



Introduction

| | |
|---|----------|
| 1.1 About Receiver Operating Characteristic Curves | 1 |
| 1.2 Summary of Chapters | 3 |

1.1 About Receiver Operating Characteristic Curves

This book describes how to analyze receiver operating characteristic (ROC) curves using SAS software. A *receiver operating characteristic curve* is a statistical tool to assess the accuracy of predictions. It is often abbreviated as ROC curve or ROC chart, the latter being used more often in data mining literature.

Making predictions has become an essential part of every business enterprise and scientific field of inquiry. A simple example that has irreversibly penetrated daily life is the weather forecast. Almost all news sources, including newspaper, radio, and television news, provide detailed weather forecasts. There is even a dedicated television channel for weather forecasts in the United States.

Of course, the influence of a weather forecast goes beyond a city dweller's decision to pack an umbrella. Inclement weather has negative effects on many vital activities such as transportation, agriculture, and construction. For this reason, collecting data that help forecast weather conditions and building statistical models to produce forecasts from these data have become major industries.

It is important for the consumers of these forecasts to know their accuracy. This helps them to incorporate these predictions into their future plans. It also helps them decide between competing providers. Similarly, it is important for forecast providers to assess the accuracy of their forecasts since accuracy is a direct indicator of the quality of their product. Assessing accuracy is also important when providers decide to invest in technologies to improve the forecasts. An improvement in the forecast is intrinsically linked to an improvement in accuracy.

2 Analyzing Receiver Operating Characteristic Curves with SAS

Credit scoring is another example of making predictions: When a potential debtor asks for credit, creditors assess the likelihood of default to decide whether to loan the funds and at what interest rate. Accurately assessing a debtor's chance of default plays a crucial role for creditors to stay competitive. For this reason, the prediction models behind credit scoring systems remain proprietary. Nevertheless, their predictive power needs to be continuously assessed for creditors to remain profitable.

A final example concerns the field of medical diagnostics. The word *prediction* rarely appears in this literature, but a diagnosis is a prediction of what might be wrong with a patient exhibiting certain symptoms and complaints. Most diseases elicit a response that increases levels of a substance in the blood or urine. However, there might be other reasons for such elevated levels; hence, assessing blood or urine levels alone can create a misdiagnosis. It is critical to understand how accurate such diagnoses are because they influence subsequent evaluations and treatments. It is also common to have multiple diagnostic markers or tools available, and a fair assessment of them involves comparing their accuracies. These diagnostic options may differ in their cost and risk to the patient, in which case a decision analysis can be performed where the value of each tool is quantified based on the accuracy of the diagnoses made by the tool.

All these examples have a common theme. A prediction is made before the value of the entity that is predicted is known. We need a method to evaluate the accuracy of these predictions. As these examples make clear, it would be helpful if the method could compare the accuracy of several predictions.

Various conventions are used to name the predictions and the outcome. Table 1.1 summarizes the most commonly used names.

Table 1.1 Common Nomenclature for the Elements of an ROC Curve

| Variable That Predicts | Variable to Be Predicted | Values of the Variable to Be Predicted |
|-------------------------------|---------------------------------|---|
| Predictor | Outcome | Case/Control |
| Marker | Status | Diseased/Non-Diseased |
| Score | Gold Standard | Positive/Negative |
| Forecast | Indicator | Present/Absent Event/Non-Event |

ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. They are widely applicable, regardless of the source of predictions. You can also compare the accuracy of different methods of generating predictions by comparing the ROC curves of the resulting predictions. Therefore, it may come as a surprise to realize that ROC curves are generally ignored during the education and training of statisticians. Most statisticians learn about ROC curves on the job, as needed, and struggle through some of the unusual features of this type of analysis.

To make matters worse for SAS users, very few direct methods are available for performing an ROC analysis in SAS. However, many existing procedures can be tailored with little effort to produce ROC curves. SAS Institute also provides a macro to perform some of the calculations. This book describes how to produce ROC curves with the available features in SAS and expands on further analyses using other SAS procedures.

1.2 Summary of Chapters

Methods for evaluation of accuracy depend on the nature of the predictor. Chapter 2, “Single Binary Predictor,” and Chapter 3, “Single Continuous Predictor,” introduce appropriate methods for binary and continuous variables. These two chapters discuss material that is used repeatedly in subsequent chapters, so you must have a good grasp of these concepts before reading further. If you are already familiar with these statistical concepts but are more interested in learning the capabilities of SAS with respect to ROC curves, skip the parts introducing and discussing these concepts. Most of the SAS code in this book is presented within the context of examples, so it will be sufficient for those readers to have a cursory reading of Chapters 2 and 3 to familiarize themselves with notation and then carefully follow the examples to master the SAS code.

