Meetings Courses Seminars Tips Conferences
Books Publications Training Software & More

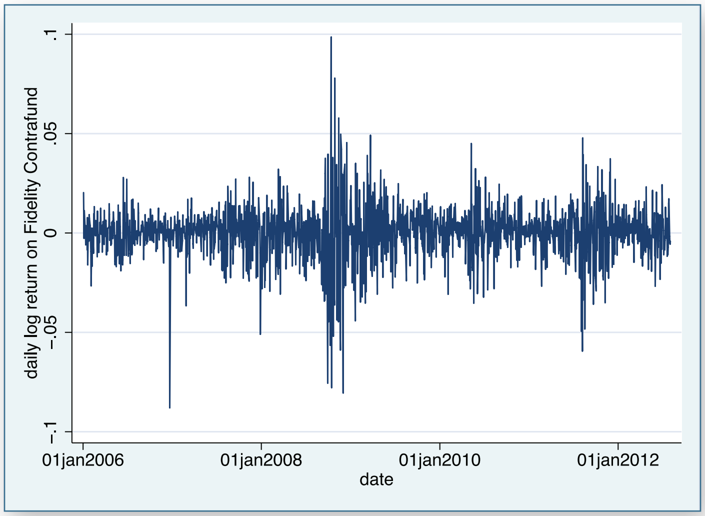
The Stata® News

Statistics Graphics Data Management & Analysis

Jul/Aug/Sept
Vol 27 No 3

In the spotlight: mgarch

You may be familiar with Stata's **arch** command, which fits univariate volatility models, also known as generalized autoregressive conditional heteroskedasticity (GARCH) models. In a GARCH framework, the conditional variance of a series is assumed to be a function of the prior volatility of the series. GARCH models are commonly applied to stock returns. To the right is a plot of log returns of a widely held mutual fund.



We can see that the returns exhibit a significant amount of volatility and that periods of high and low volatility tend to cluster. Techniques for modeling the volatility of a series are widely used in options trading, risk management, and asset allocation.

Multivariate GARCH (MGARCH) models generalize the univariate GARCH model and allow for relationships between volatility processes of multiple series. We want to know how changes in the volatility of one security affect the volatility of some other security. Those relationships can be parameterized in many different ways. Stata's **mgarch** command implements four commonly used parameterizations: the diagonal vech model (**mgarch dvech**), the constant conditional correlation model (**mgarch ccc**), the dynamic conditional correlation model (**mgarch dcc**), and the time-varying conditional correlation model (**mgarch vcc**).

Example

I use data on daily log returns of three Fidelity mutual funds—Intermediate Bond Fund (**bond**), Contrafund (**contra**), and Blue Chip Growth Fund (**bchip**). You could use Stata's **arch** command to analyze each series individually. For example, to fit a simple GARCH(1,1) model for the bond series, I type:

```
. arch bond, noconstant arch(1) garch(1)
```

Alternatively, you could use any of the conditional correlation MGARCH models. Here I use **mgarch dcc**:

```
. mgarch dcc (bond =, noconstant arch(1) garch(1))
```

Notice how easy it is to cast the **arch** syntax in terms of the **mgarch** syntax. I enclose the whole equation in parentheses and add the = sign after the dependent variable. Now you can guess how to fit a multivariate GARCH model without even looking at the help file of **mgarch dcc**:

```
. mgarch dcc (bond =, noconstant arch(1) garch(1))
             (contra =, noconstant arch(1) garch(1))
              (bchip =, noconstant arch(1) garch(1))
```

The Stata News
Executive Editor.............Karen Strope
Production Supervisor...Annette Fett

However, all the equations have the same ARCH and GARCH terms, so I can save on typing and "factor out" the common terms:

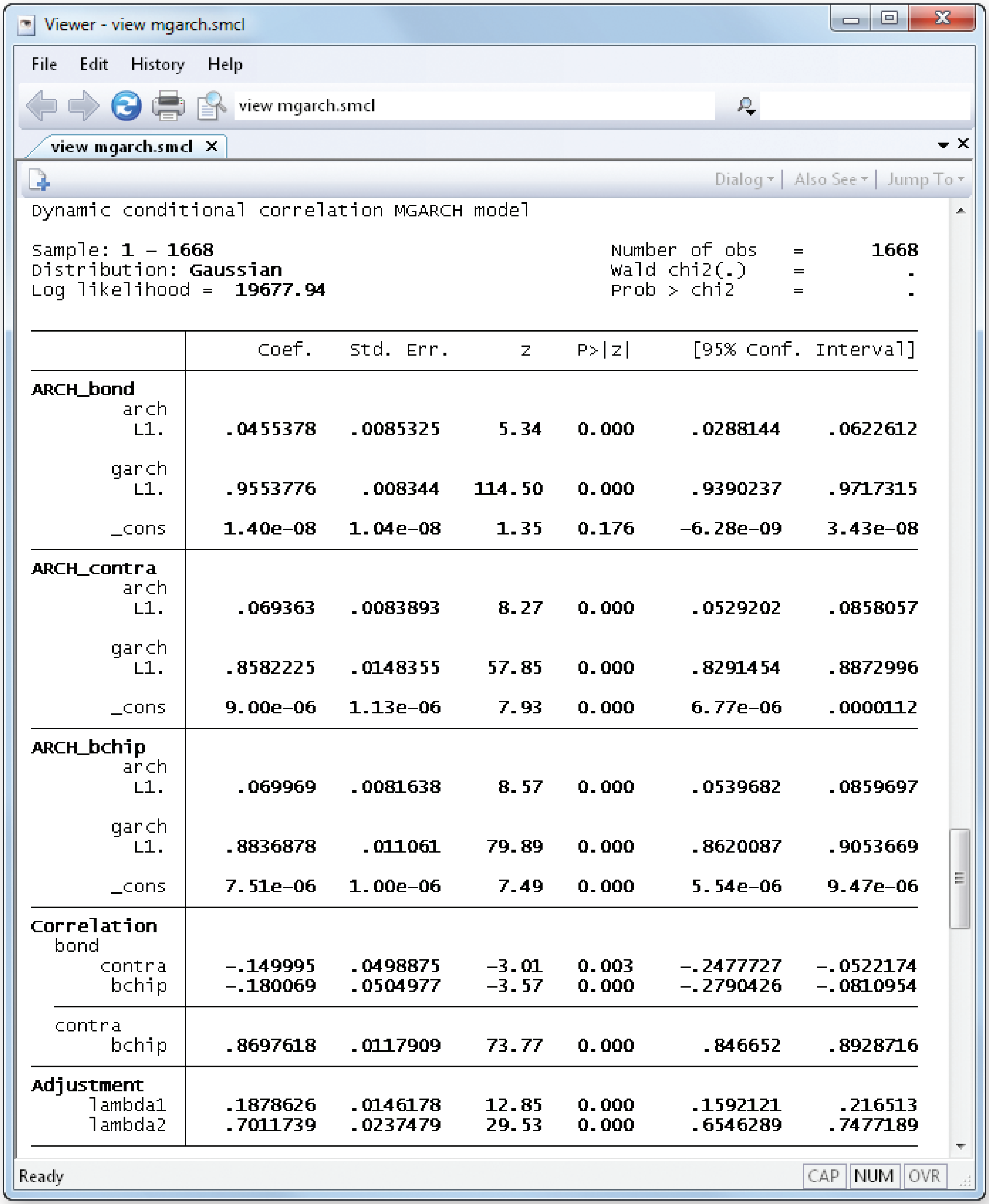
```
. mgarch dcc (bond =, noconstant)
            (contra =, noconstant)
             (bchip =, noconstant) , arch(1) garch(1)
```

