regression includes the intercept and for this example equals $\mathrm{E}(Lpres|Temp) = \beta_0 + \beta_1 Temp$. Although not relevant for this data, you can fit the regression through the origin by pressing the OPTIONS button and unchecking the `Include constant in equation` option. At this point if you click OK without making any changes to the options STATISTICS, PLOTS, or SAVE, you will simply receive the values discussed in ALR[2.6] and ALR[2.7].

In the designated `Viewer` window this analysis will produce four tables, though we will currently focus on the last two. The third table is the analysis of variance table shown in Figure 2.2. The first two lines of this table should match the values in ALR[T2.4].

The final table presented in the default output is the coefficients table given in Figure 2.3. These results should match the ones given in ALR except for the additional "Standardized Coefficients" column, which you can ignore. The "Sig." column is the p-value given in ALR. SPSS has rounded these values



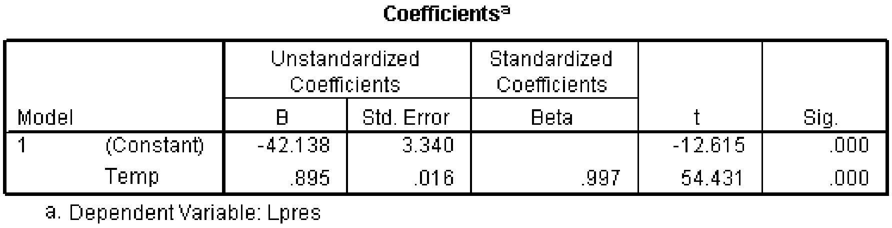
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	425.639	1	425.639	2962.785	.000[a]
	Residual	2.155	15	.144		
	Total	427.794	16			

a. Predictors: (Constant), Temp

b. Dependent Variable: Lpres

Fig. 2.2 SPSS Analysis of Variance table.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-42.138	3.340		-12.615	.000
	Temp	.895	.016	.997	54.431	.000

a. Dependent Variable: Lpres

Fig. 2.3 SPSS Coefficients table.

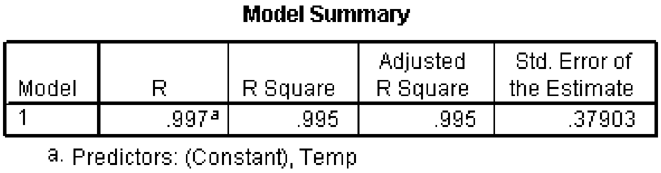
to three decimal places and a value equal to .000 means that the calculated p-value was "approximately zero".

You can edit any of these tables by double clicking on a table. Then by right clicking anywhere on the selected table, you can edit the properties or looks of the table. By right clicking on a certain cell and selecting Cell Properties you can edit the value in that particular cell. For example, if you would like to view the p-value before it was rounded, select the format `#.##E+##` under the Value tab and press Apply.

2.7 THE COEFFICIENT OF DETERMINATION, R^2

SPSS The value of R^2 can be calculated from the values obtained using the sums of squares from Section 2.2 or it can be read from the output from the linear regression fit. For the Forbes data, this table is given in Figure 2.4.

For simple regression, the column "R" is equal to the sample correlation between the predictor and response and "R Square" is the square of this value and equal to R^2. The "Adjusted R Square" can be ignored. The "Std. Error of the Estimate" is $\hat{\sigma}$.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.995	.995	.37903

a. Predictors: (Constant), Temp

Fig. 2.4 SPSS Regression Summary table.

2.8 CONFIDENCE INTERVALS AND TESTS

Confidence intervals and tests can be computed using the formulas in ALR[2.8], in much the same way as the previous computations were done.

SPSS To obtain 95% confidence intervals for the parameter estimates, we must tell SPSS to show this information as we are building the regression model. Select Analyze → Regression → Linear, and select the dependent (response) variable and independent (predictor) variable as before. Now, push the button STATISTICS then check `Confidence intervals` under the Regression Coefficients options and press Continue. This will add the upper and lower bounds to the Coefficients output table. Selecting `Covariance matrix` from this list will produce the estimated covariance matrix for the coefficients *without* the intercept. This matrix is not useful for simple linear regression. To obtain the covariance matrix for all terms in the regression model, choose SAVE from the regression dialog and then check `Coefficient statistics:` and select a file name. This will store the coefficient estimates, standard errors, and covariance matrix in a data table.

If you would like to calculate a confidence interval at a level other than 95% or test a different hypothesis, you must do so by using a calculator and the estimates and standard errors provided in the SPSS output tables. You can get the correct t multiplier needed for a, say, 90% confidence interval by selecting Transform → Compute then entering

```
IDF.T(.95,15)
```

into the "numeric expression" box, where 15 is the residual degrees of freedom. Give a name to the variable in the "Target variable" box, and then select OK. This will create a column variable whose value equals the correct multiplier for the interval. Similarly, you can obtain the p-value for a two-sided hypothesis by entering the following into the expression box:

```
(1-CDF.T(2.137,15))*2
```

where 2.137 is the test statistic for the intercept of the Forbes data calculated in ALR[2.8.1]. The value for a one-sided test can be found by deleting "∗2" from the command. For all these transformations, you may need to increase the number of decimals shown for the variable in the data table by editing that column in the Variable View.

Prediction and fitted values

SPSS Just as we told SPSS to compute confidence intervals for the parameters as we were building the model, we must also tell it to compute the predicted and fitted values. Start by selecting the button Save in the regression dialog window. By checking `Unstandardized` and `S.E of mean predictions` under the Predicted Values list you will save the predicted value and the standard error of the fitted values for each observation. The values of these new variables will be added to the data table and by pointing to the column header you will get a description of that variable. The description of these standard errors is deceiving, as "Standard Error of Predicted Values" is actually the standard error for the fitted values called "sefit" in ALR[2.8.4].

Confidence intervals are obtained by checking `Mean` and `Individual` under the Prediction Intervals section of the Save dialog. You can also modify the level of the intervals by editing the percent level. After running the analysis, columns corresponding to the lower and upper bounds of each interval are added to the data table for each observation.

Both mean and individual prediction confidence intervals can be added to the regression scatterplot. First add the regression line to the plot, then select the line and right click. Choose Regression Parameters from the popup menu. In the dialog you can select one or both of the intervals as well as the confidence level.

SPSS does provide an easy way to calculate the predicted value and its standard error (and hence confidence interval) for a new value of the predictor. In the SPSS data editor, simply add new rows to the data sheet with the values x_* where you would like to do predictions. Leave the value of the response blank, and then SPSS will treat this as a missing value. Obtain predictions and standard errors as described above. Only fully observed cases are used in the calculations, but predictions and standard errors are computed for all rows for which the predictors are entered.

2.9 THE RESIDUALS

SPSS Details on how to draw residual plots were first given in Section 1.1 for the Forbes residual plot ALR[F1.3B]. Residual plots are drawn by saving the unstandardized residual, then plotting them against the independent variable.

The final note for this section is about deleting a case from the analysis, for example, case 12 was deleted from the Forbes' data. SPSS includes a function for *selecting cases*, not for deleting them, and so you must create a description that *excludes* the case(s) you want to delete. This can be done by choosing Data → Select Cases and checking `If condition is satisfied` and pressing the IF button. Then enter a condition to select everything except the case you wish to filter, which means you want to leave the case in the data, but not use it in fitting models. For example, to filter case 12 from the Forbes data enter

```
(Temp ~= 204.60) & (Lpres ~= 142.44)
```

Any analysis run will not include this filtered case. Deleting several cases can be very tedious.

Problems

2.2.

SPSS Problem 2.2.5. is possible to do in SPSS by a transformation of *Lpres* and the saved predicted values. The correct standard errors are found from the saved standard errors for the fitted values (sefit) by using the function $\text{sepred} = (\hat{\sigma}^2 + (\text{sefit})^2)^{1/2}$.

Problem 2.2.6 is not as simple. Using the estimated mean function from the Hooker data, you need to transform u_1 to find the predicted values for the seventeen cases in the Forbes data. The z-scores are a difficult transformation because SPSS does not provide an easy way to calculate the standard error of predicted values for these new seventeen cases. This can only be done by transforming the new predictors using ALR[E2.26].

2.7.

SPSS You can remove the intercept from the fitted mean function by selecting the button OPTIONS and unchecking `Include constant in equation`.

2.10.

SPSS Remember, to select the cases with *HamiltonRank* $\leq$ 50, choose Data → Select Cases and select cases according to the following condition:

```
HamiltonRank <= 50
```

3

Multiple Regression

3.1 ADDING A TERM TO A SIMPLE LINEAR REGRESSION MODEL

SPSS Added variable plots are easily obtained in SPSS by selecting the PLOTS button in the linear regression dialog window, then checking `Produce all partial plots`. When the linear regression model is fit, the plots will be added to the output and, though named differently, are the added variable plots. The dependent (response) variable will always be the label for the vertical axis while added-variable plot term is given on the horizontal axis.

3.2 THE MULTIPLE LINEAR REGRESSION MODEL

3.3 TERMS AND PREDICTORS

SPSS The summary statistics in ALR[3.3] for the data file `fuel2001.txt` can be easily obtained after transforming the predictors into the terms used for the multiple regression model. These transformations were done in Section 1.6 to draw the scatterplot matrix of terms, so please refer back to that section for details on obtaining the variables *Dlic*, *Income*, *logMiles*, and *Fuel*.

ALR[T3.1] can be formed by selecting Analyze → Tables → Custom Tables. This command gives you a wide variety of ways to make a summary table, but the basic table shown in ALR[3.3] can be made by dragging the five variables to the vertical "Rows" bar. As you do this you can see the table take shape with the mean as the only summary statistic. Additional statistics can be added by clicking on SUMMARY STATISTICS and dragging the desired statistic to the display box then pressing APPLY TO SECTION. You can format the table as you wish, and after pressing OK the table will be added to the `Viewer` window.

The correlation matrix ALR[T3.2] and the covariance matrix found in ALR[3.4.5] can be found as discussed in Section 2.2. Begin by selecting Analyze → Correlate → Bivariate and adding the five terms to the variable box. Clicking OK now would only produce a correlation table, to add the covariances press OPTIONS and check `Cross-product deviations and covariances`.

3.4 ORDINARY LEAST SQUARES

SPSS The sample covariance matrix for fuel data can be found as described in Section 3.3.

The OLS fit of a multiple linear regression model is best done using the built in commands in SPSS. It is possible to use the SPSS command language to calculate values like $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{X}'\mathbf{Y}$ used to obtain coefficient estimates and other summaries, but we will not provide the details.

