The STATA News®

Statistics Graphics Data Management & Analysis

STATA/MP®

See how fast Stata/MP can be

Everyone knows that computers now come with multiple cores. Basically, every computer processor now has several cores integrated on it. Even most phones now have multicore processors.

Everyone also knows that Stata has superb support for multicore computers. The multicore version of Stata—Stata/MP—runs over half of Stata's estimation commands at least 1.8 times faster on dual-core machines, at least 2.8 times faster on quad-core machines, and at least 4 times faster on computers with 8 cores. You don't have to do anything to your commands or your do-files. Everything works just as it does in every other version of Stata, it just works faster. Even estimators and other features performed from dialog boxes run faster.

Not everyone knows that these performance benefits are not automatic. Stata/MP has its own versions of key computational code to efficiently distribute computations across multiple cores. Sometimes, this is impossible, meaning some algorithms cannot be made faster.

If you want to know just how fast Stata/MP runs the things you do frequently, we have the answer. Every estimator and data management facility in Stata 14 is analyzed in the updated Stata/MP white paper. We even include a section on Mata, which is highly parallelized—that is to say, much faster on Stata/MP.

Take a look at **stata.com/statamp** for an overview, or go straight to the white paper at **stata.com/statamp/statamp.pdf**.
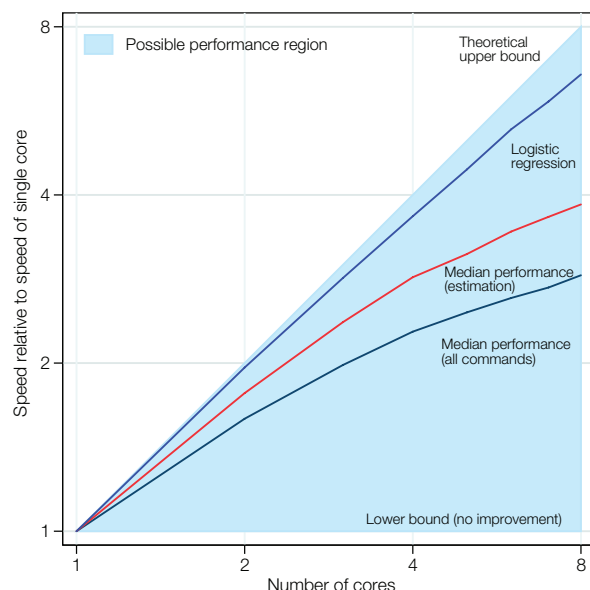
In the spotlight:
Easy-to-interpret, flexible survival-time treatment effects

Introduction

Is smoking bad for men who have already had a heart attack? You may already have an idea about the answer. But if you are a researcher studying smoking and heart disease, you know that this question is too general. One question with a statistical answer is "By how much will smoking reduce the time to a second heart attack among men aged 45–55 who have already had a heart attack? And, by the way, you clearly cannot require men to smoke".

This latter question highlights (1) that we must be precise when estimating effects, (2) that the survival analysis of nonnegative, right-censored data is essential for quantifying the effect of this treatment, and (3) that we frequently must use observational data for ethical reasons.

We use observational data out of necessity, but the randomized experiment defines the treatment effect. For our example, the treatment effect is a population measure of a comparison between what happens when everyone smokes instead of when no one smokes. Because each individual either smokes or does not smoke, we see only one of the two potential outcomes. When the treatment is randomized, the missing potential outcome is missing completely at random. With observational data, we assume that the potential outcome is missing at random, after conditioning on covariates.

In this spotlight, I reiterate the well-known point that the effect estimable from the Cox model parameters is difficult to interpret for nontechnical audiences. I illustrate why the Cox model is less flexible for estimating treatment effects than many researchers believe. Finally, I show how to use the new **stteffects** command to flexibly estimate an easy-to-interpret treatment effect.

What Cox can and cannot tell us

We have simulated data on the time to second heart attack (**atime**), a binary failure indicator (**fail**), a binary indicator for whether the man smokes (**smoke**), the man's age in decades (**age**), an exercise index (**exercise**), and a diet index (**diet**). These data have already been **stset** with **atime** as the time to event and **fail** as the failure indicator.

Many researchers would begin a survival analysis by estimating the parameters of a Cox model. The probability that an event will occur in the next moment, given that it has not yet happened, is the hazard function. The Cox model parameterizes the effect of covariates on the hazard function as a multiplicative factor.