"Factoring out" mean equation terms is not allowed, but I can combine them inside the parentheses:

```
. mgarch dcc (bond contra bchip =, noconstant) , arch(1) garch(1)
```

Now you see why the = sign is required; it separates dependent variables from covariates. The last syntax is the shortest, but the first syntax is the most flexible because it allows different mean and variance specifications for each equation.

Either form will produce the following estimates:



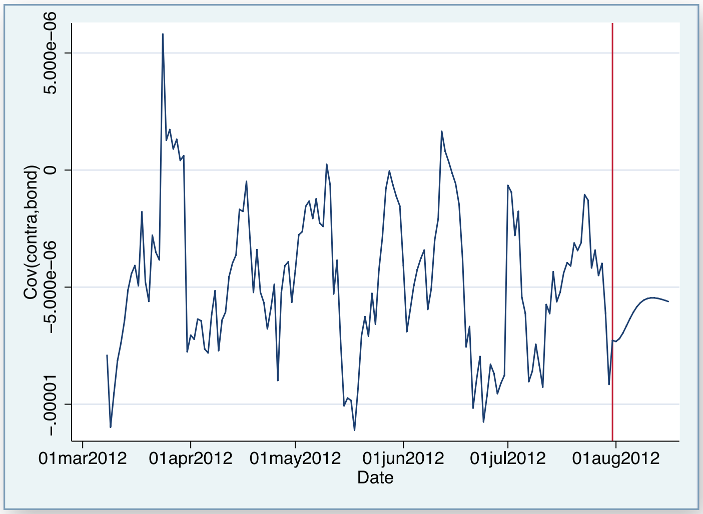
```
Dynamic conditional correlation MGARCH model

Sample: 1 - 1668                                Number of obs     =      1668
Distribution: Gaussian                          Wald chi2(.)      =         .
Log likelihood =  19677.94                      Prob > chi2       =         .

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ARCH_bond    |
        arch |
         L1. |   .0455378   .0085325     5.34   0.000     .0288144    .0622612
             |
       garch |
         L1. |   .9553776    .008344   114.50   0.000     .9390237    .9717315
             |
       _cons |   1.40e-08   1.04e-08     1.35   0.176    -6.28e-09    3.43e-08
-------------+----------------------------------------------------------------
ARCH_contra  |
        arch |
         L1. |    .069363   .0083893     8.27   0.000     .0529202    .0858057
             |
       garch |
         L1. |   .8582225   .0148355    57.85   0.000     .8291454    .8872996
             |
       _cons |   9.00e-06   1.13e-06     7.93   0.000     6.77e-06    .0000112
-------------+----------------------------------------------------------------
ARCH_bchip   |
        arch |
         L1. |    .069969   .0081638     8.57   0.000     .0539682    .0859697
             |
       garch |
         L1. |   .8836878    .011061    79.89   0.000     .8620087    .9053669
             |
       _cons |   7.51e-06   1.00e-06     7.49   0.000     5.54e-06    9.47e-06
-------------+----------------------------------------------------------------
Correlation  |
  bond       |
      contra |   -.149995   .0498875    -3.01   0.003    -.2477727   -.0522174
       bchip |   -.180069   .0504977    -3.57   0.000    -.2790426   -.0810954
  -----------+----------------------------------------------------------------
  contra     |
       bchip |   .8697618   .0117909    73.77   0.000      .846652    .8928716
-------------+----------------------------------------------------------------
Adjustment   |
     lambda1 |   .1878626   .0146178    12.85   0.000     .1592121     .216513
     lambda2 |   .7011739   .0237479    29.53   0.000     .6546289    .7477189
------------------------------------------------------------------------------
```

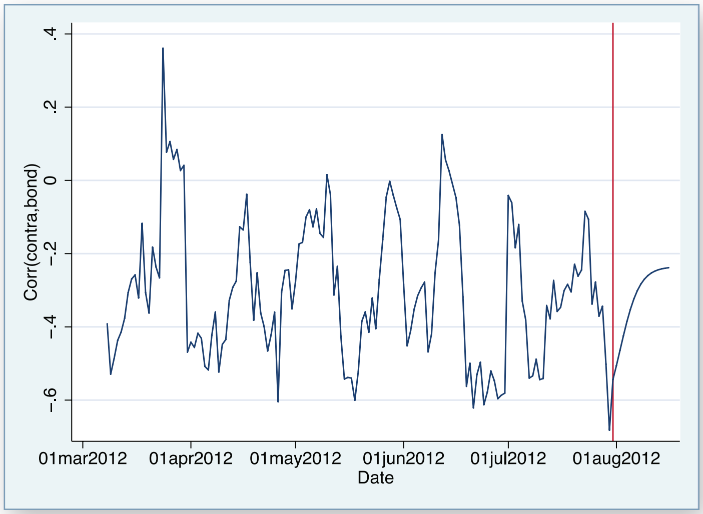
The interpretation of the variance parameters is the same for a univariate GARCH model. The new part is the estimated parameters that measure the dynamics between the volatilities. The estimated conditional quasicorrelation between the volatilities of the two stock funds, **contra** and **bchip**, is high and positive. This means that high volatility in the Contrafund is associated with high volatility in the Blue Chip fund and vice versa. The estimated conditional quasicorrelations between the volatility of the bond fund and the volatilities of the stock funds are low and negative. This means that increased volatility in the bond fund is associated with somewhat decreased volatility in the stock funds. This is all useful information if I want to minimize the volatility of a portfolio.