A multiple regression model can be fit the same way that the simple linear model was fit in Section 2.6 using Analyze → Regression → Linear, but now you add all the terms to the Independent box in the linear regression dialog. As before, select the statistics and plots you would like to see or save. If you do not modify these settings, you will simply get the tables for model summary (R^2), analysis of variance and coefficient estimates. See Section 2.8 for details on confidence intervals and tests for the coefficients.

3.5 THE ANALYSIS OF VARIANCE

SPSS The analysis of variance table from the multiple regression analysis detailed in Section 3.4 corresponds to ALR[T3.4].

You have two options to compare models with or without one or more terms, as was investigated in ALR[3.5.2]. If the larger model was already fit, simply repeat the analysis for the smaller model.

If you have yet to fit any models you can obtain both model fits with the same analysis by specifying different *blocks* for analysis. In the linear regression dialog, after entering the dependent variable, in the independent box enter the terms for only the smaller model, *Dlic*, *Income*, and *logMiles*. This will form the first block of terms added to the model and by keeping the Method as `Enter` they will be fit at the same time. Click the button NEXT to move to the second block of terms and enter *Tax* into the now empty Independent box, keeping the method as `Enter`. Selecting OK will result in SPSS fitting two mean functions: (1) the regression of *Fuel* on only the first block of terms and (2) the regression of *Fuel* on both the first and second blocks of terms. For both mean functions, the output will be tables of model summary, *ANOVA*, and coefficient estimates. The sum of squares used in ALR[3.5.2] will be provided in the two *ANOVA* tables. A table for the "Excluded variables," corresponding to the terms in the second block, provide standardized coefficient estimate (which are not discussed in ALR) and t-statistic which would be obtained if the term were added to the previous block of terms. For this example, the only variable in this table is *Tax* and the significance value is about 0.043, the same as the level obtained in ALR[3.5.2] for the difference between excluding and including *Tax*.

The mean function is fit in the order the terms are entered into the independent variable box. You cannot obtain a sequential analysis of variance table exactly like the tables in ALR[T3.5] using the using the Analyze → Regression → Linear command. Should you want this table for some reason, you can obtain it using Analyze → General Linear Model → Univariate. Select the response as the dependent variable as usual, but now put the continuous terms in the Covariates list, and if you have any factors, they go in the Factor(s) list. Press the MODEL button, and at the bottom of the resulting dialog, in select Sum of squares → Type I. The SPSS default is Sum of squares → Type III; we recommend that you never use Type III. Press Continue and then press OK to get the *ANOVA* table.

3.6 PREDICTIONS AND FITTED VALUES

SPSS Predictions, fitted values, and the standard errors for fitted values for multiple regression can be obtained by saving them before running the regression fit, just as was done for simple linear regression in Section 2.8. Confidence intervals for the mean or individuals can also be saved.

Problems

4
Drawing Conclusions

The first three sections of this chapter do not introduce any new computational methods; everything you need is based on what has been covered in previous chapters. The last two sections, on missing data and on computationally intensive methods introduce new computing issues.

4.1 UNDERSTANDING PARAMETER ESTIMATES

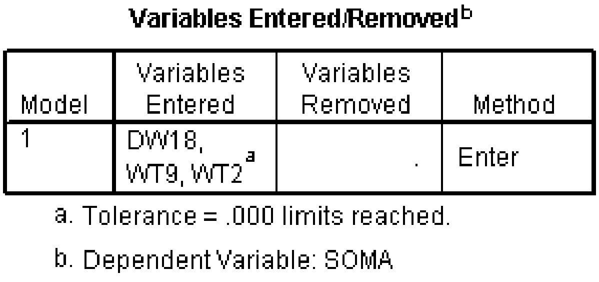
4.1.1 Rate of change

4.1.2 Sign of estimates

4.1.3 Interpretation depends on other terms in the mean function

4.1.4 Rank deficient and over-parameterized models

SPSS Over-parameterized models are recognized in SPSS by checking each variable's *tolerance* level as it is added the the model. Small tolerance values mean the variable contributes little information to the model, possibly due to collinearity with other variables already in the model. A variable will not be entered into the regression model if its tolerance is below 0.0001 or if adding it will cause the tolerance of variables already in the model to drop below 0.0001. The first table presented in the regression output will tell you if variables have been excluded because of this limit. Figure 4.1 shows an example of this table where the superscript "a" denotes that the tolerance limit has been reached



Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	DW18, WT9, WT2[a]	.	Enter

a. Tolerance = .000 limits reached.

b. Dependent Variable: SOMA

Fig. 4.1 This SPSS table gives the variables fit in a regression model. The superscript "a" tells that some regression terms have not been added due to collinearity.

for some variable (presumably, "a" is short for *aliased.* When this occurs, an Excluded Variable table will tell you the tolerance level of the variables not included in the model.

Consider the Berkeley Guidance Study example from ALR[4.1.3]. The variables *DW9* and *DW18* are linear combinations of the terms *WT2*, *WT9*, and *WT18*. If we entered the terms in this order in the Independent variables box, we would expect, according to the discussion in ALR[4.1.4] that the last two terms in the list, *WT9* and *WT18*, would be marked as aliased, and if we change the order, then different terms would be marked as aliased. This is in fact *not* the case in SPSS, as it seems to use some other algorithm for fitting terms, and seems to report the same terms as aliased for any order. While there may be a computational argument in favor of this approach, it seems to be a poor idea based on statistical ideas. You can get answers similar to those in ALR if you use Analyze → General Linear Model → Univariate, with Type I sums of squares, to do the fitting; see Section 3.5.

4.2 EXPERIMENTATION VERSUS OBSERVATION

4.3 SAMPLING FROM A NORMAL POPULATION

4.4 MORE ON R^2

4.5 MISSING DATA

The data files that are included with ALR use "NA" as a place holder for missing values. Some packages may not recognize this as a missing value indicator, and so you may need to change this character using an editor to the appropriate character for your program.

SPSS A text file containing missing values denoted by "?" (or "NA") can be imported into SPSS the usual way discussed in Section 0.2. The missing value indicator used in the SPSS data editor is a period, "."

You can get a total of the number of missing values for any variable by building a custom table, as done in Section 3.3, and including the statistic "Missing". Another option for analyzing how values are missing in an incomplete data set is to use the procedure Analyze → Missing Value Analysis. This option allows you to analyze the pattern of the missing data and estimate standard statistics for a list of variables using only complete cases from the list. This procedure also offers regression or EM methods for filling in missing values. Details about this procedure can be found by opening its dialog window and pressing the Help button.

A regression model will not use any cases which have missing values in one or more of the variables in the model. This is called "listwise exclusion" and there are other exclusion options available which can be found by clicking the OPTIONS button in the regression dialog. The listwise exclusion method will be the default option checked and is the appropriate method to use for the regression done in ALR; the other methods available in this dialog require strong assumptions to be useful, and they should be generally avoided.

If you would like to compare two models you may run into problems if you fit each model separately in SPSS. Consider the data file `sleep1.txt`. The regression fit of *SWS* on *BodyWt*, *Life*, *GP* will be based on the 42 complete cases of this list of variables, while a separate fit of *SWS* on *BodyWt* and *GP* will be based on the 44 complete cases of this smaller list of variables. You can't then compare these two fits with analysis of variance because they are based on different cases. The solution to this problem is to compare the two models with the second method described in Section 3.5. Recall that this method puts the smaller model predictors in the first block of terms and adds the remaining predictors to the second block of terms, then runs the regression. Both fits for this regression will then be based on the 42 complete cases for the whole list of variables.

4.6 COMPUTATIONALLY INTENSIVE METHODS

SPSS Computation of the bootstrap or other computationally intensive methods are possible with SPSS, but require using the SPSS programming language. We have not worked out how to do this with SPSS, but would be glad to see the scripts developed by others for this purpose. See also the on-line help for the entry "bootstrapping" in SPSS, Version 12.

5
Weights, Lack of Fit, and More

5.1 WEIGHTED LEAST SQUARES

SPSS Weighted least squares in SPSS works as suggested in ALR by specifying a variable in the data to be the weights in Analyze → Regression → Linear. In the physics data from ALR[5.1], define the weights w by the transformation

```
1/SD**2
```

WLS is computed by placing w into the WLS Weight box in the linear regression dialog, then proceeding with the usual OLS steps. The prediction intervals you obtain from the Save step (see Section 2.8) will use the correct standard error, $(\hat{\sigma}^2/w_i + \text{sefit}(y \mid X = \mathbf{x}_i))^{1/2}$, for each case in the data. Unlike OLS, you can't use the trick of adding an additional case (and case weight) to the data to get predictions for a new case.

SPSS includes several kinds of residuals that can be saved. The Unstandardized residuals, are $y - \hat{y}$, which are useful for OLS, but not WLS. SPSS has a second type of residual called a Standardized residual that it incorrectly says is equivalent to the Pearson residuals that will be discussed later in ALR; the formula that SPSS uses is incorrect for WLS. The correct residuals to use with WLS are defined by $\sqrt{w}(y - \hat{y})$. To get these residuals, you must save the Unstandardized residuals, and then do the multiplication yourself using a transformation.

The other types of residuals available in SPSS will be discussed in Chapters 8–9. These are correctly computed for both OLS and WLS.

5.1.1 Applications of weighted least squares

5.1.2 Additional comments

5.2 TESTING FOR LACK OF FIT, VARIANCE KNOWN

SPSS Any polynomial regression can be fit in SPSS by first transforming existing variables to obtain their exponential forms used in the mean function, then using them in the standard regression procedure. For the physics data this involves defining *x2* with the transformation `x**2`. The WLS fit will use the same weights defined in Section 5.1 with the independent variables x and *x2*. This will fit the quadratic mean function ALR[E5.12] via WLS and give the summary, analysis of variance, and coefficient tables seen in ALR[T5.3].

