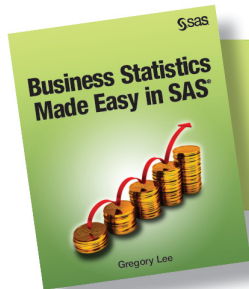


Business Statistics Made Easy in SAS®



Gregory Lee



From *Business Statistics Made Easy in SAS*®.
Full book available for purchase [here](#).

Contents

<i>Preface</i>	<i>ix</i>
<i>About the Author</i>	<i>xv</i>
<i>Acknowledgments</i>	<i>xvii</i>
Chapter 1 • Introduction to the Central Textbook Example	1
Introduction	1
The Company	2
Current Research Needs of the Company	2
Your Brief for the Case Example	5
Extended Analytical Skills Needed in the Project	6
Chapter 2 • Introduction to the Statistics Process	9
Introductory Case: Big Data in the Airline Industry	9
Introduction to the Statistics Process	11
Step 1: Your Needs & Requirements	12
Step 2: Getting Data	13
Step 3: Extracting Statistics from the Data	15
Step 4: Understanding & Decision Making	17
Summary: Challenges in the Statistics Process	17
Advice to the Statistically Terrified	18
Chapter 3 • Introduction to Data	21
Introductory Case: Royal FrieslandCampina	21
Brief Introduction to Samples, Populations & Data	23
Basic Characteristics of Variables	27
Chapter 4 • Data Collection & Capture	33
Introduction	33
Correct Sampling	34
Choose Constructs and Variable Measurements	35
Initial Data Capture: Which Package?	43
Dealing with Data Once It Has Been Captured	43

Database & Data Analysis Software	48
Some Complications in Datasets	48
Chapter 5 • Introduction to SAS®	51
Introductory Vignette: SAS On Top of the Analytics World	51
Brief Introduction to SAS	52
Introduction to the Textbook Materials	53
Getting Started with SAS 9 or SAS Studio	53
Chapter 6 • Basics of SAS Programs, Data Manipulation, Analysis & Reporting	69
Introduction	70
The Running Data Example	70
The Pre-Analysis Data Cleaning & Preparation Steps	72
Overview of the Three Big Tasks in Business Statistics	73
Basic Introduction to SAS Programming	73
Major Task #1: Data Manipulation in SAS	77
Major Task #2: Data Analysis	83
Major Task #3: SAS Reporting through Output Formats	84
The Visual Programmer Mode in SAS Studio	86
Conclusion	88
Chapter 7 • Descriptive Statistics: Understand your Data	89
Introductory Case: 2007 AngloGold Ashanti Look Ahead	90
Introduction	91
End Outcome of a Descriptive Statistics Analysis	91
Getting Descriptive Statistics in SAS	92
Statistics Measuring Centrality	94
Basic Statistics Assessing Variable Spread	97
Assessing Shape of a Variable's Distribution	99
Conclusion on Descriptive Statistics	104
Appendix A to Chapter 7: Basic Normality Statistics	104
Chapter 8 • Basics of Associating Variables	109
Introduction	109

What is Statistical Association?	110
Association Does Not Mean Causation	110
Overview of Associations for Different Variable Types . .	111
Relating Continuous or Ordinal Data:	
Correlation & Covariance	112
Relating Categorical Variables	119
Chapter 9 • Using Basic Statistics to Check & Fix Data	123
Introduction	123
Inappropriate Data Points	124
Dealing Practically with Missing Data	126
Checking Centrality & Spread	127
Strange Variable Distributions	128
Dealing Practically with Multi-Item Scales	128
Chapter 10 • Introduction to Graphing in SAS	135
Introduction	135
Major Graphing Procedures in SAS	136
The PROC SGPLOT Routine in SAS	138
Multiple Plots Simultaneously through	
PROC SGPANEL	143
Business Dashboards through PROC GKPI	143
Geographical Mapping Using PROC GMAP	145
PROC SGSCATTER for Multiple Scatterplots	146
Conclusion on SAS Graphing	147
Chapter 11 • The Statistics Process: Fitting Models to Data	149
Introduction	149
Look for Patterns in the Data (Fit)	151
Step 3: Interpret the Pattern	164
Summary of the Statistics Process	168
Chapter 12 • Key Concepts: Size & Accuracy	171
Illustrative Case: Pharmaceuticals I –	
AstraZeneca’s Crestor	172
Introduction	173

Issue # 1: Size of a Statistic	173
Issue # 2: Accuracy of Statistics	177
The Aspects of Inaccuracy	179
Putting Statistical Size and Accuracy Together	200
Conclusion	202
Appendix A to Chapter 12: More on Accuracy (optional)	203

Chapter 13 • Introduction to Linear Regression	211
Illustrative Case: West Point	212
Introduction	213
The Core Textbook Case Example for Chapter 13	213
Introduction to Linear Regression	215
A Pictorial Walk through Regression	217
Implementing Multiple Regression in SAS	226
Step 1: Collect, Capture and Clean Data	227
Step 2: Run an Initial Regression Analysis	231
Step 3: Assess Fit and Apply Remedies If Necessary . .	233
Step 4: Interpret the Regression Slopes	257
Step 5: Reporting a Multiple Regression Result	265
Other Statistical Forms	266
Conclusion	267

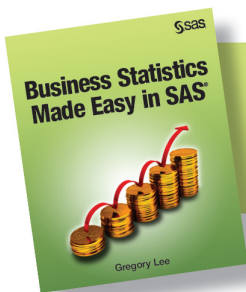
Chapter 14 • Categories Explaining a Continuous Variable:

Comparing Two Means	269
Introduction to Comparison of Categories	270
Features of the Continuous Variable to Compare Across Categories	270
Two Types of Categories to Compare	271
Numbers of Categories to Compare: Two vs. More than Two	272
Data Assumptions and Alternatives when Comparing Categories	273
Comparing Two Means: T-Tests	275

Comparing Means for More than Two Categories: ANOVA	284
Chapter 15 • Categorical Data Distributions & Associations	285
Introduction	285
Repeat: One-Way Categorical Distributions	286
Repeat: Linking Categorical Variables Together	287
Further Statistical Questions about Categorical Data	287
Assessing One-Way Frequencies	288
Tests of Categorical Variable Association	293
Conclusion on Categorical Data Analysis	298
Chapter 16 • Reporting Business Analytics	299
Reminder - Your Brief for the Textbook Case Study	299
Your Tasks in the Analytics and Reporting Stages	300
Background Analyses Versus Displayed Reports for the CEO	300
Conclusion on Business Statistics Reporting	308
Chapter 17 • Business Analysis from Statistics: Introduction	309
Case Study: Oracle South Africa	310
Introduction	311
Overall Financial Extrapolation Process	312
Step 1: Statistics Gives Level of or Change in Focal Variables	313
Step 2: Financial Estimates of Revenue or Cost of One Unit	314
Step 3: Combine Statistics with Per-Unit Financial Values	318
Step 4: Include Scope	319
Steps 5 and 6: Net Profitability Calculations	319
Some Simple Examples of Business Extrapolation	321
Conclusion of Statistical Business Extrapolation	323
Chapter 18 • Miscellaneous Business Statistics Topics	325
Introduction	326

Big Data	326
Data Warehousing	330
Machine Learning & Algorithms	335
Simulation in Business Situations	336
Bayesian Statistics	340
Conclusion	342
Chapter 19 • Bibliography	343
Books and Articles	343
<i>Index</i>	351

From *Business Statistics Made Easy in SAS®*, by Gregory John Lee. Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Business Statistics Made Easy in SAS*®.
Full book available for purchase [here](#).

69

6

Basics of SAS Programs, Data Manipulation, Analysis & Reporting

Introduction	70
The Running Data Example	70
Reminder of the Main Textbook Case Study	70
Reminder of Your Brief for the Case Example	71
The Pre-Analysis Data Cleaning & Preparation Steps	72
Overview of the Three Big Tasks in Business Statistics	73
Basic Introduction to SAS Programming	73
Running SAS Tasks through Point-and-Click Windows	73
Doing SAS Tasks through Programming Code (Syntax)	74
Major Task #1: Data Manipulation in SAS	77
Introduction to Data Manipulation	77
Creating New Datasets in SAS	78
Creating Temporary Datasets in the Work Library	79
Create New Variables or Manipulate Current Variables in SAS	80
Combining Datasets	82
Major Task #2: Data Analysis	83
Major Task #3: SAS Reporting through Output Formats	84
Introduction	84
Different ODS Outputs in SAS Studio	84
Different ODS Outputs in SAS 9	85

<i>The Visual Programmer Mode in SAS Studio</i>	86
<i>Conclusion</i>	88

Introduction

This chapter begins the many sections of this book that teach the practical implementation of statistical techniques through SAS. We start in this chapter with an overview of SAS programs and programming, data manipulation, the basics of SAS statistical analysis, and different types of documentary reports in SAS.

The Running Data Example

Reminder of the Main Textbook Case Study

To facilitate the discussion of the next few chapters, I will continue to work with the Accu-Phi case study from Chapter 1, specifically with the following variables (see Figure 6.1 on page 71 below for a reminder of the initial data format):

- *Sales*: Measured as actual services sales in dollars in the first year of sales.
- *License*: A description of what license the customer has (“Freeware” or “Premium”).
- *Size*: A description of the size of the customer by turnover, with the character values “Small,” “Medium,” or “Big.”
- *Trust*: The trust the customer has in your product and company. You have measured trust through four questions in an online survey, on a 0-100 point sliding scale.
- *Customer satisfaction*: Measured through four questions in an online customer survey, but from 1-7.
- *Enquiries*: The average number of enquiries about the core software product logged with the call center or online help by customers, per month, since starting use of the product.

