Announcing Stata Release 12

Stata 12 ships July 25. Order now at www.stata.com

SEM
Structural Equation Modeling

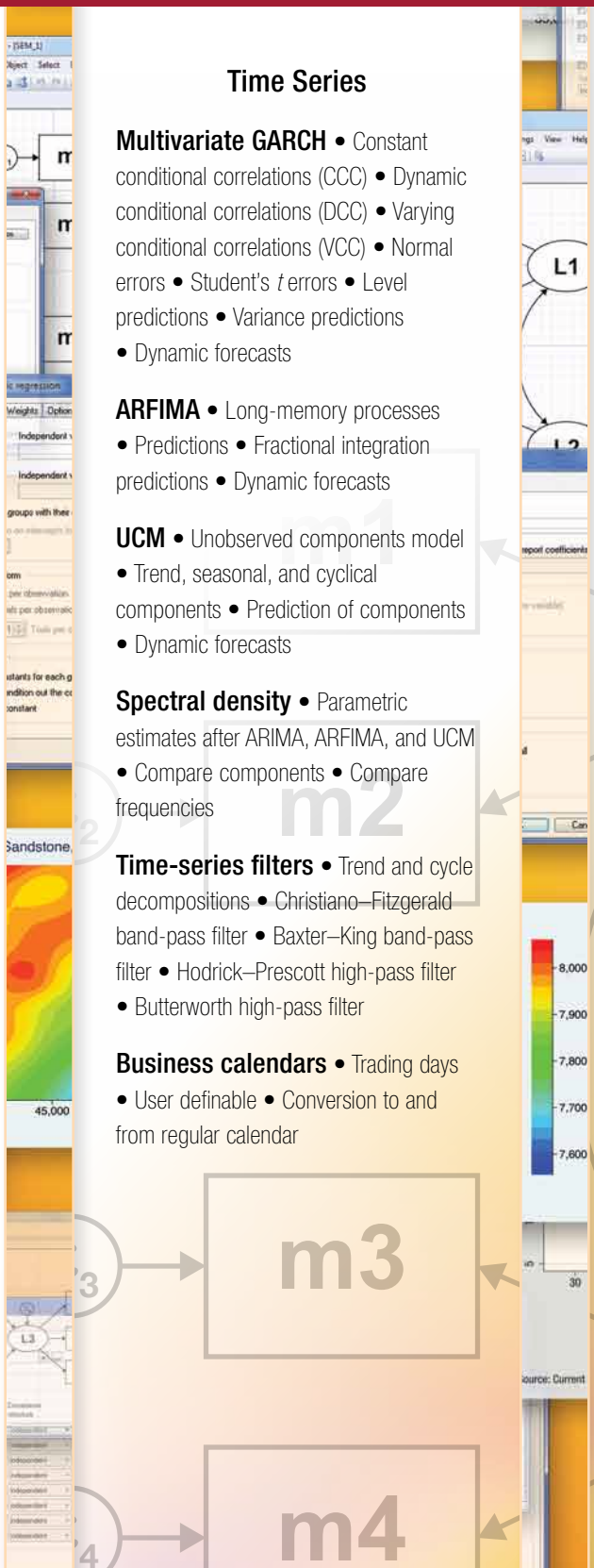
• Path diagrams • Standardized and unstandardized estimates • Modification indices • Direct and indirect effects • Score tests • Wald tests • Factor scores and other predictions • Goodness of fit • Estimation with groups and tests of invariance • Implementation for big datasets and big models • Flexible extension of multivariate regression, seemingly unrelated regression (SUR), and instrumental variables • FIML and LIML estimation of simultaneous systems • Maximum likelihood estimation • GMM estimation • Clustered data • Survey data • Missing at random (MAR) data

Contrasts and
Pairwise Comparisons

• Linear and nonlinear models • Compare means, intercepts, or slopes • Compare adjacent categories • Compare with reference category • Compare with grand mean • Treatment effects • Potential outcomes • Orthogonal polynomials • Adjustments for multiple comparisons • Graphs

Multiple Imputation (MI)

• Chained equations • Imputation of continuous, ordinal, cardinal, and count variables • Conditional imputation • Imputation separately within group • Panel data and multilevel models • Linear and nonlinear predictions



Time Series

Multivariate GARCH • Constant conditional correlations (CCC) • Dynamic conditional correlations (DCC) • Varying conditional correlations (VCC) • Normal errors • Student's *t* errors • Level predictions • Variance predictions • Dynamic forecasts

ARFIMA • Long-memory processes • Predictions • Fractional integration predictions • Dynamic forecasts

UCM • Unobserved components model • Trend, seasonal, and cyclical components • Prediction of components • Dynamic forecasts

Spectral density • Parametric estimates after ARIMA, ARFIMA, and UCM • Compare components • Compare frequencies

Time-series filters • Trend and cycle decompositions • Christiano–Fitzgerald band-pass filter • Baxter–King band-pass filter • Hodrick–Prescott high-pass filter • Butterworth high-pass filter