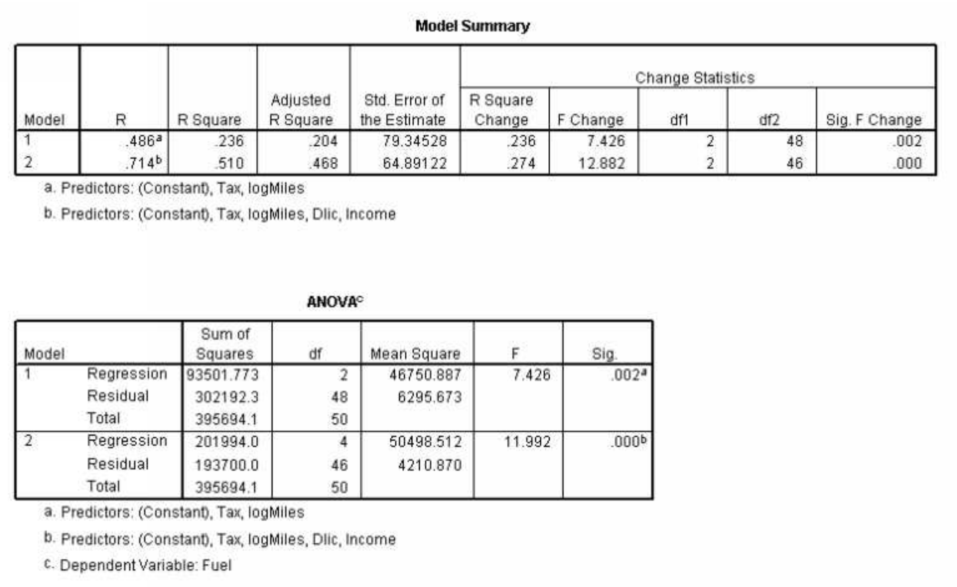
The scatterplot in ALR[F5.1] is difficult to draw in SPSS because the linear and quadratic regression fits are from WLS. For the OLS fit of the linear and quadratic mean functions, the procedure Analyze → Regression → Curve Estimation can be used to fit and plot two regression lines on a scatterplot of the data. Other estimated mean functions from an OLS fit of data can also be added. For more details on the kind of fits available, use the help button in the procedure's dialog.

5.3 TESTING FOR LACK OF FIT, VARIANCE UNKNOWN

SPSS The test for lack-of-fit is done with a generalized linear model (GLM) procedure, rather than the regression procedure in SPSS. To fit the data in ALR[T5.4], first enter the variable names with the Variable View tab of the Data Editor, then add the data values with the Data View tab. Select Analyze → General Linear Model → Univariate and add y as the dependent variable and x as the Covariate. Press OPTIONS and check `Lack of fit` and click continue. After selecting OK, a lack-of-fit table will be produced, which for this example is

```
Lack of Fit Tests
Dependent Variable: y
Source  Sum of Squares  df  Mean Square     F       Sig.
Lack of Fit     1.858   2       .929        2.364   .175
Pure Error      2.358   6       .393
```

A second analysis of variance table is provided which contains sums of squares for the GLM, and from this you can obtain the sum of squares for regression and residuals given in ALR[T5.5]. The F-value for the regression term x in this SPSS table is not the one given in ALR[T5.5]. This value, $F = 11.62$, must be calculated by hand.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.486[a]	.236	.204	79.34528	.236	7.426	2	48	.002
2	.714[b]	.510	.468	64.89122	.274	12.882	2	46	.000

a. Predictors: (Constant), Tax, logMiles

b. Predictors: (Constant), Tax, logMiles, Dlic, Income

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	93501.773	2	46750.887	7.426	.002[a]
	Residual	302192.3	48	6295.673		
	Total	395694.1	50			
2	Regression	201994.0	4	50498.512	11.992	.000[b]
	Residual	193700.0	46	4210.870		
	Total	395694.1	50			

a. Predictors: (Constant), Tax, logMiles

b. Predictors: (Constant), Tax, logMiles, Dlic, Income

c. Dependent Variable: Fuel

Fig. 5.1 Two output tables for testing the hypothesis in Section 5.4.

5.4 GENERAL F TESTING

SPSS The general F-tests described in ALR[5.4] can be calculated by entering terms into separate blocks; see Section 3.5.

Consider the fuel consumption data again. After the appropriate transformations, enter the response *Fuel* in the dependent variable box. Enter the terms in common for both the small and large model in the first Independent variable Block, click NEXT and then enter the terms found only in the larger model. To obtain the F-value in ALR[E5.16] and its p-value, select STATISTICS and check `R squared change`. This statistic is the model 2 “F change” in the Model Summary report.

For example, suppose we wish to test

$$\begin{aligned} \text{NH:}\quad & E(\mathit{Fuel}\,|\,X) = \beta_0 + \beta_1 \mathit{logMiles} + \beta_2 \mathit{Tax} \\ \text{AH:}\quad & E(\mathit{Fuel}\,|\,X) = \beta_0 + \beta_1 \mathit{logMiles} + \beta_2 \mathit{Tax} + \beta_3 \mathit{Dlic} + \beta_4 \mathit{Income} \end{aligned}$$

Block 1 in the regression dialog will contain the terms *logMiles* and *Tax* while block 2 will contain *Dlic* and *Income*. The model summary and analysis of variance tables which result from this fit are given in Figure 5.1.

In the model summary table, the “F Change” of model 2 equals 12.882 and is the F-statistic for the hypothesis. The “Sig. F Change” is the p-value for this statistic. The ANOVA table gives the analysis of variance for fitting the null and alternative models.

5.5 JOINT CONFIDENCE REGIONS

SPSS Confidence regions for parameter estimates are not available in SPSS.

Problems

5.3.

The bootstrap used in this problem is different from the bootstrap discussed in ALR[4.6] because rather than resampling *cases* we are resampling *residuals*. Here is the general outline of the method:

1. Fit the model of interest to the data. In this case, the model is just the simple linear regression of the response y on the predictor x. Compute the test statistic of interest, given by ALR[E5.23]. Save the fitted values $\hat{y}$ and the residuals $\hat{e}$.

2. A bootstrap sample will consist of the original x and a new y^*, where $y^* = \hat{y} + e^*$. The ith element of e^* is obtained by sampling from $\hat{e}$ with replacement.

3. Given the bootstrap data (x, y^*), compute and save ALR[E5.23].

4. Repeat steps 2 and 3 B times, and summarize results.

6
Polynomials and Factors

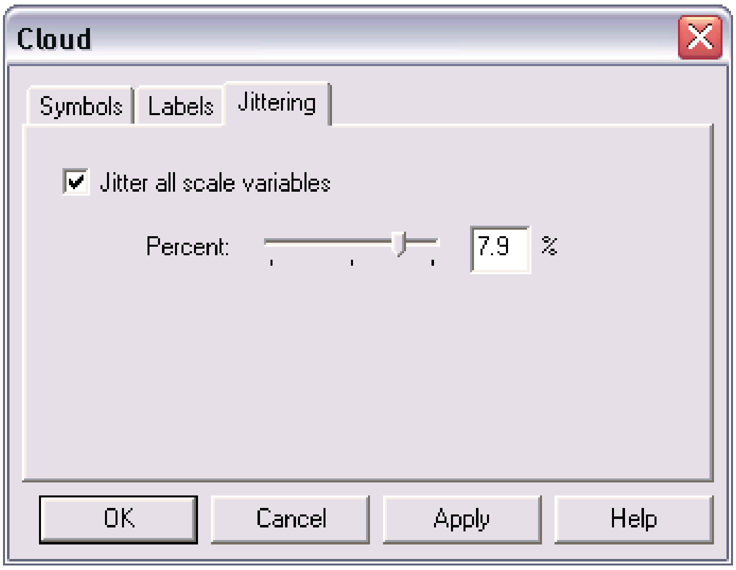
6.1 POLYNOMIAL REGRESSION

SPSS ALR[F6.2] is a plot of the design points in the cakes data, with the center points slightly jittered to avoid overprinting. SPSS allows jittering of a scatterplot when the plotted variables are of type scale. After importing `cakes.txt`, change the measure of *X1* and *X2* to scale, then use Graphs → Interactive → Scatterplot to draw the interactive scatterplot of *X2* versus *X1* (if you have forgotton how to do this, look at Section 1.1). With the Chart Manager, or Format → Graph elements → Cloud, select the Cloud element and press EDIT. In the Jitter tab of the cloud dialog, check the box shown in Figure 6.1. This will activate the slide bar that controls the amount of jittering added to the plot. As you change this percent, press APPLY to see the effect it has on the scatterplot.

SPSS does not seem to have any special tools for working with polynomial mean functions with more than one predictor. Polynomial models generally require creating many terms that are functions of a few base predictors. As discussed in Section 5.2, these higher-order terms must be defined via a transformation, then the polynomial mean function can be fit with the linear regression procedure.

6.1.1 Polynomials with several predictors

SPSS To fit the second-order mean function in ALR[E6.4] begin by transforming the predictors *X1* and *X2* to obtain the higher-order terms. The variables we defined were



Cloud

☑ Jitter all scale variables

Percent: 7.9 %

Fig. 6.1 The Cloud dialog from the Chart Manager allows jittering of scale variables by adding a user specified amount of random noise to the data points.

```
X1.2 = X1 ** 2
X2.2 = X2 ** 2
X1X2 = X1 * X2
```

where the first two terms give the quadratic values of the individual predictors and the last term defines the interaction between *X1* and *X2*. In the linear regression dialog, enter Y as the dependent variable and *X1*, *X2*, *X1.2*, *X2.2*, and *X1X2* as the independent variables. When the regression model is fit, the fitted mean function is the same as ALR[E6.7].

The plots shown in ALR[F6.3] are hard to obtain in SPSS without writing a program for them. The following (tedious) technique will produce similar plots using the fitted mean equation in ALR[E6.7]. To make ALR[F6.3A], define a new variable *X2cat* which had fifty values equal to 340, fifty values equal to 350, and fifty values equal to 360. To make this variable define the variable name with the Variable View tab, then enter 340 in the first row of *X2cat* and copy the cell value. Next, highlight cells 2 through 50 in this column, right click and copy the value 340 into all the highlighted cells. Enter the fifty values of 350 and 360 similarly. This variable will be used as the value of *X2* in the three fitted mean equations corresponding to the three curves in ALR[F6.3A]. It will also be used as the legend variable when plotting the fitted values and *X1*.

