Essential Statistics Using SAS® University Edition™

Geoff Der · Brian S. Everitt

Contents

Chapter 1: Statistics and an Introduction to the SAS University Edition 1

1.1 Introduction 1
1.2 Measurements and Observations 2
1.3 Nominal or Categorical Measurements 2
1.4 Ordinal Scale Measurements 3
1.5 Interval Scales 3
1.6 Ratio Scales 4
1.7 Populations and Samples 4
1.8 SAS University Edition 5
  1.8.1 Starting the SAS University Edition 6
  1.8.2 The SAS University Edition User Interface 6
  1.8.3 The Navigation Pane 7
  1.8.4 The Work Area 8
  1.8.5 Tasks and Task Settings 8
  1.8.6 The Data Tab 9
  1.8.7 The Model Tab 10
  1.8.8 The Options Tab 11
  1.8.9 The Output Tab 12
  1.8.10 The Information Tab 12
  1.8.11 Abbreviations of Task Settings Used in This Book 13
  1.8.12 The Results Pane 13
  1.8.13 Options and Preferences 14
  1.8.14 Setting Up the Data Used in This Book 15

Chapter 2: Data Description and Simple Inference 17

2.1 Introduction 17

2.2 Summary Statistics and Graphical Representations of Data 18

2.2.1 Initial Analysis of Room Width Guesses Using Simple Summary Statistics and Graphics 18

2.3 Testing Hypotheses and Student's t-Test 22

2.3.1 Applying Student's t-Test to the Guesses of Room Width 23

2.3.2 Checking the Assumptions Made When Using Student's t-Test and Alternatives to the t-Test 29

2.4 The t-Test for Paired Data 31

2.4.1 Initial Analysis of Wave Energy Data Using Box Plots 31

2.4.2 Wave Power and Mooring Methods: Do Two Mooring Methods Differ in Bending Stress? 34

2.4.3 Checking the Assumptions of the Paired t-Tests 36

2.5 Exercises 37

Chapter 3: Categorical Data 41

3.1 Introduction 41

3.2 Graphing and Analysing Frequencies: Horse Race Winners 42

3.2.1 Looking at Horse Race Winners Using Some Simple Graphics: Bar Charts and Pie Charts 42

3.2.2 Chi-Square Goodness-of-Fit Test: Does Starting Stall Position Predict Horse Race Winners? 47

3.3 Two-By-Two Tables 48

3.3.1 Chi-Square Test: Breast Self-Examination 48

3.3.2 Odds and Odds Ratios 50

3.3.3 The Cochrane-Mantel-Haenszel Test for Multiple Related 2 X 2 Tables 52

3.4 Larger Cross-Tabulations 54

3.4.1 Tabulating the Brain Tumour Data Into a Contingency Table 56

3.4.2 Do Different Types Of Brain Tumours Occur More Frequently at Particular Sites? Chi-Squared Test 56

3.5 Fisher's Exact Test 59

3.5.1 How Is Baiting Behaviour at Suicides Affected by Season? Fisher's Exact Test .. 60

3.5.2 Fisher's Exact Test for Larger Tables 63

3.6 McNemar's Test Example 64

3.6.1 Juvenile Felons: Where Should They Be Tried? McNemar's Test 65

3.7 Exercises 67

Chapter 4: Bivariate Data: Scatterplots, Correlation, and Regression....... 71

4.1 Introduction 71

4.2 Scatter Plots, the Correlation Coefficient, and Simple Linear Regression..... 72

4.2.1 The Scatter Plot..... 72

4.2.2 The Correlation Coefficient 74

4.2.3 Simple Linear Regression..... 77

4.2.4 A Further Example of Linear Regression 80

4.2.5 Checking Model Assumptions: The Use of Residuals..... 85

4.3 Adjusting the Scatter Plot To Show Patterns in the Data..... 91

4.3.1 Plotting the Birthrate Data: The Aspect Ratio of a Scatter Plot 92

4.4 Exercises..... 96

Chapter 5: Analysis of Variance 105

5.1 Introduction 105

5.2 One-Way ANOVA..... 106

5.2.1 Initial Examination of Teaching Arithmetic Data with Summary Statistics and Box Plots 106

5.2.2 Teaching Arithmetic: Are Some Methods for Teaching Arithmetic Better than Others? 108

5.3 Two-Way ANOVA 113

5.3.1 A First Look at the Rat Weight Gain Data Using Box Plots and Numerical Summaries 114

5.3.2 Weight Gain in Rats: Do Rats Gain More Weight on a Particular Diet?..... 116

5.4 Unbalanced ANOVA..... 119

5.4.1 Summarizing the Post-Natal Depression Data 121

5.4.2 Post-Natal Depression: Is a Child's IQ Affected? 122

5.5 Exercises..... 127

Chapter 6: Multiple Linear Regression 133

6.1 Introduction 133

6.2 Multiple Linear Regression 134

6.2.1 The Ice Cream Data: An Initial Analysis Using Scatter Plots..... 135

6.2.2 Ice Cream Sales: Are They Most Affected by Price or Temperature? How to Tell Using Multiple Regression..... 137

6.2.3 Diagnosing the Multiple Regression Model Fitted to the Ice Cream Consumption Data: The Use of Residuals 139

6.2.4 A More Complex Example of the Use of Multiple Linear Regression..... 142

6.2.5 The Cloud Seeding Data: Initial Examination of the Data Using Box Plots and Scatter Plots 143
6.2.6 When Is Cloud Seeding Best Carried Out? How to Tell Using Multiple Regression Models Containing Interaction Terms 150
6.3 Identifying a Parsimonious Regression Model 155
6.4 Exercises 164
Chapter 7: Logistic Regression 169
7.1 Introduction 169
7.2 Logistic Regression 169
7.2.1 Intra-Abdominal Sepsis: Using Logistic Regression to Answer the Question of What Predicts Survival after Surgery 171
7.2.2 Odds 172
7.2.3 Applying the Logistic Regression Model with a Single Explanatory Variable 173
7.2.4 Logistic Regression with All the Explanatory Variables 178
7.2.5 A Second Example of the Use of Logistic Regression 179
7.2.6 An Initial Look at the Caseness Data 180
7.2.7 Modeling the Caseness Data Using Logistic Regression 181
7.3 Logistic Regression for 1:1 Matched Studies 185
7.4 Summary 191
7.5 Exercises 191
Chapter 8: Poisson Regression and the Generalized Linear Model 199
8.1 Introduction 199
8.2 Generalized Linear Model 200
8.2.1 Components of Generalized Linear Models 200
8.2.2 Using the Generalized Linear Model to Apply Multiple Linear Regression and Logistic Regression 201
8.3 Poisson Regression 205
8.3.1 Example 1: Familial Adenomatous Polyposis (FAP) 205
8.3.2 Example 2: Bladder Cancer 208
8.4 Overdispersion 212
8.5 Summary 215
8.6 Exercises 215
References 219
Index 225

Chapter 1: Statistics and an Introduction to the SAS University Edition

1.1 Introduction **1**
1.2 Measurements and Observations **2**
1.3 Nominal or Categorical Measurements **2**
1.4 Ordinal Scale Measurements **3**
1.5 Interval Scales **3**
1.6 Ratio Scales **4**
1.7 Populations and Samples **4**
1.8 SAS University Edition **5**
- 1.8.1 Starting the SAS University Edition 6
- 1.8.2 The SAS University Edition User Interface 6
- 1.8.3 The Navigation Pane 7
- 1.8.4 The Work Area 8
- 1.8.5 Tasks and Task Settings 8
- 1.8.6 The Data Tab 9
- 1.8.7 The Model Tab 10
- 1.8.8 The Options Tab 11
- 1.8.9 The Output Tab 12
- 1.8.10 The Information Tab 12
- 1.8.11 Abbreviations of Task Settings Used in This Book 13
- 1.8.12 The Results Pane 13
- 1.8.13 Options and Preferences 14
- 1.8.14 Setting Up the Data Used in This Book 15

1.1 Introduction

Statistics is a general intellectual method that applies wherever data, variation, and chance appear. It is a fundamental method because data, variation, and chance are omnipresent in modern life. It is an independent discipline with its own core ideas,, rather than, for example, a branch of mathematics. Statistics offers general, fundamental, and independent ways of thinking (Moore, 1998).