Business calendars • Trading days • User definable • Conversion to and from regular calendar

And More

Contour plots • See back cover

Multilevel/mixed models
• Estimation with complex survey data • Frequency and sample weights • Robust and clustered SEs • Overall weighting and weighting at each level

ROC analysis • Parametric and non-parametric • Adjustments for covariates • Case–control regression models • Bootstrap and model-based SEs • Comparative graphs • Area under the curve (AUC) and partial AUC

Data management • Excel® import and export • EBCDIC • ODBC connection strings • PDF export of results and graphs

Installation Qualification • Report for submission to regulatory agencies

More support for multicore computers • More estimators • Up to 64 cores supported

Interface • See page 7

Automatic memory management
• Up to 1 terabyte of memory

Contrasts and pairwise comparisons

Contrasts, pairwise comparisons, and margins plots are about understanding results from your model. And communicating them. How does a covariate affect the response? Is the effect nonlinear? Does the effect depend on other covariates?

Linear models

Consider blood pressure and its response to age, sex, and weight. We might estimate a fully interacted linear model,

```
regress pressure agegroup##sex
```

which we could equally well estimate using **anova**. To obtain the estimated cell means, we type

```
margins agegroup#sex
```

and to obtain the graph, we type

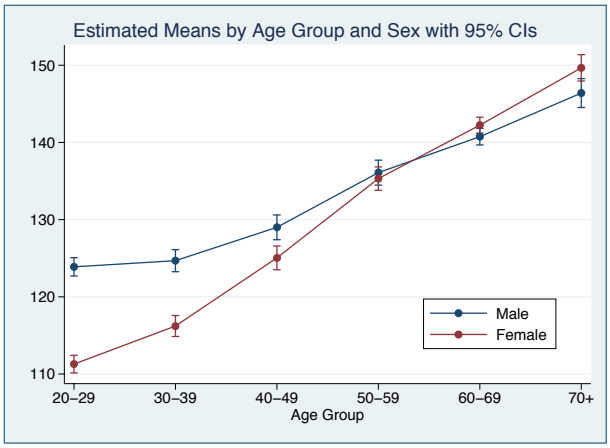
```
marginsplot
```

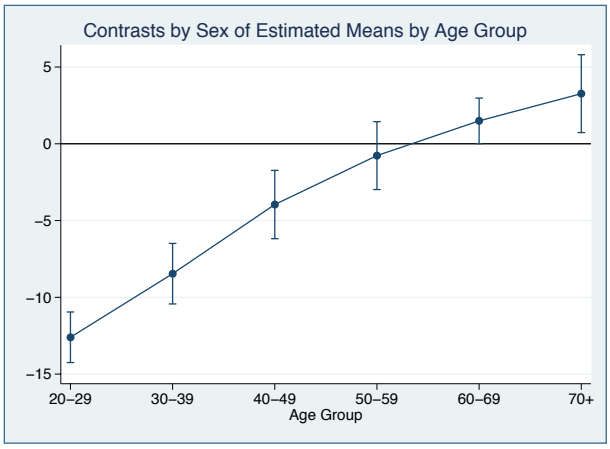
The estimates clearly differ by gender. Is that difference significant? We could do a contrast to find out by typing

```
contrast r.sex@agegroup
```

which compares males and females at each level of **agegroup**. We can graph those results by typing

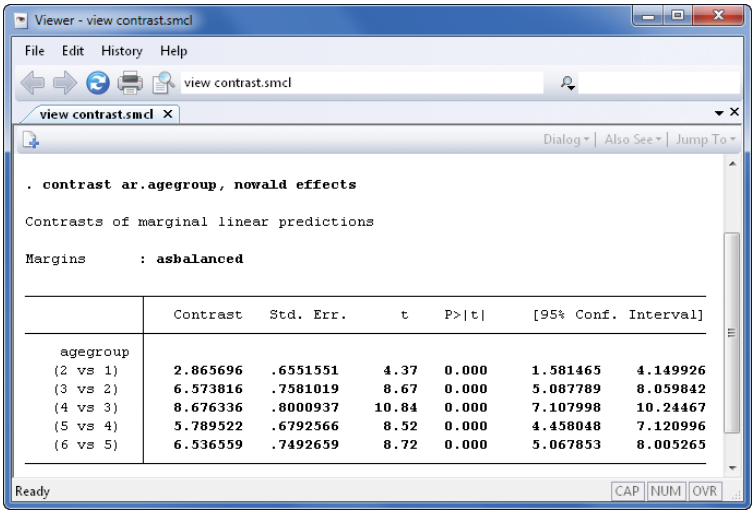
```
margins r.sex@agegroup
marginsplot, yline(0)
```





Adjacent comparisons

Can we combine any of the age groups? It's easy to obtain the answer to that question by using the new command **contrast**.