Below, we estimate the parameters of a Cox model for our data.

```
. stcox smoke age exercise diet
Cox regression -- no ties

No. of subjects =         5,000              Number of obs    =         5,000
No. of failures =         2,969
Time at risk    =   10972.84266
                                             LR chi2(4)       =        271.77
Log likelihood  =     -21963.163             Prob > chi2      =        0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
smoke	1.540071	.0764791	8.70	0.000	1.397239	1.697505
age	2.024237	.1946491	7.33	0.000	1.676527	2.444062
exercise	.5473001	.0454893	-7.25	0.000	.465026	.6441304
diet	.4590354	.0379597	-9.42	0.000	.3903521	.5398037

The output indicates that smoking increases the hazard of a second heart attack by a factor of 1.5. Because **smoke** is not interacted with the other covariates, the hazard ratio is constant.

Though researchers become accustomed to the hazard ratios, it is notoriously difficult to explain to patients, policy makers, and other nontechnical people. I, as a technical person, can back out how bad a factor of 1.5 might be, but I would prefer an effect measured in the time to failure.

When it is constant, the hazard ratio from these observational data has the same interpretation as the hazard ratio we would obtain if the treatment **smoke** was randomized, given that the missing potential outcome is missing at random

conditional on the covariates. When the treatment **smoke** is interacted with the other covariates, the hazard ratio varies over covariate patterns, and the Cox model parameters cannot be used by themselves to recover the hazard ratio when smoking is randomized.

Below, we estimate the parameters of a Cox model in which **smoke** is interacted with the other covariates.

```
. stcox ibn.smoke#c.(age exercise diet)
Cox regression -- no ties

No. of subjects =         5,000              Number of obs    =        5,000
No. of failures =         2,969
Time at risk    =   10972.84266
                                             LR chi2(6)       =       223.11
Log likelihood  =    -21987.493             Prob > chi2      =       0.0000


          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]

 smoke#c.age |
   Nonsmoker |   1.714749    .1751413     5.28   0.000     1.403655    2.094791
      Smoker |   3.979649    1.110035     4.95   0.000     2.303673    6.874936

      smoke# |
  c.exercise |
   Nonsmoker |   .5514891    .0476827    -6.88   0.000     .4655224    .6533309
      Smoker |   .2839313    .0822003    -4.35   0.000     .1609844    .5007752

 smoke#c.diet|
   Nonsmoker |   .4461597    .0389598    -9.24   0.000     .3759769    .5294433
      Smoker |   .6908017    .1785842    -1.43   0.152      .416201    1.146578
```

The slope parameters of **smoke** differ across levels of **exercise** and the **diet** index. The formal test reported below supports this conclusion.

```
. contrast smoke#c.age smoke#c.exercise smoke#c.diet, overall
Contrasts of marginal linear predictions

Margins        : asbalanced


                      |     df        chi2     P>chi2

          smoke#c.age |      1        8.05     0.0045

     smoke#c.exercise |      1        4.83     0.0280

         smoke#c.diet |      1        2.58     0.1083

              Overall |      3       21.91     0.0001
```

The overall test rejects that the effect of smoking is independent of **age**, **exercise**, and **diet**. We cannot obtain the randomized hazard–ratio effect from these Cox model parameters because the hazard ratio is not constant.

Easy to interpret and easy to estimate

So instead of estimating the difficult-to-interpret hazard ratio, we estimate the average difference in the time to second heart attack if everyone smoked instead of if no one smoked. This effect is known as the average treatment effect (ATE), and it is defined as

$$\text{ATE} = E[t_{smoke} - t_{not\_smoke}]$$

where t_{smoke} is the time to second heart attack when a person smokes and t_{not_smoke} is the time to second heart attack when that person does not smoke. The expected value computes the mean of these individual–level differences in the population.

We can use a model for the outcome, a model for the treatment, or both models to account for the missing potential outcome, because either t_{smoke} or t_{not_smoke} is missing at random, conditional on covariates.

The regression-adjustment (RA) estimator uses a model for the outcome to account for the missing potential outcome. The data lost to censoring are handled in the log-likelihood function. Below, we use **stteffects ra** to estimate the ATE.

```
. stteffects ra (age exercise diet) (smoke)
Survival treatment-effects estimation          Number of obs    =      5,000
Estimator      : regression adjustment
Outcome model  : Weibull
Treatment model: none
Censoring model: none
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf. Interval]	
ATE						
smoke (Smoker vs Nonsmoker)	-1.520671	.2011014	-7.56	0.000	-1.914822	-1.126519
POmean						
smoke Nonsmoker	4.057439	.1028462	39.45	0.000	3.855864	4.259014

The average time to a second heart attack would be 1.5 years earlier when everyone smokes than the 4.0 years that would be observed if no one smoked.

These estimates are much easier to interpret than the hazard ratio, and the underlying model implicitly interacts the treatment with all the covariates. Instead of the default Weibull distribution, we could have used the gamma distribution or the log normal distribution to model the outcome. **stteffects ra** allows us to parameterize the second parameter in each of these distributions to produce flexible outcome models.