Figure 6.1 First lines of initial dataset

	Respondent	License	Size	Trust01	Trust02	Trust03	Trust04
1	1	Freeware	Small	60	60	55	65
2	2	Premium	Big	100	100	80	100
3	3	Freeware	Big	70	64	84	83
4	4	Freeware	Big	67	75	77	70
5	5	Freeware	Bigg	70	55	55	56
6	6	Premium	Medium

Below is the second half of variables: they are separated here due to width

Satisfaction01	Satisfaction02	Satisfaction03	Satisfaction04	Enquiries	Sales
6	6	6	5	16	\$58,346.00
5	5	55	5	19	\$144,175.00
4	5	5	6	16	\$88,764.00
6	6	6	6	21	\$81,777.00
6	5	6	5	12	\$84,403.00
.	.	.	.	14	\$110,458.00

The download available on the course website in the “Textbook Materials” folder, gives this initial dataset (“Data01_Initial”).

Reminder of Your Brief for the Case Example

Let us say that your CEO wants you to analyze the data and answer the following questions which are important to the company:

- How did the first-year sales go?
- Are our customers satisfied and to what extent do they trust us?
- How many enquiries do customers make?
- Do sales, satisfaction, trust or enquiries differ depending on whether the customer has a premium or freeware contract, and depending on customer size?
- What is the distribution of licenses between the levels of size?
- Is sales seemingly substantially associated with any of the other variables?

The Pre-Analysis Data Cleaning & Preparation Steps

Before actually analyzing data to answer questions such as the CEO's queries above, you will need to assess the data for integrity, clean any obvious errors and mistakes, and prepare the data for final analysis. These checks may include:

- 1** *Initial data assessment and cleaning.* Notice the following in Figure 6.1 on page 71:
 - a** Size of the fifth respondent is captured as "Bigg," obviously a typographical error.
 - b** The "Satisfaction03" score for Respondent 2 is captured as a "55", but this is supposed to be a 1-7 scale.
 - c** These are data entry mistakes. While easy to spot in such a small with the eye, you'll not see this in a bigger table easily. Mis-entered data can seriously impact any analysis.
- 2** *Missing data:* There is missing data; we need to assess and possibly deal with this as discussed in Chapter 4.
- 3** *Multi-item scales assessing trust and satisfaction:* We need to assess and aggregate these into single measures of the variables if possible.

We need to pre-assess and clean our data. We usually do these sorts of assessments through basic descriptive statistics and variable associations. Therefore, the next four chapters will sequentially discuss the following:

- Chapter 6 discusses how to create, change and manipulate data, as well as give an overview of some other topics. To do things like create aggregated variables from multi-item scales, we'll need these skills.
- Chapter 7 discusses the essential descriptive statistics we use for single variables.
- Chapter 8 discusses basic measures of variable association.
- Chapter 9 discusses using these analyses in an initial set of steps for the purposes of data checking, cleaning, and preparation.

Overview of the Three Big Tasks in Business Statistics

Having been introduced to the SAS products in the previous chapter, we now turn our attention to a basic introduction to the three major types of tasks you may wish to perform in SAS:

- 1 *Data manipulation* tasks are those where you wish to change or add to your current data set. For instance, you may wish to sort your current dataset by some variable, or add a new column of data that is the sum of three other columns. Appendix A to this chapter gives you some lessons on how to do such tasks, including manipulating data and creating new datasets.
- 2 *Data analysis* involves generating representative numbers or pictures of the data that tell you something you wish to know about the data. This could range from an analysis as simple as the average of a variable to complex analysis of the relationships between many variables.
- 3 *Reporting* obviously means formatting your findings into a useful report that will be appropriate and engaging for the user.

The next sections introduce each of these major steps in greater or less detail, after an initial overview of SAS programming in general.

Basic Introduction to SAS Programming

Running SAS Tasks through Point-and-Click Windows

You can use various point-and-click windows to perform tasks in SAS. This method is relatively simple to use, and favored by many people. If you were using the point-and-click options you could open and use SAS products that work like this, such as *SAS Enterprise Guide* or *JMP*. *SAS Studio* also has a version of this sort of approach built in, called the "Visual Programmer."

Point-and-click has serious disadvantages, however, because there are often a great number of check boxes and options, and SAS does not remember your settings. Therefore, every time you re-start a certain section of SAS you have to re-enter many check box options. For this

reason, we will not use the point-and-click options very much in this book, as they are very slow and inefficient.

Doing SAS Tasks through Programming Code (Syntax)

Advantages of Programming Code

Instead of point-and click, SAS usually uses programming code in the SAS 9 Editor window or the SAS Studio Code window to input keywords that tell SAS what you want. Note the following about programming code in general:

- 1 *Programming is efficient:* The programming code input method is very efficient and advantageous. It is far quicker than using point-and-click. You can save programming code for later use more easily than you can in many point-and-click programs. Finally, point-and-click takes a lot of time to go through if you are in a classroom teaching situation, whereas opening and running a programming code file is quick.
- 2 *Saving and re-using programming code:* You can save your programming code files and re-use them time and time again (see for instance the programming code files in the “Textbook Materials” folder). Generally, once you have the programming files you like to use, the only thing you have to do is change the names of the datasets and variables.
- 3 *This book mostly uses programming code:* Because of the advantages of programming code, I will mostly use and teach this input method in this book. You will not have to learn what programming to use; the textbook comes with pre-written programming code files (see the “Textbook Materials” folder at <http://support.sas.com/publishing/authors/lee.html>). Each time we run an analysis, you will be directed to open and run a pre-existing file as described below.

First Lessons on SAS Programming

Programming can be a daunting task for many people. However, it is actually a very easy language simply composed of a few keywords, as well as a basic structure to which you need to stick.

For instance, take a look at Figure 6.2 on page 75, which shows an example of SAS code in either the Editor window of SAS 9 or the Code window of SAS Studio. Here, you can see various keywords and variable names that tell SAS what dataset to analyze, which variables to analyze, and what statistical analysis to do on these variables.

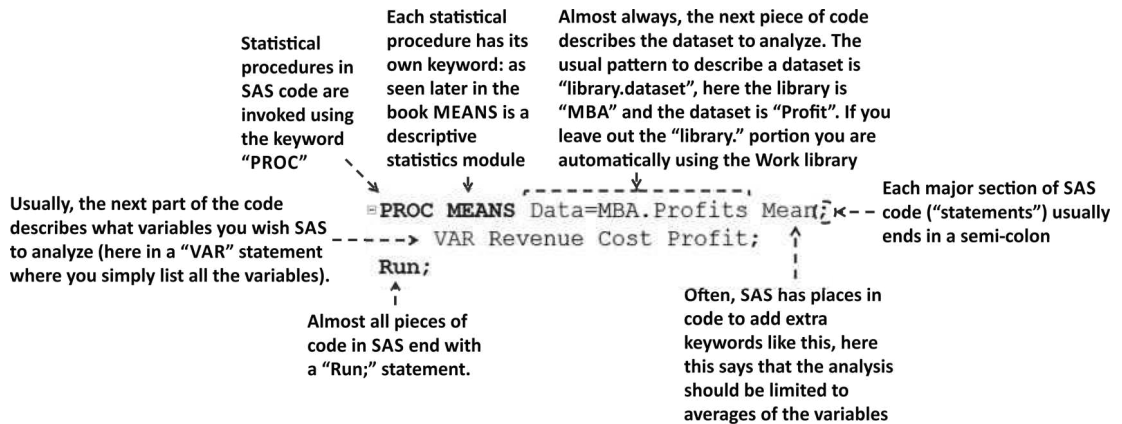
Figure 6.2 Example of programming code in a SAS Editor or Code window

Figure 6.2 on page 75 is a specific type of code that runs a statistical analysis. We can see the following in this figure:

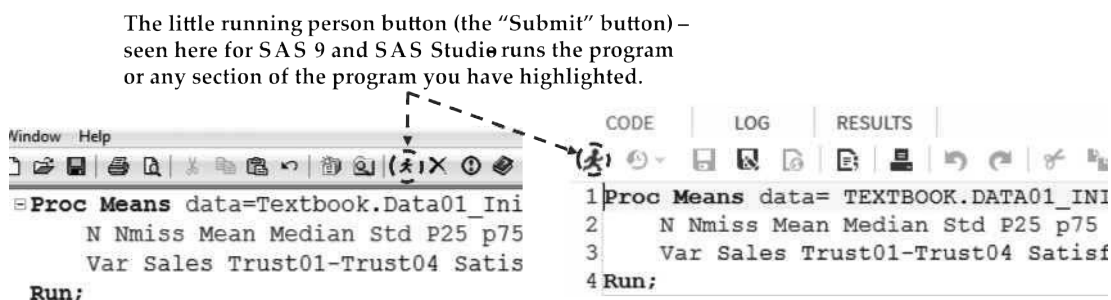
- 1 To run a SAS statistical procedure, you usually start with the keyword PROC followed by a specific keyword that identifies which particular statistical analysis you want. For instance, in Figure 6.2 on page 75 the keyword MEANS asks SAS to do basic descriptive statistics on variables, as described in later chapters.
- 2 When running procedures, we next usually identify the dataset to be analyzed by its library and then its dataset name, i.e. the general structure is "<Name of the library>.<Name of the dataset>." In Figure 6.2 on page 75, the dataset to be analyzed is the "Profits" dataset within the "MBA" library, as identified by the "`Data=MBA.Profits`" part of the code.
- 3 There are often extra keywords to identify further statistical options.
- 4 Usually, the middle section of SAS procedure code contains a description of the variables to be analyzed. In the simple example in Figure 6.2 on page 75, we simply list the variables to be analyzed after the keyword VAR. In more complex procedures that are mostly beyond the scope of this book, we sometimes also have to tell SAS how the variables are related.

There are also certain general SAS programming rules that can be seen in the example in Figure 6.2 on page 75:

- 1 *Capitalization of words in SAS code:*
 - a SAS programs usually do not care about capitalization of words. For instance, in Figure 6.2 on page 75, keywords such as "Proc Means" could easily be spelled "PROC means" or any combination of lower and uppercase, as can the dataset names.