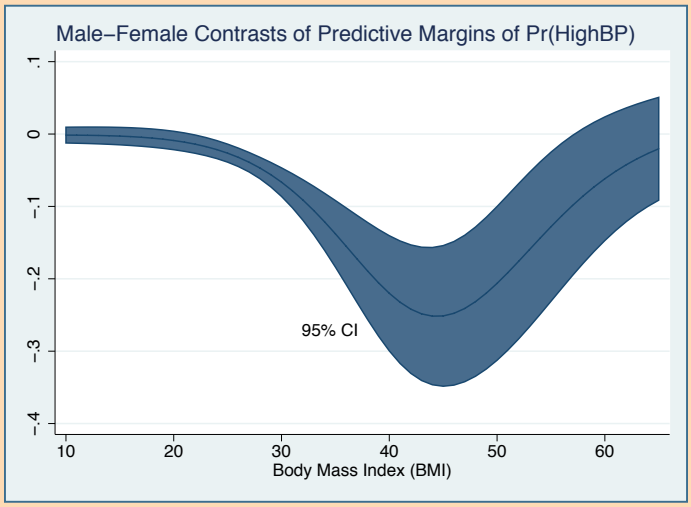
ar. asks for the reverse-adjacent comparisons, which is to say that it compares age group 1 with 2, 2 with 3, and so on. We specified some options to shorten the output.

We can do a lot more with contrasts.

By the way, pairwise comparisons are contrasts where each level of a variable is compared with every other level. They are performed by **pwcompare**.

Nonlinear models

We can use these tools to analyze continuous variables and nonlinear models. For example, here is a graph of the difference between males and females evaluated at 65 levels of body mass index after fitting a logistic regression model of high blood pressure on sex, age group, and the continuous variable body mass index. The graph was produced by **marginsplot**.



Structural Equation Modeling (SEM)

What is SEM?

SEM stands for structural equation modeling. SEM is a notation for specifying structural equations, a way of thinking about them, and methods for estimating their parameters.

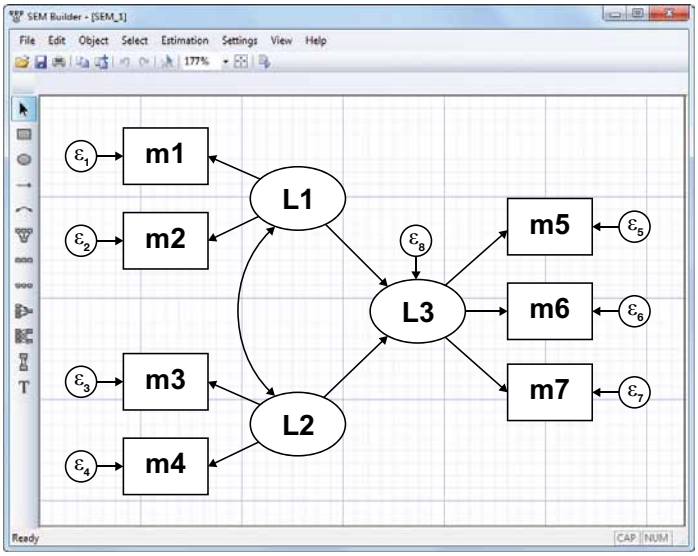
SEM encompasses a broad array of models from linear regression to measurement models to simultaneous equations, including along the way confirmatory factor analysis (CFA), correlated uniqueness models, latent growth models, and multiple indicators and multiple causes (MIMIC).

Features

- Use of GUI or command language to specify your model
- Standardized or unstandardized results
- Direct and indirect effects
- Goodness-of-fit statistics
- Tests for omitted paths and tests of model simplification including modification indices, score tests, and Wald tests
- Predicted values and factor scores
- Linear and nonlinear tests of estimated parameters
- Linear and nonlinear combinations of estimated parameters with CIs
- Estimation across groups
  - › As easy as adding **group(sex)** to the command
  - › Test for group invariance
  - › Easily add or relax constraints across groups
  - › May use raw or summary statistic data

Specifying models

Enter your models graphically



Or use the command syntax

```
sem (L1 -> m1 m2)
    (L2 -> m3 m4)
    (L3 <- L1 L2)
    (L3 -> m5 m6 m7)
```

It's the same model either way.

Stata's GUI uses standard path notation. In command syntax, you can type the arrows in either direction, **(L1 -> m1 m2)** or **(m1 m2 <- L1)**. It doesn't matter. You can specify paths individually, **(L2 -> m3) (L2 -> m4)**, or combined, **(L2 -> m3 m4)**.