Next we need values of *X1* between 32 and 38 which can be used to fit the mean function for the fixed values of *X2* defined in *X2cat*. We obtained these

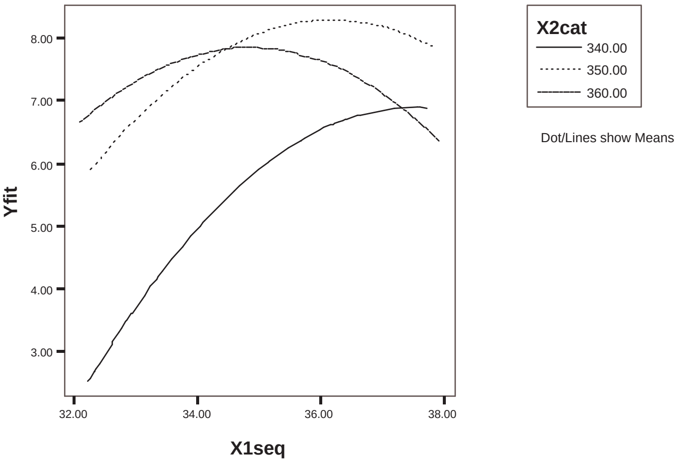


Fig. 6.2 The SPSS version of ALR[F6.3A].

values by using the transformation function `UNIFORM(max)` which generates uniform random number between 0 and `max`. By defining the transformation

```
X1seq = 32 + UNIFORM(4)
```

the variable *X1seq* will contain 150 random values between 32 and 38.

The fitted values, *Yfit*, for each pair in *X1seq* and *X2cat* can be found using ALR[E6.7] as the transformation function:

```
Yfit = -2204.485 + 25.9176 * X1seq + 9.9183 * X2cat + -0.156875 *
       X1seq ** 2 + -0.01195 * X2cat ** 2 + -0.041625 * X1seq * X2cat
```

It is important to keep as many significant digits as possible, so we suggest you activate the Coefficient pivot table and select and copy the coefficient estimates from each cell, then paste the values in the appropriate spot in the transformation expression box.

Finally, select Graphs → Interactive → Line from a menu. Place *Yfit* on the vertical axis, *X1seq* on the horizontal axis, and *X2cat* as the Style legend variable. Since the legend variable must be categorical, you will be reminded of this and given the option of converting *X2cat*, select CONVERT to do so. Press OK and the resulting plot should be similar to Figure 6.2. ALR[F6.3B] can be obtained in a similar manner, though to define the random sequence of *X2* values between 335 and 365 use the transformation `335 + UNIFORM(30)`.

6.1.2 Using the delta method to estimate a minimum or a maximum

6.1.3 Fractional polynomials

6.2 FACTORS

Factors are a slippery topic because different computer programs will handle them in different ways. In particular, while SAS and SPSS use the same default for defining factors, JMP, R and S-Plus all used different defaults. A factor represents a qualitative variable with say a levels by $a-1$ (or, if no intercept is in the model, possibly a) dummy variables. ALR[E6.16] describes one method for defining the dummy variables, using the following rules:

1. If a factor A has a levels, create a dummy variables $U_1, \ldots, U_a$, such that U_j has the value one when the level of A is j, and value zero everywhere else.

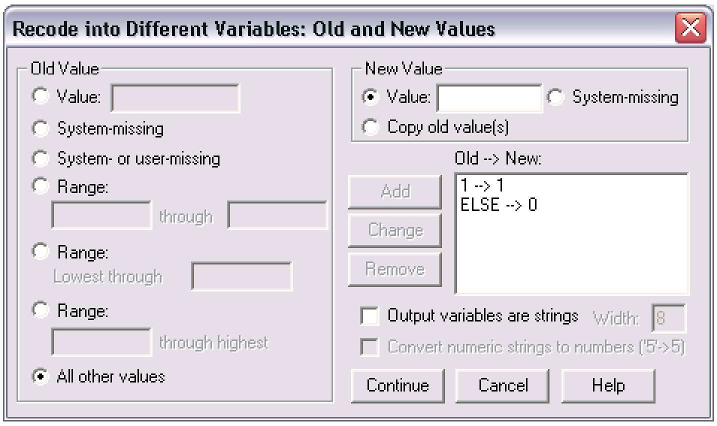
2. Obtain a set of $a-1$ dummy variables to represent factor A by dropping one of the dummy variables. For example, using the default coding in R, the first dummy variable U_1 is dropped, while in SAS and SPSS the last dummy variable is dropped.

3. JMP and S-Plus use a completely different method.

Most of the discussion in ALR assumes the R default for defining dummy variables.

SPSS SPSS will not recognize factors when using Analyze → Regression → Linear. There are two options for fitting a regression model with factors: create dummy variables for the factor and fit the linear regression, or use Analyze → General Linear Model → Univariate which will correctly recognize the factor. Using dummy variables is OK for problems with few factors and few interactions, but otherwise this can be very tedious. However, the range of options available in the regression procedure, such as added-variable plots, and predictors, is reduced with the GLM procedure. You will probably want to learn to use both methods.

Creating a set of dummy variables from a factor is straightforward, but it is tedious using the SPSS graphical user interface if the factor has several levels. For illustration, consider the factor D from the sleep data. To create the dummy variables $U_1, \ldots, U_5$ defined in ALR[E6.14] select Transform → Recode → Into Different Variables. To define U_1, select D and add it to the Input/Output variable box in the Recode dialog. Next, give the output variable the name `U1` and press CHANGE then select OLD AND NEW VALUES. The dialog which is produced by this button in shown in Figure 6.3. First, set the Old Value equal to 1 and the New Value also equal to 1, then press ADD. Since U_1 equals zero for all other values of D, next press `All other values` under Old Value and set the New Value equal to 0, then press ADD. Figure 6.3



Recode into Different Variables: Old and New Values

Old Value

○ Value:
○ System-missing
○ System- or user-missing
○ Range: through
○ Range: Lowest through
○ Range: through highest
● All other values

New Value

● Value:
○ System-missing
○ Copy old value(s)

Old --> New:

```
1 --> 1
ELSE --> 0
```

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

Fig. 6.3 The dialog for specifying a new variable using the Recode procedure.

shows what the dialog should look like after these steps. Press CONTINUE to return to the Recode dialog and press OK to add U_1 to the data table. Repeat these steps to create the four other dummy variables, changing only the value of D which gets assigned the value one in the new variable.

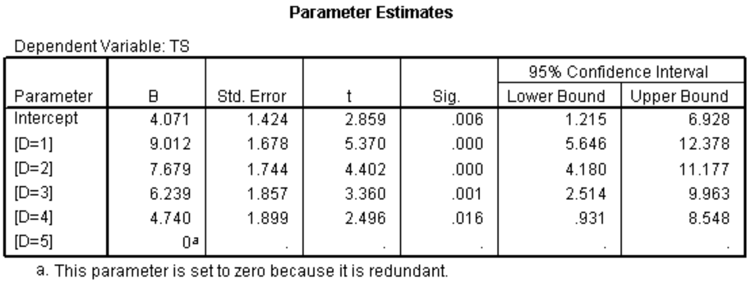
To make D into a factor, use the Variable View tab of the Data Editor, and change the Method for D to either nominal or ordinal.

6.2.1 No other predictors

SPSS Using the dummy variables for D defined above, ALR[T6.1A] can be obtained in SPSS by fitting the linear regression through the origin of TS on $U_1, \ldots, U_5$. Recall that the intercept can be removed from the fit using the OPTIONS button in the linear regression dialog. ALR[T6.1B] is obtained by including the intercept in the linear fit, though depending on the ordering of the dummy variables, SPSS may not remove the indicator removed in ALR[T6.1B].

Both tables can also be produced using the procedure Analyze → General Linear Model → Univariate, which provides regression analysis for continuous and categorical terms. In the dialog for this procedure, continuous terms are called Covariates. For sleep data, enter TS as the Dependent Variable and D as the Fixed Factor. The dummy variables for D are not needed for this procedure. Next, press OPTIONS and check the display option `Parameter estimates` and press Continue. To obtain an ALR[T6.1B] press OK after this step.

The parameter estimates which SPSS gives for this fit are in Figure 6.4. These estimates are different than those given in ALR[T6.1B] because SPSS drops the last level rather than the first level. To obtain ALR[T6.1B], remove



Parameter Estimates

Dependent Variable: TS

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.071	1.424	2.859	.006	1.215	6.928
[D=1]	9.012	1.678	5.370	.000	5.646	12.378
[D=2]	7.679	1.744	4.402	.000	4.180	11.177
[D=3]	6.239	1.857	3.360	.001	2.514	9.963
[D=4]	4.740	1.899	2.496	.016	.931	8.548
[D=5]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Fig. 6.4 The parameter estimates obtained for the GLM fit in SPSS of ALR[E6.16] for the sleep data.

the intercept from the GLM fit by selecting MODEL from the GLM dialog and uncheck the intercept option.

The Contrast button on the GLM dialog allows you to select a different way to define the contrasts for factors. This option is complex and not completely intuitive, since it does *not* change the parameterization for the factor, but it *does* provide tests and estimates as if the parameterization were changed; see the SPSS documentation if you think this option might be useful to you.

6.2.2 Adding a predictor: Comparing regression lines

SPSS To obtain the model fits from ALR[6.2.2] use the GLM procedure and the predictors *D* and *logBW*, the log transformation of *Body Wt.* Each model can be obtained by specifying a different combination of main effect and interactions using the MODEL button in the GLM dialog.

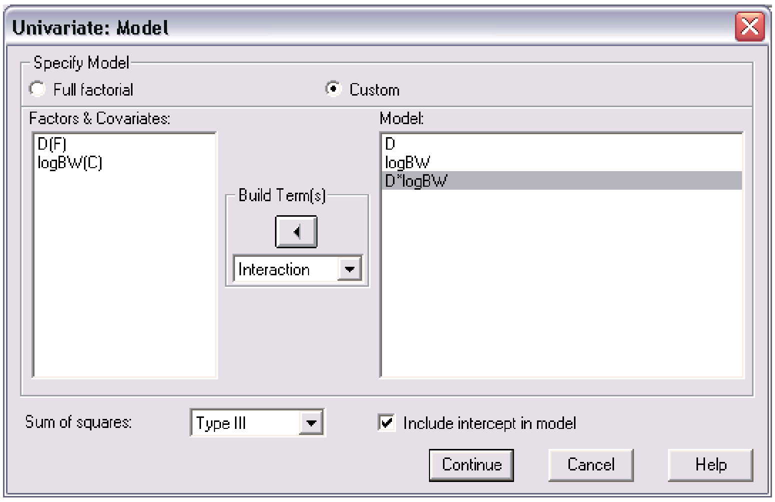
Model 1 Use the model terms `D`, `logBW`, `D*logBW`. In the model dialog, shown in Figure 6.5, check `Custom`, then highlight *D* and *logBW* and select “Main effects” from the build terms list. Use the arrow button to add them to the model. Highlight the terms again, and choose “Interaction” from the list to add `D*logBW` to the model. Press continue to return to the GLM dialog and press OK to fit the model.

Model 2 Use the model terms `D` and `logBW`.

Model 3 Use the model terms `logBW` and `D*logBW`.