- b** Almost the only time that SAS cares about capitalization is if you are referring to specific text data within a dataset. For instance if “Gregory Lee” is a field in a dataset, then if you need to refer to this data in code, you must get the exact capitalization correct.
- 2 Spacing, lines and tabs in SAS code:**
- a** It does matter that you keep at least one space between different keywords of SAS programming (e.g. you can’t put “PROCMEANS” above).
 - b** However, other than that, SAS does not mind where in the code or editor window you place code so long as the basic statements are in the right order. You can place different statements on different lines, run them together without line breaks, or use multiple spaces or tabs between pieces of code, etc.
- 3 Semicolons as the key for endings of sections:** Sections of SAS programs end with a semicolon (“;”). If you try to run a SAS program and find that it does not work, it is often because you have failed to add the semicolon at the end of a section.
- 4 The Run command as the key for the end of a program:** SAS programs usually end with a “Run;” command.
- 5 Running a SAS program:** To actually make the program run, you click the little running person icon in the SAS 9 or SAS Studio toolbar, as seen in Figure 6.3 on page 76 below.

Figure 6.3 Running a SAS Program



One cardinal rule is to always check the SAS log after running code to see if the program has worked and to determine if there are errors (e.g. misspelling the dataset name). In such cases, SAS will warn you in the log with red error sections. This is particularly easy in SAS Studio, which lists any errors at the top of the log section.

Finally, note that “PROC”-type code to invoke SAS statistical analyses are not the only form of programming. Notably, the very important DATA keyword is used to create and manipulate datasets, as described below in “Major Task #1: Data Manipulation in SAS” on page 77.

Opening Existing SAS Code Files

As I have discussed above, this book does not expect the reader to become a SAS programmer immediately. All the analyses taught in the book are given to you as pre-written programming code files that you simply have to open and run to get the results. As you work with these files, you will quickly see how the underlying programs work, and soon be able to apply them to your own datasets and variables with little change.

Even if you were to write your own programs from scratch, you would usually save the code files and then re-open and run them later when you wish to recreate the analysis.

To open existing programming code files like those in the “Textbook Materials” folder, do the following:

- *In SAS 9*, go to File > Open Program and navigate to where the file is stored on your hard drive.
- *In SAS Studio*, go to the Server Files and Folders section, and open the code file by double clicking on it (for instance, see the many code files in the “Textbook Materials SAS Studio” folder).

As mentioned in the chapter introduction, there are three big tasks in SAS, namely, data manipulation, data analysis, and report generation. The following sections discuss these steps further.

Major Task #1: Data Manipulation in SAS

Introduction to Data Manipulation

Data manipulation – in other words, changing data or creating new data – is one of the most important tasks in practical business statistics. After capturing data, it is rarely the case that the initial sheet or database query is completely perfect for analysis. Often, changes need to be made, for various reasons such as:

- Imperfections in the original data that need to be fixed
- The need to add new data
- The need to combine multiple datasets

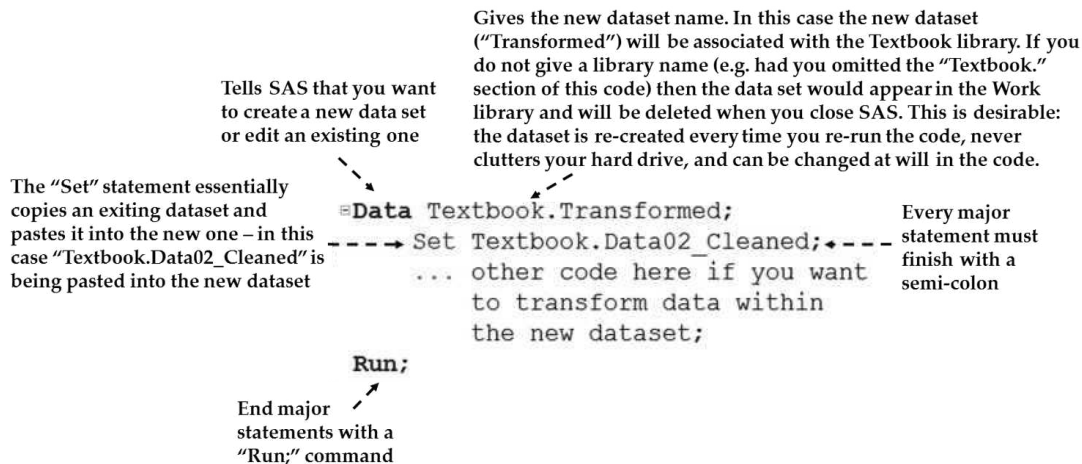
While you can manipulate data in more basic spreadsheet programs like Microsoft Excel, you can also do so in SAS, and far more simply, flexibly and reliably. This book cannot cover much of the SAS data manipulation universe, which is enormous and world-leading. The next few sections can cover only a few salient topics. For a broader introduction to these topics, the reader should consult texts such as Delwiche & Slaughter (2012).

Creating New Datasets in SAS

As a first topic, we often create new datasets in SAS programming code. This section discusses the basics of doing this.

Of course, one way to create new datasets in SAS is to import them from elsewhere, such as importing Microsoft Excel files. Chapter 5 describes how to do this. This chapter is more interested in dealing with data once it is in SAS. To create or manipulate data in SAS we use a “DATA” statement. Figure 6.4 on page 78 shows the outline of a data step for creating a new dataset SAS.

Figure 6.4 *Creating a new dataset in SAS*



As seen in Figure 6.4 on page 78, if you wish to create a new dataset you do the following:

- 1** *Start with the keyword DATA*, which tells SAS that you wish to create a new dataset.
- 2** *Name the new dataset*. Note the following:
 - a** Specify the name of a library and a dataset name, separated by a period (e.g. “Textbook.Transformed” in Figure 6.4 on page 78). The new SAS dataset will appear in the physical folder you have associated with this library. Of course, you have to have associated this library name with the folder beforehand, as described in Chapter 5.
 - b** There are basic rules for naming SAS datasets. This can be any name – in the code above we used the name “Transformed” – so long as it follows these rules:
 - i** A SAS name can contain from one to 32 characters.
 - ii** The first character must be a letter or an underscore (_).

- iii Subsequent characters must be letters, numbers, or underscores.
 - iv Blanks cannot appear in SAS names. If you want to separate parts of the dataset name, use underscores, e.g. "Dataset_03."
 - c If you leave out the library name and give only a dataset name (e.g. the "Data Transformed;" line in Figure 6.5 on page 81 below) then the new dataset will be created in the special "Work" library. In other words, calling the dataset "MyData" is the same as calling it "Work.MyData". The "Work" library is automatically created as part of the SAS installation, and I explain it in more detail in the next section. Using this option is often desirable.
 - d If you choose the same name and library as an existing dataset, then you will overwrite (i.e. replace) the original version of the dataset.
- 3** *Populate the new dataset with initial data.* There are two main choices here:
- a *Populate the new dataset with data from another dataset.* We frequently base the new dataset on the data from an existing dataset. Think of this as a copy and paste, i.e. you are copying data from an existing dataset into your new dataset. As seen in Figure 6.4 on page 78, we can do this by putting the line "*SET <name of existing dataset>;*" into a DATA step. In Figure 6.4 on page 78, we are using the SET statement to copy all the contents of the "Data02_Cleaned" dataset into the new "Transformed" dataset. In this code, both datasets are located in the "Textbook" library.
 - b *Enter raw data directly into SAS.* You can also enter data literally in SAS in a DATA step. This book will not cover this direct data input option. I personally advocate importing initial raw data from a spreadsheet program such as Microsoft Excel.
- 4** *If desired, manipulate the data.* In the DATA step, we can manipulate the data in a great number of ways. "Create New Variables or Manipulate Current Variables in SAS" on page 80 below describes more on such steps.
- 5** *Other programming notes:* As seen in Figure 6.4 on page 78, do not forget to place semicolons between major statements and add a "Run;" statement at the end before running.

Creating Temporary Datasets in the Work Library

The previous section noted that if you do not give a library name as part of a dataset name then you are automatically linking the dataset with the special "Work" folder (so specifying "Profits" is the same as saying "Work.Profits").

The Work library has a special property: all datasets contained within it are deleted when you close SAS. This is desirable in many cases for two major reasons:

- Datasets created in the Work folder do not clutter your hard drive or server, as they are deleted once you close SAS. However, because you can save the code used to create

them, these datasets can be re-created every time you re-run the code. Programming code takes up far less space on a computer than data.

- If you keep your original data and copy it to a Work library dataset, then changes you make to the new dataset do not affect the original data, which means you are never at risk of harming your original dataset.

This method of creating datasets out of programming code only for the duration of your session – and analyzing the temporary data as you need - is highly efficient and often used by SAS analysts.

On the other hand, giving a SAS library name other than Work causes the dataset to be stored permanently in the folder associated with that library. This is, of course, desirable in cases where you do wish to maintain a permanent copy.

Create New Variables or Manipulate Current Variables in SAS

There are many situations in business statistics where you wish to create a new variable that is, in effect, a transformation of an existing variable's data. Here are some initial examples:

- Creating an index such as a financial ratio (such as creating a price/earnings ratio from two columns containing price and earnings data, respectively).
- Creating mathematical transformations of variables, such as a new variable that is the square root or log of another variable.
- Using the birthdates of people to create a new column that, on a consistently updating basis, calculates their ages.

In addition, you can change and manipulate existing variables in SAS.

In our main textbook example, so far we have two major types of such tasks:

- 1 *Creating new variables that reverse the data of reverse-worded survey questions.* Specifically, Satisfaction04 is a reverse-worded survey item (see Chapter 4 and Chapter 9 for more on this), which required us to create a new variable that reverses its data.
- 2 *Creating two new factor variables, which are the aggregation of multi-item scales.* Trust and satisfaction ultimately needed to be created as factors which are an average of the individual multi-item scores. (Of course, we can't do this step without having assessed internal reliability. Again, see Chapters 4 and 9).

