Building Regression Models with SAS®

A Guide for Data Scientists

Robert N. Rodriguez

The correct bibliographic citation for this manual is as follows: Rodriguez, Robert N. 2023. *Building Regression Models with SAS®: A Guide for Data Scientists*. Cary, NC: SAS Institute Inc.

Building Regression Models with SAS®: A Guide for Data Scientists

ISBN 978-1-955977-94-4 (Hardcover)
ISBN 978-1-63526-155-4 (Paperback)
ISBN 978-1-63526-190-5 (Web PDF)
ISBN 978-1-951684-00-6 (EPUB)
ISBN 978-1-951684-01-3 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

February 2023

Contents

Quick Guide to Key Procedures

Table 1 SAS 9 Procedures for Building Regression Models

Procedure	Model	Introduction	Example
GLMSELECT	General linear models including least squares regression	page 50	page 56
QUANTSELECT	Quantile regression	page 143	page 146
HPLOGISTIC	Logistic regression	page 163	page 174
HPGENSELECT	Generalized linear models	page 197	page 201
GAMPL	Generalized additive models	page 228	page 230
HPSPLIT	Classification and regression trees	page 274	page 276
ADAPTIVEREG	Multivariate adaptive regression splines	page 302	page 304

Table 2 SAS Viya Procedures for Building Regression Models

Procedure	Model	Introduction	Example
REGSELECT	General linear models including least squares regression	page 77	page 81
QTRSELECT	Quantile regression	page 155	page 156
LOGSELECT	Logistic regression	page 183	page 187
GENSELECT	Generalized linear models	page 215	page 216
GAMMOD	Generalized additive models	page 239	page 239
GAMSELECT	Generalized additive models	page 244	page 246
PHSELECT	Proportional hazards models	page 259	page 261
TREESPLIT	Classification and regression trees	page 282	page 283

Preface

Highway 53 in Cibola National Forest, New Mexico

If you travel in the western mountains of the United States, you will eventually encounter the Continental Divide. When a thunderstorm drops its contents on the divide, a portion flows eastward to the Mississippi River and then to the Atlantic Ocean; the other portion flows westward to the Pacific Ocean. During the 1800s, the Great Divide, as it is known, was the highest hurdle faced by settlers trekking across the American frontier until the construction of railways.

Great divides are also encountered in scientific fields, where philosophical differences impede practical applications until they are eventually resolved—often by breakthroughs in technology. In the field of statistics, the great divide of the 20th century was the disagreement between proponents of frequentist and Bayesian approaches. Today, objective Bayesian methods are widely accepted due to computational advances in the 1990s.

Machine learning has created a new divide for the practice of statistics, which relies heavily on data from well-designed studies for modeling and inference. Statistical methods now vie with algorithms that learn from large amounts of observational data. In particular, the new divide influences how regression models are viewed and applied. While statistical analysts view regression models as platforms for inference, data scientists view them as platforms for prediction. And while statistical analysts prefer to specify the effects in a model by drawing on subject matter knowledge, data scientists rely on algorithms to determine the form of the model.

This book equips both groups to cross the divide and find value on the other side by presenting SAS procedures that build regression models for prediction from large numbers of candidate effects. It introduces statistical analysts to methods of predictive modeling drawn from supervised learning, and at the same time it introduces data scientists to a rich variety of models drawn from statistics.

Throughout, the book uses the term *model building* because the procedures provide far more than sequential methods for model selection such as stepwise regression. The procedures also provide shrinkage methods, methods for model averaging, methods for constructing spline effects, and methods for building trees.

Motivation for the Book

The need for this book originated some years ago with the introduction of SAS/STAT procedures that were specifically designed to build regression models for prediction. The first was the GLMSELECT procedure, which builds general linear models (Cohen 2006). It not only equips analysts with modern methods for prediction but also provides the scalability that is essential in data mining and business analytics, where the number of observations can be in the millions and the number of potential predictors can be in the tens of thousands.

The GLMSELECT procedure was followed by a series of procedures that build other types of models. For instance, the HPLOGISTIC procedure builds logistic regression models, and the HPGENSELECT procedure builds generalized linear models.

Naturally, with so many new tools to choose from, SAS users began to ask questions such as the following:

How is the GLMSELECT procedure different from the GLM and REG procedures?

What methods are available for model validation?

Should I switch from the GAM procedure to the GAMPL procedure?