Model 4 Use the model term `logBW`. This can also be fit using the linear regression procedure.

ALR[F6.6A] is made by drawing an interactive scatterplot of *TS* and *logBW* with *D* as the legend color or style variable. The five lines regression lines can be added by adding a regression line “Fit for Subgroups”. To obtain ALR[F6.6D], follow the same steps but choose “Fit for Total”. The remaining



Specify Model
○ Full factorial ● Custom

Factors & Covariates:
D(F)
logBW(C)

Model:
D
logBW
D*logBW

Interaction

Sum of squares: Type III

☑ Include intercept in model

Fig. 6.5 The GLM Model dialog for fitting model 1 in alr6.2.2.

two plots, ALR[F6.6BC], are not easy to draw in SPSS so we will not discuss how to obtain them.

6.3 MANY FACTORS

6.4 PARTIAL ONE-DIMENSIONAL MEAN FUNCTIONS

ALR[F6.8] is much more compelling in color, and is shown here as Figure 6.6.

SPSS The partial one-dimensional mean function is fit in SPSS using the nonlinear regression procedure. Select Analyze → Regression → Nonlinear. The dialog which appears, shown in Figure 6.7, is similar to the transformation dialog. This dialog allows you to specify the nonlinear mean function, which, for the Australian Institute of Sport data, is given in ALR[E6.26]. To define this function, first create the six parameters it contains by pressing the PARAMETERS button and defining `b0`, . . . , `b4` and `eta1`, and assigning one to each starting value. Enter *LBM* as the dependent variable, then use the predictors and the newly defined parameters to define the Model Expression. Figure 6.7 shows the correct expression of ALR[E6.26]. Press OK to run the model.

The output for this procedure includes details on the iterations needed for convergence, analysis of variance, and parameter estimates and correlations. The ANOVA and estimates for this example are

```
Nonlinear Regression Summary Statistics     Dependent Variable LBM

  Source                 DF   Sum of Squares   Mean Square
```

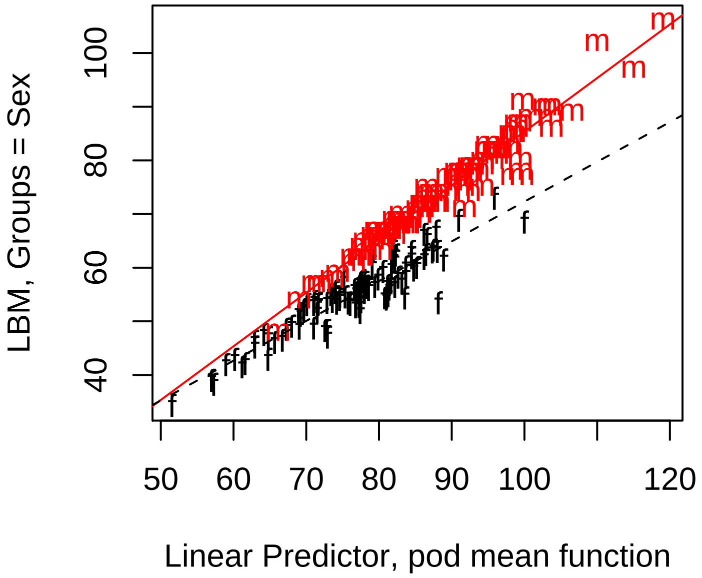


Fig. 6.6 ALR[F6.8] in color.

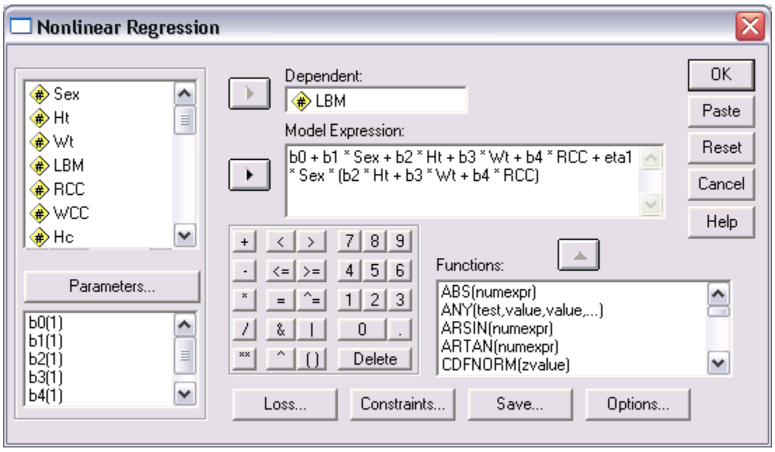
```
Regression              6   883287.85162   147214.64194
Residual              196     1185.91108        6.05057
Uncorrected Total     202   884473.76270

(Corrected Total)     201    34336.84112

R squared = 1 - Residual SS / Corrected SS =     .96546

                                         Asymptotic 95 %
                        Asymptotic     Confidence Interval
Parameter   Estimate    Std. Error     Lower         Upper

b0        -14.65640475  6.464485340 -27.40528276 -1.907526732
b1        12.847199164  3.763419978  5.425203491 20.269194837
b2          .146263801   .034243613   .078730561   .213797042
b3          .709342087   .024163903   .661687458   .756996716
b4          .724760698   .585401803  -.429734328  1.879255725
```



Nonlinear Regression

Dependent: LBM

Model Expression: b0 + b1 * Sex + b2 * Ht + b3 * Wt + b4 * RCC + eta1 * Sex * (b2 * Ht + b3 * Wt + b4 * RCC)

Parameters...

b0(1)
b1(1)
b2(1)
b3(1)
b4(1)

Fig. 6.7 Nonlinear regression dialog for defining the partial one-dimensional mean function ALR[E6.26].

```
eta1        -.258749127   .034463628   -.326716269   -.190781986
```

The only way to draw ALR[6.8] in SPSS is to define the linear transformation of *Ht*, *Wt*, and *RCC* using the estimates of β_2, β_3, and β_4. Make a scatterplot of this variable and *LBM*, choosing *Sex* as the legend variable. Under the Fit tab, select a regression line to add, but choose to fit by `Subgroups` instead of `Total`. Press OK and the plot drawn will have separate regression lines for each sex.

6.5 RANDOM COEFFICIENT MODELS

SPSS Random coefficient models can in principle be fit in SPSS using Analyze → Mixed models → Linear, but we were unable to get the procedure to work.

7 Transformations

7.1 TRANSFORMATIONS AND SCATTERPLOTS

7.1.1 Power transformations

7.1.2 Transforming only the predictor variable

SPSS A plot similar to ALR[F7.3] can be obtain in SPSS using the Curve Estimation procedure. Using the Upper Flat Creek data in `ufcwc`, select Analyze → Regression → Curve Estimation and enter *Height* as the dependent variable and *Dbh* as the independent variable. Check the boxes for `Linear`, `Logarithmic`, and `Inverse` to obtain fits for the power transformations $\lambda = 1, 0$, and -1; other powers are not available, but these are the three most important choices. Check the box for `Display ANOVA Table` to obtain the *RSS* values for each fit and click OK.

The plot is shown in Figure 7.1. A printed summary of each of the regressions includes the *RSS*, which is smallest for the log transformation.

7.1.3 Transforming the response only

SPSS The method described ALR[7.1.3] requires the steps: (1) fit the model with the response untransformed, and predictors transformed; (2) draw the inverse plot with fitted values on the horizontal axis, and the untransformed response on the vertical axis; (3) estimate the transformation from among the inverse, logarithmic, and untransformed, as outlined in Section 7.1.2. Use the Regression → Curve Estimation procedure to visually estimate the best trans-

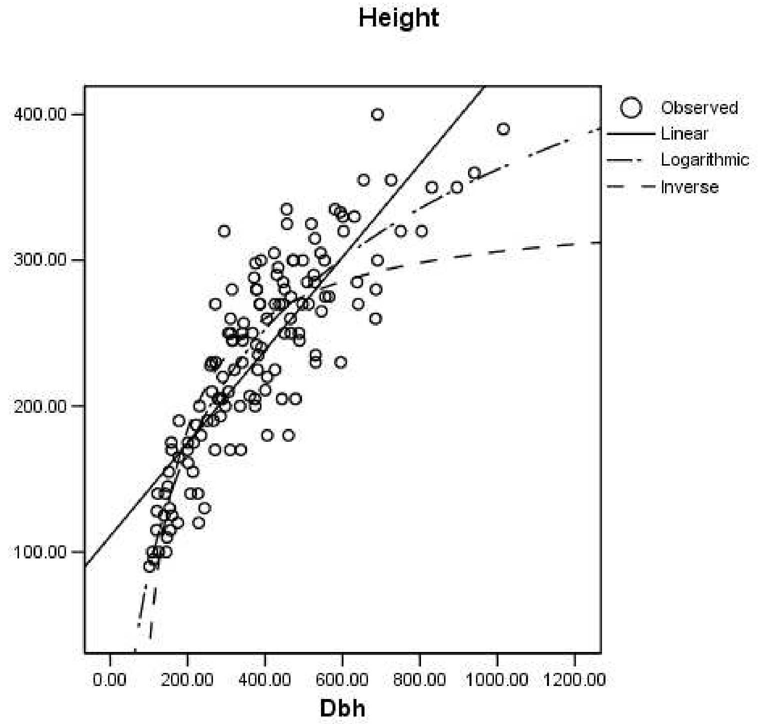


Fig. 7.1 SPSS's version of ALR[F7.3].

formation of a predictor because SPSS does not provide a scaled power transformation procedure. We will follow the same steps to transform the response, but now we will consider the response variable as the predictor and the fitted mean values as the response.

To get the fitted values, fit the regression of *Rate* on the transformed predictors log(*Len*), log(*ADT*), log(*Trks*), *Slim*, *Shld* and *logSigs1*, making sure to check the Predicted Values option `Unstandardized` from the SAVE dialog. The predictor transformations were determined by the multivariate method described in ALR[7.2.2], where the terms log(*Len*), log(*ADT*), and log(*Trks*) are the log transformations of the appropriate variable, and *logSigs1* is equal to the function $logSigs1 = \log((Len \times Sigs + 1)/Len)$.