Having estimated the model, you may be interested in forecasting the series or their volatilities. Here I obtain predictions of the conditional variances and covariances for all the series. First, I use **tsappend** to extend the data, and then I use **predict** to obtain the predictions:

```
. tsappend, add(14)
. predict H*, variance dynamic(td(31jul2012))
```

The conditional predictions up to 31 July 2012 are in-sample one-step-ahead forecasts; after that they become dynamic out-of-sample forecasts. Below, on the left is a graph showing the predicted conditional covariance between the Contrafund and the bond fund. The dynamic forecasts are plotted to the right of the vertical line. It may be easier to interpret the relationship between the volatilities on a correlation scale; the graph on the right shows that.





The graphs show that the conditional correlation between the two volatilities varies over time. This behavior is a property of the dynamic and time-varying conditional correlation MGARCH models. In a constant conditional correlation MGARCH model, correlations do not vary over time and the line would be flat.

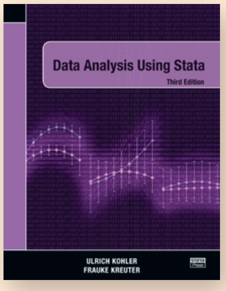
Summary

Multivariate GARCH models are used to analyze dynamic relationships between volatility processes of multiple series. Stata's **mgarch** command provides easy access to some of the commonly used parameterizations.

— Rafal Raciborski
Econometrician

New from Stata Press

Data Analysis Using Stata, Third Edition



Authors:	Ulrich Kohler and Frauke Kreuter
Publisher:	Stata Press
Copyright:	2012
ISBN-13:	978-1-59718-110-5
Pages:	497; paperback
Price:	$56.00

Data Analysis Using Stata, Third Edition has been completely revamped to reflect the capabilities of Stata 12. This book will appeal to those just learning statistics and Stata as well as to the many users who are switching to Stata from other packages. Throughout the book, Kohler and Kreuter show examples using data from the German Socioeconomic Panel, a large survey of households containing demographic, income, employment, and other key information.

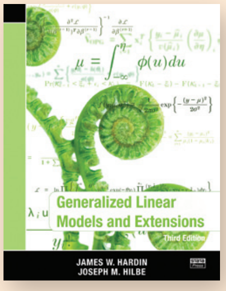
Kohler and Kreuter take a hands-on approach, first showing how to use Stata's graphical interface and then describing Stata's syntax. The core of the book covers all aspects of social science research, including data manipulation, production of tables and graphs, linear regression analysis, and logistic modeling. The authors describe Stata's handling of categorical covariates and show how the new **margins** and **marginsplot** commands greatly simplify the interpretation of regression and logistic results. An entirely new chapter discusses aspects of statistical inference, including random samples, complex survey samples, nonresponse, and causal inference.

The rest of the book includes chapters on reading text files, writing programs and ado-files, and using Internet resources such as the **search** command and the SSC archive.

Data Analysis Using Stata, Third Edition has been structured so that it can be used as a self-study course or as a textbook in an introductory data analysis or statistics course. It will appeal to students and academic researchers in all the social sciences.

You can find the table of contents and online ordering information at **stata-press.com/books/data-analysis-using-stata**.

Generalized Linear Models and Extensions, Third Edition



Authors:	James W. Hardin and Joseph M. Hilbe
Publisher:	Stata Press
Copyright:	2012
ISBN-13:	978-1-59718-105-1
Pages:	455; paperback
Price:	$58.00

Generalized linear models (GLMs) extend linear regression to models with a non-Gaussian, or even discrete, response. GLM theory is predicated on the exponential family of distributions—a class so rich that it includes the commonly used logit, probit, and Poisson models. Although one can fit these models in Stata by using specialized commands (for example, **logit** for logit models), fitting them as GLMs with Stata's **glm** command offers some advantages. For example, model diagnostics may be calculated and interpreted similarly regardless of the assumed distribution.

This text thoroughly covers GLMs, both theoretically and computationally, with an emphasis on Stata. The theory consists of showing how the various GLMs are special cases of the exponential family, showing general properties of this family of distributions, and showing the derivation of maximum likelihood (ML) estimators and standard errors. Hardin and Hilbe show how iteratively reweighted least squares, another method of parameter estimation, is a consequence of ML estimation by using Fisher scoring. The authors also discuss different methods of estimating standard errors, including robust methods, robust methods with clustering, Newey–West, outer product of the gradient, bootstrap, and jackknife. The thorough coverage of model diagnostics includes measures of influence such as Cook's distance, several forms of residuals, the Akaike and Bayesian information criteria, and various R^2-type measures of explained variability.

After presenting general theory, Hardin and Hilbe then break down each distribution. Each distribution has its own chapter that explains the computational details of applying the general theory to that particular distribution. Pseudocode plays a valuable role here, because it lets the authors describe computational algorithms relatively simply. Devoting an entire chapter to each distribution (or family, in GLM terms) also allows for including real-data examples showing how Stata fits such models, as well as presenting certain diagnostics and analytical strategies that are unique to that family. The chapters on binary data and on count (Poisson) data are excellent in this regard. Hardin and Hilbe give ample attention to the problems of overdispersion and zero inflation in count-data models.

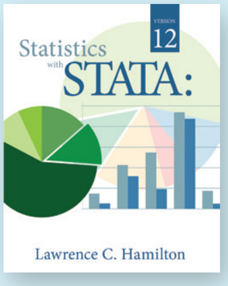
The final part of the text concerns extensions of GLMs, which come in three forms. First, the authors cover multinomial responses, both ordered and unordered. Although multinomial responses are not strictly a part of GLM, the theory is similar in that one can think of a multinomial response as an extension of a binary response. The examples presented in these chapters often use the authors' own Stata programs, augmenting official Stata's capabilities. Second, GLMs may be extended to clustered data through generalized estimating equations (GEEs), and one chapter covers GEE theory and examples. Finally, GLMs may be extended by programming one's own family and link functions for use with Stata's official **glm** command, and the authors detail this process.