Can I trust the *p*-values and confidence intervals I get when I select a model?

Because SAS is a global company, I had opportunities to answer these questions in presentations at customer sites, user events, and conferences on six continents. During this odyssey, I discussed the new procedures with business analysts, data scientists, researchers, university faculty, and students. As I listened to their questions, I realized they could benefit from a book that served as a guide to the procedures. I also recognized that there were two audiences for this book.

Audiences for the Book

Many of the people I met during my odyssey were SAS users with experience in applied statistics. They were familiar with the GLM and GENMOD procedures for analyzing general linear models and generalized linear models. However, they were now working on projects where they had to build predictive models from large databases—and that was unfamiliar territory.

For this audience, adopting the model building procedures is mostly a matter of learning the concepts of predictive modeling, which differ considerably from the concepts of statistical modeling. The syntax of the procedures is not a challenge because they implement MODEL and CLASS statements similar to those of the GLM procedure.

I also met SAS users who had graduated from programs in business analytics and data science where machine learning is emphasized. They knew the basics of linear regression and logistic regression but were not acquainted with more specialized regression models. Nonetheless, they understood the concepts of predictive modeling because they were familiar with supervised learning.

For this second audience, adopting the model building procedures involves learning about regression models that are considerably more versatile than standard linear and logistic regression. With quantile regression and generalized additive models, for instance, it is possible to gain insights that cannot be obtained with conventional regression models.

Knowledge Prerequisites for the Book

This book assumes you know the basics of regression analysis. It uses standard matrix notation for regression models but explains the concepts and methods behind the procedures without mathematical derivations.

For readers who want to dive into the technical aspects of concepts and algorithms, explanations are given in appendices which use calculus and linear algebra at the level expected by master of science programs in data science and statistics.

The book also assumes you know enough about SAS to write a program that reads data and runs procedures. If you are new to SAS or need a refresher, an excellent primer is *The Little SAS Book* by Delwiche and Slaughter (2019).

Software Prerequisites for the Book

This book covers regression model building procedures in SAS 9 and SAS Viya. In order to run the examples that illustrate the procedures in SAS 9, you need SAS 9.4 with SAS/STAT installed. In order to run all the examples, you need SAS Viya with SAS Visual Statistics installed.

If you have questions about the software in this book, contact SAS Technical Support at https://support.sas.com. You can enter a request or problem at Submit a Support Request. Other questions related to this book can be directed to the book website at https://support.sas.com/rodriguez.

What the Book Does Not Cover

Regression analysis is a vast subject. This book does not cover the use of regression models for statistical inference, nor does it cover the basics of regression analysis. It is not a substitute for introductory textbooks such as Rawlings, Pantula, and Dickey (1998); Sheather (2009); Montgomery, Peck, and Vining (2012); and Fox (2016).

Furthermore, this book does not cover predictive or prognostic models in clinical research, biostatistics, and epidemiology. In those areas, the data typically come from carefully designed studies, the model variables are specified based on scientific knowledge, and the models are crafted by checking model assumptions, transforming variables, imputing missing data, and applying diagnostic techniques. The book *Regression Modeling Strategies* by Harrell (2015) gives a thorough exposition of these methods, and it discusses problems with automatic model selection that are echoed here. Another recommended book is *Clinical Prediction Models* by Steyerberg (2019).

Moreover, this book does not cover the breadth of methods now available for supervised learning or statistical learning. For the latter, the definitive text is *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman (2009). Another useful text is *An Introduction to Statistical Learning* by James et al. (2021).

Finally, this book is not a substitute for the procedure documentation in *SAS/STAT User's Guide* and *SAS Visual Statistics: Procedures*. The documentation contains details of features, methods, and options that are not covered in this book. SAS documentation is available at https://support.sas.com/en/documentation.html.

Acknowledgments

During the writing of this book, I benefited extensively from technical reviews provided by SAS developers who are highly knowledgeable about different aspects of regression modeling. I thank Ralph Abbey, Weijie Cai, Fang Chen, Bob Derr, Bruce Elsheimer, Gordon Johnston, David Kessler, Michael Lamm, Warren Kuhfeld, Pushpal Mukhopadhyay, Ying So, Clay Thompson, Randy Tobias, Yingwei Wang, and Yonggang Yao.