There are, of course, less worthy and less formal statements of what the field of statistics is about;

> *"There are two kinds of statistics, the kind you look up and the kind you make up. (Rex Stout, Death of a Doxy)*
>
> *Do not put your faith in what statistics say until you have carefully considered what they do not say. (William W. Watt)"*

Quintessentially, statistics is about solving problems. Data (measurements, observations) relevant to these problems are collected and statistical analyses are used to provide (hopefully) useful answers. But the path from data collection to analysis and interpretation is often not straightforward. Most real-life applications of statistical methodology have one or more nonstandard features. In practice, this means that there are few routine statistical questions, although there are questionable statistical routines. Many statistical pitfalls lie in wait for the unwary. Indeed, statistics is perhaps more open to misuse than most other subjects, particularly by the non-statistician with access to powerful statistical software. Because this book is aimed essentially at users of statistics who might not be expert statisticians and who are using the SAS University Edition, we shall try to make clear in subsequent chapters what is and what is not good statistical practice so that readers can avoid such pitfalls.

Although as mentioned in the Preface, we do not intend this book to be an introduction to statistics *per se*, we will in this chapter cover very briefly some basic concepts that will hopefully be useful to readers in the chapters to come.

1.2 Measurements and Observations

The basic material, the *data* that are the foundation of all scientific investigation no matter in what field, are the *measurements* taken and the *observations* made on the things of interest to the investigator (for example, subjects perhaps in psychology, patients in medicine, artefacts in archaeology, or animals in zoology). These measurements will, of course, *vary* between the subjects, patients, artefacts, and so on, and so are usually all referred to as *variables.* For any defendable conclusions to arise from these measurements, they need to be *objective*, *precise,* and *reproducible* (Fleiss, 1999). Measurements come in a variety of types and the type of measurement will, in part at least, determine the appropriate method of statistical analysis. It is time to say a little about *scales of measurement*.

1.3 Nominal or Categorical Measurements

Nominal measurements allow subjects, patients, and so on to be classified with respect to some characteristic. Examples of such measurements are categories such as marital status (single, married, divorced), sex (male, female), and blood group (A, O, AB, Other). The properties of a nominal scale are

- The categories are mutually exclusive (an individual can belong to only one category).

- The categories have no logical order—numbers can be assigned to categories but merely as convenient labels.

With nominal scaled measurements, we are essentially confined to counting the number of subjects (or whatever we are dealing with) in each category of the scale.

1.4 Ordinal Scale Measurements

The next level of measurement is the *ordinal scale*. This scale has one additional property over those of a nominal scale—a logical ordering of the categories. With such measurements, numbers assigned to the categories can be used to indicate the amount of a characteristic each variable possesses. A psychiatrist might, for example, grade patients on an anxiety scale as “not anxious,” “mildly anxious,” “moderately anxious,” or “severely anxious” and use the numbers 0, 1, 2, and 3 to label the categories, with lower numbers indicating less anxiety. The psychiatrist cannot infer, however, that the difference in anxiety between patients with scores of 0 and 1 is the same as the difference between patients assigned scores of 2 and 3. The scores on an ordinal scale do, however, allow patients to be *ranked* with respect to the characteristic being assessed.

The following are the properties of an ordinal scale:

- The categories are mutually exclusive.
- The categories have some logical order.
- The categories are scaled according to the amount of a particular characteristic that they indicate.

1.5 Interval Scales

The third level of measurement is the *interval scale*. Such scales possess all the properties of an ordinal scale plus the additional property that equal differences between category levels, on any part of the scale, reflect equal differences in the characteristic being measured. An example of such a scale is temperature on the Celsius (C) or Fahrenheit (F) scale; the difference between temperatures of 80 °F and 90 °F represents the same difference in heat as that between temperatures of 30 °C and 40 °C on the Celsius scale. An important point to make about interval scales is that the zero point is simply another point on the scale; it does *not* represent the starting point of the scale or the total absence of the characteristic being measured. This implies that quoting ratios of such variables is not valid.

The properties of an interval scale are as follows:

- The categories are mutually exclusive.
- The categories have a logical order.
- The categories are scaled according to the amount of the characteristic they indicate.

- Equal differences in the characteristic are represented by equal differences in the numbers assigned to the categories.
- The zero point is completely arbitrary.

1.6 Ratio Scales

The final level of measurement is the *ratio scale*. This type of scale has one further property in addition to those listed for interval scales: it possesses a true zero point that represents the absence of the characteristic being measured. Consequently, statements can be made about both the differences on the scale and the ratio of points on the scale. An example is weight, where not only is the difference between 100 kg and 50 kg the same as between 75 kg and 25 kg, but an object weighing 100 kg can be said to be twice as heavy as one weighing 50 kg. Celsius and Fahrenheit temperatures are not ratio scales, so, for example, a weather reporter who says that today with a temperature of 30 °C is twice as hot as the corresponding day last week when the temperature was 15 °C is wrong. But temperature measured on the Kelvin scale, which does have a true zero point (absolute zero or −273 °C) is a ratio scale, so converting the two temperatures from Celsius to Kelvin by simply adding 273 to each to give 288 K and 303 K lets us say, correctly, that a day with a temperature of 30 °C is 303/288=1.05 times as hot as a day with a temperature of 15 °C.

The properties of a ratio scale are

- The categories are mutually exclusive.
- The data categories have a logical order.
- The categories are scaled according to the amount of the characteristic they possess.
- Equal differences in the characteristic being measured are represented by equal differences in the numbers assigned to the categories.
- The zero point represents an absence of the characteristic being measured.

An awareness of the different types of measurement that might be encountered when collecting data is important because the appropriate method of statistical analysis to use can often depend on the type of measurement involved, a point that we shall consider where necessary in the subsequent chapters.

A further classification of variable types is into *response* or *dependent* variables and *explanatory* variables; a variety of statistical techniques are used to investigate the effects of the latter on a response variable of interest (for example, sex and age on IQ). We shall consider such techniques in Chapters 4 and 6 (linear regression) and in Chapter 7 (logistic regression).

1.7 Populations and Samples

Most statistical methods seek to help the investigator draw conclusions (*inferences*) about a *population* of interest based on a *sample* of observations from that population. For example, we might be interested the height of men of age 70 born in the county of Essex in the United Kingdom in 1944; this is our population. And we might then begin our investigation by measuring the heights of 100 men from this population; these 100 men constitute our sample. (The sample is usually considered to have been taken at *random*; that is, each member of the population has the same chance of being included in the sample, but this is a detail that we will not elaborate on further.) We can now use the sample values of height to say something about the average height in the population. This could be formulated in terms of a hypothesis about the population average, that it is 6 feet, and a suitable *significance test* applied to see whether the sample values of height suggested evidence against the hypothesis. Or we might use the sample values to *estimate* the population average or, better, to produce a range of likely values for the population average, a range know as a *confidence interval*. Neither a significance test nor a confidence interval can give you certainty because, of course, they are based on only a sample of the values from the population; however, the larger the size of the sample, the more credence you can give to the results. (This is probably a good time to remind ourselves of that old Chinese proverb: "*To be uncertain is uncomfortable, but to be certain is to be ridiculous.*")

A variety of significance tests and confidence intervals will appear throughout subsequent chapters.