One of the many things SAS is brilliant at is data manipulation. You can manipulate data by using the SAS point-and-click interfaces like SAS Enterprise Guide, but it is quicker and easier to use code in programs like SAS 9 or SAS Studio. The DATA step in SAS not only creates new datasets or edits existing ones, but manipulates data columns or rows.

Figure 6.5 on page 81 shows a sample SAS data step in which the new dataset is created based on an existing dataset (specifically, we create a dataset called "Transformed" in the Work library because no library is specified, and we copy and paste everything from the Textbook.Data02_Cleaned dataset using the SET statement).

Figure 6.5 Example of creating new variables in the SAS DATA step

```

Data Transformed;
  Set Textbook.Data02_Cleaned;
  /*Some basic transformations*/
  Rev_Satisfaction04 = 8-Satisfaction04;
  Trust = mean(of Trust01-Trust04);
  Satisfaction = mean(of Satisfaction01-Satisfaction03);
  /*Mathematical transformations*/
  LogSales = Log(Sales);
  SalesSqu = Sales**2;
  /*New variables created based on values in old variables*/
  if License = "Premium" then Premium = 1;Else Premium = 0;
  if Size = "Small" then Small = 1;Else Small = 0;
  if Size = "Medium" then Medium = 1;Else Medium = 0;
Run:
  
```

Annotations:

- Creates new variable that is the reverse of Satisfaction04** → `Rev_Satisfaction04 = 8-Satisfaction04;`
- IF statements identify the condition for something to occur** → `if License = "Premium" then Premium = 1;Else Premium = 0;`
- Note: SAS does not care about capitalization unless you are trying to refer to character-based values within the data itself in quotation marks like this – in that case you must use the same capitals as the original values**
- In this case the new Transformed dataset will exist in the Work library (because you only give a dataset name no library name): it will delete when you close SAS** (points to `Set Textbook.Data02_Cleaned;`)
- Creating new variables that are the average (mean) of other variables** (points to `Trust = mean(of Trust01-Trust04);` and `Satisfaction = mean(of Satisfaction01-Satisfaction03);`)
- New variables as mathematical transformations of others: the natural log and square of Sales respectively** (points to `LogSales = Log(Sales);` and `SalesSqu = Sales**2;`)
- Here the new variable is being created and values chosen depending on what Size is in the row. Here, the Medium column will have values of "1" for all medium-sized firms and 0 otherwise** (points to `if Size = "Medium" then Medium = 1;Else Medium = 0;`)

Then, each subsequent line creates a new variable:

- We create a new variable called "Rev_Satisfaction04" that takes the data from the existing variable Satisfaction04 and reverses it using the principles discussed in Chapter 4.
- We create new variables called "Trust" and "Satisfaction" that are averages of some of the individual currently existing multi-item scale columns. Note the way the average works. Also, note here that I have only averaged the values for Satisfaction01-Satisfaction03; see Chapter 9 a little later for why.
- We create two new mathematical transformations of the Sales variable, one the natural log and one for the square (each Sales number to the power of two).
- We create several *conditional* variables using the IF-THEN concept, where the new variable only takes on a certain value if a given condition is true. In the first of these, we create a new variable called "Premium" that will have the value 1 whenever the currently existing License variable contains the value "Premium" in a row, and takes the value 0 for all rows where License is not "Premium."

Take note of the following programming notes about this sort of programming:

- Take another look at the IF-THEN statements in Figure 6.5 on page 81. Note here that this is the only situation in which capitalization counts in SAS. Take the example of the *if License = "Premium"* section of the code. Here, we are asking SAS to go look in the dataset for all rows where this exact condition is true *including the exact capitalization of*

“Premium,” and then apply the result only in those rows. If there are also entries in the License column spelled “premium” then the above condition will not identify these rows. So, be careful of capitalization in these situations only.

- As always, note that all statements are separated by semicolons and the entire set ends with a “Run;” statement.

You could do so much more. For instance, you could create a new variable that is the sum of other variables (replace MEAN in the above code with SUM). You can identify rows to delete based on certain rules. SAS has an almost endless set of possible variable manipulations – see the SAS helpfiles (notably SAS/STAT 13.2 User’s Guide) for more.

Once you have told SAS what you want to do, submit the code using the Run button as seen above. Once you have done so, always check the log for errors and always open the new dataset to check that it is right. (And then close it: an open dataset in SAS cannot be replaced).

You can see the code from this section in the textbook resources files, under “Code06 Manipulating data example.”

Combining Datasets

Often in the business world, we need to combine two or more datasets together. You can combine datasets side-by-side, one on top of the other, merge them based on a match in a certain variable, and so on.

Let us look at one of the most common examples: match merging. Imagine you are an organization with the following two datasets:

- 1 A database of customer account data, where each customer is identified by a unique customer number.
- 2 A different database of customer satisfaction survey data. Again, each customer’s survey responses are identified by the customer number. Typically, only a limited subset of customers would have filled in the survey.

Now, let us say that you wish to combine these two datasets so that you can link the data. Each row needs to be matched up by customer number. You can do this in SAS using the MERGE statement. See the following example:

Example Code 6.1 *Example of merge matching data in SAS*

```
Data Customers.Merged;
    Merge Customers.Accounts Customers.Survey2016;
    By Customer_ID;
Run;
```

There are many nuances and complexities to combining datasets – for instance, to match merge by a common variable as I show above, both datasets must be sorted by the common matching variable (i.e. you would have to sort both of the above datasets by Customer_ID). For more on combining datasets, reference the SAS helpfiles or books such as Delwiche & Slaughter (2012).

This basic understanding of SAS data manipulation will help us in various parts of the rest of the book, since data manipulation is frequently required in statistical analysis.

Major Task #2: Data Analysis

“Basic Introduction to SAS Programming” on page 73 above discussed the basics of programming a PROC step in SAS, which is the foundation of SAS statistical analyses. The rest of this book gives various examples of core SAS statistical analyses in the context of business.

Just a few more general points apply to thinking about SAS data analyses:

- Knowing which analysis is the appropriate one for your situation is obviously critical. This book discusses many introductory analyses to help you begin this journey. However, especially when you are entering into more complex modelling, you should first carefully investigate the general ideas behind what the correct analysis is. Thereafter, you can read up on how SAS implements that specific analysis through code.
- You can easily find prior examples of SAS code for your desired analysis in the SAS helpfiles, online through SAS User Group articles or the like, or in books like this one. Then, you can copy the code developed in those sources and simply change the names of the dataset and variables for your particular analysis. In a similar vein, SAS Studio has pre-written code in the Tasks section.
- Often, in the same SAS program, we will first manipulate data and then – immediately below the DATA step – place the PROC step that references and analyses the dataset created above. We can then run the set together, change the data or analysis steps again if required, and so on. Example Code 6.2 on page 83 is an example.

Example Code 6.2 *Example of running DATA and PROC steps together*

```
Data Transformed;
    Set MBA.Profits;
    LogRevenue = Log(Revenue);
Run;
Proc Means data=Transformed;
    Var LogRevenue Cost Profit;
Run;
```

Major Task #3: SAS Reporting through Output Formats

Introduction

In the early days of its development, SAS reproduced statistical reports in very simple, old-fashioned listing type format, which was designed for line printers. How times have changed!

Now, modern SAS technologies work with their proprietary Output Delivery System (ODS) system, which allows you to tell SAS to output reports like tables and graphs in multiple different formats. For instance, SAS can put output into:

- 1 *Attractive HTML files.* This is set up as the default in newer SAS versions, and we have already seen in Chapter 5 how to change the automatic settings of how this output will look. You can save the automatic output as an HTML file in either SAS 9 or SAS Studio.
- 2 *Rich Text Files*, which will open as Microsoft Word or similar files.
- 3 *PDF files*, which will open in Adobe Acrobat or other PDF readers.
- 4 *Datasets created from output.* These, in turn, can be exported to spreadsheet or database programs such as Microsoft Excel or Access.
- 5 Several more.

These output delivery options are incredibly flexible, easy, and attractive. How to get these different formatted outputs differs between SAS Studio and SAS 9.

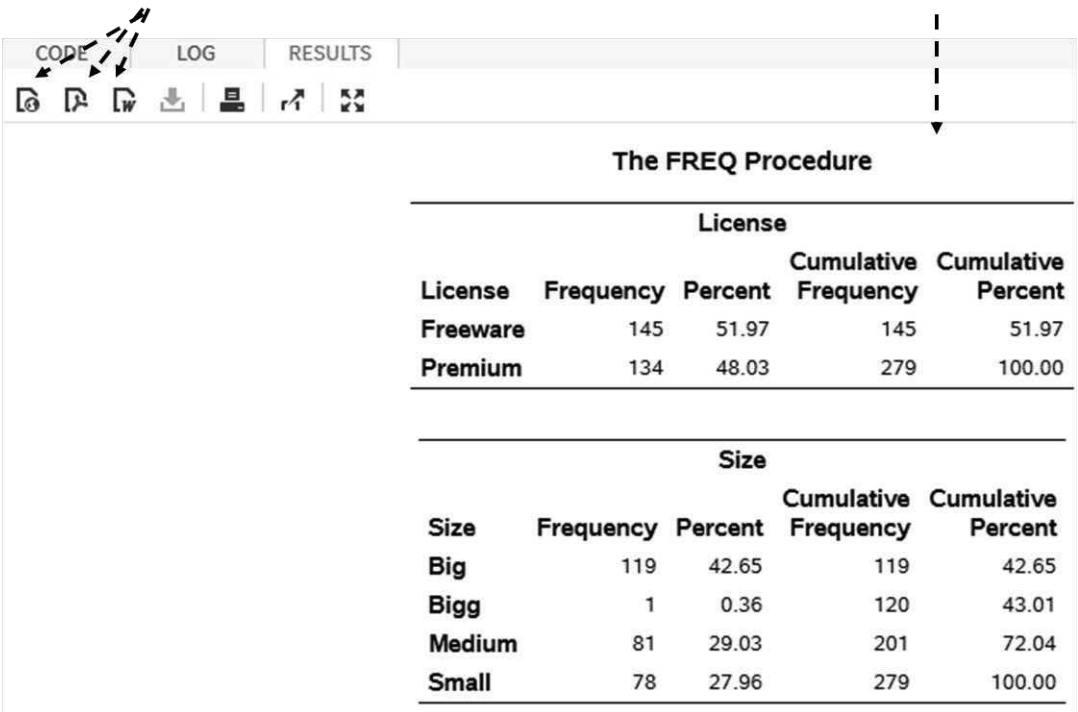
Different ODS Outputs in SAS Studio

In SAS Studio, you can download results in HTML (web browser), PDF (Acrobat or similar) or RTF (Microsoft Word or similar) formats at the click of a button, as seen in Figure 6.6 on page 85.