I also received expert advice from SAS staff who work in the areas of Technical Support (Cyrus Bradford, Phil Gibbs, and David Schlotzhauer), the Output Delivery System (David Kelley and Dan O'Connor), and ODS Graphics (Dan Heath, Prashant Hebbar, and Lingxiao Li).

My thanks also go to Joseph Gardiner (Michigan State University), Aric LaBarr (North Carolina State University), Simon Sheather (University of Kentucky), Besa Smith (ICONplc and Analydata), and Tyler Smith (National University), who read the manuscript and sent me detailed comments that led to significant improvements.

Tim Arnold, Bob Derr, Warren Kuhfeld, and Ed Porter patiently answered questions about the LaTeX system with which I typeset the book. I am also grateful to Ed Huddleston, who meticulously formatted the references so that I could manage them with BibTeX, and to Jennifer Evans and Karissa Wrenn, who pointed me to online research tools that were indispensable.

I also extend thanks to Gary McDonald, who showed me the merits of ridge regression at a time when it had not yet gained acceptance; Richard Cutler, who taught me much about classification and regression trees; and Robert Cohen, who introduced me to model averaging and to parallel computing techniques that underpin the scalability of the procedures.

This book would not have come about without the guidance of Catherine Connolly, my editor at SAS Press, nor could it have been completed without the expertise of Suzanne Morgen, my copy editor.

It is a special pleasure to acknowledge assistance from three members of my family. My daughter-in-law Kayla provided technical advice for one of the examples. My daughter Susan, who is a champion of correct English usage, often pointed out how I could improve my writing. And my wife Sandra listened carefully whenever I mentioned a problem with the book and invariably helped me think about it in ways that proved useful.

Chapter 1
Introduction

Contents

This book is about SAS procedures that use algorithmic methods to build a variety of regression models for prediction. The following introduction explains the origins of the methods, the benefits of the procedures, and how to read the book.

Model Building at the Crossroads of Machine Learning and Statistics

The procedures in this book combine regression models from the field of statistics with predictive modeling methods from the field of machine learning. The two fields take very different approaches to model building:

- In machine learning, models are computational algorithms. Large amounts of data are used to train predictive models that can have hundreds of thousands of parameters. Models are assessed by how well they generalize—in other words, how well they predict new data not seen during training. The internal complexity of machine learning models cannot be grasped by the human mind but enables them to excel at prediction by capturing intricate relationships among myriad variables. Although these models lend themselves to automation, their reliability rests squarely on the quality of the data and the selection of model features.
- In statistics, models are assumptions about the process that generates the data. Statistical modeling relies on scientific and business knowledge to decide which variables and effects to include in the model. Statistical modeling employs algorithms to estimate model parameters, to determine how well the model agrees with the data, and to make inferences about the process. Unlike machine learning models, which attempt to capture reality in its entirety, statistical models are simplified descriptions of reality; this makes them valuable for understanding which variables affect the response. Furthermore, statistical models distinguish between signal and noise, and they quantify uncertainty in their results by incorporating a probability distribution for noise.

Statistical Learning: A Blend of Two Cultures

The approaches followed by statistics and machine learning remained far apart in practice until data mining drew them together in the 1990s. Companies in the retail, insurance, financial, and telecommunications sectors began using machine learning methods to find patterns in large customer and transactional databases. Concurrently, companies began using statistical models for regression and classification to predict outcomes and make business decisions. Banks, for example, adopted this combination of approaches for credit and market risk analysis, fraud detection, and gaining insights about customers.

In a 2001b paper written for statisticians, Leo Breiman at the University of California, Berkeley, referred to the statistical approach as the "data modeling culture" and the machine learning approach as the "algorithmic modeling culture." Over time, the paper convinced many statisticians that they too needed a broader set of methods to solve real world problems in settings where large amounts of data are available and the goal is prediction rather than inference.

Breiman's thinking predated the growth of business analytics and the advent of data science. Today, the methods for algorithmic modeling that he advocated—in particular, random forests—are widely accepted by statisticians and are standard tools for data scientists.

During the last 20 years, powerful new methods have emerged from the area of statistical learning, which combines the goals of statistical modeling with concepts from supervised and unsupervised learning (Hastie, Tibshirani, and Friedman 2009; James et al. 2021; Hastie, Tibshirani, and Wainwright 2015).