Statistical methods often require extensive calculations and so require some friendly software for them to be applied to data. And many statistical methods are quintessentially graphical and so require the software to have the ability to construct a variety of plots. Cue the SAS University Edition.

1.8 SAS University Edition

SAS University Edition is a powerful statistical package provided free of cost by SAS Institute to universities.

The user interface for SAS University Edition (known as SAS Studio) is a browser interface that you access using the web browser of your computer. The common browsers —Internet Explorer, Safari, Firefox, and Chrome—are all supported. An advantage of a browser interface is that the computer doing the processing could be anywhere on the web. Usually, it will be the same PC as the browser is running on but could equally well be a remote computer or a cloud computing service, such as Amazon Web Services. In the latter case, the browser could be running on a mobile device such as a tablet.

Installing SAS University Edition on a Windows PC is covered in a document entitled "SASUniversityEditionInstallGuideWindows.pdf", which can be found with a web search but is also contained in the online materials for this book. Some tips on installation and setup are also given there.

On a PC, SAS University Edition runs under a virtual machine and requires virtualisation software to be installed. For Windows, there is a choice of Oracle VM VirtualBox and VMware Player.

An important part of the installation is the creation of shared folders. We recommend using **`c:\SASUniversityEdition\myfolders`** as the path and `myfolders` as the name. Take care to match the case exactly.

1.8.1 Starting the SAS University Edition

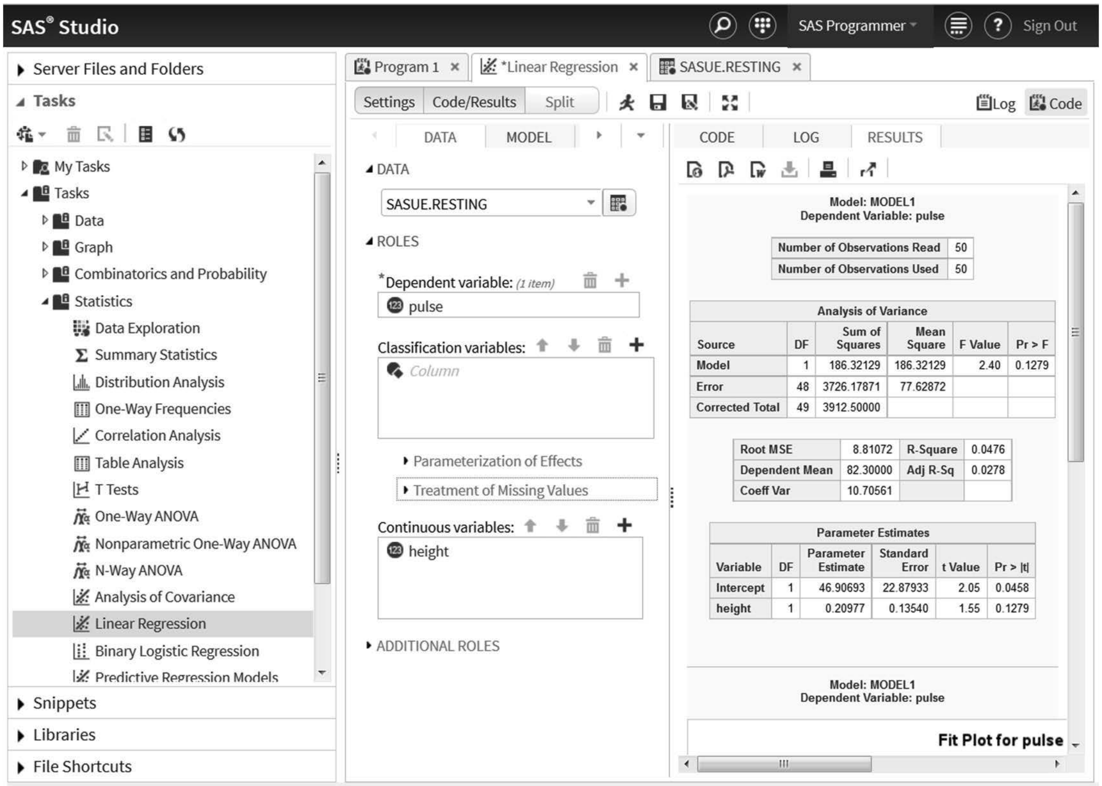
To start the SAS University Edition under Windows:

1. Start VirtualBox or VMware Player.
2. Start (or play) the SAS University Edition virtual machine. It takes a little time to start and then shows an address to enter in your browser. For VirtualBox, this is likely to be http://localhost:10080; for VMware Player, it will be an IP address such as http://172.16.49.136.
3. Note this address.
4. Minimize VirtualBox or VMware Player (do not close it).
5. Start your browser and enter the address above in the address box (the initial http:// might be optional).
6. At the welcome screen, click **Start SAS Studio**.

1.8.2 The SAS University Edition User Interface

The SAS University Edition interface is shown in Display 1.1 and consists of a navigation pane on the left and a work area on the right. In the view shown in Display 1.1, the work area is itself split into two parts: the settings pane and the results pane. Display 1.1 is based on an example in Chapter 4, Section 4.2.3.

Display 1.1: The SAS University Edition Interface



1.8.3 The Navigation Pane

The navigation pane has the following areas: Server Files and Folders, Tasks, Snippets, Libraries, and File Shortcuts. Clicking on the title of an area expands it to show its contents. In Display 1.1, the Tasks area has been expanded. The Tasks area contains menus for data manipulation, graphics, and statistical analyses, and these form the main focus of this book. In Display 1.1, the **Statistics** tasks are shown within the Tasks area and the **Linear Regression** task is selected. The Libraries area allows the contents of SAS data sets to be viewed. The Snippets area contains saved pieces of SAS syntax (also known as *code*) that can be edited. Code snippets are typically used to do things that are not available as tasks. Usually, these require an understanding of SAS syntax, which is beyond the limit of this book, although interested readers can consult one of our other books. The Server File and Folders area is for manipulation of files in a way that will be familiar to most readers. Perhaps its most important use is to verify that files are visible to SAS University Edition in the **myfolders** folder. File Shortcuts are the SAS equivalent of Windows shortcuts.

1.8.4 The Work Area

The work area is where the settings for a task are entered and where the results are shown when the task has run. In Display 1.1, the work area covers the two-thirds of the screen on the right. The middle section, the settings pane, shows some of the settings for the linear regression task and the right-hand section, the results pane, shows the results of the analysis.

The work area is also where the contents of SAS data sets can be displayed. At the top of the work area in Display 1.1 are three tabs: **Program 1**, **Linear Regression**, and **sasue.resting.** The **Program 1** tab is there by default when the software is started. The tab for **sasue.resting** shows that the sasue.resting data set—the one being analysed—has been opened, but its contents are hidden behind those for the linear regression analysis. Given that the data set has already been opened, to view the contents, simply click on the tab, which brings it to the front.

Below the three tabs at the top of the work area is a second line with three more tabs and some icons. The three tabs are labelled **Settings**, **Code/Results**, and **Split**, and they determine what is shown in the work area. If **Settings** is selected, the whole work area is reserved for the task settings. If **Code/Results** is selected, the work area is reserved for code (SAS syntax), and/or the task results. The **Split** tab, selected in Display 1.1, splits the work area into two panes, one on the left for task settings and the other on the right for results. The righthand pane can also show the code and/or log by clicking on the icons on the right-hand end of the line containing the three tabs.

We recommend using the split view for most purposes.

Sometimes more space is needed for the task settings than is available in the standard split view, although this does depend on the size of the screen. One option here would be to temporarily switch to the settings view, but a better option is to maximize the work area by clicking the **Maximize View** button () or pressing Alt-F11, which enlarges the work area to cover the whole screen. Clicking the same button returns the work area to its previous size.