Figure 6.6 Downloading ODS results in different formats in SAS Studio

Simply press these buttons to download HTML, PDF and RTF versions of the output respectively that will open in a web browser, programs like Acrobat and programs like Word respectively – see Chapter 5 for how to change their look

The HTML output is default



The screenshot shows the SAS Studio interface with the 'RESULTS' tab active. Above the results, there are icons for downloading the output in different formats: HTML (web browser), PDF (Acrobat), and RTF (Word). The HTML output is selected and displayed. The results are titled 'The FREQ Procedure' and contain two tables. The first table is for the 'License' variable, and the second table is for the 'Size' variable. Both tables show frequency, percent, cumulative frequency, and cumulative percent.

The FREQ Procedure				
License				
License	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Freeware	145	51.97	145	51.97
Premium	134	48.03	279	100.00

Size				
Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Big	119	42.65	119	42.65
Bigg	1	0.36	120	43.01
Medium	81	29.03	201	72.04
Small	78	27.96	279	100.00

This is a major advantage for SAS Studio. Note also that the PDF output contains a menu allowing you to navigate between different sections of a longer report.

Different ODS Outputs in SAS 9

In SAS 9, you need to program the ODS outputs. Luckily, this is mostly very easy. For instance, say you wish to create a rich text output of various tables and graphs that you have created with SAS. Then, merely enter code like that in Example Code 6.3 on page 85 which will open your default program for processing rich text (like MS Word) and create a new file containing your SAS output (as you can see, you stipulate a filename and location for it to be saved to):

Example Code 6.3 *Example: Output in a rich text format that will open in MS Word or similar*

```
ODS RTF file='c://Output.rtf';
<Insert SAS code here to create output like statistical tables & graphs>
```



```
ODS RTF close;
```

The ODS formats need to be studied by the dedicated user, but they all mostly work as simply as the above example. The following are further examples:

Example Code 6.4 *Example of changing HTML output style*

```
ODS HTML style =HTMLBlue;  
<Insert SAS code here to create output like statistical graphs>  
ODS HTML style = Journal2;
```

The above example changes the HTML output style – which will usually open in SAS when you run anything – to a specific style called HTMLBlue. I set your style to Journal2 above because it produces clean black-and-white tables, however, in Chapter 10 later we will do graphing which is often best done in color. In the above code, you change to HTMLBlue which allows color output, then change back to Journal2.

Example Code 6.5 *Example: Writing SAS output to a PDF that will open in Acrobat or similar*

```
ODS PDF file='c://Output.pdf';  
<Insert SAS code here to create output like statistical tables & graphs>  
ODS PDF close;
```

Once again, this will save and open a PDF file of your output.

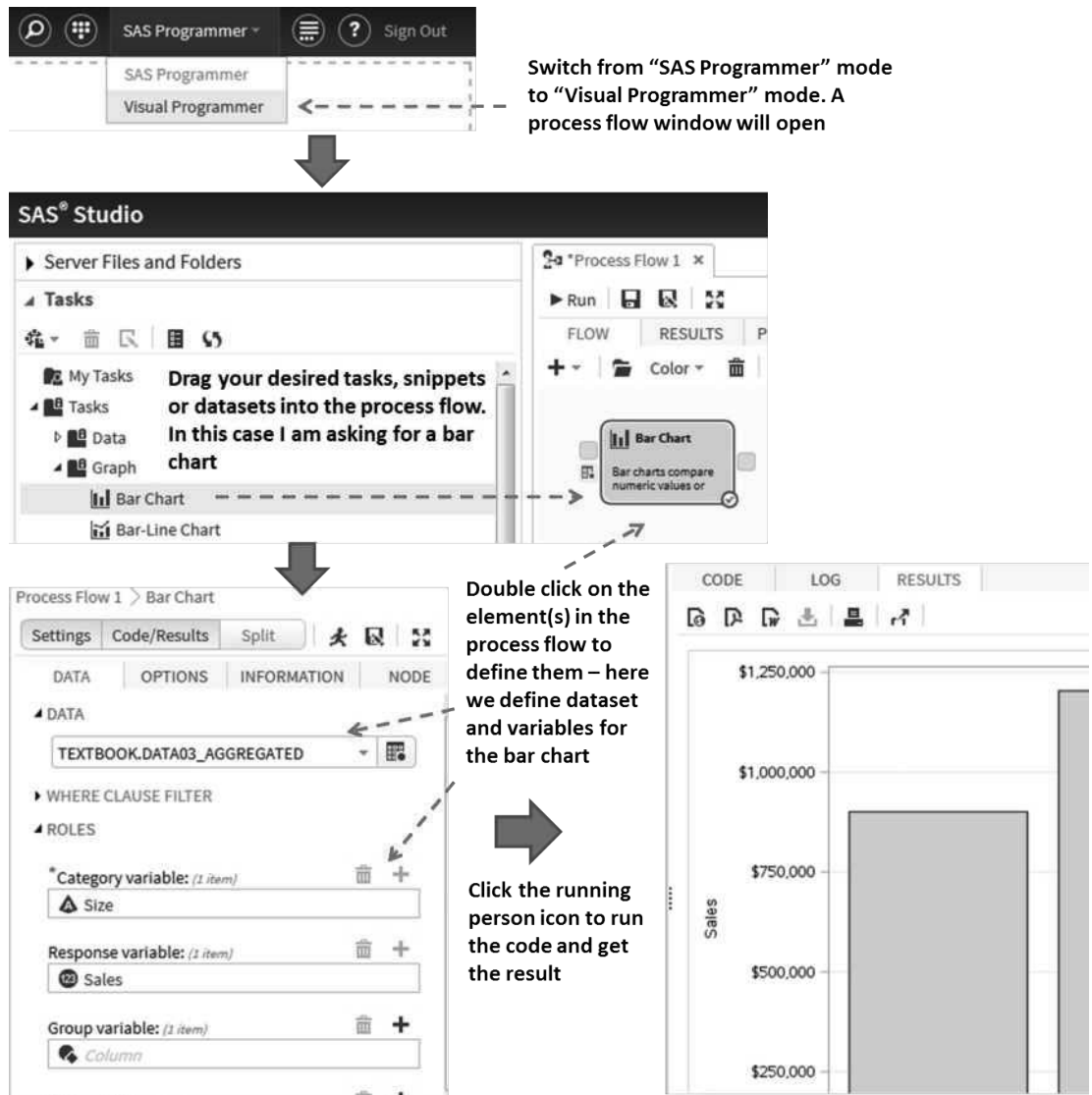
SAS ODS is an incredibly powerful system for crafting your SAS output. Any time you want to say “*hey, I’m creating such-and-such analysis in SAS and I would want it to look like that and come out in such-and-such a format,*” then ODS can usually do it for you.

The Visual Programmer Mode in SAS Studio

So far, I have demonstrated programming in SAS. As much as I have argued for using programming as the most efficient way of achieving analysis and teaching statistics in many cases, SAS Studio has created a clever way of generating your programs that allows you the comfort of a point-and-click type approach that works with SAS programming. This is known as the Visual Programmer mode.

In the SAS Studio Visual Programmer mode, you can define your dataset, task and variables for SAS Studio using easy-to-understand drag-and-drop methods. As an example of the use of this mode, see Figure 6.7 on page 87 below.

Figure 6.7 Example of using the Visual Programmer mode in SAS Studio



In this example, I have generated a bar chart simply by doing the following easy steps:

- Initiate SAS Studio Visual Programmer mode by switching from SAS Programmer mode at the top right. This opens a process flow window.
- Drag a pre-defined task from the Task window (in this case the Graphs > Bar Chart task) to the process flow.
- Double click the resulting Bar Chart process piece gives the settings. Here I define the dataset and variables using easy drop-down fields.

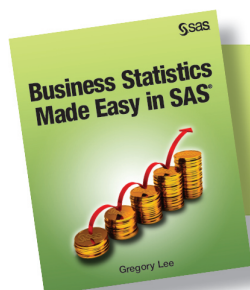
- Click the “running person” icon to get the results. Note: To see the graph in color, switch to the HTMLBlue results style in Preferences.

There are many other Tasks and what are called “Snippets” (pieces of code that can be used in various places). You should browse through these – perhaps after reading the book and acquainting yourself with the field of basic statistics – to see what Visual Programmer has to offer. It is an intuitive and pleasing way to generate simple tasks, but has other disadvantages of point-and-click modes, such as lack of the full functionality SAS programming can offer.

Conclusion

This chapter has introduced data manipulation in SAS, the absolute basics of analysis, and it has shown us how to create results in various formats. The rest of this book discusses a variety of analyses and principles that – used correctly - will launch you on a productive and profitable business statistics path.

From *Business Statistics Made Easy in SAS®*, by Gregory John Lee. Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Business Statistics Made Easy in SAS®*.
Full book available for purchase [here](#).