Estimation methods

- Maximum likelihood (ML) and asymptotic distribution free (ADF)
- ADF is generalized method of moments (GMM)
- Robust estimate of standard errors available
- Estimates of standard errors for clustered samples
- Support for survey data, including sampling weights, stratification and poststratification, and clustered sampling at one or more levels
- Missing at random (MAR) data supported via FIML

What can you do with `sem`?

sem can fit linear regressions

```
sem (y1 <- x1 x2 x3)
```

It can fit multivariate regression

```
sem (y1 y2 <- x1 x2 x3)
```

It can fit seemingly unrelated regression

```
sem (y1 <- x1 x2 x3) (y2 <- x1 x4), cov(e.y1*e.y2)
```

It can fit simultaneous systems

```
sem (y1 <- y2 x1 x2) (y2 <- y1 x1 x3 x4), cov(e.y1*e.y2)
```

It can fit factor models (capital letters indicate latent [unobserved] variables)

```
sem (L -> m1 m2 m3 m4)
```

It can fit measurement error models

```
sem (y <- X) (X -> x1 x2 x3)
```

And you can combine all the above to create true SEMs

```
sem (m1 m2 <- L1)      //   measurement piece
    (m2 m3 <- L2)      //   measurement piece
    (L3 -> m4 m5)      //   measurement piece
    (L1 <- L3)         //   structural piece
    (L2 <- L1 L3)      //   structural piece
```

or to create multilevel CFAs

```
sem (M -> m1 m2 m3 m4)
    (N -> n1 n2 n3 n4)
    (O -> o1 o2 o3 o4)
    (P -> p1 p2 p3 p4)
    (L -> M  N  O  P )
```

And, without showing examples, **sem** can fit

- path analysis models
- single-factor measurement models
- multiple-factor measurement models
- MIMIC models
- latent growth models
- correlated uniqueness models
- and more

Statistical summary data

sem works with raw or statistical summary data (SSD) of covariances or correlations, variances or standard deviations, and means. The new Stata feature **ssd** makes it easy to create, enter, and use SSD. SSD are stored as standard Stata datasets.

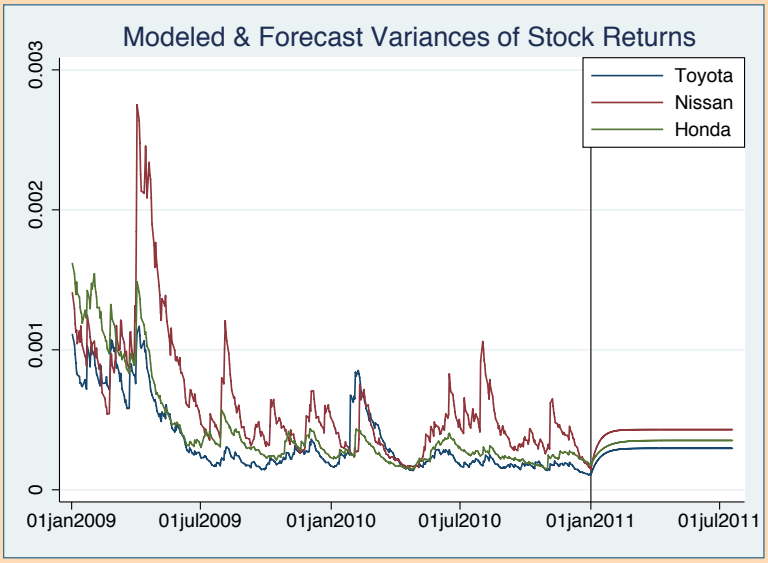
Multivariate GARCH

Multivariate GARCH stands for multivariate generalized autoregressive conditional heteroskedasticity and deals with models of time-varying volatility in multiple series. These models allow the conditional covariance matrix of the dependent variables to follow a flexible, dynamic structure and allow the conditional mean to follow a vector-autoregressive (VAR) structure.

The figure is based on an ARCH(1) GARCH(1) constant conditional correlation (CCC) model. The vertical line separates the one-step-ahead modeled variances from the dynamic forecasts of variances.

In addition to CCC, also available are dynamic conditional correlations (DCC), varying conditional correlations (VCC), and diagonal VECH (DVECH).

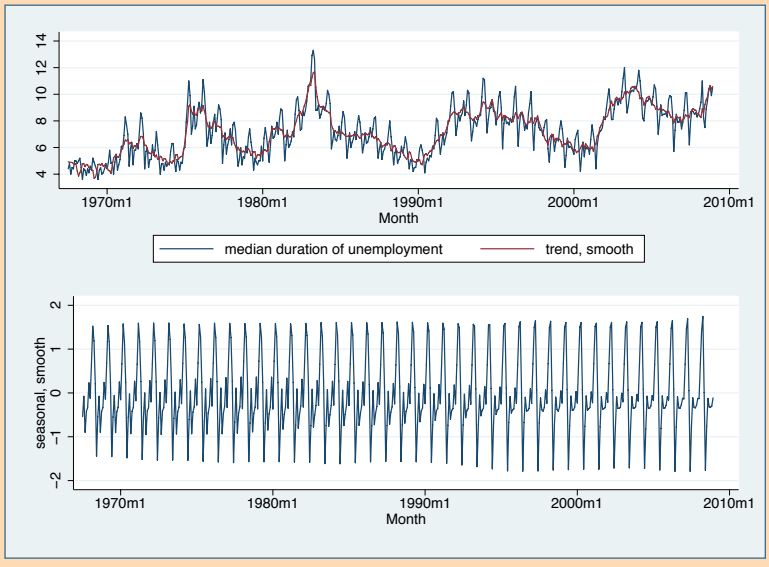
The new **mgarch** command fits multivariate GARCH models.