The other important icon on this line is the **Run** button (); clicking this button runs the task to produce the results. Pressing the F3 function key has the same effect.

Further details of the SAS University Edition interface can best be illustrated by an example task. As in Display 1.1, we will use the example from Chapter 4, Section 4.2.3.

1.8.5 Tasks and Task Settings

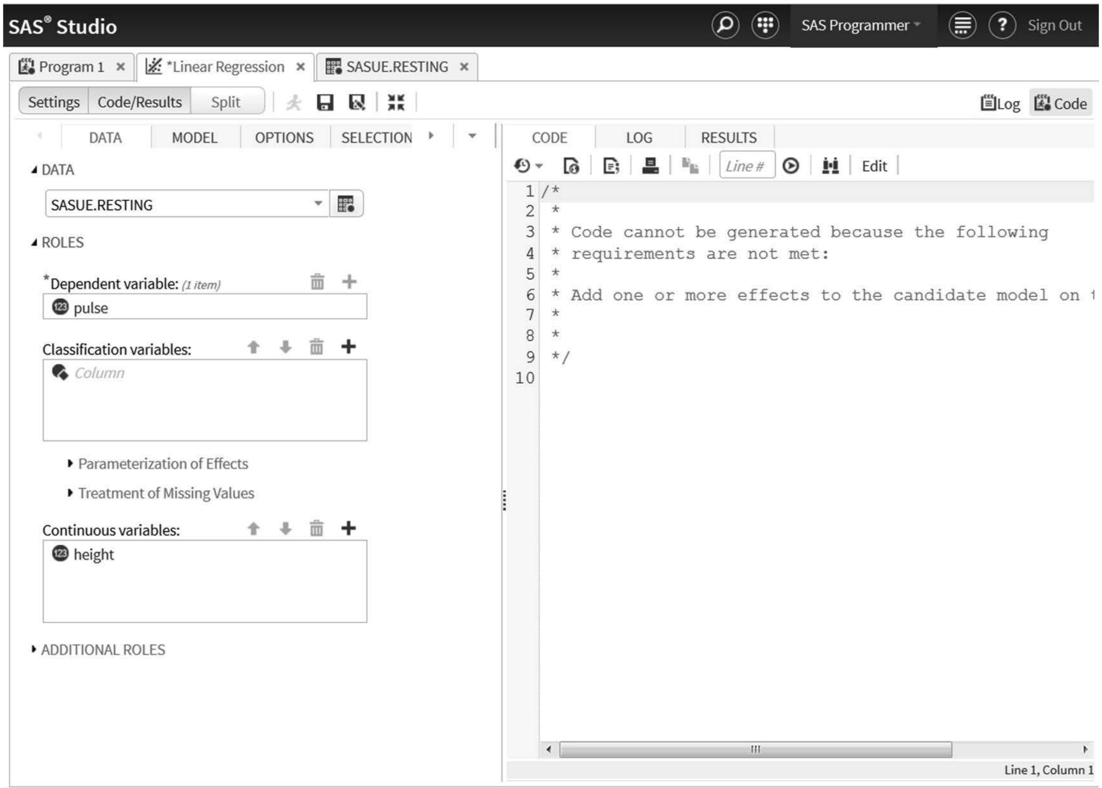
Opening a Task

The first step in using a task is to open the task in the work area. In this book, we will use tasks for statistical analysis, graphics, and data manipulation. Those tasks are found in the navigation pane under **Statistics**, **Graph**, or **Data**, respectively. In Display 1.1, the Statistics area of the navigation pane has been expanded and the Linear Regression task is highlighted. To expand an area of the navigation pane, click on the triangle () on its left; to collapse it, click on the triangle again (which now looks like this: ◢). To open a task in the work area, either double-click on it, right-click it and select **Open**, or drag it across to the work area.

Entering the Task Settings

Having opened a task, the next step is to enter the task settings. The settings that are shown and that need to be entered vary according to the task. Display 1.2 shows the same example task as Display 1.1 as we begin to enter the settings, but in this case the work area has been expanded and the settings pane enlarged (by dragging the pane border). We can see that the Linear Regression task has tabs for **Data**, **Model**, **Options**, **Selection**, **Output**, and **Information**.

Display 1.2: The SAS University Edition Interface with the Work Area Maximized



1.8.6 The Data Tab

The most important is the **Data** tab; nearly all tasks have a **Data** tab and many will only require settings within the **Data** tab. Within it, the first setting specifies the data set that is to be analysed. In Display 1.2, we can see that the data set named in the box is **sasue.resting**. To select a data set for analysis, click on the icon by the box ( ); a popup window then shows the available libraries, and a data set can be selected from one of these by double-clicking on it.

The **Data** tab is also the place where the variables to be analysed are specified, and these will be assigned roles depending on the task. In a regression type task, there will usually be one dependent variable and one or more predictor variables. The predictor variables might be subdivided into classification variables or continuous variables. Classification variables are nominal scale variables and continuous variables are interval or ratio scale variables. The treatment of ordinal variables depends on the analysis. In Display 1.2, Pulse is the dependent variable and Height a continuous predictor variable.

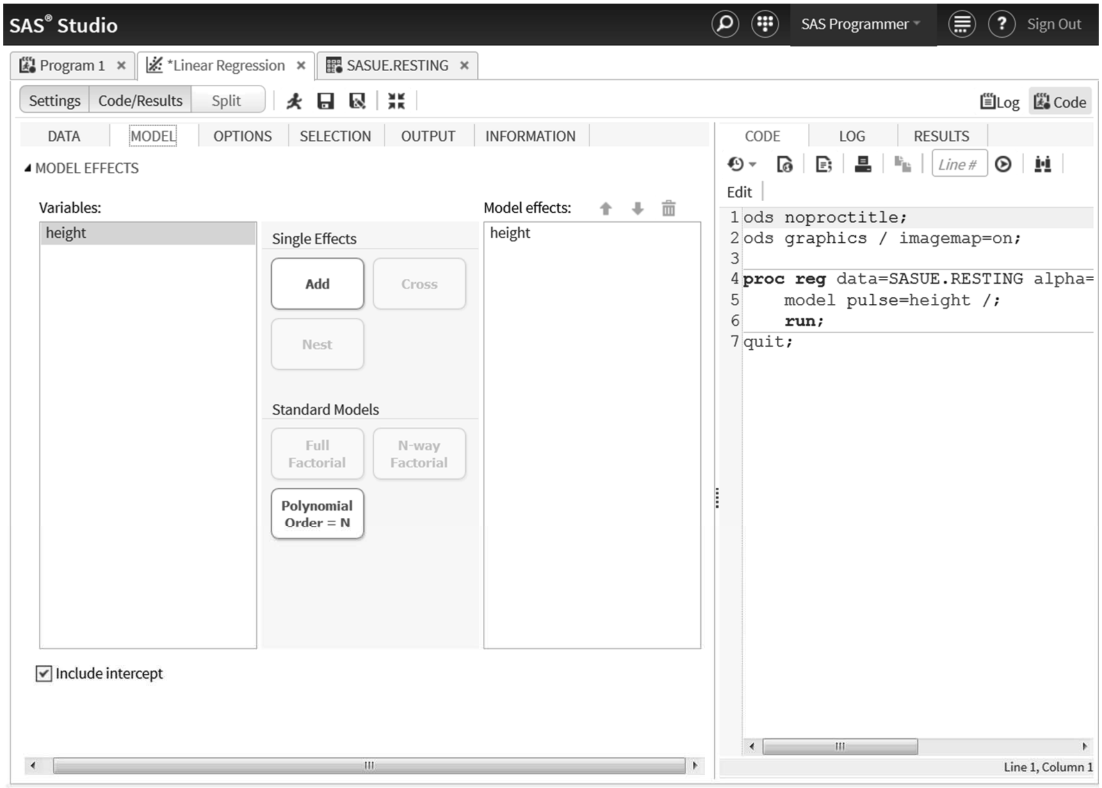
To assign a variable to a particular role, click the **Add** button ( ) and select it from the popup list of variables in the data set. To remove a variable, select it and click the **Delete** button ( ).