In addition to other enhancements—for example, a new section on marginal effects—the third edition contains several new extended GLMs, giving Stata users new ways to capture the complexity of count data. New count models include a three-parameter negative binomial known as NB-P, Poisson inverse Gaussian (PIG), zero-inflated generalized Poisson (ZIGP), a rewritten generalized Poisson, two- and three-component finite mixture models, and a generalized censored Poisson and negative binomial. This edition has a new chapter on simulation and data synthesis but also shows how to construct a wide variety of synthetic and Monte Carlo models throughout the book.

You can find the table of contents and online ordering information at **stata-press.com/books/generalized-linear-models-and-extensions**.

New from the Stata Bookstore

Statistics with Stata: Version 12, Eighth Edition



Author:	Lawrence C. Hamilton
Publisher:	Cengage
Copyright:	2013
ISBN-13:	978-0-8400-6463-9
Pages:	512; paperback
Price:	$79.00

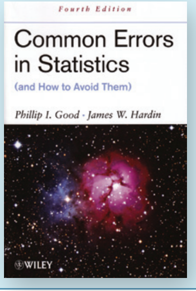
Statistics with Stata: Version 12 is the latest edition in Professor Lawrence C. Hamilton's popular *Statistics with Stata* series. Intended to bridge the gap between statistical texts and Stata's own documentation, *Statistics with Stata* demonstrates how to use Stata to perform a variety of tasks.

The first three chapters cover getting started in Stata, data manipulation, and graphics. Hamilton then introduces many statistical procedures available within Stata. These include summary statistics and tables, ANOVA, linear regression (and diagnostics), robust methods, nonlinear regression, regression models for limited dependent variables, complex survey data, survival analysis, factor analysis, cluster analysis, structural equation modeling, multiple imputation, time series, and multilevel mixed-effects models. The final chapter provides an introduction to programming.

The organization of this book makes it ideal for those who are new to statistics, experienced statisticians who are new to Stata, and Stata users wishing to explore Stata's capabilities in a new field. A series of example commands with brief descriptions at the beginning of each chapter demonstrates the Stata syntax for topics discussed in the chapter. For those already familiar with the statistical technique but not with the corresponding Stata commands, this example section may be all that is needed to begin an analysis using Stata. Following the example sections, Hamilton addresses each topic in more detail with descriptions of statistical procedures, examples using real data, and interpretation of the Stata output.

You can find the table of contents and online ordering information at **stata.com/bookstore/statistics-with-stata**.

Common Errors in Statistics (and How to Avoid Them), Fourth Edition


Fourth Edition
Common Errors in Statistics
(and How to Avoid Them)
Phillip I. Good · James W. Hardin
WILEY

Authors:	Phillip I. Good and James W. Hardin
Publisher:	Wiley
Copyright:	2012
ISBN-13:	978-1-118-29439-0
Pages:	336; paperback
Price:	$49.75

Common Errors in Statistics (and How to Avoid Them), Fourth Edition, by Phillip I. Good and James W. Hardin, contains a wealth of advice on how to improve experimental design, produce informative tables and graphs, and effectively analyze data. This book is not a treatise on statistical theory. Rather, it provides information on how to best apply that theory to real-world applications and obtain informative results. As the title implies, the book provides many examples of poorly executed analyses and then explains in detail how those examples can be improved.

The authors begin by discussing foundational issues of statistical analysis, including sources of error, data collection, and hypothesis formation. Chapter 2, on hypotheses, has been completely rewritten and now emphasizes the importance of formulating a null hypothesis and all the alternatives, including the conclusions that you would draw based on the outcome you later obtain. The chapter also discusses traditional Neyman–Pearson testing as well as decision making.

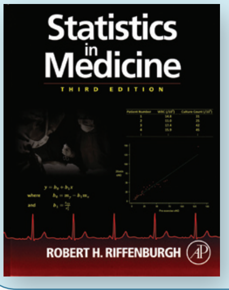
The second part of the book focuses on hypothesis testing and parameter estimation. Here the authors examine the statistical evaluation of the data as well as the strengths and limitations of various statistical procedures. Chapter 8, on how to report results, has been updated to reflect the strengths and weaknesses of *p*-values and confidence intervals as well as to show the important distinction between the statistical significance and the practical significance of results. Chapter 10 discusses how to make effective graphs and includes a list of 11 helpful rules to follow.

The last part of the book shows how to build a model, including linear and nonlinear regression, quantile regression, count models, and panel (longitudinal) data models. The final chapter discusses model validation.

The applied exposition in *Common Errors in Statistics (and How to Avoid Them), Fourth Edition* will be useful to experienced practitioners, and its many examples and careful explanations make it a helpful supplemental textbook for students of statistics.

You can find the table of contents and online ordering information at **stata.com/bookstore/common-errors-statistics**.

Statistics in Medicine, Third Edition



Author:	Robert H. Riffenburgh
Publisher:	Academic Press
Copyright:	2012
ISBN-13:	978-0-12-384864-2
Pages:	690; hardcover
Price:	$64.25

Statistics in Medicine, Third Edition, by Robert H. Riffenburgh, is an excellent book, useful as a reference for researchers in the medical sciences and as a textbook. It focuses largely on understanding statistical concepts rather than on mathematical and theoretical underpinnings. Riffenburgh covers both introductory statistical techniques and advanced methods commonly appearing in medical journals.

Riffenburgh begins with a discussion related to planning studies and writing articles to report results. Following this, he introduces statistics that would typically be covered in an introductory biostatistics course. These include summary statistics, distributions, two-way tables, confidence intervals, and hypothesis tests. In addition, he gives an overview of a variety of more sophisticated statistical techniques such as advanced regression, survival analysis, equivalence testing, Bayesian analysis, and time-series analysis.

