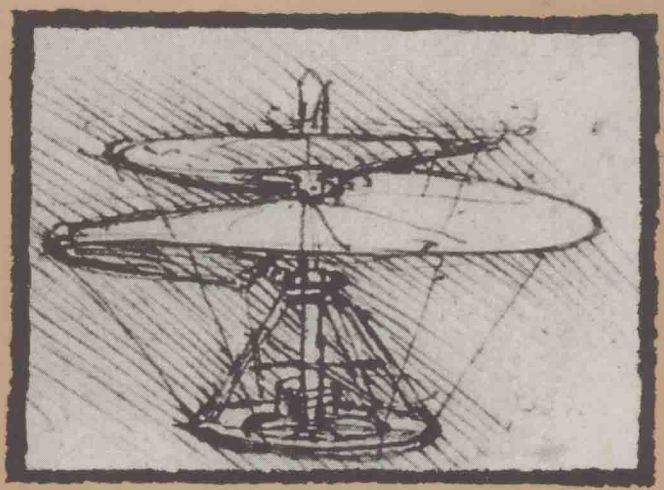
Applied Statistics and the SAS® Programming Language

Fourth Edition



RONALD P. CODY

JEFFREY K. SMITH

FOURTH EDITION

Applied Statistics and the SAS Programming Language

Ronald P. Cody
ROBERT WOOD JOHNSON MEDICAL SCHOOL

Jeffrey K. Smith
RUTGERS UNIVERSITY

PRENTICE HALL, Upper Saddle River, New Jersey 07458

Library of Congress Cataloging–in–Publication Data

Cody, Ronald P.
Applied statistics and the SAS programming language / Ronald P. Cody, Jeffrey K. Smith – 4th ed.
p. cm.
Includes bibliographical references (p. –) and index.
ISBN 0–13–743642–4 (pbk.)
1. SAS (Computer file) 2. Mathematical statistics–Data processing. I. Smith, Jeffrey K. II. Title.
QA276.4.C53 1997
519.5'0285'5369–dc21 97–1737
CIP

Acquisition editor: Ann Heath
Editorial assistant: Mindy Ince
Editorial director: Tim Bozik
Editor-in-chief: Jerome Grant
Assistant vice president of production and manufacturing: David W. Riccardi
Editorial/production supervision: Nicholas Romanelli
Managing editor: Linda Mihatov Behrens
Executive managing editor: Kathleen Schiaparelli
Manufacturing buyer: Alan Fischer
Manufacturing manager: Trudy Pisciotti
Marketing manager: Melody Marcus
Marketing assistant: Jennifer Pan
Creative director: Paula Maylahn
Cover designer: Jayne Conte

Cover illustration: From the manuscripts of Leonardo DaVinci; published by Charles Raviasson-Mollien, 6 vols., Paris, 1881-91.
Back cover photo of Ron Cody by Russ Cody.

Printed in the United States of America

10 9 8 7 6 5 4

ISBN: 0-13-743642-4

Prentice-Hall, International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada, Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Applied Statistics and the SAS Programming Language

To our parents,

Ralph and Bettie Smith

and

Philip and Margaret Cody

Preface to the Fourth Edition

When we began creating this fourth edition, several facts were clear: First, SAS software continues to evolve and improve. Second, our programming techniques have also improved. Third, several statistical techniques (such as logistic regression) have become popular and required coverage in this edition.

We have met many readers of earlier editions at meetings and conferences and were delighted to hear good things and constructive criticisms of the book. These we have taken to heart and attempted to improve old material and add relevant new topics. This fourth edition is the result of such reader reaction.

Most researchers are inherently more interested in the substance of their research endeavors than in statistical analyses or computer programming. Yet, conducting such analyses is an integral link in the research chain (all too frequently, the weak link). This condition is particularly true when there is no resource for the applied researcher to refer to for assistance in running computer programs for statistical analyses. *Applied Statistics and the SAS Programming Language* is intended to provide the applied researcher with the capacity to perform statistical analyses with SAS software without wading through pages of technical documentation.

The reader is provided with the necessary SAS statements to run programs for most commonly used statistics, explanations of the computer output, interpretations of results, and examples of how to construct tables and write up results for reports and journal articles. Examples have been selected from business, medicine, education, psychology, and other disciplines.

SAS software is a combination of a statistical package, a data-base management system, and a high-level programming language. Like SPSS, BMDP, Systat, and other statistical packages, SAS software can be used to describe a collection of data and produce a variety of statistical analyses. However, SAS software is much more than just a statistical package. Many companies and educational institutions use SAS software as a high-level data-management system and programming language. It can be used to organize and transform data and to create reports of all kinds. Also, depending on which portions of the SAS system you have installed in your computer (and what type of computer system you are running), you may be using the SAS system for interactive data entry or an on-line system for order entry or retrieval.