As settings are entered on the right of the work area, SAS code is generated on the left. In Display 1.2, no code has been generated. Instead, there are comments to say why not. The message is not fully visible in Display 1.2, but it reads: `Add one or more effects to the candidate model on the MODEL tab`. It is also worth noting that the **Run** button ( ) is greyed out at this point, indicating that the analysis is not ready to be run.

1.8.7 The Model Tab

Display 1.3 shows the next step in entering the settings for this example. The **Model** tab has been selected and the settings pane enlarged further by dragging the dividing bar. The predictor variable, Height, has been added to the **Variables** box on the left automatically by virtue of being assigned that role under the **Data** tab. It has been added to the **Model effects** by selecting it and clicking the **Add** button. There are now sufficient settings entered to generate SAS code, which is shown on the right and the **Run** button is no longer greyed out, indicating that the analysis is ready to run.

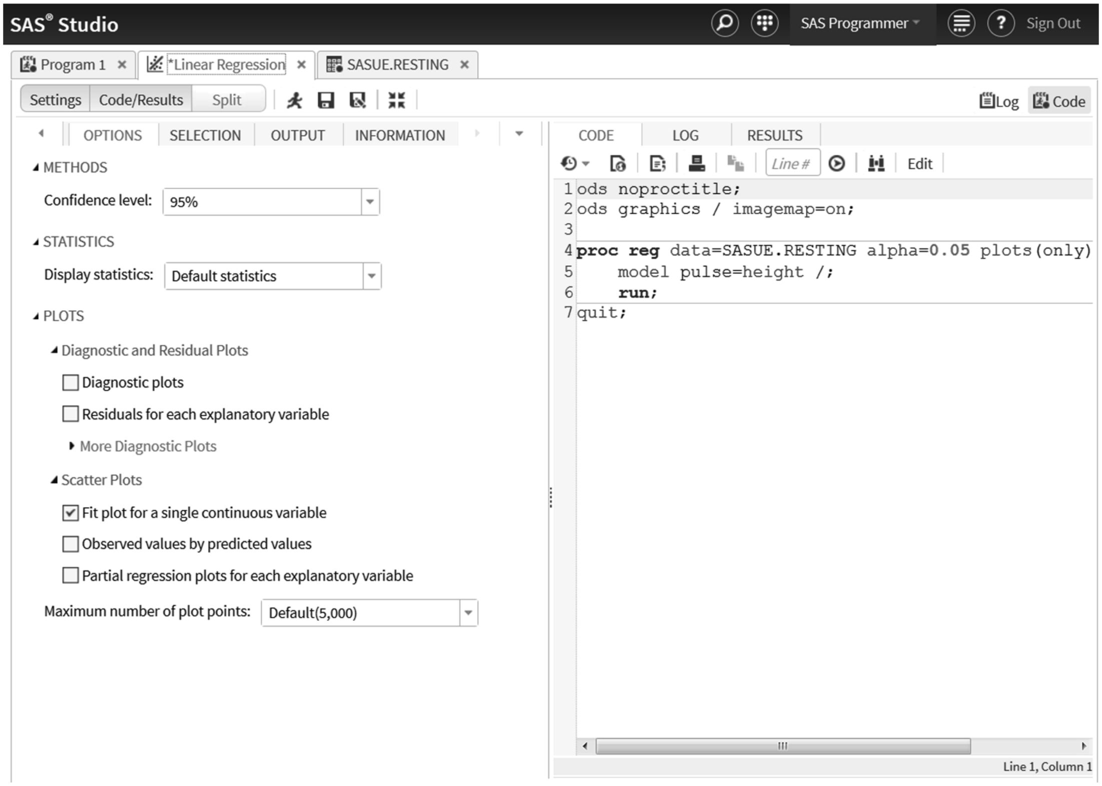
Display 1.3: Entering the Model Effect Settings



1.8.8 The Options Tab

Even though the task is now ready to run, we might want to alter some of the remaining settings, which have been set by default. In the example given in Chapter 4, we do make some changes. Display 1.4 shows these: the **Options** tab has been selected and in the **Plots** section, the **Fit plot for a single continuous variable** has been selected. **Diagnostic plots**, **Residuals for each explanatory variable**, and **Observed values by predicted values** had been selected by default, but we deselected them. Most tasks have an **Options** tab, which typically contains sections for plots and statistics. These allow you to select additional plots and statistics as well as to deselect the default ones when they are not required.

Display 1.4: Entering Additional Task Options



1.8.9 The Output Tab

Some tasks have the ability to produce new data sets as part of their results. These are usually a copy of the input data set with new variables added for such things as predicted values and residuals. The **Output** tab controls whether such a data set is produced and allows it to be named and the additional variables that it contains specified.

1.8.10 The Information Tab

The **Information** tab gives a brief description of the task and the underlying SAS procedures that it uses.

Other Tabs

The **Data**, **Options**, and **Information** tabs are the most common, followed by the **Output** tab. The Linear Regression task also has a **Selection** tab. This is specific to certain regression tasks that have the facility for selecting a subset of predictors from a larger number.

1.8.11 Abbreviations of Task Settings Used in This Book

When describing the settings needed for a particular analysis, we use an abbreviated form of instructions. The example below is the abbreviated form of the instructions for the example described in more detail above:

1. Open **Tasks** ▶ **Statistics** ▶ **Linear Regression**.
2. Under **Data** ▶ **Data**, enter **sasue.resting**.
3. Under **Data** ▶ **Roles** add **pulse** as the **Dependent variable** and **height** to the **Continuous variables** box.
4. Under **Model** ▶ **Model Effects**, select **height** and click the **Add** button.
5. Under **Options** ▶ **Plots** ▶ **Scatter plots**, select **Fit plot for a single continuous variable** and deselect the other plots.
6. Click the **Run** button.

The first instruction is to open the required task. In this case, it is the Linear Regression task and is to be found in the Statistics group of tasks. Tasks are located in the navigation pane so that pane must be visible. Occasionally, a set of instructions will begin with an instruction to reopen the task. This will be the case where a task has been run and we want to rerun it after changing some settings. In this case, all that is needed is to click on the task's tab in the work area.

The second instruction is shorthand for "enter **sasue.resting** in the Data area under the **Data** tab". Similarly, the third describes the settings for the variable roles in the **Data** tab and the fourth and fifth describe settings in the **Model** and **Options** tabs.

The final instruction is to run the task (that is, to click on the Run button or press F3).

1.8.12 The Results Pane

In Display 1.1, the pane on the right shows the results of running the task. The scroll bars on the edges of the pane indicate that the results will not fully fit within the pane at that size. Rather than using the scroll bars, a better option for viewing results is to have the work area maximised via the **Maximize View** button ( ) mentioned above. Another option is to display the results in a new browser tab by clicking the right-hand button ( ).

Each time a task is run, the contents of the results pane are replaced. Displaying the results in separate browser tabs is helpful when comparing two or more sets of results. Results can also be viewed, or saved to a file, in any of three formats, HTML ( ), PDF ( ), or RTF ( ), by clicking the appropriate button. RTF– rich text format – is a format designed for importing into word processing programmes. Settings for controlling the format and appearance of results are described below. Results can also be printed ( ).