Overview of Procedures for Building Regression Models

The procedures in this book build the following regression models:

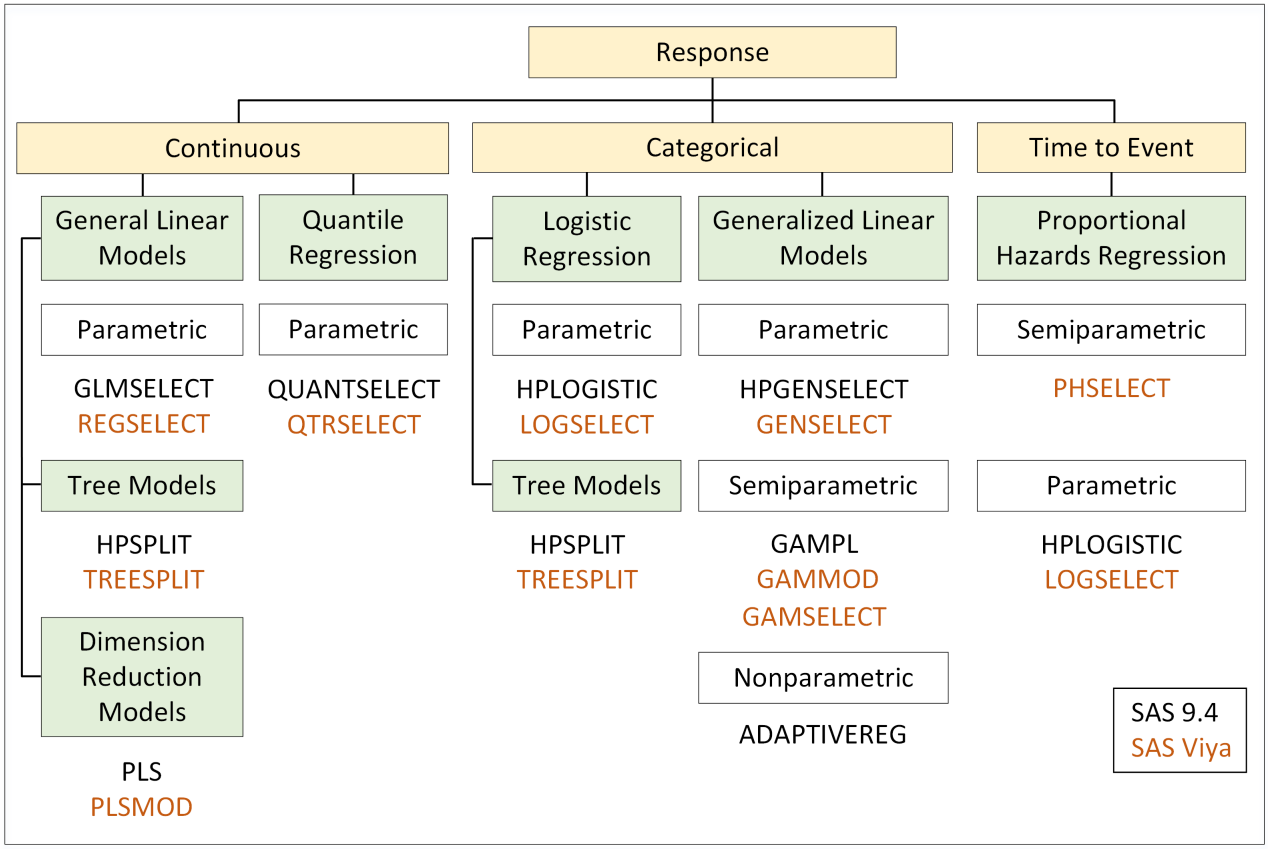
- standard regression models
- general linear models
- quantile regression models
- logistic regression models
- generalized linear models
- generalized additive models
- proportional hazards regression models
- tree models for classification and regression
- models based on multivariate adaptive regression splines

The procedures provide algorithmic methods for building predictive and explanatory regression models when there are many candidate predictors to choose from. These methods include the following:

- sequential selection methods such as forward and stepwise regression
- shrinkage methods such as the lasso
- dimension reduction methods such as principal components regression
- model averaging
- recursive partitioning for constructing tree models
- multivariate adaptive regression splines

Figure 1.1 is a high-level view of the procedures in this book categorized by the models they build. Each model is explained by a chapter in the book, and for most models there is a procedure in SAS 9 and a procedure in SAS Viya. For instance, Chapter 5 explains how to build general linear models with the GLMSELECT procedure in SAS 9 and the REGSELECT procedure in SAS Viya.