Index

A

a-priori power 196
 absenteeism 318
 accuracy
 about 203
 assessing 203
 of regression slopes 260
 of statistics 177
 statistical power 192
 statistics size and 200
 Acemoglu, D. 163
 advanced statistical analysis packages 43
 agreement tests 297
 airline industry, big data in 9
 analysis of level 313
 analytical skills 6
 analytics and reporting stages, tasks in 300
 AngloGold Ashanti Look Ahead
 See descriptive statistics 91
 ANN (artificial neural networks) 336
 annotations, placing in graphs 137
 ANOVA 284
 ANOVA F-Statistic 256
 answer formats 37, 39
 Apache Hadoop® 328
 artificial neural networks (ANN) 336
 associating variables
 about 109
 causation and 110
 continuous data 112
 correlation 112
 correlation coefficients 113
 covariance 112
 ordinal data 112
 relating categorical variables 119

statistical association 110
 variable categories 111

AstraZeneca case 172
 autocorrelation 251
 average 15, 94, 95

B

background analyses, versus displayed reports 300
 bar graphs, in SGPLOT procedure 141
 Barr, J. 51
 Bauer, H.H. 316
 Bayesian statistics
 about 340, 341
 classical statistics 341
 final answer (posterior) 342
 pre-existing guesses (proprs) 342
 sample data 342
 BCA (Bias Corrected and Accelerated) 210
 Becker, G. 163
 Berengueres, J. 10
 Bias Corrected and Accelerated (BCA) 210
 big data
 about 326, 330
 characteristics of 327
 in airline industry 9
 solutions for 328
 bimodal distribution 102
 binomial data 291
 binomial proportions, assessing categories through 291
 black-and-white graphs, versus color graphs 137
 Blattberg, R.C. 10

- Boom, A. 163
 - bootstrapped confidence intervals 245
 - bootstrapping 190, 208, 210, 249, 274, 280
 - Boudreau, J.W. 317
 - box-and-whisker plots, in SGPLOT
 - procedure 141
 - breakeven 319
 - Burmeister, S. 90
 - business analysis
 - about 311
 - combining statistics with per-unit financial values 318
 - examples of business extrapolation 321
 - financial estimates of revenue or cost of one unit 314
 - financial extrapolation process 312
 - focal variables 313
 - net profitability 319
 - scope 319
 - business statistics
 - interpretation of 7
 - reporting 308
 - tasks in 73
-
- C**
- CALIS procedure 234, 253
 - capitalization, in SAS code 75
 - Cascio, W.F. 317
 - categorical data
 - about 30, 285, 298
 - linear regression and 227
 - linking categorical variables 287
 - one-way categorical distributions 286
 - statistical questions about 287
 - categorical predictors 227
 - categorical variables
 - about 111, 119
 - associating 293
 - centrality for 95
 - crosstabs 119
 - FREQ procedure for associating 294
 - linking 119, 120, 287
 - relating 119
 - spread for 99
 - testing general association between 295
 - testing possibilities in association 297
 - categories
 - assessing through binomial proportions 291
 - comparing 270, 272
 - comparing continuous variables across 270
 - comparing means for more than two 284
 - comparing means with related categories 281
 - CATMOD procedure 296, 298
 - causation
 - associating variables and 110
 - between independent variables 234
 - central tendency 91
 - centrality
 - about 94
 - as a variable characteristic 31
 - checking 127
 - for categorical variables 95
 - for continuous variables 94
 - for ordinal variables 95
 - change analysis 314
 - change situations, static situations and 313
 - character (text) data, versus numerical data 25
 - chart modules 103
 - Cherrier, J. 10
 - Chi-Square 295
 - classical statistics 341
 - CLV (customer lifetime value) 316
 - Cochran-Mantel-Haenszel Statistics 296
 - code files
 - capitalization in 75
 - opening existing 77
 - Code window (SAS Studio) 62
 - coefficients, implications of 165
 - color graphs, versus black-and-white graphs 137
 - comparison
 - of categories 270

- of dependent variables 271
- of independent variables 271
- of means 271, 281, 284
- of means for more than two categories 284
- of means with related samples or categories 281
- of more than two categories 272
- of related categories 272
- of two categories 272
- of two means 275
- computers, versus math 16
- computing power and speed, growth in 327
- concepts, measuring relationships
 - between 15
- condition indices 236
- conditional variables 81
- confidence intervals 184, 259
- confirmatory factor analysis 133
- constellations 157
- constructs
 - about 35
 - choosing 35
 - control 37
 - defined 163
 - focal 36
 - importance of 35
 - predictor 36
- context 16, 223
- Contingency Coefficient 295
- contingency tables 119
- continuous (ratio or interval) data 29, 39, 111, 112
- continuous variable spread 97
- continuous variables
 - centrality for 94
 - comparing across categories 270
 - interquartile range for 98
 - linking to categorical variables 120
- control constructs, data and 37
- convergent validity 133
- Cook's D 247
- CORR procedure 115, 130
- correlation analysis 117
- correlation coefficients 113
- correlation tables 116
- correlations
 - as back-up diagnostics 235
 - between independent variables 237
 - calculating 115
 - compared with covariance 117
 - sizes of 116
 - types of 115
- cost of one unit, financial estimates of 314
- covariance
 - about 117
 - compared with correlation 117
- Cramer's V statistic 295
- Crestor case 172
- Cronbach alpha 130, 132
- crosstabs 119
- customer lifetime value (CLV) 316
- customer satisfaction, as a variable 3

D

- data
 - about 21
 - assumptions about 273, 278
 - binomial 291
 - capturing 33, 34, 43, 227
 - characteristics of variables 27
 - checking for mistakes in 43
 - cleaning 72, 227
 - collecting 227
 - continuous (ratio or interval) 29, 39, 111, 112
 - control constructs and 37
 - defined 163
 - dichotomous 291
 - entering 231
 - existing 38
 - extracting statistics from 15
 - fitting 155
 - fitting complex mathematical equations to 161
 - focal constructs and 36
 - forming data tables 24
 - gathering 13, 33, 34, 37
 - importance of in statistics 13

- importing 231
- initial assumptions about 233
- interval 29, 39, 111, 112
- issues with 23, 227
- manipulating 73, 77
- modeling preconceived ideas about 153
- multi-row 49
- objects 23
- observations 23
- ordinal 30, 111, 112
- populations 23
- post-capturing issues of 43
- ratio 29, 39, 111, 112
- real-time 38
- samples 23
- See also big data 326
- See also categorical data 285
- See also data patterns 151
- See also errors 123
- See also missing data 44
- See descriptive statistics 91
- shape issues with 237
- testing for normal distributions 159
- testing for straight line shapes 158
- data analysis
 - about 73, 83
 - in data warehousing 333
 - software for 48
- data architecture skills 7
- DATA keyword 76
- data management
 - combining datasets 82
 - creating datasets 78
 - creating temporary datasets in Work library 79
 - creating variables 80
 - manipulating current variables 80
- data mining
 - about 336
 - compared with theory-based analysis 154
 - patterns and 154
 - theory versus 153
- data patterns
 - about 151
 - comparing theory-based analysis and data mining 154
 - defined 163
 - fitting mathematical models 155
 - forcing 162
 - multivariate patterns 152
 - plots versus statistical fit measures of 155
 - See also interpreting patterns 164
 - single variable patterns 151
 - theory versus data mining 153
 - troubleshooting 163
- data points, inappropriate 124
- data tables, forming 24
- data warehousing
 - about 330, 335
 - issues and alternatives in 333
 - steps in traditional 331
- database software 48
- dataset analysis 252
- datasets
 - combining 82
 - complex types of 49
 - complications in 48
 - creating 78, 79
 - creating in Work library 79
 - dispersed 330
 - incongruent 330
 - integrating 26
 - longitudinal 49
 - multi-level 49
 - primary 26
 - secondary 26
 - vulnerable 330
- dates, capturing 48
- Davenport, T.H. 327
- decision-making, in statistics process 17
- deep learning 336
- Delwiche, L.D. 77
- dependent variables
 - characteristics of 255
 - comparison of 271
 - missing 45
 - transforming 273, 280
- descriptive statistics
 - about 91, 104
 - assessing distribution 103

- centrality 94
- end outcome of analysis of 91
- getting in SAS 92
- shape 99
- spread 97
- dichotomous data 291
- discriminant validity 133
- dispersed datasets 330
- displayed reports, versus background analyses 300
- distributed computing, improved storage and processing through 328
- distribution, assessing 103
- Dull, T. 334
- dummy variables 227, 264
- Durbin-Watson statistic 251
- Dyché, J. 327

E

- Editor window (SAS 9) 54
- Efimov, D. 10
- Ellison, L. 310
- employee stocks 316
- employee-related variables, value of 316
- employees
 - movement of 316
 - performance of 317
 - reductions in expensive behaviors 317
 - turnover of 317
- endogeneity 234
- enquiries
 - as a variable 3
 - of customers 303
- Enterprise Resource Programs (ERPs) 48
- equivalence, testing for 293
- ERPs (Enterprise Resource Programs) 48
- errors
 - about 123
 - checking centrality and spread 127
 - inappropriate data points 124
 - missing data 126
 - multi-item scales 128

- residuals and 222
- strange variable distributions 128
- ETL (Extract-Transform-Load) 334
- examples
 - about 1
 - brief 5, 299
 - company 2
 - correlation analysis 117
 - current research needs 2
 - of business extrapolation 321
 - of interpreting when patterns are not found 167
 - of SGPLOT procedure graphs 138
 - of simulation 337
- existing data 38
- exploratory factor analysis 133
- Explorer window (SAS 9) 54
- Extract-Transform-Load (ETL) 334
- extracting
 - statistics from data 15
 - to data marts 333

F

- face-to-face interviews 38
- Facebook 326
- feedback loops 234
- FIML (full information maximum likelihood) 127, 253
- final statistic parameters and coefficients, intermediate fit statistics versus 161
- financial extrapolation process 312
- financial profitability 311
- financial variables, values of 318
- fit
 - about 151, 155, 222
 - assessing 233
 - steps in 233
 - troubleshooting 224, 257
- fitting models
 - See statistics process 149
- focal constructs, data and 36
- focal variables 313
- folders and files, linking with 62, 63
- follow-up recommendations 308