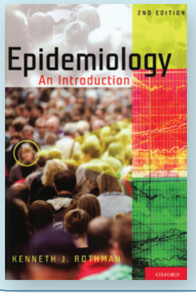
You can find the table of contents and online ordering information at **stata.com/bookstore/statistics-medicine**.

New from the Stata Giftshop

Bookmarks: Series 4

$2.25 in North America (price includes shipping)
$3.50 elsewhere (price includes shipping)

Epidemiology: An Introduction, Second Edition



Author:	Kenneth J. Rothman
Publisher:	Oxford University Press
Copyright:	2012
ISBN-13:	978-0-19-975455-7
Pages:	268; paperback
Price:	$33.50

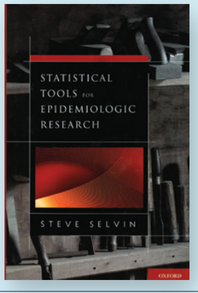
The study of epidemiology requires a deep understanding of statistics. Many texts focus on formulas, calculations, and software rather than on the concepts you need to know. It's easy to get lost in all the jargon.

This text is ideal for beginners in epidemiological statistics to learn the terminology and understand how and when to use statistical tools. The text focuses on concepts, not on mathematics, and discusses statistical techniques in the context of the real problems they can solve. This text bridges the gap between what is taught in an introductory statistics text and what you need to be an effective researcher and analyst.

The second edition has two new chapters, chapter 2 and chapter 6. Chapter 2 provides a historical overview of the early contributions by the pioneers of epidemiology and public health. Chapter 6 presents a summary of key concepts in infectious disease epidemiology. The text remains current with continually evolving epidemiological concepts and is now accompanied by a website (**oup.com/us/epi**), where readers can participate in discussions of the concepts presented in the book.

You can find the table of contents and online ordering information at **stata.com/bookstore/epidemiology-introduction**.

Statistical Tools for Epidemiologic Research



Author:	Steve Selvin
Publisher:	Oxford University Press
Copyright:	2011
ISBN-13:	978-0-19-975596-7
Pages:	494; hardcover
Price:	$56.00

This is a new book from the author of the acclaimed *Statistical Analysis of Epidemiologic Data*. It is intended for a second course in statistical methods. Calculus is used occasionally, but most of the text requires only a knowledge of elementary mathematics. Stata datasets and output are available at the book's website.

The book is mostly devoted to regression modeling, with separate chapters on logistic regression, Poisson regression, and conditional logistic regression. The chapters emphasize confounding, interaction, and connections to the two-by-two table. There are also chapters on topics not often covered at this level, including misclassification, longitudinal data analysis, and smoothing.

You can find the table of contents and online ordering information at **stata.com/bookstore/statistical-tools-for-epidemiologic-research**.

Visit us at APHA 2012

San Francisco, California, October 27–31

The 2012 American Public Health Association (APHA) annual meeting will take place in San Francisco, California, from October 27–31. For more information, visit **apha.org/meetings/AnnualMeeting**.

Stop by booth #2000 to visit with the people who develop and support Stata software.

Public training courses

Course	Dates	Location	Cost
Multilevel/Mixed Models Using Stata	October 4–5, 2012	Washington, DC	$1,295
Panel-Data Analysis Using Stata	October 16–17, 2012	Washington, DC	$1,295
Structural Equation Modeling Using Stata NEW	October 25–26, 2012	Washington, DC	$1,295
Survey Data Analysis Using Stata	October 23–24, 2012	Washington, DC	$1,295
Using Stata Effectively: Data Management, Analysis, and Graphics Fundamentals	October 2–3, 2012 November 1–2, 2012	Washington, DC San Francisco, CA	$950

Multilevel/Mixed Models Using Stata

This course is an introduction to using Stata to fit multilevel/mixed models. Mixed models may contain more than one level of nested random effects, and hence, these models are also referred to as multilevel or hierarchical models, particularly in the social sciences. The course will be interactive, use real data, offer ample opportunity for specific research questions, and provide exercises to reinforce what you learn.

Panel-Data Analysis Using Stata

This course provides an introduction to the theory and practice of panel-data analysis. After introducing the fixed-effects and random-effects approaches to unobserved individual-level heterogeneity, the course covers linear models with exogenous covariates, linear models with endogenous variables, dynamic linear models, and some nonlinear models. An introduction to the generalized method of moments estimation technique is also included. Concepts are extensively illustrated using exercises and examples worked in Stata.

Structural Equation Modeling Using Stata NEW

This course covers the use of Stata for structural equation modeling (SEM). The course introduces a variety of models, including path analysis, confirmatory factor analysis, full structural equation models, latent growth curves, and more. Examples demonstrate the **sem** command as well as the SEM Builder—a graphical interface for building path diagrams that can be used to illustrate, specify, and estimate SEM models.

Survey Data Analysis Using Stata

This course covers how to use Stata for survey data analysis assuming a fixed population. The course covers the sampling methods used to collect survey data and how they affect the estimation of totals, ratios, and regression coefficients as well as Stata's support for many survey variance estimators, including linearization, balanced and repeated replications (BRR), and jackknife. Each topic will be illustrated with one or more examples using Stata.

Using Stata Effectively: Data Management, Analysis, and Graphics Fundamentals

Become intimately familiar with all three components of Stata: data management, analysis, and graphics. This two-day course, taught by Bill Rising (StataCorp's Director of Educational Services), is aimed at both new Stata users and those who want to use Stata more effectively. You will learn to use Stata efficiently and to make your work reproducible and self-explanatory. As a result, collaborative changes and follow-up analyses will become much simpler.

We offer a 15% discount for group enrollments of three or more participants. Contact us at **training@stata.com** for details. For course details, or to enroll, visit **stata.com/public-training.**

NetCourse® schedule

Enroll by visiting stata.com/netcourse.

NetCourse 101, Introduction to Stata

This introductory course is designed to take smart, knowledgeable people and turn them into proficient interactive users of Stata. It covers several detailed techniques and tricks to make you a powerful Stata user.

Dates:	October 19–November 30, 2012
Enrollment deadline:	October 18, 2012
Price:	$95

stata.com/netcourse/intro-nc101

NetCourse 151, Introduction to Stata Programming

This introductory course on Stata programming deals with what most statistical software users mean by programming, namely, the careful performance of reproducible analyses.