This book concentrates on the use of the SAS system for the statistical analysis of data and the programming capabilities of SAS software most often used in educational and research applications.

The SAS system is a collection of products, available from the SAS Institute in Cary, North Carolina. The major products available from the SAS Institute are:

Base SAS ®	The main SAS module, which provides some data manipulation and programming capability and some elementary descriptive statistics
SAS/STAT ®	The SAS product that includes all the statistical programs except the elementary ones supplied with the base package
SAS/GRAPH ®	A package that provides high-quality graphs and maps. Note that "line graphics" (the graphs and charts that are produced by normal character plots) are available in the base and SAS/STAT packages. SAS/GRAPH adds the ability to produce high-quality camera-ready graphs, maps, and charts.
SAS/FSP ®	These initials stand for the Full Screen Product. This package allows you to search, modify, or delete records directly from a SAS data file. It also provides for data entry with sophisticated data checking capabilities. Procedures available with FSP are FSBROWSE, FSEDSIT, FSPRINT, FSLIST, and FSLETTER.
SAS/AF ®	AF stands for the SAS Applications Facility. This product is used by data processing professionals to create "turn key" or menu systems for their users. It is also used to create instructional modules relating to the SAS system.
SAS/ETS ®	The Econometric and Time Series package. This package contains specialized programs for the analysis of time series and econometric data.
SAS/OR ®	A series of operations research programs.
SAS/QC ®	A series of programs for quality control.
SAS/IML ®	The Interactive Matrix Language module. The facilities of IML used to be included in PROC MATRIX in the version 5 releases. This very specialized package allows for convenient matrix manipulation for the advanced statistician.

SAS software now runs on a variety of computers, from personal computers to large multimillion dollar mainframes. The original version of the SAS system was written as a combination of PL/1 and IBM assembly language. Today, SAS software runs under Windows and Windows 95 on IBM compatible minicomputers, on most Macintosh computers, under UNIX on a large number of minicomputers and workstations, on IBM computers under a variety of operation systems, on Digital Equipment VAX computers, and others too numerous to mention. The major reason for the ability of SAS software to run on such a variety of machines is that all SAS software was rewritten in C and designed so that most of the code was system-independent. The conversion of the entire system to the C programming language was one of the largest (and most successful) programming projects ever undertaken. To migrate the SAS system to another computer or operating system, only a small system-dependent portion of code needs to be rewritten. The result is that new versions of SAS software

are made available for all computers very quickly, and the versions of SAS systems from one computer to another are much alike.

Learning to program is a skill that is difficult to teach well. While you will find our examples "reader-friendly" and their logic easy to follow, learning to write your own programs is another matter. Only through lots of practice will your programming skills develop. So, when you finish a chapter, please spend the time doing as many problems as you can. We wish you happy programming.

We express our gratitude to two colleagues who reviewed the manuscript and made many comments beneficial to this revision of the text. Our thanks, therefore, to Sylvia Brown and Robert Hamer, at the Robert Wood Johnson Medical School. Our sincere thanks also to Ann Heath, acquisitions editor for statistics at Prentice Hall, for her encouragement and support, and to Nicholas Romanelli for his superb editing, patience, and good cheer.

RON CODY
JEFFREY SMITH

Contents

Applied Statistics and SAS Software

SAS Programming

CHAPTER

1

A SAS Tutorial

A. Introduction

For the novice, engaging in statistical analysis of data can seem as appealing as going to the dentist. If that pretty much describes your situation, perhaps you can take comfort in the fact that this is the fourth edition of this book—meaning that the first three editions sold pretty well, and this time we may get it right. Our purpose for this tutorial is to get you started using SAS software. The key objective is to get one program to run successfully. If you can do that, you can branch out a little bit at a time. Your expertise will grow.

The SAS System is a combination of programs originally designed to perform statistical analysis of data. Other programs you may have heard of are SPSS, BMDP, or SYSTAT. If you look at personal computer magazines, you might run across other programs, primarily designed to run on personal computers. Since its inception, the SAS system has grown to where it can perform a fairly impressive array of nonstatistical functions. We'll get into a little of that in later chapters. For now, we want to learn the most basic rudiments of the SAS system. If you skipped over it, the Preface to the fourth edition contains some history of SAS software development and a more complete overview of the capabilities of SAS software.