In Display 1.1, the results pane also has tabs for **Code** and **Log**. These tabs can be toggled on and off with the **Log** () and **Code** () buttons. For the most part, we will be relying on the tasks and task settings to generate the necessary code and can therefore ignore the Code tab. However, as seen, the code pane can provide useful information if we have not entered all the settings necessary for a task. Occasionally, we will edit the code generated by the task in order to access options or settings that are not available in the task settings pane. The **Log** tab is useful for showing whether any errors or warnings were issued when the task was run. Errors are relatively rare when using the tasks, but it is worth checking the log when in doubt.

1.8.13 Options and Preferences

Various aspects of the interface can be configured via the application options button () on the title bar. Clicking this button opens a drop-down menu with sections for **Edit Autoexec File, View, Preferences, Tools**, and **Reset SAS Session**. The autoexec file contains SAS code that is run every time the SAS University Edition is started and can safely be ignored by most users. The View section allows areas of the navigation pane to be hidden. For example, once the data files have been set up for this book, the only areas of the navigation pane that will be needed are the tasks and libraries areas, so Folders, Snippets, and File Shortcuts could all be hidden.

Preferences

The **Preferences** sub-menu opens a popup window with four groups of options: **General**, **Editor**, **Results**, and **Tasks**. The Results section with its default settings is shown in Display 1.5. There are two main points to note: PDF and RTF format output can be turned off and the output style can be chosen separately for each format of output.

To turn off PDF or RTF output, deselect the **Produce PDF output** box or the **Produce RTF output** box. The corresponding button (or) in the Results pane will then be greyed out. The HTML output cannot be turned off because the software's interface is a browser interface, so HTML format is its native format. In practice, most users will want to leave RTF output on as this is the most convenient way to transfer results to a word processor document.

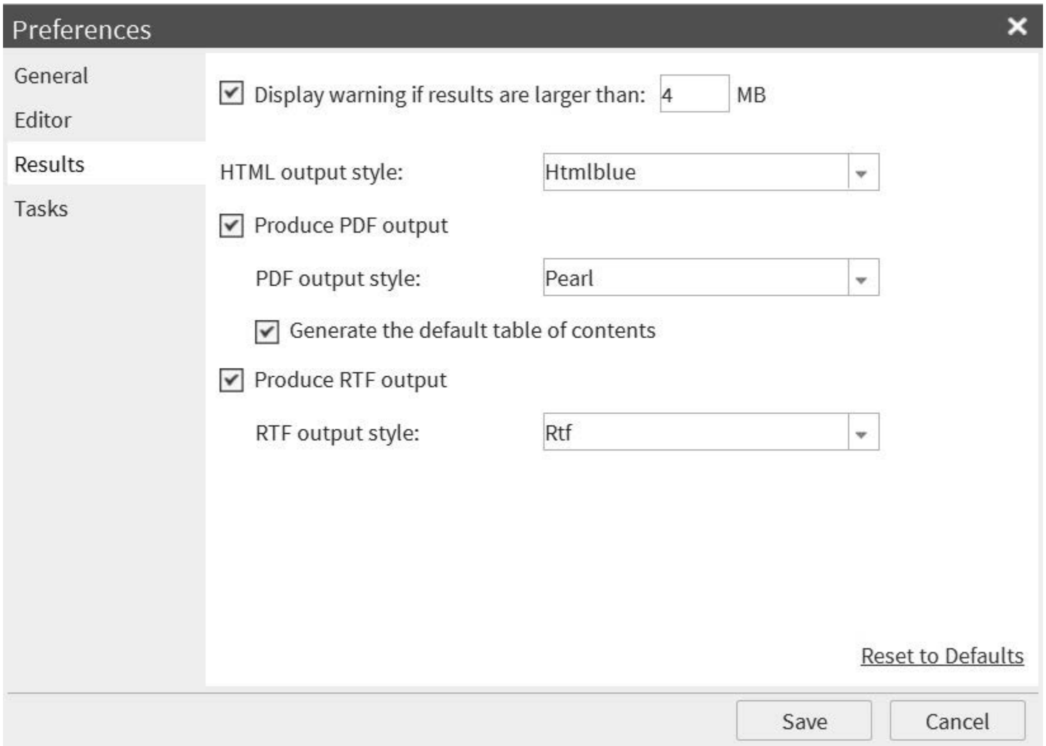
Under **Preferences ▶ Tasks**, one setting we recommend is to deselect **Generate header comments for task code.**

Output Styles

The output style determines the appearance of the output (that is, the fonts, colours, and layout of the output). Clicking the drop-down button (▼) shows the styles available. The choice of output style is largely a matter of personal preference. The **Htmlblue** style, which is the default for the Results pane, is a good general choice, as is the **Statistical** style. If black and white output is required, one of the **Journal** styles might be suitable.

As Display 1.5 shows, HTML, PDF, and RTF each have a different default style. However, there are practical advantages to having them set to the same style if, for example, results are being copied to a word processor document.

Display 1.5: The Results Section of the Preferences Menu



1.8.14 Setting Up the Data Used in This Book

Download the online material for this book, which is available at either of our author pages on the support.sas.com web site:

https://support.sas.com/publishing/authors/der.html

http://support.sas.com/publishing/authors/everitt.html

The ZIP file will contain a folder named sasdata with numerous files within it having the extension .sas7bdat. These are the SAS data sets used in this book.

The instructions for installing SAS University Edition included setting up a shared folder. The recommended path for this was `C:\SASUniversityEdition\myfolders` and the name for the shared folder `myfolders`. (If you did not create a shared folder at installation, you will need to do it now.)

1. Start SAS University Edition.

2. Under **Server files and folders ▶ My Folders**, click the **New** button ( ), select **Folder** from the drop-down menu, type **sasdata** for the name, and click **Save**.
3. Select the newly created **sasdata** folder and click the **Upload Data** ( ) button.
4. Click **Choose files**, and navigate to the place where the downloaded online materials were saved.
5. Select all sasdata files (that is, those with the extension .sas7bdat), and click **Upload**.

   (On a PC, it is also possible to simply copy the downloaded **sasdata** folder to `C:\SASUniversityEdition\myfolders`.)

Having copied or uploaded the SAS data files, we now need to assign them to a library.

1. In the navigation area under **Libraries**, click the **New Library** button ( ).
2. In the box for the name, enter **sasue**.
3. For the path, enter **/folders/myfolders/sasdata.**
4. Select **Re-create this library at start-up.**

Index

A

Add button 10
adjusted residuals 58
Agresti, A. 50, 52, 164
AIIC 158
Aitkin, M. 126
Akaike's information criteria (AIC) 158
Altman, D.G. 23, 30, 77, 78
analysis of variance (ANOVA)
  about 105
  exercises 127–132
  one-way 106–113
  two-way 113–119
  unbalanced 119–127
analysis of variance table 79
ANOVA
  See analysis of variance (ANOVA)
aspect ratio, of scatter plots 92–96

B

backward elimination 156
baiting behaviour, effect of season on, at suicides 60
bar charts 42–46
Barnett, A. 200, 201
binary response variable 171–172
binary variables 48
birthrate data, plotting 92–96
bivariate data 71
bivariate normal distribution 77
bladder cancer, Poisson regression example of 208–212
box plots
  about 19–22
  analysing cloud seeding data using 143–150
  analysing rat weight gain using 114–116
  initial analysis of wave energy data using 31–34
  teaching arithmetic data with 106–108
Bruce, R.A. 156