Dates:	October 19–November 30, 2012
Enrollment deadline:	October 18, 2012
Price:	$125

stata.com/netcourse/programming-intro-nc151

NetCourse 152, Advanced Stata Programming

This advanced course teaches you how to create and debug new commands that are indistinguishable from the commands in Stata. It is assumed that you know why and when to program and to some extent how. You will learn how to parse both standard and nonstandard Stata syntax by using the intuitive syntax command, how to manage and process saved results, how to process by-groups, and more.

Dates:	October 12–November 30, 2012
Enrollment deadline:	October 11, 2012
Price:	$150

stata.com/netcourse/programming-advanced-nc152

NetCourse 461, Introduction to Univariate Time Series with Stata

This course provides an introduction to univariate time-series analysis that emphasizes the practical aspects most needed by practitioners and applied researchers. The course is written to appeal to a broad array of users, including economists, forecasters, financial analysts, managers, and anyone who encounters time-series data.

Dates:	October 12–November 30, 2012
Enrollment deadline:	October 11, 2012
Price:	$295

stata.com/netcourse/univariate-time-series-intro-nc461

NetCourseNow

Would you prefer to choose the time and set the pace of a NetCourse? Want to have a personal instructor? NetCourseNow for you: stata.com/netcourse/ncnow.

Save the date: Stata Conference

NEW ORLEANS
JULY 18–19, 2013

Venue:	Hyatt French Quarter New Orleans 800 Iberville Street New Orleans, Louisiana 70112 frenchquarter.hyatt.com
Chair:	R. Carter Hill Louisiana State University

stata.com/new-orleans13

2012 Polish Stata Users Group meeting

Date:	October 19, 2012
Venue:	Faculty of Economic Sciences Warsaw University ul. Długa 44/50 PL-00241 Warsaw
Cost:	60 PLN regular; 30 PLN student
Details:	stata.com/meeting/poland12

At the 2012 Polish Stata Users Group meeting, users will have the opportunity to share their scientific achievements and the results of their empirical research using Stata. In addition, there will be a master lecture addressed to both experienced Stata users and non-Stata users.

Organizers

Scientific organizers

- Leszek Morawski (*Katedra Statystyki i Ekonometrii*)
- Jerzy Mycielski (*Katedra Statystyki i Ekonometrii*)
- Paweł Strawiński (*Katedra Statystyki i Ekonometrii*)

Logistics organizer

- Timberlake Consultants Ltd.
  biuro@timberlake.pl

For additional information, visit stata.com/meeting/poland12.

In the spotlight: Receiver operating characteristic curves

Suppose a doctor is trying to decide whether to diagnose an individual with pancreatic cancer. Several diagnostic tests may be administered that classify an individual as having cancer (case) or not having cancer (control) based on one of the individual's biological attributes. The measured attribute is the classification variable for the test. The individual's levels of the tumor marker CA 19-9 and the protein CA-125 are examples of classification variables that may be used to diagnose pancreatic cancer.

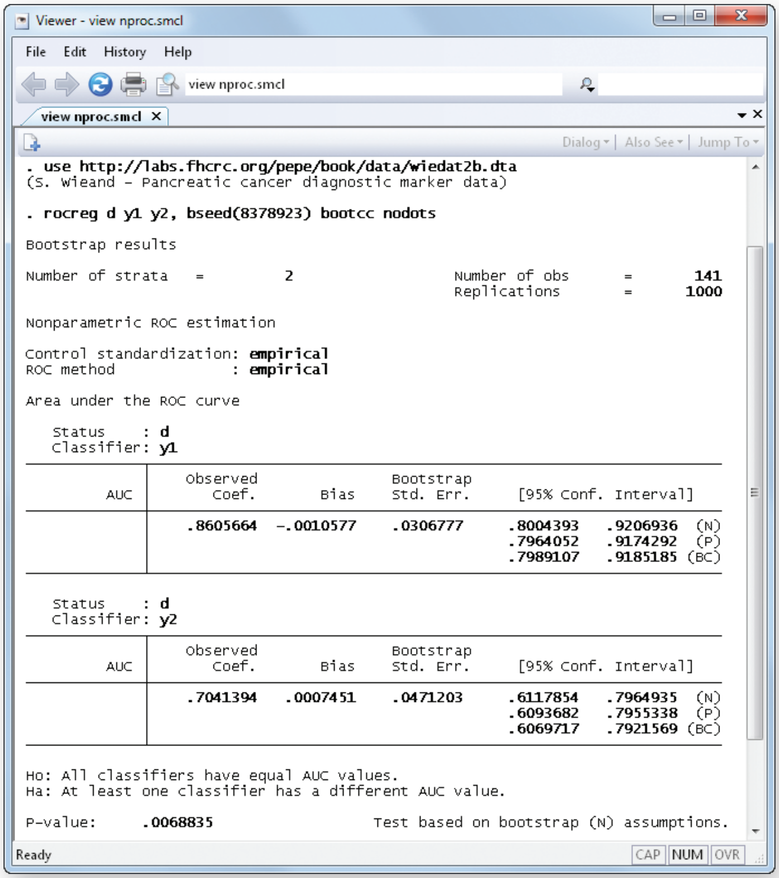
In this spotlight, we will see how receiver operating characteristic (ROC) curves can be used to evaluate diagnostic tests in Stata. In our first example, we use ROC curve analysis to evaluate the diagnostic tests for pancreatic cancer based on classifiers CA 19-9 and CA-125.

The user of the diagnostic test may choose different values of the classifier as "cutoffs" for diagnosing an individual. Those falling below the cutoff are classified as controls, while those above or at the cutoff are classified as cases. Each cutoff value of the classifier corresponds to a proportion of correctly classified subjects with the condition (true-positive rate) and to a proportion of incorrectly classified subjects without the condition (false-positive rate). An ideal test has cutoffs with high true-positive rates and low false-positive rates.

When the true conditions of the individuals are determined, the true-positive rates and false-positive rates can be estimated. The ROC curve is a graph of the estimated true-positive rate (ROC value) versus the estimated false-positive rate. ROC curves can be used to evaluate and compare diagnostic tests. In addition to the true-positive and false-positive rates, the area under the ROC curve (AUC) is a widely used summary statistic for ROC curves and may also be used to compare diagnostic tests.