If *Pred1* is the column name of the saved predicted values, fit the linear, logarithmic, and inverse regressions of *Pred1* on *Rate* using the Regression → Curve Estimation procedure, checking the option `Display ANOVA Table` to get the *RSS* values of each fit. The inverse response plot with the three fitted lines is given in Figure 7.2. The *RSS* for the inverse, log, and linear fits are, respectively, 34.72, 30.73, and 32.46. From the inverse response plot

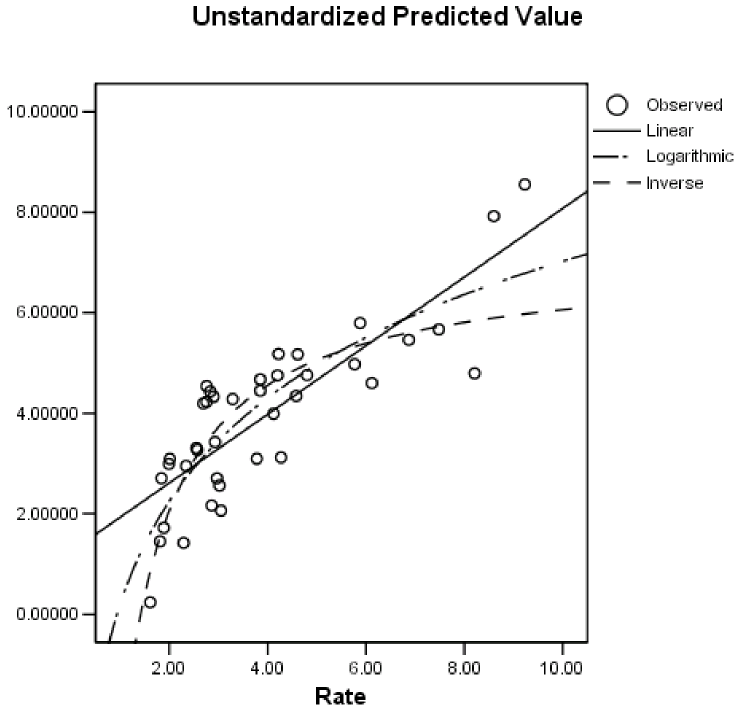


Fig. 7.2 Transforming the response in the highway data.

and the *RSS* values, we can conclude that the log transformation is the best choice of transformation.

7.1.4 The Box and Cox method

SPSS SPSS does not provide the Box-Cox method for transforming the response for normality. A useful project for students would be to write an SPSS program that will fit the Box-Cox method.

Lacking this procedure suggest using the method described in Section 7.1.3 to transform the response for linearity.

7.2 TRANSFORMATIONS AND SCATTERPLOT MATRICES

The scatterplot matrix is the central graphical object in learning about regression models. You should draw them all the time; all problems with many continuous predictors should start with one.

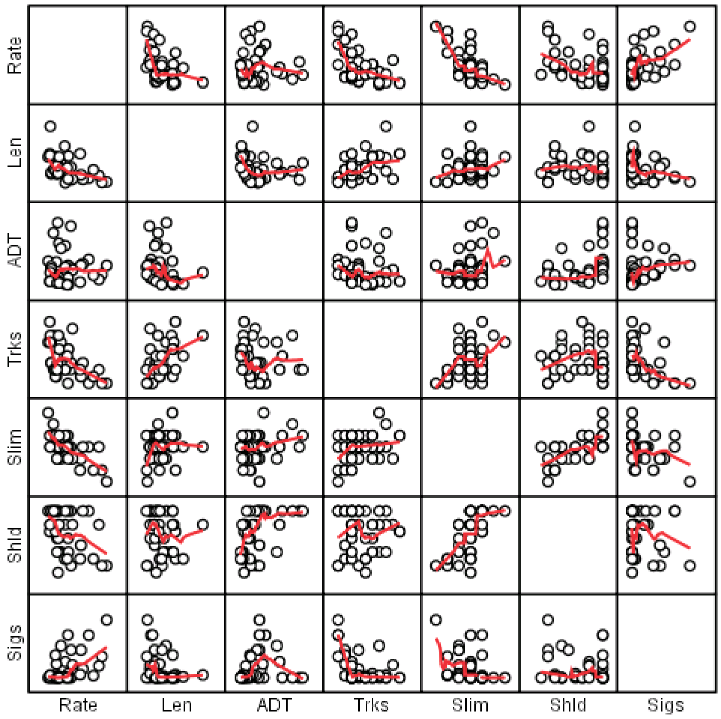


Fig. 7.3 The SPSS version of ALR[F7.5] with loess curves.

SPSS The scatterplot matrix in ALR[F7.5] can be duplicated in SPSS using the standard scatterplot matrix described in Section 1.6. Select Graphs → Scatterplots and choose the `Matrix` plot type. Enter the variables used in ALR[F7.5]. You can change the plotting symbol color by adding a variable to the Set Markers by option of this dialog. For the highway data, it may be useful to color the symbols according to the value of the variable *Hwy*. Press OK and the matrix will be made. Regression lines and loess curves can be added to each plot by double-clicking on the matrix. In the Chart Editor, highlight the data cloud and select Chart → Add Chart Element → Fit Line at Total. Figure 7.3 shows this scatterplot with loess curves added and colored red.

7.2.1 The 1D estimation result and linearly related predictors

7.2.2 Automatic choice of transformation of the predictors

SPSS SPSS does not have a multivariate extension to the Box and Cox method that can be used to automatically transformation multiple predictors. To handle data with many predictors, we suggest first viewing the scatterplot matrix and making appropriate transformations using the log and range rules discussed in ALR[7.1.1]. If predictors in the scatterplot matrix still look nonlinear after applying these rules, try finding transformations using individual scatterplots, as done in Section 7.1.2. Once the predictors are adequately transformed, use the method from Section 7.1.3 to determine if a transformation of the response is needed.

7.3 TRANSFORMING THE RESPONSE

SPSS See Section 7.1.3 above for the examples in this section.

7.4 TRANSFORMATIONS OF NON-POSITIVE VARIABLES

SPSS SPSS does not provide a Yeo-Johnson method for transforming non-positive variables.

8

Regression Diagnostics: Residuals

8.1 THE RESIDUALS

SPSS SPSS has two types of fitted values and five types of residuals, as defined in Table 8.1. Although SPSS (2003) refers to the residuals saved by the `Standardized` option as *Pearson* residuals, they are not equal to ALR[E8.13] and are *not* the same as ALR's Pearson residuals.

8.1.1 Difference between $\hat{\mathbf{e}}$ and $\mathbf{e}$

8.1.2 The hat matrix

SPSS "Centered" leverages, equal to $h_{ii} - 1/n$, can be saved by checking `Leverages values` in the Save dialog, see Table 8.1. If the intercept is not included in the mean function, then the leverages are labelled as centered, but they are, in fact, not centered.

8.1.3 Residuals and the hat matrix with weights

As pointed out in ALR[8.1.3], the residuals for WLS are $\sqrt{w_i} \times (y_i - \hat{y}_i)$. Whatever computer program you are using, you need to check to see how residuals are defined.

SPSS You need to compute these residuals yourself by first saving the unstandardized residuals, and then using a transformation to multiply them by

Table 8.1 Values available in the SAVE dialog for SPSS linear regression. Only SAVE options discussed in ALR are listed.

SPSS	ALR
Predicted Values	
`Unstandardized`	ALR[E8.2], fitted values for OLS or WLS.
`Adjusted`	The values $\hat{\mathbf{Y}}_{(i)}$ used in ALR[E9.7].
Residuals	
`Unstandardized`	ALR[E8.4], the usual residuals for OLS.
`Standardized`	These residuals are not discussed in ALR. They are equal to ALR[E8.4] divided by the estimated OLS standard deviation, $\hat{\mathbf{e}}/\hat{\sigma}$.
`Studentized`	ALR[E9.3], often called standardized residuals.
`Deleted`	The *PRESS* residuals, $y_i - \hat{y}_{(i)}$, used in ALR[E10.10].
`Studentized Deleted`	ALR[E9.4], Studentized residuals.
Distances	
`Cook's`	ALR[E9.6], Cook's distance.
`Leverage Values`	"Centered" leverages equal to $h_{ii} - 1/n$, where h_{ii}, ALR[E8.11], is the ith diagonal of the hat matrix.
Influence Statistics	
`DfBeta(s)`	The difference between parameter estimates defined by ALR[E3.9] and ALR[E9.5], $\hat{\beta} - \hat{\beta}_{(i)}$. Values are computed for all coefficients, including the intercept.
`DfFit`	The differenct between fitted values in ALR[E8.2] and $\mathbf{Y}_{(i)}$.

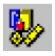
the square root of the weights. SPSS apparently returns missing values for the standardized residuals when weights are present. However, the Studentized and Studentized Deleted residuals, using the SPSS nomenclature, are correctly computed with weights present.

8.1.4 The residuals when the model is correct

8.1.5 The residuals when the model is not correct

8.1.6 Fuel consumption data

SPSS The plots in ALR[F8.5] must be made separately in SPSS by saving the residuals and fitted values, then plotting them with the appropriate variable. When using an interactive scatterplot, individual points can be labelled by creating the usual plot and activating it with a double-click. Once the plot is activated, any point can be identified with its case number by right-clicking on it and selecting Symbol Label. If you would like another identifier, first click the following icon:



In the resulting dialog, select the Cases tab then drag the identifier variable to the `Identify Points by:` box. Close the dialog and the new identifying variable will be used as the point label.

8.2 TESTING FOR CURVATURE

SPSS You check for curvature by adding squared terms to the model and using the usual t-test. To do Tukey's test, save the unstandardized predicted values, use a Transformation to square them, and then refit the regression with the squared fitted values as an additional predictor. The t-statistic for this added variable is Tukey's test. It should be compared to a standard normal distribution to get significance levels, not a t-distribution.

You can get all the curvature tests at once. For example, consider the `UN2.txt` data. First use Analyze → Regression → Linear to fit the mean function with response log(*Fertility*) and predictors log(*PPgdp*) and *Purban*. Save the Unstandardized fitted values from this regression.

Next, use the Transform → Compute item to compute $(\log(PPgdp))^2$, $Purban^2$ and *Tukey*, the squares of the fitted values you just saved. Return to the regression dialog, and press the button marked NEXT near the Independent variable (predictor) list, and put the three terms you just created in block number 2, and then press OK. The output table labelled "Excluded Variables" will contain the t-statistics for adding each quadratic term individually after the first block. These t-values are the lack-of-fit values for log(*PPgdp*) and *Purban* given in ALR[T8.2].