C

caseness data
  about 180–181
  modeling using logistic regression 181–185
categorical data
  about 41–42
  analysing frequencies 42–46
  exercises 67–69
  Fisher's Exact test 59–64
  graphing frequencies 42–46
  larger cross-tabulations 54–59
  McNemar's test example 64–66
  two-by-two tables 48–54
categorical measurements 2–3
chi-square goodness-of-fit test 47–48
chi-squared test 48–50, 56–59
cloud seeding 142–155
Cochrane-Mantel-Haenszel Test (CMH), for multiple related 2 x 2 tables 52–54
Code button 14
Collett, D. 187, 191, 195, 215
conditional likelihood function 187
confidence interval 5, 24
contingency tables
  defined 48
  tabulating brain tumor data into 56
Cook, R.D. 92
Cook's distance 140
correlation coefficient 72, 74–77
cross-tabulations 54–59

D

data
  See also categorical data
  adjusting scatter plots to show patterns in 91–96
  bivariate 71
  defined 2
  exercises 37–39
  graphical representations of 18–22
  paired 31–37

setting up 15–16
transformation of 110
Data tab 9–10
Delaney, H.D. 126, 129
Delete button 10
Der, G. 49, 50, 51, 52, 54, 140, 156, 172, 178
diagnostic plots 140
diffogram 110
Dobson, A.J. 200, 201

E

error distribution 200
Everitt, B.S. 29, 49, 50, 51, 52, 54, 57–58, 60, 65, 109, 110, 138, 140, 150, 151, 156, 158, 172, 178
examples
logistic regression 179–180
multiple linear regression 142–143
exercises
analysis of variance (ANOVA) 127–132
categorical data 67–69
data description 37–39
logistic regression 191–197
multiple linear regression 164–167
scatter plots 96–103
simple inference 37–39
explanatory variables
applying logistic regression model with single 173–177
logistic regression with all 178–179
exponential family 201

F

factorial design 116
Familial Adenomatous Polyposis (FAP), Poisson regression example of 205–208
Fisher, Ronald Aylmer 109
Statistical Methods for Research Workers 23
Fisher's Exact test 59–64
five-number summary 20
forward selection 156
Freedman, M. 101
frequencies, analysing and graphing 42–46
F-test 109–110, 118, 138

G

gamma distribution 216
General Health Questionnaire (GHQ) 179
generalized linear model (GLM)
about 199–200
applying multiple linear regression and logistic regression using 201–205
components of 200–201
GLM
See generalized linear model (GLM)
Goldberg, D. 179
graphical representations, of data 18–22
Greenwood, M. 213

H

Harberman, S.J. 58, 195
Hardy, R.J. 197
histogram 21
Hosmer, D.W. 186, 191, 194
Howell D.C. 69, 126
Htmlblue style 14
Hurvich, C.M. 158
hypotheses, testing 22–23

I

ice cream data
analysing sales 137–139
analysing using scatter plots 135–137
independent samples *t*-test 26
Information tab 12
interaction plot 118
interaction terms 150–155
interquartile range 20
interval scales 3–4

J

Journal style 14

K

Kapor, M. 128
Kelsey, J.L. 197
kinesiology 80

L

larger tables, Fisher's Exact test for 63–64
least squares 138
least squares estimation 77
Lemeshow, S. 186, 191, 194
Levene's test 110
leverage 140
Lewis, T. 18
linear model 83

linear regression, example of 80–85
linear relationship 75
Liu, Z. 52
Log button 14
logistic regression
  about 169–171, 191
  with all explanatory variables 178–179
  applying using generalized linear model (GLM) 201–205
  applying with single explanatory variable 173–177
  caseness data 180–181
  example of 179–180
  exercises 191–197
  modeling caseness data using 181–185
  odds 172
  for 1:1 matched studies 185–191
  predicting survival after surgery with 171–172
logit transformation 172

M

Mann, L. 59
matched case-control design 185
matched pairs 64–66
Maximize View button 8, 13
Maxwell, S.E. 126, 129
McCullagh, P. 200, 201, 213
McNemar's test 64–66
mean square 138
measurements
  categorical 2–3
  nominal 2–3
  ordinal scale 3
Mehta, C.R. 60
Miles, J. 100
Model tab 10–11
models
  checking assumptions 85–91
  multiple regression 139–142
mooring methods 34–36
multiple comparison test 110
multiple linear regression
  about 133–135
  analysing cloud seeding data using box plots and scatter plots 143–150
  analysing ice cream data using scatter plots 135–137
  analysing ice cream sales 137–139
  applying using generalized linear model (GLM) 201–205
  diagnosing models 139–142
  example of 142–143
  exercises 164–167
  identifying parsimonious regression models 155–164
  interaction terms and cloud seeding data 150–155

N

Navigation pane 7
negative binomial distribution 213
Nelder, J.A. 126, 200, 201, 213
New button 19
Newton metres 31
nominal measurements 2–3
nonlinear model 83
non-overlapping sums of squares 125
normal probability plot 140
null hypothesis 23
numerical summaries, analysing rat weight gain using 114–116
N-Way 116–117

O

odds/odds ratios 50–51, 172
1:1 matched studies, logistic regression for 185–191
one-sample *t*-test 24, 34–35
one-way ANOVA
  methods of teaching arithmetic 108–113
  teaching arithmetic data with summary statistics and box plots 106–108
options 14–15
Options tab 11–12
ordinal scale measurements 3
orthogonal sums of squares 125
outliers 19
output styles 14
Output tab 12
overdispersion 212–215

P

paired data, *t*-test for 31–37
paired t-test 34–35
paired t-test, checking assumptions of 36–37
parsimonious regression model, identifying 155–164
Patel, N.R. 60, 63
patterns, adjusting scatter plots to show in data 91–96

PDF output, turning off 14
Pearson residuals 57–58
Pearson's correlation coefficient 74
Peterson, D.R. 68
Piantadosi, S. 205
pie charts 42–46
Poisson regression
  about 200
  bladder cancer example 208–212
  Familial Adenomatous Polyposis (FAP) example 205–208
  overdispersion 212–215
populations 4–5
post-natal depression data
  IQ and 122–127
  summarizing 121–122
preferences 14–15
probability plot 30
product-moment correlation coefficient 74
Pugh, M.D. 69
p-value 23

Q

quadratic effect 83

R

ratio scales 4
Rawlings, J.O. 140
regression
  See also logistic regression; multiple linear regression
  simple linear 72, 77–80
  stepwise 156
regression coefficients 137–138
residuals
  adjusted 58
  defined 57
  standardized 58, 140
  studentized 140
  use of 85–91
  using 139–142
results page 13–14
Ripley, B.D. 193
R-square 79, 152, 158
RTF output, turning off 14
Run button 8, 10

S

samples 4–5
SAS University Edition
  about 5–6
  starting 6
  user interface 6–7
scales of measurement 2
scatter plots
  about 72
  adjusting to show patterns in data 91–96
  analysing cloud seeding data using 143–150
  analysing ice cream data using 135–137
  aspect ratio of 92–96
  exercises 96–103
Schwarz Bayesian Criterion (SBC) 151
season, effect of in baiting behaviour at suicides 60
Seeber, G.U.H. 208
Selection tab 12
Senie, G.U.H. 48
SGPLOT procedure 45
Shevlin, M. 100
significance test 5, 23
simple inference 37–39
simple linear regression 72, 77–80
skewed 20
Skrondal, A. 29, 151, 158
Snippet button 46
split view 8
standardized residuals 58, 140
statistical inference 22
Statistical Methods for Research Workers (Fisher) 23
Statistical style 14
statistics
  about 1–2
  summary 18–22
stepwise regression 156
studentized residuals 140
Student's t-Test
  about 22–23
  alternatives to 29–30
  applying 23–29
  checking assumptions of 29–30
styles, output 14
suicides, effect of baiting behaviour affected by season at 60

summary statistics
  about 18–22
  teaching arithmetic data with 106–108
support (web site) 15