ARFIMA

ARFIMA stands for autoregressive fractionally integrated moving average and handles long-memory processes. ARFIMA generalizes the ARMA and ARIMA models. ARMA models assume short memory; after a shock, the process reverts back to its trend relatively quickly. ARIMA models assume that shocks are permanent and memory never fades. ARFIMA provides a middle ground in the length of the process's memory.

The new **arfima** command fits ARFIMA models. In addition to one-step and dynamic forecasts, Stata's **arfima** can predict fractionally integrated differences in a series.



UCM

UCM stands for unobserved components model and decomposes a time series into trend, seasonal, cyclic, and idiosyncratic components after controlling for optional exogenous variables. UCM provides a flexible and formal approach to smoothing and decomposition problems.

In the figure, we use monthly data on the median duration of U.S. unemployment that has not been seasonally adjusted. We decompose the data into trend (top panel) and seasonal (bottom panel) components. The careful reader will note that a piece of the seasonal component is trending over time. Look at not the lowest but the second-lowest peak in each year. It is smoothly trending up.

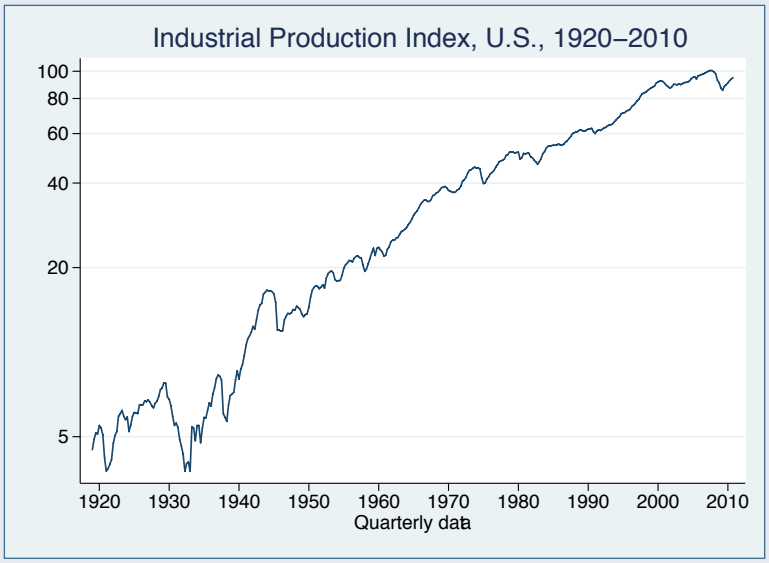
The new **ucm** command fits UCM models.

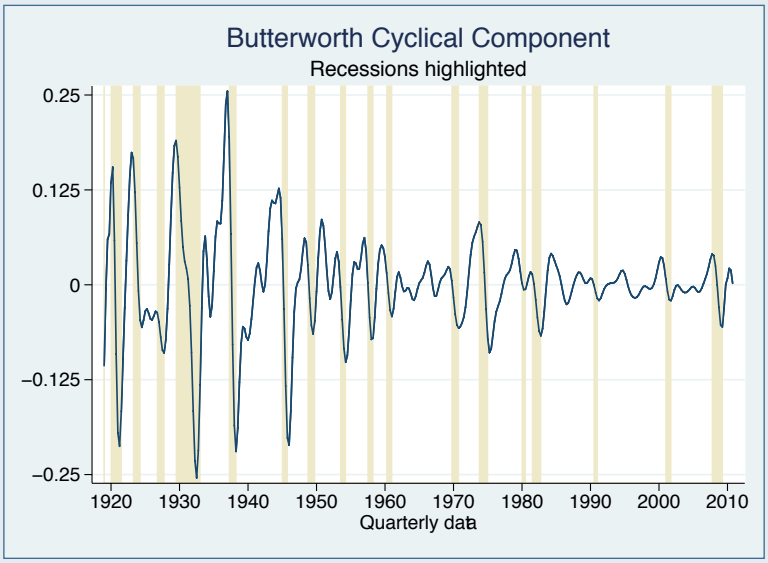
series

Time-series filtering

The new command **tsfilter** allows you to filter a time series and keep only selected periodicities (frequencies). One use of a filter is to separate a time series into trend and cyclical components. In such applications, the cyclical part is assumed to be driven by shocks within a specified range of periods. The method has been used, for instance, to estimate the business cycle component of indices of industrial production.