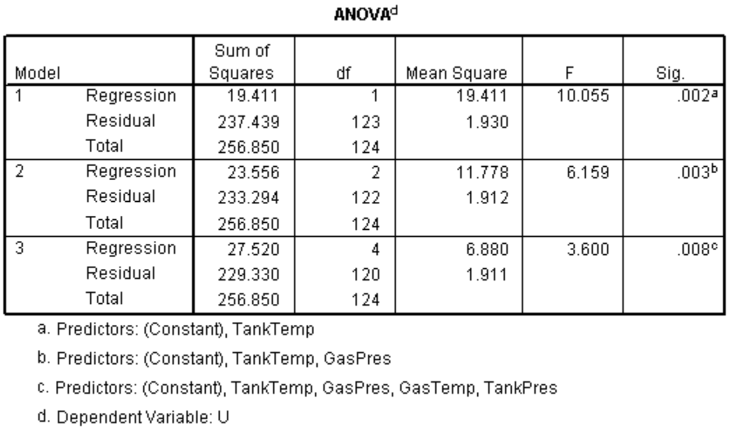
8.3 NONCONSTANT VARIANCE

8.3.1 Variance Stabilizing Transformations

8.3.2 A diagnostic for nonconstant variance

SPSS The score test of nonconstant variance can be done in SPSS by following the four steps in ALR[8.3.2]. Consider the test for the snow geese data. Begin by saving the `Standardized` residuals from the regression fit of *photo* on *obs1*. From Table 8.1, we know these residuals, named say *ZRE*, are equal to $\hat{\text{e}}/\hat{\sigma}$. Using the equation $u_i = n\hat{e}_i^2/[(n-p')\hat{\sigma}^2]$, create the variable U with the transformation `ZRE ** 2 * 45/42` where $n = 45$ and $p' = 2$. Next, compute the regression of U on *obs1*. The score test is equal to 1/2 times the sum of squares for regression of this model, or $162.83/2 = 81.41$. The transforation function `SIG.CHISQ(81.41,1)` will compute the p-value of this statistic.

For the sniffer data, follow the steps above to calculate U from the residuals from the full regression fit of all four predictors. To calculate the first four score statistics in ALR[T8.4], two regressions must be fit:



ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19.411	1	19.411	10.055	.002[a]
	Residual	237.439	123	1.930		
	Total	256.850	124			
2	Regression	23.556	2	11.778	6.159	.003[b]
	Residual	233.294	122	1.912		
	Total	256.850	124			
3	Regression	27.520	4	6.880	3.600	.008[c]
	Residual	229.330	120	1.911		
	Total	256.850	124			

a. Predictors: (Constant), TankTemp

b. Predictors: (Constant), TankTemp, GasPres

c. Predictors: (Constant), TankTemp, GasPres, GasTemp, TankPres

d. Dependent Variable: U

Fig. 8.1 ANOVA for the fit of U on blocked terms from the sniffer data.

- U on the three blocks (1) *TankTemp*, (2) *GasPres*, and (3) *GasTemp*, *TankPres*
- U on *GasPress*

The first blocked fit will produce the ANOVA table in Figure 8.1, from which to score statistics 5.50, 11.78, and 13.76 can be calculated using the appropriate *RSS* values. The second regression fit will produce the *RSS* value used to obtain the statistic 9.71.

8.3.3 Additional comments

8.4 GRAPHS FOR MODEL ASSESSMENT

8.4.1 Checking mean functions

SPSS There is no procedure in SPSS which will produce the marginal plots displayed in ALR[F8.13]. We suggested saving the fitted values from the regression so the two plots in ALR[F8.12] can be drawn. Compare the smoother fits of both plots to determine whether the mean function is adequate. Use a spline smoother for interactive plots and a loess smoother for standard plots.

To make a random linear combination of two predictors, say *x1* and *x2*, make the linear transformation `RV.UNIFORM(0,1) * x1 + RV.UNIFORM(0,1) * x2`. This new variable can be used to make plots similar to the one in ALR[F8.13D].

8.4.2 Checking variance functions

SPSS Standard deviation lines cannot be added to loess and spline smoothes.

9

Outliers and Influence

9.1 OUTLIERS

9.1.1 An outlier test

SPSS As summarized in Table 8.1, SPSS's `Studentized` residuals are ALR's standardized residuals, r_i, from ALR[E9.3] and SPSS's `Studentized Deleted` residuals are ALR's studentized residuals, t_i, from ALR[E9.4]. Thus, use the `Studentized Deleted` residuals to test for outliers.

9.1.2 Weighted least squares

9.1.3 Significance levels for the outlier test

SPSS Significance levels for the outlier test can be obtained by saving the `Studentized Deleted` residuals and finding the level of the appropriate t probability. This is done by transforming the residuals to their absolute values using the function `Abs()`. The maximum value of these absolute values can be found using the procedure Analyze → Descriptive Statistics → Descriptives. Enter the name of the absolute values in the Variables box and click OK. The case number for value is be found by searching the column or by plotting the absolute values again the case numbers. For any scatterplot, the case numbers are always given in the variable list as *Case*[*\$case*].

Suppose we found the largest of the absolute values was 2.85, with $n = 65$ and $p' = 5$. We will use a t-distribution with df$= 65 - 5 - 1 = 59$ to calculate the Bonferroni bound. Using the transformation function, this bound is equal to `65 * 2 * (1-CDF.T(2.85,59))`. By subtracting the CDF from one, we get the upper tail probability. Multiplying this value by two will give a two-tailed test which is multiplied by n to get the Bonferroni bound.

Another approach is to save the `Studentized Deleted` residuals, and then use Transform → Compute to create a new variable named, say `Outlierp` defined by

$$\texttt{outlierp} = \texttt{min}(n{*}2*(\texttt{1} - \texttt{CDF.T}(\texttt{abs}(\texttt{SDR\_1}),n - p' - 1)), \texttt{1})$$

where `SDR_1` is the name that SPSS gives to the Studentized deleted residuals, n is the number of cases in the data, and $n - p' - 1$ is the df for the outlier test. This will compute the Bonferroni p-values for every case, most of which will be equal to one.

9.1.4 Additional comments

9.2 INFLUENCE OF CASES

SPSS Table 8.1 shows the influence and distance options available in SPSS. The `DfBeta(s)` option from the SAVE dialog will save $\hat{\beta} - \hat{\beta}_{(i)}$ for each data case and each parameter estimate. For instance, with the UN data used to construct ALR[F9.1], this option will save these differences for the three parameter estimates in the model. ALR[F9.1] can be drawn in SPSS by making a scatterplot of these differences. The scale of this plot will be different than the scale of ALR[F9.1] because each $\hat{\beta}_{(i)}$ is subtracted from the undeleted estimate $\hat{\beta}$, but the information contained in both plots will be the same.

9.2.1 Cook's distance

SPSS Cook's distance is saved by checking the `Cook's` distance option. See Table 8.1.

9.2.2 Magnitude of D_i

SPSS The plots in ALR[F9.3] can be drawn in SPSS by saving the `Studentized Deleted` residuals, `Leverage Values`, and `Cook's` distances. Each column can be plotted by selecting Graphs → Interactive → Line, and placing the statistic on the vertical axis and the variable *Case*[*$case*] on the horizontal axis. Click on the Dots and Lines tab, check `Dots`, and press OK.

9.2.3 Computing D_i

9.2.4 Other measures of influence

SPSS Added-variable plots are discussed in Section 3.1.

9.3 NORMALITY ASSUMPTION

SPSS The graphic Q-Q can be used to make normal probability plots. To draw either plot in ALR[F9.5], fit the regression model and save the `Unstandardized` residuals. Select Graphs → Q-Q and place the residuals in the variable box and press OK.

10
Variable Selection

10.1 THE ACTIVE TERMS

The first example in this chapter uses randomly generated data. This can be helpful in trying to understand issues against a background where we know the right answers. Generating random data is possible in most statistical packages, though doing so may not be easy or intuitive.

SPSS We could not find an easy way to generate a data set like those discussed in the text using SPSS; if you know how to do it, let us know.

You can duplicate the example in ALR by generating data using a different program such as Microsoft Excel, and then importing the data into SPSS or analysis.

10.1.1 Collinearity

SPSS The variance inflation factors, defined following ALR[E10.5], can be obtained by checking the `Collinearity` option in the STATISTICS dialog.

10.1.2 Collinearity and variances

10.2 VARIABLE SELECTION

10.2.1 Information criteria

The information criteria ALR[E10.7]–ALR[E10.9] depend only on the *RSS*, p', and possibly an estimate of σ^2, and so if these are needed for a particular model, they can be computed from the usual summaries available in a fitted regression model.

SPSS The criteria in ALR[10.2] are not available for linear regression models in SPSS.

10.2.2 Computationally intensive criteria

Computation of *PRESS*, ALR[E10.10], is not common in regression programs, but it is easy to obtain given the residuals and leverages from a fitted model.

SPSS The *PRESS* statistic is not available in SPSS, although the it would be easy to compute by saving the deleted residuals, squaring them and adding them up.

10.2.3 Using subject-matter knowledge

10.3 COMPUTATIONAL METHODS

SPSS SPSS does subset selection without reference to a criterion statistic like C_p or *AIC* for selecting terms. Rather, SPSS is based on an older idea of adding or removing terms based on the value of a t-statistic (which SPSS squares and calls an F-statistic). Selection methods in SPSS are available only in problems with no factors or with all factors replaced by sets of dummy variables.

Suppose that you have a current mean function that includes a set of terms, say $X_{\mathcal{I}}$ with k terms. If using a forward selection method, SPSS will essentially compute all subsets that include $X_{\mathcal{I}}$ plus one additional term, and it will select the term to add that has the largest t-value, if the t is large enough. This is equivalent to using one of the information criteria to find the best subset of $k + 1$ terms with $X_{\mathcal{I}}$ included. Depending on the choice of the "F to enter" value, this enlarged subset may or may not improve over the current k-term mean function. Backward elimination is similar, except that we consider removing a term from $X_{\mathcal{I}}$. When using either forward or backward selection, changing the "F to enter" or "F to remove," using the Options button in the regression dialog, will only change the stopping rule, but it will not change the subsets selected. If you use the hybrid Stepwise method in SPSS, changing the values of these setting can change the subsets considered. Regardless of the values of the settings you chose, there is no guarantee that mean functions considered by these methods will include the functions that optimize an information criterion of interest.

SPSS allows five methods for entering blocks of terms into a regression mean function. The Enter method, used for all previous problems, enters all terms in a block in a single step. The Remove method removes all terms in a block in a single step. The Forward and Backward methods are as described in ALR, except they use the F to enter and F to remove as a stopping criterion; in ALR, we stop based on an information criterion. The Stepwise method allows for entering or deleting terms at each step.