- formats
 - answer 37, 39
 - question 37, 38
- formatting, in SGPLOT procedure 142
- FREQ procedure 92, 95, 120, 287, 294, 296, 298
- full information maximum likelihood (FIML) 127, 253
- effects of 239
 - in residual plots 243
 - remedies for 244
- Hoeffding Dependence Cpefficient 115
- Hong, S.J. 10
- HTML files 84
- hypothesis testing 184

G

- Garbage in, Garbage out (GIGO) 14
- geographical mapping, using GMAP procedure 145
- GIGO (Garbage in, Garbage out) 14
- GKPI procedure 136, 143
- global fit, troubleshooting 257
- GMAP procedure 136, 145
- good fit 221
- Goodnight, Jim 51
- GPLOT procedure 136
- graphing
 - about 135, 147
 - black-and-white versus color 137
 - flexibility in 136
 - GKPI procedure 136, 143
 - GMAP procedure 136, 145
 - modules for 136
 - placing annotations in graphs 137
 - procedures for 136
 - SGPANEL procedure 143
 - SGPLOT procedure 138
 - SGSCATTER procedure 146
- groups, comparing 271

H

- Hammerschmidt, M. 316
- Hats (leverage scores) 247
- Heath, D. 142
- Helwig, J. 51
- heteroscedasticity
 - about 237

I

- IF-THEN concept 81
- in-memory processing, improved
 - processing through 329
- inaccuracy, faces of 179
- inappropriate data points 124
- incongruent datasets 330
- independent variable slopes 259
- independent variables
 - causal relationships between 234
 - comparison of 271
 - correlations between 237
- influence, defined 247
- influential outliers 245
- initial phase, in data warehousing 332
- inputs, costs of 315
- integration phase, in data warehousing 332
- intercept 258
- intermediate fit statistics, versus final
 - statistical parameters and coefficients 161
- interpretations 308
- interpreting patterns
 - about 164
 - implications of model and coefficients 165
 - steps in 164
- interquartile range, for continuous and ordinal variables 98
- interval data 29, 39, 111, 112
- issues 23

J

Jackofsky, E.F. 153
 Janmaat, E. 22
 JIPSA (Joint Initiative for Priority Skills Acquisition) 90
 JMP® 53, 73
 Joint Initiative for Priority Skills Acquisition (JIPSA) 90

K

Kendall's Tau 115
 knowledge 7
 Kuhfeld, W. 142
 kurtosis 105, 106, 160

L

Lawrence, R.D. 10
 Lehrer, J. 162
 leverage scores (Hats) 247
 libraries
 creating in SAS 9 55
 creating in SAS Studio 63
 licenses
 as variable 3
 distribution of 304
 variables analyzed by 303
 Likelihood Ratio Chi-Square 295
 Likert-type scale 39, 231
 line plots, in SGPLOT procedure 138
 linear regression
 about 213, 215
 aim of 216
 applying remedies 233
 assessing fit 233
 categorical predictors 227
 core textbook example 213
 defined 215
 implementing multiple regression 226
 initial data issues 227

interpreting regression slopes 257
 ordinal predictors 227
 reporting multiple regression results 265
 running regression analysis 231
 simplest case of 217
 single Likert-type scale items 231
 variables in 216
 variables in multiple regression 216
 linearity 112
 lines, in SAS code 76
 loading, in data warehousing 333
 Log window
 SAS 9 54
 SAS Studio 62
 logic, importance of 235
 lognormal distribution 100
 longitudinal datasets 49

M

machine learning 335, 336
 magnitude
 See size, of statistics 173
 Malthouse, E.C. 10
 Mantel-Haenszel Chi-Square test 296
 Mardia score 176
 marketing outcomes 316
 Matange, S. 142
 math, versus computers 16
 mathematical models, fitting 155
 mathematical simulations 338
 means
 about 94
 comparing 271
 comparing for more than two categories 284
 comparing to population benchmarks 283
 comparing two 275
 comparing with related samples or categories 281
 MEANS procedure 92, 93, 95, 121
 measurement error 223
 measurement, growth in 327

- medians 94, 95
 - MI procedure 253
 - MIANALYZE procedure 253
 - Miner, Bob 310
 - Mining Qualifications Authority (MQA) 90
 - missing data
 - as a diagnostic issue 252
 - assessing in observations 126
 - assessing in variables 127
 - dealing with 44
 - diagnosis of in regression 252
 - in observations 253
 - in variables 253
 - linear regression and 227
 - remedies for 252
 - steps 126
 - mode 95
 - model fitting
 - See statistics process 149
 - MODEL statement 232
 - models
 - structures of 234
 - theoretical and practical implications of 166
 - modules, for graphing 136
 - MQA (Mining Qualifications Authority) 90
 - multi-item assessment 41
 - multi-item scales
 - about 45, 128
 - aggregating multiple items into summary variables 132
 - assessing internal reliability of each 129
 - dealing with 47
 - linear regression and 227
 - reversed items 128
 - tasks in preparing 47
 - multi-item variables 127
 - multi-level datasets 49
 - multi-row datasets 49
 - multicollinearity 235, 236
 - multiple imputations 127, 253
 - multiple regression
 - implementing 226
 - reporting results of 265
 - multivariate patterns 152
- ## N

 - needs, for statistics process 12
 - negative linearity 112
 - net profitability
 - about 319
 - basic profit 319
 - breakeven 319
 - return on investment (ROI) 320
 - New Import Data wizard 65
 - Nirmalanof, G. 161
 - non-linearity
 - about 237
 - effects of 239
 - in residual plots 242
 - remedies for 244
 - noninferiority tests 298
 - nonparametric statistics 274
 - nonparametric T-test 280
 - normal distribution 100, 159
 - normality 273
 - normality statistics 104
 - Normalized Multivariate Kurtosis score 176
 - numerical data, versus text (character) data 25
- ## O

 - Oates, E. 310
 - objects 23
 - observations
 - about 23
 - loss of 252
 - missing data in 126, 253
 - odds ratio test, homogeneity of 297
 - ODS Graphics engine 136
 - ODS outputs
 - in SAS 9 85
 - in SAS Studio 84
 - one-way categorical distributions 286
 - one-way frequencies
 - about 288

- assessing categories through binomial proportions 291
 - assessing distribution of 289
 - online slider scale 40
 - operational time, value of 315
 - operational variables, costs and revenues of 315
 - Oracle South Africa case study 310
 - Oracle VirtualBox 60
 - ordinal data 30, 111, 112
 - ordinal predictors
 - single Likert-type scale items as 231
 - special treatment of 227
 - ordinal variables
 - about 228
 - centrality for 95
 - interquartile range for 98
 - testing trend in 296
 - outlier weighting 249
 - outlier, defined 247
 - output formats, reporting through 84
 - overtime 318
- P**
- Phi coefficient 295
 - physical simulations 337
 - Pischke, J-S. 163
 - plots, versus statistical fit measures of
 - patterns 155
 - point-and-click 73
 - populations 23, 283
 - positive linearity 112
 - post-capturing 43
 - POWER procedure 196
 - pre-analysis data cleaning and
 - preparation 72
 - pre-existing guesses (proprs) 342
 - predictor constructs 36
 - primary datasets 26
 - process simulations 337
 - products, value of 315
 - programming code
 - about 73
 - advantages of 74
 - doing tasks through 74
 - lessons on 74
 - running 76
 - protocols, in data analysis software 49
 - psychometric measures 41

- p-value 186, 205, 256, 259
 - paired samples 282
 - parabola 225
 - paradigms, patterns and 153
 - parametric 273, 274
 - parametric approach
 - about 204
 - p-value 205
 - standard error 205
 - test statistic 205
 - patterns
 - implications of 154
 - over time 15
 - reasons for 154
 - See also data patterns 151
 - PDF files 84, 137
 - Pearson correlations 115
 - people variables 316
 - per-unit financial values, combining
 - statistics with 318

Q

- question formats 37, 38

R

- R-Sq statistics
 - about 254
 - interpreting size of 255
 - random patterns 154
 - ratio data 29, 39, 111, 112
 - raw data records 25
 - raw datasets 332
 - real-time data 38
 - REG procedure 232, 252
 - regression 119
 - See also linear regression 231

- regression analysis, running 231
 - regression parameters
 - about 257
 - independent variable slopes 259
 - intercept 258
 - regression slopes
 - about 119, 259, 264
 - interpreting 257
 - process for interpreting 259
 - significance and accuracy of 260
 - size of significance and accuracy of 262
 - reliability output, assessing 131
 - remedies, applying 233
 - REPORT procedure 92
 - reporting
 - about 73
 - skills for 6
 - through output formats 84
 - representativity 34
 - requirements, for statistics process 12
 - residual plots
 - about 240
 - diagnosing data shape issues with 240
 - heteroscedasticity in 243
 - non-linearity in 242
 - residuals
 - about 273
 - error and 222
 - normality of 250
 - Results window
 - SAS 9 54
 - SAS Studio 62
 - return on investment (ROI) 320
 - returns 316
 - revenue, financial estimates of 314
 - reverse-worded items 42
 - reversed items, dealing with 128
 - Rich Text Files 84, 137
 - robust regression 248
 - ROBUSTREG procedure 248
 - ROI (return on investment) 320
 - Royal FrieslandCampina example
 - See data 21
 - Run command 76
- S**
- sales 305, 316
 - Sall, J. 51
 - sample size 34
 - samples and sampling 23, 34
 - SAS
 - about 51, 52
 - website 60
 - SAS® 9
 - about 52
 - creating libraries in 55
 - importing data into 58
 - installing 53
 - ODS outputs in 85
 - opening 53
 - opening code files in 77
 - setting options 59
 - setting up 53
 - SAS® Enterprise Guide® 52, 73
 - SAS® Enterprise Miner 52
 - SAS® LASR 329, 334
 - SAS® Studio
 - about 52, 60, 61, 73
 - creating libraries 63
 - importing data 65
 - installing 60
 - linking libraries with folders 63
 - linking with folders and files on computers 62
 - ODS outputs in 84
 - opening 60
 - opening code files in 77
 - setting options 67
 - setting up 60
 - Visual Programmer mode 86
 - SAS® Text Miner 329
 - SAS® University Edition 52, 60
 - SAS® Visual Analytics 53
 - satisfaction, of customers 302
 - scalability 329
 - scatter graphs, in SGPLOT procedure 139
 - scatterplots, SGSCATTER procedure for multiple 146
 - scope 319