For instance, we could start with industrial production data for the U.S. from 1920 to 2010 (shown below left) and choose filters to extract the cyclical components (shown below right). On the cyclical graph, we shaded recessionary periods.





Spectral density

The new command **psdensity** estimates the spectral density of a stationary process using the parameters of a previously estimated parametric model.

In the time domain, the dependent variable evolves over time because of random shocks. The autocovariances of a covariance-stationary process specify its variance and dependence structure, and the autocorrelations provide a scale-free measure of its dependence structure. For instance, the autocorrelation at lag j specifies whether realizations at time t and realizations at time $t + j$ are positively related, unrelated, or negatively related.

In the frequency domain, the dependent variable can be thought of as being generated by an infinite number of random components that occur at the frequencies 0 to π. The spectral density specifies the relative importance of these random components. The area under the spectral density in any interval is the fraction of the variance of the process that can be attributed to the random components that occur at the frequencies in the interval.

Either way, it's the same information.

Business calendars

Stata's new business calendar facilities allow you to define your own calendars so that dates display correctly and lags and leads work as they should.

You could create file **lse.stbcal** that recorded the days the London Stock Exchange is open (or closed), and then Stata would understand format **%tblse** just as it understands the usual date format **%td**.

Once you define a calendar, Stata deeply understands it. You can, for instance, easily convert between **%tblse** and **%td** values.

Multiple Imputation (MI)

- **mi impute** now supports chained equations, conditional imputation, and imputation by group. And it's faster.
- **mi estimate** now supports multilevel and panel-data models, so you can use **mi estimate** with **xtmixed** or **xtreg**.
- **mi estimate** now allows you to measure the amount of simulation error in your final model, so you can decide whether you need more imputations.
- **mi predict** and **mi predictnl** create linear and nonlinear predictions. These predictions are in the original data, and not just for the complete observations, but also for the observations with missing values among the explanatory variables.

Chained equations

Chained equations are used to impute missing values when variables may be of different types and missing-value patterns are arbitrary. The first variable could be imputed using logit, the second using linear regression, and the third using multinomial logistic regression. Use of appropriate imputation methods will more accurately reflect the structure of the data.

Nine imputation methods are provided: linear regression, predictive mean matching, interval regression, truncated regression, logistic regression, ordered logistic regression, multinomial logistic regression, Poisson regression, and negative binomial regression.

Conditional imputation

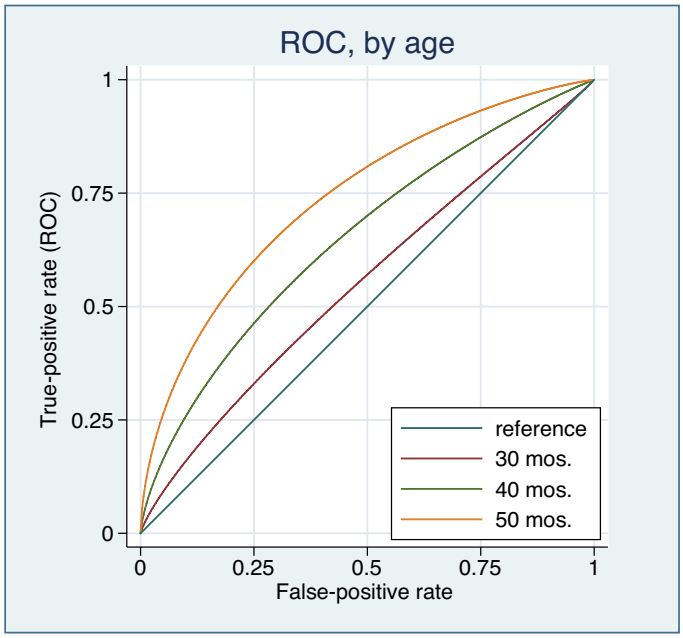
Conditional imputation is customized imputation within group when group itself might be imputed. You can restrict imputation of number of pregnancies to females even when **female** itself contains missing values and so is being imputed.

Imputation can also be performed by group. Thus Australians could have their missing values imputed using only data from other Australians.

Receiver Operating Characteristics (ROC)

You can now model ROC curves that control for covariates. Think of it like regression for ROC.

In a neonatal audiology study of hearing impairment, a hearing test was applied to children aged 30 to 53 months. It is believed the test is more accurate at older ages. In Stata 12, we can use **rocreg** with these data to model how sensitivity and specificity of this test depends on age. Then we can use new command **rocregplot** to compare ROC at various ages:



Area under the curve (AUC) increases with age.