Figure 1.1 Procedures in SAS 9 and SAS Viya for Building Regression Models

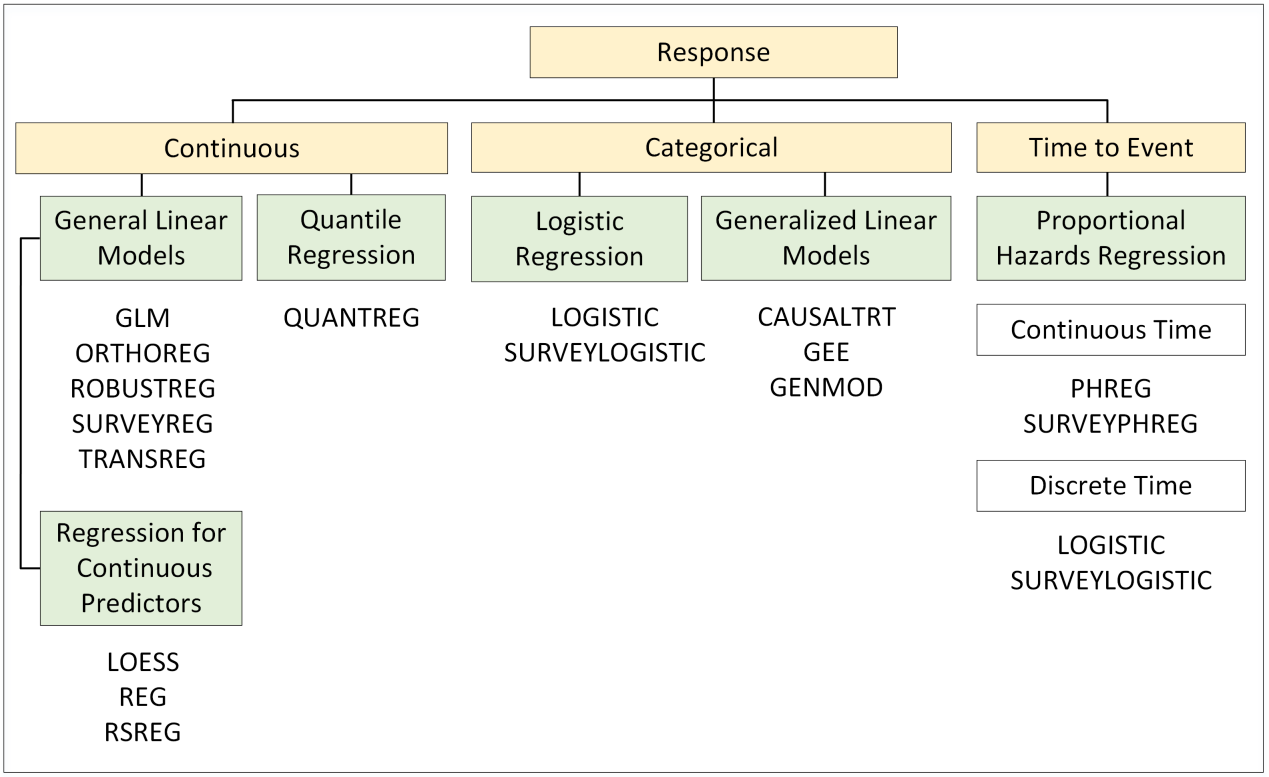


Practical Benefits

Both data scientists and statistical analysts will benefit from the models and procedures in this book:

- If you are a data scientist, you are undoubtedly familiar with linear regression and logistic regression as basic models for supervised learning. The specialized models presented here give you valuable alternatives for prediction. For instance, by applying quantile regression to customer lifetime value, you can predict the 10th, 50th, and 90th percentiles of the response, which can give you important insights about customer behavior.
- If you are a statistical analyst, you might be familiar with the SAS/STAT regression procedures in Figure 1.2 (not covered in this book), which fit models for statistical analysis. The model building procedures share various features with the statistical analysis procedures, which will help you get started with predictive modeling. For instance, if you have used the GENMOD procedure to analyze generalized linear models, you will encounter the same basic syntax in the HPGENSELECT procedure, which builds generalized linear models for prediction.