- SD (standard deviation) 97
- secondary datasets 26
- semantic differential 40
- semicolons 76
- Server Files and Folders (SAS Studio) 61
- services, value of 315
- SGPANEL procedure 136, 143
- SGPLOT procedure
 - about 136, 138
 - examples of graphs 138
 - graphing options and formatting in 142
- SGSCATTER procedure 136, 146
- shapes
 - about 99
 - bimodal distribution 102
 - fitting data to exact mathematical 155
 - lognormal distribution 100
 - normal distribution 100
 - testing data for straight line 158
 - uniform distribution 101
- significance, of regression slopes 260
- simple imputations 127, 253
- simulation
 - about 336, 340
 - example of 337
 - types of 337
- single accuracy estimates 186
- single data points 25
- single variable patterns 151
- size
 - as variable 3
 - levels of 304
 - of correlations 116
 - of R-Sq statistics 255
 - of significance and accuracy of regression slopes 262
 - of statistics 173, 174, 200
 - variables analyzed by 303
- skewness 105, 106, 160
- skills
 - data architecture 7
 - extending your 6
 - reporting 6
- Slaughter, S.J. 77
- slow to access data 331
- Snippets 88
- social media 326
- software
 - data analysis 48
- spacing, in SAS code 76
- Spearman correlations 115
- specification error 223
- spread
 - about 97
 - as a variable characteristic 31
 - calculating variables spread 99
 - checking 127
 - continuous variable 97
 - for categorical variables 99
 - interquartile range for continuous and ordinal variables 98
- Sreekumar, K.P. 161
- staging phase, in data warehousing 332
- standard deviation (SD) 97
- standard error, parametric approach and 205
- standardized slopes 259, 263
- standardized statistics 175
- static situations, change situations and 313
- statistical association 110
- statistical effect 313
- statistical extrapolation
 - about 323
 - examples of 321
 - means-based example of 321
 - regression-based example of 322
- statistical power
 - about 192
 - before and after testing 195
 - elements of 194
 - measurement of 192
 - problems with 198
 - understanding 192
- statistical significance
 - about 183
 - bootstrapping 190
 - confidence intervals 184
 - single inaccuracy estimates and p-values 186
- statistical tests of distribution 103
- statistics
 - about 15
 - accuracy of 15, 177

- advice on 18
- classical 341
- combining with per-unit financial values 318
- extracting from data 15
- generating 16
- importance of data in 13
- meaning of 15
- nonparametric 274
- normality 104
- See also descriptive statistics 91
- standardized 175
- statistics process
 - about 9, 149, 168
 - challenges in 17
 - decision-making 17
 - extracting statistics from data 15
 - getting data 13
 - needs and requirements for 12
 - patterns in data 151
 - understanding 17
- storage, growth in 327
- strikes 318
- structural equation modeling 234
- Studentized Residual 247
- subgroups, comparing 271
- summary variables 132
- superiority tests 298
- supervised learning algorithms 336
- surveys 38
- SYSLIN procedure 234
- doing through programming code (syntax) 74
- in analytics and reporting stages 300
- running through point-and-click 73
- test statistic, parametric approach and 205
- testing
 - assessing power before and after 195
 - for statistical significance 183
- text (character) data, versus numerical data 25
- textbook materials 53
- textual analysis 329
- theory
 - defined 163
 - importance of 235
 - versus data mining 153
- theory-based analysis, compared with data mining 154
- times
 - capturing 48
 - changes in 317
- traditional parametric t-test, versions of 279
- transformations 244
- trust
 - as variable 3
 - of customers 302
- TTEST procedure 278
- Twitter 326
- two-stage least squares regression 234
- type, as a variable characteristic 28

T

- T-tests
 - about 275
 - assessing data assumptions 278
 - end-point of 276
 - implementing nonparametric 280
 - related data 283
 - running initial 278
 - versions of traditional parametric 279
- tabs, in SAS code 76
- TABULATE procedure 92
- tasks

U

- understanding, in statistics process 17
- unequal variances t-test 280
- uniform distribution 101
- UNIVARIATE procedure 92, 93, 95, 136
- unstandardized slopes 259, 262
- unstructured data, growth in 327
- unsupervised learning 336

V

- value, of big data 328
- variable distribution 91
- variables
 - about 35
 - analyzed by license and size 303
 - assessing missing data in 127
 - calculating spread 99
 - categories of 111
 - characteristics of 27
 - choosing 32
 - choosing the right 154
 - conditional 81
 - continuous 94, 98, 120, 270
 - creating 80
 - dependent 45, 255, 271, 273, 280
 - dummy 227, 264
 - focal 313
 - importance of types 30
 - in linear regression 216
 - independent 234, 237, 271
 - manipulating 80
 - missing data in 253
 - ordinal 95, 98, 228, 296
 - sales and 305
 - See also associating variables 109
 - See also categorical variables 119
 - specifying 130
 - strange distributions of 128
 - summary 132
- variance inflation factors (VIFs) 236

- variances 97, 216, 273
- variety, of big data 328
- velocity, of big data 328
- veracity, of big data 328
- Viewers (SAS 9) 54
- VIFs (variance inflation factors) 236
- virtualization program 60
- Visa 326
- Visual Programmer mode 73, 86
- VMWare Player 60
- volume, of big data 327
- vulnerable datasets 330

W

- wage bill, changes in 317
- Walmart 326
- weak relationship 221
- weighted regression 245
- West Point
 - See linear regression 213
- Windows folders, linking SAS library to 55
- Work library, creating datasets in 79
- workforce numbers, changes in 317

Z

- zero relationship 221

About the Author



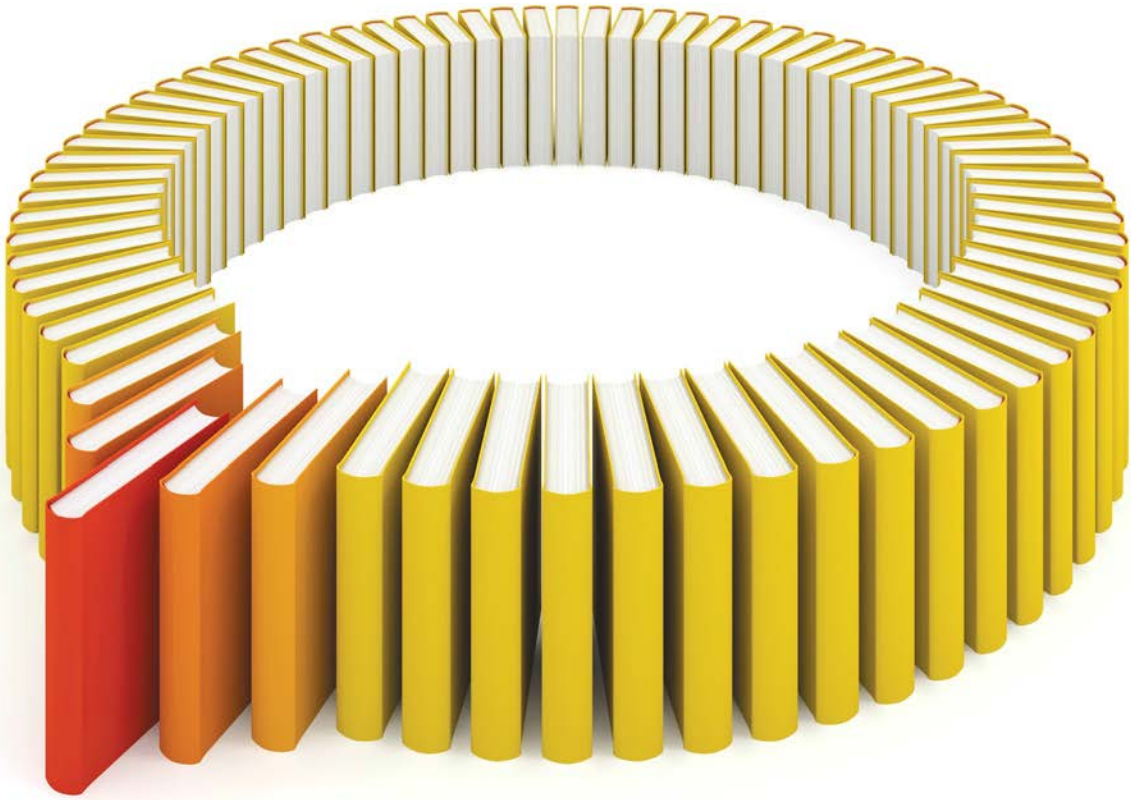
Professor Gregory Lee is currently the Research Director and an Associate Professor in Research Methodology and Decision Sciences at the AMBA-rated Wits Business School.

He has published prior books on human resources (HR) metrics, and has many article publications in the international arena such as the *Human Resource Management Journal*, *European Journal of Operational Research*, *Scientometrics*, *Journal of Business-to-Business Marketing*, *The International Journal of Human Resource Management*, *International Journal of Manpower*, *Review of Income & Wealth*, *Journal of Human Resource Costing & Accounting* and many others.

He focuses on issues in human resource management, notably HR metrics (in which he has established himself as a leading expert) and other areas such as training, employee turnover and the employee-customer link.

He has served in many capacities within the international academic field. He has sat on the Graduate Management Admissions Council (GMAC®) advisory council, the editorial boards of the *Journal of Organizational and Occupational Psychology*, and engages in frequent reviewing for many journals.

In addition, he is a well-known consultant, writer and speaker in the corporate and practical management arenas, notably in the area of HR metrics, but extending to other areas such as human resources strategy and foresight.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

 support.sas.com/bookstore
for additional books and resources.


THE POWER TO KNOW.®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613