To begin, SAS software runs on a wide variety of computers and operating systems (computer people call these platforms), and we don't know which one you have. You may have an IBM compatible personal computer running Windows or, perhaps, Windows 95. You may have a Macintosh computer, or you may be connected to a network or a mainframe computer by way of a modem or network

connection. You may only have a sophisticated VCR, which you think is a computer. If you are unsure of what platform you are using or what version of SAS software you are using, ask someone. As a matter of fact, right now would be a good time to invite the best computer person you know to lunch. Have that person arrive at your office about an hour before lunch so you can go over some basic elements of your system. You need to find out what is necessary on your computer to get the SAS system running. What we can teach you here is how to use the SAS system, and how to adapt to your computer system.

If you are running on a mainframe, you may well be submitting what are called "batch" jobs. When you run batch jobs, you send your program (across a phone line or a network from your personal computer or terminal) to the computer. The computer runs your program and holds it until you ask for it or prints out the results on a high-speed printer. You may even need to learn some Job Control Language (which you have to get from your local computer folks), and then you can proceed.

If you are running on a personal computer, or running in what is called interactive mode on a minicomputer or mainframe, then you need to learn how to use the SAS Display Manager. The look and feel of SAS once you are in the Display Manager is pretty much the same whatever platform you are using. If you are undaunted, take a deep breath and plunge into the real content in the next section. If you are already daunted, take a minute and get that lunch scheduled, then come back to this.

B. Computing with SAS Software: An Illustrative Example

SAS programs communicate with the computer by SAS "statements." There are several kinds of SAS statements, but they share a common feature—they end in a semicolon. A semicolon in a SAS program is like a period in English. Probably the most common error found in SAS programs is the omission of the semicolon. This omission causes the computer to read two statements as a run-on statement and invariably fouls things up.

SAS programs are comprised of SAS statements. Some of these statements provide information to the system, such as how many lines to print on a page and what title to print at the top of the page. Other statements act together to create SAS data sets, while other SAS statements act together to run predefined statistical or other routines. Groups of SAS statements that define your data and create a SAS data set are called a DATA step; SAS statements that request predefined routines are called a PROC (pronounced "prock") step. DATA steps tell SAS programs about your data. They are used to indicate where the variables are on data lines, what you want to call the variables, how to create new variables from existing variables, and several other functions we mention later. PROC (short for PROCEDURE) steps indicate what kind of statistical analyses to perform and provide specifications for those analyses. Let's look at an example. Consider this simple data set:

SUBJECT NUMBER	GENDER (M or F)	EXAM 1	EXAM 2	HOMEWORK GRADE
10	M	80	84	A
7	M	85	89	A
4	F	90	86	B
20	M	82	85	B
25	F	94	94	A
14	F	88	84	C

We have five variables (SUBJECT NUMBER, GENDER, EXAM 1, EXAM 2, and HOMEWORK GRADE) collected for each of six subjects. The unit of analysis, people for this example, is called an observation in SAS terminology. SAS software uses the term "variable" to represent each piece of information we collect for each observation. Before we can write our SAS program, we need to assign a variable name to each variable. We do this so that we can distinguish one variable from another when doing computations or requesting statistics. SAS variable names must conform to a few simple rules: They must start with a letter, be not more than eight characters (letters or numerals) in length, and cannot contain blanks or certain special characters such as commas, semicolons, etc. (The underscore character (_) is a valid character for SAS variable names and can be used to make variable names more readable.) Therefore, our column headings of "SUBJECT NUMBER," or "EXAM 1" are not valid SAS variable names. Logical SAS variable names for this collection of data would be:

```
SUBJECT   GENDER   EXAM1   EXAM2   HWGRADE
```

It is prudent to pick variable names that help you remember which name goes with which variable. We could have named our five variables VAR1, VAR2, VAR3, VAR4, and VAR5, but we would then have to remember that VAR1 stands for "SUBJECT NUMBER," and so forth.

To begin, let's say we are interested only in getting the class means for the two exams. In reality it's hardly worth using a computer to add up six numbers, but it does provide a nice example. In order to do this, we could write the following SAS program:

```
DATA TEST;  ①
   INPUT SUBJECT 1-2 GENDER $ 4 EXAM1 6-8 EXAM2 10-12  ②
   HWGRADE $ 14;
DATALINES;  ③
10 M  80  84 A
 7 M  85  89 A
 4 F  90  86 B
20 M  82  85 B
25 F  94  94 A
14 F  88  84 C
;
PROC MEANS DATA=TEST; ④
RUN; ⑤
```