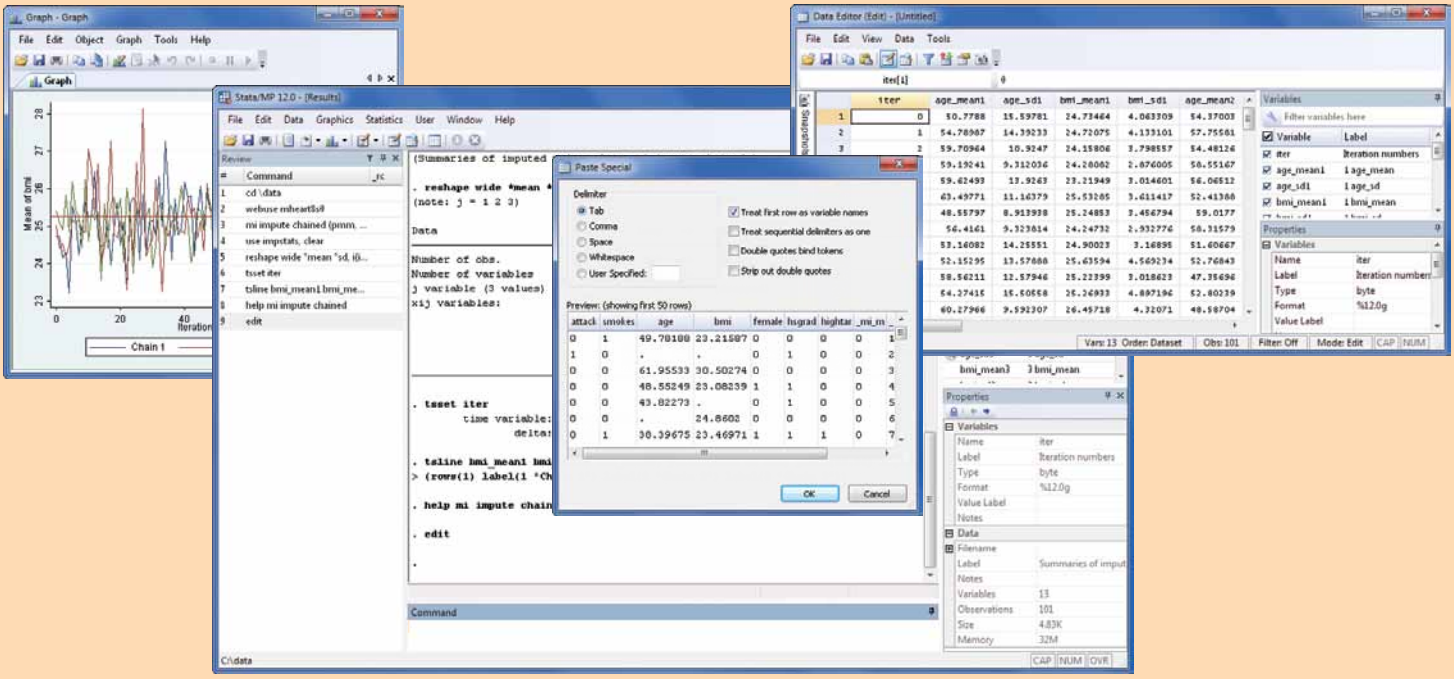
We could also test whether AUC increases with age, estimate sensitivity for a given specificity (and vice versa), and estimate partial AUC (area to a given point of false positive), all controlling for age.

Stata Conference Chicago 2011 July 14–15

The Stata Conference will be held at the University of Chicago Graduate School of Business's Gleacher Center. The program includes presentations by StataCorp developers on some of the new features listed here.

Register now and see the full program at www.stata.com/chicago11

Dates	July 14–15, 2011
Venue	Gleacher Center The University of Chicago Booth School of Business www.gleachercenter.com
Cost	$195 two-day professional; $75 student $125 single-day professional; $50 student $38 dinner (optional)



New interface

The new layout better fits wider screens. That's the new Properties window at the bottom right. It lets you manage your variables, including their names, labels, value labels, notes, formats, and storage types. The magnifying glass on the Review and Variables windows lets you filter commands and variables. Press *Ctrl+F* and you can search the Results window, too.

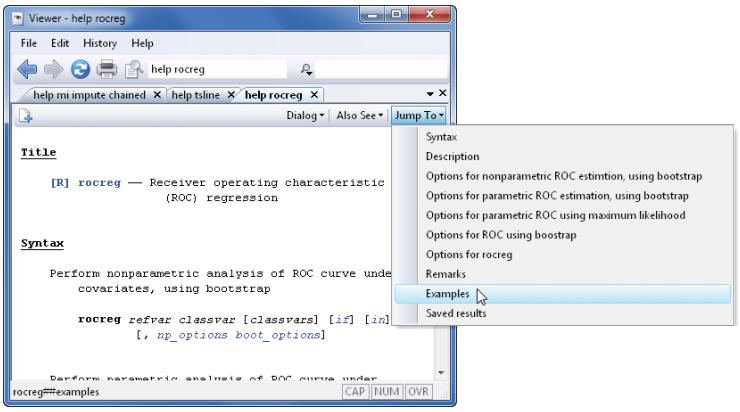
Windows, Mac, and Unix. Their interfaces are the same. Except the Mac now supports gestures.