Figure 1.2 Procedures in SAS 9 for Statistical Analysis of Fitted Regression Models



Regardless of your training, the procedures in this book provide three major benefits: scalability, versatility, and interpretability.

Scalability

The procedures are designed to build models with large data; they provide high-performance computing on a single machine by exploiting all the available cores and threads. The procedures in SAS Viya can also run on a cluster of machines that distribute the data and the computations. With this much computational power, you can readily explore different combinations of models, model building methods, and selection criteria to determine which ones provide the best predictive performance for your work.

Versatility

With the procedures in this book, you can predict variables that are continuous, categorical, or the time to an event. Most of the procedures provide the following features:

- specification of candidate effects that can be continuous variables, categorical variables, classification effects, or spline effects
- a variety of parameterizations for classification levels
- assessment of prediction error based on validation, cross validation, and information criteria
- partitioning of data into training, validation, and testing roles
- tables and graphs for understanding the model building process
- facilities for saving the final model and scoring future data

Interpretability

All of the models in this book are interpretable. Most have a linear or additive structure that describes how the inputs are combined to predict the response. Parameter estimates and model summaries can yield valuable insights, especially when interpreted with the help of business or scientific knowledge. Other models—for instance, tree models—produce rules you can readily understand.

When Does Interpretability Matter?

The interpretability of a model is essential when it is used to make critical decisions. In the biopharmaceutical, insurance, and banking industries, interpretability is a regulatory requirement.

Interpretability is also essential for avoiding algorithmic bias, an ethical concern in any endeavor that relies on black box models—models whose internal working cannot be ascertained—to make life-changing decisions. There is growing evidence that black box models can be seriously biased.

One example, reported in *The Wall Street Journal*, is a complex algorithm deployed by a large hospital to identify patients with diabetes, heart disease, and other chronic conditions who could benefit from having health care workers monitor their status. A 2019 research study discovered the algorithm was giving priority to healthier white patients over sicker black patients because it used cost to rank patients. As it turned out, health care spending was less for black patients than it was for white patients with similar conditions (Evans and Mathews 2019; Obermeyer et al. 2019).

The COVID-19 pandemic accentuated the importance of interpretability. Predictive models for case counts and hospitalizations were at the forefront of medical efforts and policy making, and the reliability of these models was hotly debated in social media. Across the globe, there was an unprecedented need for models that could be understood by the general public (Stuart et al. 2020).

Can Interpretable Models Match the Accuracy of Black Box Models?

Although machine learning algorithms are only superficially understood by humans, they are often preferred over regression models because they can achieve higher predictive accuracy. Nonetheless, it is not unusual for evaluations of machine learning methods to conclude that the accuracy of logistic regression is comparable to that of support vector machines, neural nets, and random forests.

That was the finding of a recent study in *JAMA Neurology*, which compared algorithms for predicting the likelihood of freedom from seizure for patients after their first prescribed antiseizure medication (Chiang and Rao 2022; Hakeem et al. 2022). Likewise, an editorial in *Nature Medicine* concerning misuse of machine learning in clinical research recognized that it is unlikely to improve over statistical models in problems where the effects of strong predictors, chosen on the basis of prior research, are inherently linear (Volovici et al. 2022).

Cynthia Rudin, a prominent computer scientist at Duke University, and her colleagues have provided compelling examples to demonstrate that an interpretable model is not necessarily less accurate than a black box model (Rudin and Radin 2019; Rudin, Wang, and Coker 2020). They recommend against the use of black box models for high-stakes decisions unless no interpretable model with the same level of accuracy can be constructed. They conclude, "It is possible that an interpretable model can always be constructed—we just have not been trying" (Rudin and Radin 2019, p. 7).

Explanation Models

Rudin (2019) distinguishes between models (such as regression models) that are inherently interpretable and so-called explanation models used to elucidate black box models in machine learning. An explanation model can only approximate the original model (otherwise it would not be necessary), which raises the question of which model to trust. Reliance on two models that can disagree complicates the decision-making process. And even if an explanation model makes predictions that are identical to those of the black box model, the explanation that it offers could well be incorrect in the sense that it incorporates a different set of features in the data. This can occur, for instance, if the features happen to be correlated with those of the black box model. In fact, highly correlated variables are common in large databases.