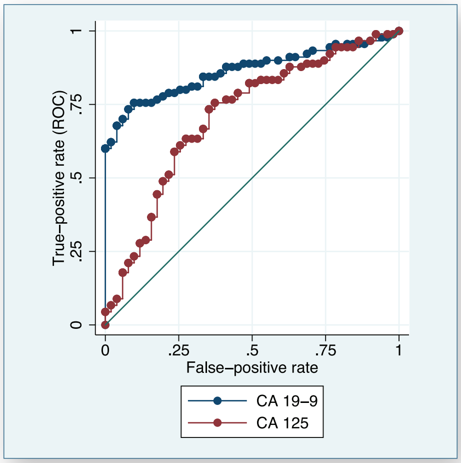
Nonparametric ROC curve estimation

We examine data from a pancreatic cancer study with two continuous classifiers, **y1** (CA 19-9) and **y2** (CA-125), for pancreatic cancer. The variable **d** indicates whether the individual truly has cancer. We use **rocreg** to make a nonparametric estimate of the true-positive and false-positive rates for the unique values of each classifier. **rocreg** also computes a nonparametric estimate of the AUC and uses the bootstrap to obtain the standard error. We use the **bootcc** option to specify that case- and control-cluster resampling be performed in the bootstrap.



```
. use http://labs.fhcrc.org/pepe/book/data/wiedat2b.dta
(S. Wieand - Pancreatic cancer diagnostic marker data)

. rocreg d y1 y2, bseed(8378923) bootcc nodots

Bootstrap results

Number of strata   =          2                  Number of obs      =       141
                                                 Replications       =      1000

Nonparametric ROC estimation

Control standardization: empirical
ROC method             : empirical

Area under the ROC curve

   Status    : d
   Classifier: y1

--------------------------------------------------------------------------------
             |    Observed                 Bootstrap
         AUC |       Coef.        Bias     Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------------------------
             |    .8605664   -.0010577     .0306777      .8004393    .9206936  (N)
             |                                           .7964052    .9174292  (P)
             |                                           .7989107    .9185185 (BC)
--------------------------------------------------------------------------------

   Status    : d
   Classifier: y2

--------------------------------------------------------------------------------
             |    Observed                 Bootstrap
         AUC |       Coef.        Bias     Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------------------------
             |    .7041394    .0007451     .0471203      .6117854    .7964935  (N)
             |                                           .6093682    .7955338  (P)
             |                                           .6069717    .7921569 (BC)
--------------------------------------------------------------------------------

Ho: All classifiers have equal AUC values.
Ha: At least one classifier has a different AUC value.

P-value:      .0068835                    Test based on bootstrap (N) assumptions.
```

The estimated AUC is higher under CA 19-9, and we reject that the two tests have equal AUC at the 0.01 level. To put these numbers in perspective, we use **rocregplot** to draw the ROC curve implied by the **rocreg** results:



The test that uses classifier CA 19-9 is clearly preferable to the test using classifier CA-125. For all but the highest false-positive rates, CA 19-9 yields a higher true-positive rate than CA-125.

Normal ROC curve estimation

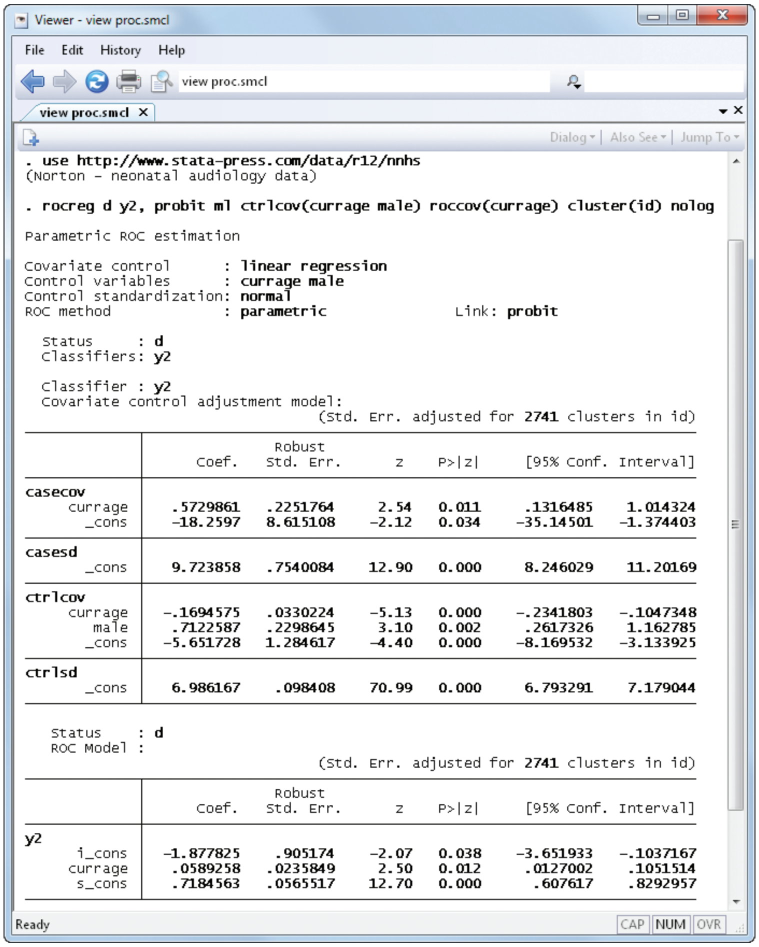
rocreg can also estimate ROC curves parametrically by using estimating equations or maximum likelihood (ML). With ML, we model the classifier **y**, using control covariates **z**, case covariates **x**, and true condition dummy variable **d** assuming that **y** is normally distributed conditional on the covariates and the condition indicator

d. Control covariates affect the classifier for both the case and the control populations, while case covariates provide an additional effect on only the case population.

When a classification variable that corresponds to this model is used, there is a different ROC curve for each case–covariate value combination. Essentially, a different classifier **y** | **x** is used for each case–covariate value combination **x**. Both the case and control covariates for each subject are used to estimate the underlying model, and then the diagnostic test using classifier **y** under case covariates is evaluated based on the covariate-specific ROC curve.