Enhanced Data Editor

Just like the new main window, there is a new Properties window at the right of the Data Editor. And just like in the main window, you can manage your variables from it. Above it is the new Variables Management Tool. Hide, show, filter, and reorder.

The new Clipboard Preview Tool lets you see and prepare your raw data before pasting.

New Viewer



The **Dialog**, **Also See**, and **Jump To** menus provide quick access to exactly what they say. The Viewer is now tabbed so you can open multiple help files and documents in a single window.

More import

Import and export Excel files. No add-ons required. And the new Excel import preview tool lets you see the data before you import them.

Import EBCDIC files.

Customizable database connections with ODBC connection strings.

Export results and graphs to PDFs!

Automatic memory management

Automatic memory management means that you no longer have to **set memory**, and never again will you be told that there is no room because you set too little! Stata automatically adjusts its memory usage up and down according to current requirements. Up to 1 terabyte.

Installation Qualification

Installation Qualification (IQ) is provided by a new tool that you can download for free. IQ produces a report for submission to regulatory agencies such as the FDA to establish that Stata is installed correctly. Visit www.stata.com/support/installation-qualification

Stata/MP

More support for multicore computers. Faster. More estimators. Up to 64 cores supported.

Stat/Transfer 11

Stat/Transfer 11 understands Stata 12 datasets. Visit www.stata.com/products/transfer.html for details.

StataCorp
4905 Lakeway Drive
College Station, TX 77845
USA

Return service requested.

Stata 12 ships July 25. www.stata.com/stata12

Contact us

Phone 979-696-4600 Fax 979-696-4601
Email service@stata.com Web www.stata.com

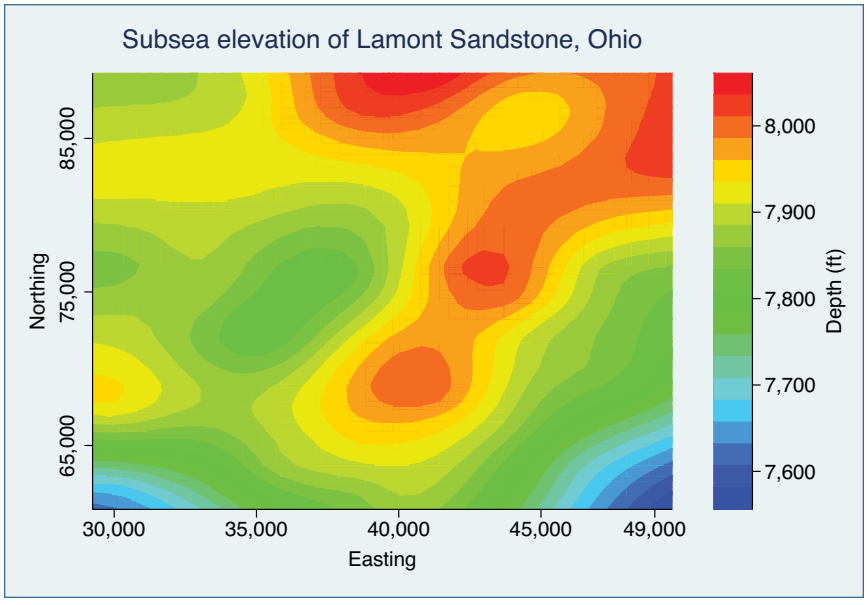
Please include your Stata serial number with all correspondence.

To locate a Stata international distributor near you, visit www.stata.com/worldwide

Find us on Facebook. **Follow us on Twitter.** **Check out our blog.**

More Stata 12 features

Did we mention contour plots?



There will not be space to do more than mention the new **rename** command for renaming groups of variables; or that probability predictions are now available after count-data models and panel count-data models; or the new functions for Tukey's Studentized range and Dunnett's multiple range distributions; or that absorbed regression is now faster; or the new estimation commands for truncated count-data models; or tabbed graphs for Mac; or file drag-and-drop for Windows; or the tabbed Do-file Editor for Mac and Unix; or syntax highlighting in the Do-File Editor for Mac, or imputation by drawing posterior estimates from bootstrapped samples; or the handling of perfect prediction during imputation of categorical variables; or that **misstable** will create summary variables of missing-value patterns; or the new goodness-of-fit statistic that is robust to censoring for survival data; or that baseline odds are now reported along with odds ratios; or that saved results can be marked as hidden or historical; or bootstrap inference for ROC regression; or setting the number of digits displayed in estimation output; or the robust and cluster-robust SEs after fixed-effects **xtpoisson**; or exponential, banded, and Toeplitz residual covariance structures for linear mixed models; or that the **matrix accum** utility now allows absorbed variables; or the improved importing of data from Haver Analytics; or the goodness-of-fit test after binary models for survey data; or the coefficient of variation for survey data; or that the new memory manager is tunable; or that time-series operators are now supported by more graph commands; or the new **getmata** and **putmata** commands for easily moving data between Stata and Mata; or that MP is even faster on 16+ cores.