Explanatory Models

The field of statistics uses the term *explanatory model*—not to be confused with explanation model!—for models that yield understanding about the process generating the data. The most valuable explanatory models are simple—they incorporate only the most relevant variables in the data, and they give you meaningful parameter estimates. On the other hand, the most valuable *predictive models* minimize prediction error when applied to future data, whether or not the parameter estimates have meaning. The two kinds of models are fundamentally different, and they require distinct approaches.

When Should You Use the Procedures in This Book?

You should use the procedures in this book when you are faced with many potential predictors and you need to build a regression model that not only generalizes well to future data but is also interpretable.

If interpretability is not a concern, you should also consider random forests, neural networks, gradient boosting, and other methods of supervised learning. These lie outside the scope of this book but are available in SAS Enterprise Miner, which runs in SAS 9, and SAS Visual Data Mining and Machine Learning, which runs in SAS Viya.

Comparison with Regression Procedures for Statistical Analysis

If you need to evaluate the statistical significance of effects in a regression model, you should not use the model building procedures because the p-values and confidence limits they compute are unadjusted for the process of model selection. Instead, you should use one of the many procedures available in SAS/STAT for statistical analysis of regression models whose effects are fully specified based on domain knowledge. Those procedures, listed in Figure 1.2, are not covered in this book.

Among the procedures for statistical analysis, the REG, LOGISTIC, and PHREG procedures provide limited features for effect selection. These features are superseded by modern model building capabilities available in the GLMSELECT, HPLOGISTIC, and PHSELECT procedures, respectively.

Table 1.1 compares regression procedures that build models with those that analyze models.

Table 1.1 Comparison of Regression Procedures for Model Building and Statistical Analysis

Characteristics	Procedures for Building Regression Models	Procedures for Fitting and Analyzing Regression Models
Goal	Prediction of future observations	Inference about model parameters
Model effects	Selected from candidates	Specified by analyst
Types of effects	Continuous, classification, polynomial, spline	Continuous, classification, polynomial, spline
Uses of training data	Estimation and model selection	Estimation
Parameter estimates	Subject to various types of bias	Unbiased if model is assumed to include true effects
p-values, confidence intervals	Unadjusted for model selection	Valid
Model fit diagnostics	No	Yes
Model validation methods	Yes	No
Model prediction	Applies to future data	Applies only to training data

How to Read This Book

The chapters of this book are organized into two parts:

- Part I is about building standard regression models and general linear models. It spans six chapters because so many methods are available for these models.
- Part II is about building specialized models. It provides a chapter for each type of model.

Part I begins by discussing concepts that play a role throughout the book. Chapter 2 explains the fundamentals of building predictive models, Chapter 3 explains issues associated with model building, and Chapter 4 explains methods for building generalized linear models.

Each of the subsequent chapters introduces the procedures that build a particular model. The chapter describes important characteristics of the model, summarizes essential procedure options, and presents examples that illustrate options and interpret the output. The examples are drawn from actual scenarios, but the data have been simplified to circumvent the many steps of data preparation that are inevitable in practice.

You can read the chapters independently of each other. However, if you are not familiar with model selection, you should start with Chapters 2, 3, and 4.

Two sets of appendices cover supplementary topics:

- Appendices in Part III cover algorithms and computational methods for readers who would like a deeper understanding of those aspects.
- Appendices in Part IV cover topics common to all of the procedures: methods for scoring data, coding schemes for categorical predictors, and the use of ODS graphics.

Programs, Data Sets, and Macros

Programs and data sets for the examples are available on the website for this book, which you can access at support.sas.com/rodriguez.

In a number of situations, the book fills gaps in procedure functionality by providing SAS macros which you can use in your own work. The macros are available on the book website.

Conventions

The book gives you many tips for running the procedures in your own programs. These are shown as follows:

Programming Tip: Suggestions in boxes like this one will save you time and work!

The SAS code for each example is presented in blocks consisting of statements that involve multiple steps and procedure options. In some situations, to focus attention on a critical option that might easily be overlooked, the option is highlighted as in the following example:

```
proc glmselect plots=coefficients data=Stores;
   class Region(param=reference ref='Midwest')
         Training(param=reference ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=lasso( adaptive  choose=sbc stop=sbc showstepL1);
run;
```

Throughout the book, you will encounter road warning signs in the margins. These are placed to alert you to unavoidable problems, common misunderstandings, and potential tripping points. When you see these signs, slow down and read carefully!