To demonstrate this model, we examine data from a neonatal audiology study. Our classification variable is **y2** (TEOAE 80 at 2kHz). The dummy variable **d** indicates whether an individual is hearing impaired. The child's age is recorded in months as **currage**, while gender is indicated by **male**. We do not think gender will have a direct effect on the ROC curve, so we expect it to be a control covariate.

We will use **rocreg** with the **probit ml** option to estimate the parameters of the ROC curve. Over 90% of the newborns were tested in both ears, so we cluster on infant ID (**id**).



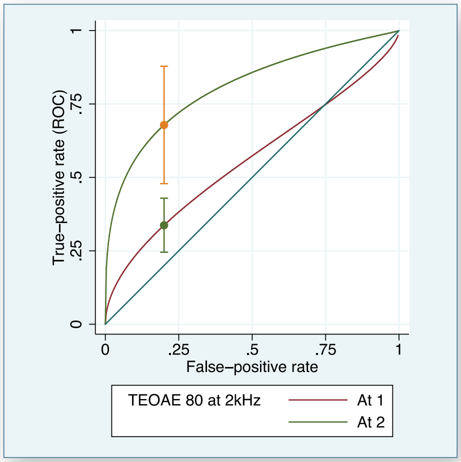
```
. use http://www.stata-press.com/data/r12/nnhs
(Norton - neonatal audiology data)

. rocreg d y2, probit ml ctrlcov(currage male) roccov(currage) cluster(id) nolog

Parametric ROC estimation

Covariate control       : linear regression
Control variables       : currage male
Control standardization: normal
ROC method              : parametric                  Link: probit

  Status     : d
  Classifiers: y2

  Classifier : y2
  Covariate control adjustment model:
                                (Std. Err. adjusted for 2741 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
casecov      |
     currage |   .5729861   .2251764     2.54   0.011     .1316485    1.014324
       _cons |   -18.2597   8.615108    -2.12   0.034    -35.14501   -1.374403
-------------+----------------------------------------------------------------
casesd       |
       _cons |   9.723858   .7540084    12.90   0.000     8.246029    11.20169
-------------+----------------------------------------------------------------
ctrlcov      |
     currage |  -.1694575   .0330224    -5.13   0.000    -.2341803   -.1047348
        male |   .7122587   .2298645     3.10   0.002     .2617326    1.162785
       _cons |  -5.651728   1.284617    -4.40   0.000    -8.169532   -3.133925
-------------+----------------------------------------------------------------
ctrlsd       |
       _cons |   6.986167    .098408    70.99   0.000     6.793291    7.179044
------------------------------------------------------------------------------

   Status    : d
   ROC Model :
                                (Std. Err. adjusted for 2741 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y2           |
      i_cons |  -1.877825    .905174    -2.07   0.038    -3.651933   -.1037167
     currage |   .0589258   .0235849     2.50   0.012     .0127002    .1051514
      s_cons |   .7184563   .0565517    12.70   0.000      .607617    .8292957
------------------------------------------------------------------------------
```

All coefficients are significant at the 0.05 level. The "Covariate control adjustment model" table indicates how the covariates directly affect the classification variable. Current age has a negative effect on the classifier within the control population, but this is moderated under the case population. Being male has a positive effect on the classifier; it does not affect the ROC curve, because we assume gender does not have an extra effect in the case population. However, including it in the model improves the model's overall accuracy.

The "ROC Model" table shows how the case covariates directly affect the ROC curve. Here we see that current age has a positive influence on the curve. Also the coefficient for the standard normal quantile of the false-positive rate, referred to as **s_cons**, is significantly lower than 1. This parameter can be viewed as a form of slope for the false-positive rate, and it corresponds to the ratio of the control population error standard deviation to that of the case population. In this example, the control population is estimated to be less variable than the case; hence, the false-positive rate has a smaller effect on the ROC curve.

We can visualize the ROC curves for different ages by using **rocregplot**. Here we contrast the curves for infants aged 35 months with those aged 50 months by using the **at()** option. We also specify the **roc()** option to estimate the true-positive rate at a false-positive rate of 0.2.

```
. rocregplot, at1(currage=35) at2(currage=50) roc(.2)
```



The classifier TEOAE 80 at 2kHz clearly performs better at age 50 months than age 35 months.

Summary

ROC curves are a useful tool for evaluating and comparing diagnostic tests. The Stata commands **rocreg** and **rocregplot** make the drawing and evaluation of ROC curves simple.

— Charles Lindsey
Statistician and Software Developer

StataCorp
4905 Lakeway Drive
College Station, TX 77845-4512
USA
Return service requested.

Contact us

979-696-4600 979-696-4601 (fax)
service@stata.com stata.com
Please include your Stata serial number with all correspondence.

Find a Stata distributor near you
stata.com/worldwide

facebook.com/StataCorp

twitter.com/Stata

blog.stata.com

New public training course: Structural Equation Modeling Using Stata

October 25–26, 2012

This course covers the use of Stata for structural equation modeling (SEM). SEM is a class of statistical techniques for modeling relationships among variables, both observed and unobserved. SEM encompasses some familiar models such as linear regression, multivariate regression, and factor analysis and extends to a variety of more complicated models.

The course will give an introduction to linear SEM. In addition, a number of models that fall within the SEM framework will be discussed with an emphasis on using Stata to fit each one. These models include confirmatory factor analysis (CFA), path analysis, nonrecursive (simultaneous) models, latent growth models, models with mediation, multiple group models, and full structural equation models with latent variables and measurement components. Stata allows for fitting structural equation models in two ways: by using the **sem** command syntax or by using the graphical user interface to draw path diagrams. Examples will demonstrate both approaches.

Course Topics

- Overview of SEM
  - Model description
  - Process of fitting and evaluating structural equation models
  - Description of path diagrams
- Stata's tools for SEM
  - Fitting models with the **sem** command
  - Building models using the GUI for SEM
  - Using the **ssd** commands to work with summary statistics
- Testing and interpreting SEM results
  - Standardized results
  - Direct, indirect, and total effects
  - Goodness-of-fit statistics
  - Modification indices
  - Score tests and Wald tests
  - Tests for multiple group models

Enrollment for this course is $1,295 per person, and space is limited to 24 seats. For more details or to register, visit stata.com/training/structural-equation-modeling-using-stata.

For a list of all public training courses, see page 8.