Most of the computations are performed using PROC FREQ, PROC LOGISTIC, or PROC NLMIXED. There are also a few macros that are very useful in plotting the ROC curve or computing the standard errors of the areas under the ROC curves. Occasional calls to PROC TRANSREG (for Box-Cox transformation) or PROC MIXED along with the use of PROC SURVEYSELECT for creating bootstrap samples are used.

Note: There is no standard mathematical notation for most of what needs to be presented here. I tried to balance my personal preference with widely accepted practices; this is why a cursory reading is recommended, even for those who feel comfortable with the statistical concepts.

Chapter 4, “Comparison and Covariate Adjustment of ROC Curves,” compares the ROC curves of several markers and adjusts them for covariates. The principle tool for this purpose is regression, which accommodates both categorical and continuous covariates. Regression methods can also be used to compare the accuracy of several markers by representing the markers with dummy variables in an ANOVA-like model. Although the mechanistic aspects of these regression models are similar to other regression models, the inclusion and interpretation of model coefficients are unique to the field of ROC curves.

Chapter 5, “Ordinal Predictors,” repeats the material in Chapters 3 and 4 for an ordinal predictor. The ideas are very similar, but the statistical techniques are slightly different, such as the use of a latent variable probit regression model, which is also commonly called the *binormal model* in ROC literature. While it is possible to study Chapter 5 with only a superficial understanding of the earlier material, I recommend mastering the concepts in Chapters 2 and 3 first.

Chapter 6, “Lehmann Family of ROC Curves,” and Chapter 7, “ROC Curves with Censored Data,” present relatively new material that has not yet made its way into other books. The Lehmann family of ROC curves, the focus of Chapter 6, uses the proportional hazards model exclusively. Proportional hazards models are routinely used in survival analysis but rarely in other applications. Chapter 6 shows how it can be used to create ROC curves and extended to regression models and clustered data using the capabilities of PROC PHREG. Most statisticians identify PROC PHREG with censored data, but Chapter 6 deals with a binary outcome that is fully observed, just like the outcome variables in Chapters 2 through 5. The problem of creating ROC curves with censored data is tackled in Chapter 7. Two methods of computing a concordance probability are provided, along with a discussion of time-dependent ROC curves.

Chapter 8, “Using the ROC Curve to Evaluate Multivariable Prediction Models,” discusses the use of ROC curves when multivariable prediction models are built and assessed on the same data set. Chapter 9 uses the same concepts in the context of data mining. Although the concepts remain the same, the primary SAS data mining engine, SAS Enterprise Miner, has a very

4 *Analyzing Receiver Operating Characteristic Curves with SAS*

different user interface and functionality than SAS/STAT software. Hence, most of Chapter 9 discusses how models are developed in SAS Enterprise Miner and how ROC curves can be produced using the built-in functionality for model assessment. Also shown are ways of exporting the data for the ROC curves so that you can create custom plots using SAS/GRAPH software.

Most of the SAS code presented consists of SAS/STAT and SAS/GRAPH software; Base SAS is used occasionally to prepare the data for analysis and plotting. SAS/STAT procedures `FREQ`, `LOGISTIC`, `MIXED`, and `NLMIXED` provide most of the required ammunition for the analyses.

You should have a basic understanding of linear models and regression analysis. This book assumes no prior experience with SAS/STAT procedures; however, if you aren't familiar with common SAS System concepts, such as BY processing or the `CLASS` statement, you may benefit from consulting a general-purpose SAS manual.

This book features a generic macro that can be used to plot ROC curves regardless of the nature or origin of the predictions. Those who find the options offered by this macro sufficient may not need any direct use of SAS/GRAPH software. However, graphical presentation involves a degree of personal style, and you might like to customize your curves and to use particular annotations. If so, you can use the intermediate data sets created by the macro and write your own SAS/GRAPH code to produce custom graphics.

It is likely that the SAS code presented in this book, especially the macros, will evolve. The code, which is available for download from the book's companion Web site at support.sas.com/gonen, will be updated routinely, so check the Web site frequently for the latest developments.