T

task settings
  abbreviations for 13
  about 8–9
  entering 9
tasks 8–9
tests
  chi-square goodness-of-fit 47–48
  chi-squared 48–50, 56–59
  Cochrane-Mantel-Haenszel Test (CMH) 52–54
  Fisher's Exact 59–64
  F-test 109–110, 118, 138
  independent samples *t*-test 26
  Levene's 110
  McNemar's 64–66
  multiple comparison 110
  one-sample *t*-test 24, 34–35
  paired t-test 34–37
  significance 5, 23
  Student's t-Test 22–30
  t-Test 31–37
  Tukey 110
  Wilcoxon Mann-Whitney 30
  Wilcoxon signed rank 36
3 x 3 contingency table 56
transformation, of data 110
Tsai, C.L. 158
t-Test, for paired data 31–37
Tufte, E.R. 72–73
Tukey test 110
two-by-two tables
  about 48
  chi-square test 48–50
  Cochrane-Mantel-Haenszel Test (CMH) for multiple 52–54
  odds/odds ratios 50–51
two-way ANOVA
  about 113–114
  rat weight gain and diet 116–119
  rat weight gain using box plots and numerical summaries 114–116
Type I sums 125, 126
Type III sums 125, 126
Type IV sums 126

U

unbalanced ANOVA
  about 119–121
  post-natal depression and IQ 122–127
  summarizing post-natal depression data 121–122
user interface, SAS University Edition 6–7

V

Van Belle, G. 68, 156, 170
variables
  binary 48
  binary response 171–172
  defined 2
  explanatory 173–179
VBAR statement 46
Venables, W.N. 193

W

Wald statistic 202
wave power methods 34–36
weather modification 142–155
Wedderburn, R.W.M. 200
Weisberg, S. 92
Wilcoxon Mann-Whitney test 30
Wilcoxon signed rank test 36
Woodley, W.L. 142
work area 8

Y

Yule, C.V. 213

About This Book

Purpose

This book teaches students of many courses and applied research workers how to analyse data using statistical methods with SAS University Edition.

Is This Book for You?

Students of a variety of courses who need to use statistical methods to complete their coursework will benefit from this book, as well as research workers who are not statisticians but need to apply the correct statistical methods to their data using SAS University Edition.

Prerequisites

This book will be most useful if you are attending or have attended an introductory course of statistics.

What's New in This Edition

SAS University Edition includes the SAS products Base SAS®, SAS/STAT®, SAS/IML®, SAS/ACCESS® Interface to PC Files, and SAS® Studio. Teachers, students, adult learners, and academic researchers can access the SAS University Edition software for noncommercial learning purposes. You can download SAS University Edition for free, directly from SAS. After you download it to your PC, Mac, or Linux workstation, SAS works locally on your machine by using virtualization software and your browser, so no Internet access is required. Or you can get SAS University Edition for free via AWS Marketplace (AWS usage fees may apply). The software runs in the cloud - all you need is a browser and an Internet connection. For more information, visit SAS University Edition at www.sas.com.

SAS University Edition features:

- **An intuitive interface** that lets you interact with the software from your PC, Mac or Linux workstation. Learn more about SAS Studio.

- **A powerful programming language** that's easy to learn, easy to use. Learn more about Base SAS.
- **Comprehensive, reliable tools** that include state-of-the-art statistical methods. Learn more about SAS/STAT.
- **A robust, yet flexible matrix programming language** for more in-depth, specialized analysis and exploration. Learn more about SAS/IML.
- **Several time series forecasting procedures** – TIMEDATA, TIMESERIES, ARIMA, ESM, UCM and TIMEID – from SAS/ETS .
- **Out-of-the-box access to PC file formats** for a simplified approach to accessing data. Learn more about SAS/ACCESS.

Additionally, see the SAS University Edition fact sheet.
http://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-university-edition-107140.pdf

Scope of This Book

The use of SAS University Edition for applying a range of statistical techniques widely used in data analysis, and the detailed mathematics behind the statistical methods.

About the Examples

Software Used to Develop the Book's Content

SAS University Edition, version 3.4.

Example Code and Data

The book concentrates almost exclusively on using the built-in tasks so that SAS code is kept to a minimum.

All the example datasets used can be downloaded from the authors' pages at
http://support.sas.com/publishing/authors/der.html

http://support.sas.com/publishing/authors/everitt.html

Look for the cover thumbnail of this book, and select Example Code and Data to download the SAS datasets that are used in this book.

If you are unable to access the datasets through the website, send an e-mail to saspress@sas.com.

SAS University Edition

If you are using SAS University Edition to access data and run your programs, then please check the SAS University Edition page to ensure that the software contains the product or products that you need to run the code: http://support.sas.com/software/products/university-edition/index.html.

Output and Graphics Used in This Book

The output and graphics used in the book were copied from the rtf download. The style used for rtf was changed to Htmlblue so that the appearance would be similar to the onscreen appearance.

Exercise Solutions

Exercises are included at the end of each chapter.

Additional Help

Although this book illustrates many analyses regularly performed in businesses across industries, questions specific to your aims and issues may arise. To fully support you, SAS Institute and SAS Press offer you the following help resources:

- For questions about topics covered in this book, contact the author through SAS Press:
  - Send questions by email to saspress@sas.com; include the book title in your correspondence.
  - Submit feedback on the author's page at http://support.sas.com/author_feedback.
- For questions about topics in or beyond the scope of this book, post queries to the relevant SAS Support Communities at https://communities.sas.com/welcome.
- SAS Institute maintains a comprehensive website with up-to-date information. One page that is particularly useful to both the novice and the seasoned SAS user is its Knowledge Base. Search for relevant notes in the "Samples and SAS Notes" section of the Knowledge Base at http://support.sas.com/resources.
- Registered SAS users or their organizations can access SAS Customer Support at http://support.sas.com. Here you can pose specific questions to SAS Customer Support; under *Support*, click *Submit a Problem*. You will need to provide an email address to which replies can be sent, identify your organization, and provide a customer site number or license information. This information can be found in your SAS logs.

Keep in Touch

We look forward to hearing from you. We invite questions, comments, and concerns. If you want to contact us about a specific book, please include the book title in your correspondence.

Contact the Author through SAS Press

- By e-mail: saspress@sas.com
- Via the Web: http://support.sas.com/author_feedback

Purchase SAS Books

For a complete list of books available through SAS, visit sas.com/store/books.

- Phone: 1-800-727-0025
- E-mail: sasbook@sas.com

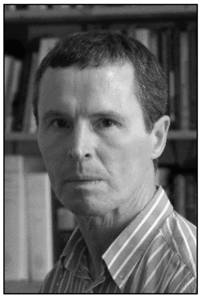
Subscribe to the SAS Training and Book Report

Receive up-to-date information about SAS training, certification, and publications via email by subscribing to the SAS Training & Book Report monthly eNewsletter. Read the archives and subscribe today at http://support.sas.com/community/newsletters/training!

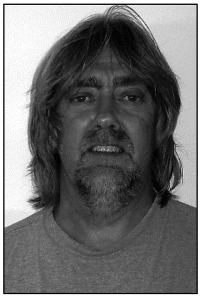
Publish with SAS

SAS is recruiting authors! Are you interested in writing a book? Visit http://support.sas.com/saspress for more information.

About These Authors



Geoff Der is a research statistician in the Social and Public Health Sciences Unit at the University of Glasgow, Scotland.



Brian Everitt is Professor Emeritus at King's College, London.

Geoff and Brian have jointly authored *Basic Statistics Using SAS® Enterprise Guide®: A Primer, Applied Medical Statistics Using SAS®, A Handbook of Statistical Analyses Using SAS®, Third Edition,* and *Statistical Analysis of Medical Data Using SAS®*.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:
http://support.sas.com/der
http://support.sas.com/everitt