To force a term like log(*Len*) for the highway data from ALR[10.3] in all mean functions, place the term in the first block and choose the Enter selection method. Then add the remaining terms to the second block and choose the Forward, Backward, or Stepwise selection method. For the highway data, the model selected using the Forward method with entry p-value of 0.1 has terms log(*Len*), *Slim*, and *Acpt*. The ANOVA tables produced with this procedure are

```
ANOVA(d)
Model           Sum of Squares  df   Mean Square     F       Sig.
1   Regression  5.537           1    5.537           17.950  .000(a)
    Residual    11.414          37   .308
    Total       16.951          38
2   Regression  10.839          2    5.419           31.920  .000(b)
    Residual    6.112           36   .170
    Total       16.951          38
3   Regression  11.439          3    3.813           24.213  .000(c)
    Residual    5.512           35   .157
    Total       16.951          38
a.   Predictors: (Constant), logLen
b.   Predictors: (Constant), logLen, Slim
c.   Predictors: (Constant), logLen, Slim, Acpt
d.  Dependent Variable: logRate
```

10.3.1 Subset selection overstates significance

10.4 WINDMILLS

10.4.1 Six mean functions

10.4.2 A computationally intensive approach

The data for the windmill example in ALR[10.4.2] is not included with the `alr3` library, and must be downloaded separately from `www.stat.umn.edu/alr`.

11 Nonlinear Regression

11.1 ESTIMATION FOR NONLINEAR MEAN FUNCTIONS

11.2 INFERENCE ASSUMING LARGE SAMPLES

SPSS The command Analyze → Regression → Nonlinear is used to fit nonlinear regression models; we have illustrated this previously in Section 6.4 to fit a partial one-dimensional mean function. To fit the mean function ALR[E11.16] using the `turk0` data, use the model expression `th1 + th2*(1 - EXP(-(th3*A)))` with starting values discussed in ALR[11.2]. The nonlinear fitted line in ALR[F11.2] cannot be added to a scatterplot in SPSS. The program can be used to save residuals and fitted values, and these could then be used in more standard SPSS graphics.

SPSS does not fit nonlinear weighted least squares. To fit the weighted models for the `turkey` data in ALR[11.2] we can apply ALR[E5.8] to get WLS estimates. We have $y = g(\theta, x) + e/\sqrt{w}$ where the e's have constant variance, so the y's have variance σ^2/w. Multiply both sides of the mean function by $\sqrt{w}$ to get $\sqrt{w}y = \sqrt{w}g(\theta, x) + e$ so we can get WLS estimates in the original problem by getting OLS estimates with $\sqrt{w}g(\theta, x)$ as the kernel mean function, and $\sqrt{w}y$ as the response. Fitting this model in SPSS requires defining *wGain* as the transformation `SQRT(m)*Gain` because the weights are equal to the number of pens, m. When $g(\theta, x)$ is equal to ALR[E11.16], the model expression for the weighted nonlinear regression is `SQRT(m)*(th1 + th2*(1 - EXP(-(th3*A))))`. Dummy variables for the factor S must be created to fit the mean functions ALR[E11.17]-ALR[E11.19].

Another useful feature of of SPSS is the ability to constrain the estimated value of some of the parameters, using the CONSTRAINTS button on the nonlinear dialog. In particular, you can force some of the parameters to be equal to specified values, which can allow quickly fitting a sequence of mean functions to the same data.

11.3 BOOTSTRAP INFERENCE

SPSS The bootstrap can be used to get standard errors for coefficient estimates in nonlinear regression by selecting OPTIONS in the nonlinear regression dialog, and then checking `Bootstrap estimates of standard error`.

11.4 REFERENCES

12

Logistic Regression

Both logistic regression and the normal linear models that we have discussed in earlier chapters are examples of *generalized linear models.* Many programs, including SAS, R, and S-Plus, have procedures that can be applied to any generalized linear model. Both JMP and SPSS seem to have separate procedures for logistic regression. There is a possible source of confusion in the name. Both SPSS and SAS use the name *general linear model* to indicate a relatively complex linear model, possibly with continuous terms, covariates, interactions, and possibly even random effects, but with normal errors. Thus the general linear model is a special case of the generalized linear models.

12.1 BINOMIAL REGRESSION

12.1.1 Mean Functions for Binomial Regression

12.2 FITTING LOGISTIC REGRESSION

SPSS To fit logistic regression with a Bernoulli response variable in SPSS use the procedure Regression → Binary Logistic. Pearson's χ^2, ALR[12.9], is not available from this procedure.

12.2.1 One-predictor example

SPSS To fit the logistic regression of y on $\log(D)$ for the blowdown data, select Analyze → Regression → Binary Logistic. The response or dependent variable is y and the Covariate is $\log(D)$. After pressing OK, SPSS will produce alot of output but the tables of interest will be in the "Block 1" section.

The first table of this section is

```
Omnibus Tests of Model Coefficients
                Chi-square  df  Sig.
Step 1  Step    200.965     1   .000
        Block   200.965     1   .000
        Model   200.965     1   .000
```

This gives the change in deviance when comparing the logistic model fit to the model containing only the intercept. The Chi-square value for the Block is the same as the change in deviance between the first two models in ALR[T12.4].

The second table of this section is

```
Model Summary
Step    -2 Log likelihood   Cox & Snell R Square    Nagelkerke R Square
1           655.242(a)          .263                        .361
a. Estimation terminated at iteration number 5 because
parameter estimates changed by less than .001.
```

The "-2 Log likelihood" value is the residual deviance from ALR[E12.8]. The other statistics are not discussed in ALR.

The next table of interest is

```
Variables in the Equation
                        B       S.E.    Wald        df    Sig.    Exp(B)
Step 1(a)   logD        2.263   .191    139.742     1     .000    9.608
            Constant    -7.892  .633    155.681     1     .000    .000
a.   Variable(s) entered on step 1: logD.
```

This table gives the parameter estimates seen in ALR[T12.1]. The Wald test statistics in this SPSS table are the square of the z-values given in ALR[T12.1]. The SPSS test statistic is compared to a χ^2 distribution to obtain the significance level.

The logistic curves seen in ALR[F12.1A] cannot be added to a scatterplot of y and $\log(D)$ in SPSS. If you check the Predicted Values option `Probabilities` from the Save dialog before the logistic regression is fit, then you can obtain a scatterplot of the regression curve by using an overlay plot. After the probabilities are saved, select Graphs → Scatter and choose the overlay plot from the four plot types. If the probabilities were saved under the variable name *pred*, then the two pairs of `Y-X` variables to plot are `y-logD` and `pred-logD`. You select one pair at a time from the variable list by holding down the control key as each variable is selected. If the pair is display in the wrong order, press the button SWAP PAIR.

ALR[F12.1B] cannot be drawn in SPSS, but histograms of log(D) for each level of y can be drawn side by side. Select Graphs → Interactive → Histogram and place log(D) in the horizontal variable box and y in the Panel variable box, then press OK.

12.2.2 Many Terms

SPSS The two models in ALR[T12.2] are fit the same way described in Section 12.2.1. Another way to get all three models described in ALR[T12.1] and ALR[T12.2] is to fit the terms log(D), S, and log(D)×S as three separate blocks in the logistic regression model. For this fit, the block "Chi-square" in the Omnibus Tests of Model Coefficients table for each block is the change in deviance for adding each term to the previous blocks. Thus the values in these tables are the same as the change in deviance in ALR[T12.4].

The plots in ALR[F12.3] can be drawn using SPSS histograms and scatterplots. The plots in ALR[F12.4] cannot be drawn in SPSS.

12.2.3 Deviance

12.2.4 Goodness of Fit Tests

SPSS The SPSS logistic procedure has several unexpected limitations. First, to include interactions in the mean function, you must precompute them using a transformation. Logistic regression requires the response to have two categories, like 0 and 1. If you have grouped binomial data, as in the Titanic example in ALR, you can't use logistic regression. You can, however, use Analyze → Regression → Probit, which is very similar to logistic regression (and not discussed in ALR). SPSS does include several generalizations of logistic regression to problems with a response with several categories, and to log-linear models for categorical data.

12.3 BINOMIAL RANDOM VARIABLES

12.3.1 Maximum likelihood estimation

12.3.2 The Log-likelihood for Logistic Regression

12.4 GENERALIZED LINEAR MODELS

Problems

References

1. Chambers, J. and Hastie, T. (eds.) (1993). *Statistical Models in S.* Boca Raton, FL: CRC Press.

2. Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman & Hall.

3. Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* New York: Wiley.

4. Cook, R. D. and Weisberg, S. (2004). Partial One-Dimensional Regression Models.

5. Dalgaard, Peter (2002). *Introductory Statistics with R.* New York: Springer.

6. Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application.* Cambridge: Cambridge University Press.

7. Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall.

8. Fox, John (2002). *An R and S-Plus Companion to Applied Regression.* Thousand Oaks, CA: Sage.

9. Fruend, R., Littell, R. and Creighton, L. (2003). *Regression Using JMP.* Cary, NC: SAS Institute, Inc., and New York: Wiley.

10. Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics*, 16, 499-511.

11. Knüsel, Leo (2005). On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, 48, 445–449.

12. Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R.* Cambridge: Cambridge University Press.

13. Muller, K. and Fetterman, B. (2003). *Regression and ANOVA: An Integrated Approach using SAS Software.* Cary, NC: SAS Institute, Inc., and New York: Wiley.

14. Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society*, A140, 48–77.

15. Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-plus.* New York: Springer.

16. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

17. Sall, J., Creighton, L. and Lehman, A. (2005). *JMP Start Statistics*, third edition. Cary, NC: SAS Institite, and Pacific Grove, CA: Duxbury. **Referred to as** JMP-START**.**

18. SPSS (2003). *SPSS Base 12.0 User's Guide.* Chicago, IL: SPSS, Inc.

19. Thisted, R. (1988). *Elements of Statistical Computing.* New York: Chapman & Hall.

20. Venables, W. and Ripley, B. (2000). *S Programming.* New York: Springer.

21. Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*, 4th edition. New York: Springer. **referred to as** VR**.**

22. Venables, W. and Smith, D. (2002). *An Introduction to R.* Network Theory, Ltd.

23. Verzani, John (2005). *Using R for Introductory Statistics.* Boca Raton: Chapman & Hall.

24. Weisberg, S. (2005). *Applied Linear Regression*, third edition. New York: Wiley. **referred to as** ALR.

25. Weisberg, S. (2005). Lost opportunities: Why we need a variety of statistical languages. *Journal of Statistical Software*, 13(1), www.jstatsoft.org.

Index