The inverse-probability-weighted (IPW) estimator uses models for the treatment assignment and the censoring process to account for the missing data. Below, we use **stteffects ipw** to estimate the ATE.

```
. stteffects ipw (smoke age exercise diet) (age exercise diet)
Survival treatment-effects estimation          Number of obs    =      5,000
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
Censoring model: Weibull
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf. Interval]	
ATE						
smoke (Smoker vs Nonsmoker)	-1.689397	.3373219	-5.01	0.000	-2.350536	-1.028258
POmean						
smoke Nonsmoker	4.200135	.2156737	19.47	0.000	3.777423	4.622848

The average time to a second heart attack would be 1.7 years earlier when everyone smokes than the 4.2 years that would be observed if no one smoked.

As for the RA estimator, other distributions and flexible parameterizations are available for the treatment model and the censoring-process model.

The inverse-probability-weighted regression–adjustment estimator implemented in **stteffect ipwra** uses models for the outcome and the treatment, and optionally the censoring process, to increase efficiency. It offers the flexibility of both the RA and the IPW estimator.

Conclusion and further reading

With survival data, many researchers default to a Cox model, even though the effect estimable from the Cox model parameters is difficult to interpret for nontechnical audiences. The Cox model is less flexible for estimating treatment effects than many researchers believe. The new **stteffects** command can flexibly estimate an easy-to-interpret treatment effect. For more information, see [TE] **stteffects**.

—David M. Drukker
Director of Econometrics

In the spotlight: What's next

Stata promotes an estimation/postestimation paradigm. Which is to say, first you estimate something and then you analyze those estimates. After a linear regression, you might test whether two parameters are the same, or predict leverage statistics, or test for heteroskedasticity, or look at diagnostics plots, or do any of about 70 other things. Actually, the number of things you might do numbers in the thousands, but they fall under 70 or so commands or facilities.

First, wow! That's a lot of different things I could do. How did I know about all that? Well, it is fully documented in [R] **regress postestimation**. But what if I had requested robust standard errors? We then lose the postestimation facilities that are based on the likelihood, because the likelihood no longer has the same interpretation. How did I know that? Again, it is documented, but you have to look closely at the documentation.

What if, instead of linear regression, I had run a logistic regression? We now have about 60 general things we can do to further analyze that regression. Some of those things are the same as the ones after linear regression, while others are new and specific to logistic regression. Again, all documented in [R] **logistic postestimation**. You can see that it becomes tricky to keep up with what is available.

So, what's next?!

Or, more accurately, what's available to do next?

If you want to know, launch the Postestimation Selector—a boring name for a nifty tool. Just click on **Postestimation Selector** in the **Statistics** menu. It will show you exactly what is available for whatever you have most recently estimated or most recently restored from previously saved estimations. Want to see a "Nonlinear expression of parameter estimates" with confidence intervals and a test against zero? Click on it, and type or build your expression.
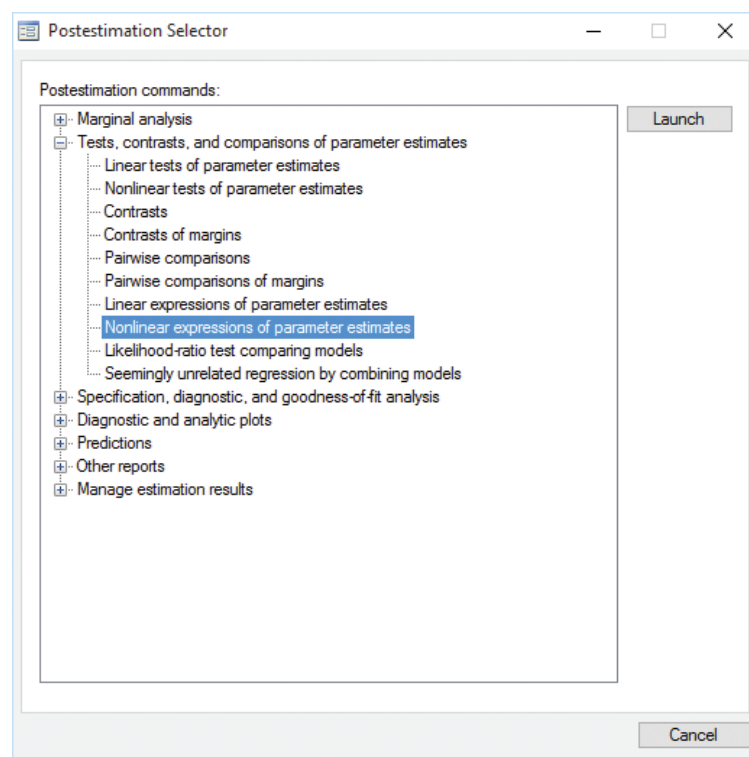
Whenever you estimate a new model, the Selector updates itself. You can drag it to an unused corner of your desktop and just leave it up as you work.

A written description of the Selector simply cannot do it justice. Luckily, my colleague Chuck Huber has created a video. Take a look: **stata.com/videos14/postestimation**.

—Vince Wiggins
Vice President, Scientific Development

New from Stata Press

Meta-Analysis in Stata: An Updated Collection from the Stata Journal, Second Edition

Editors:	Tom M. Palmer and Jonathan A. C. Sterne
Copyright:	2016
Price:	$48.00

Meta-analysis allows researchers to combine results of several studies into a unified analysis that provides an overall estimate of the effect of interest and to quantify the uncertainty of that estimate. Stata has some of the best statistical tools available for doing meta-analysis. The unusual thing about these tools is that none of them are part of official Stata. They are all created by and documented by experts in the broader research community who also happen to be proficient Stata developers.

Editors Tom Palmer and Jonathan Sterne show how each of the articles in this collection relates to others and how each fits in the overall literature of meta-analysis. For the first edition, Sterne convinced over half the authors to update their software and articles for the collection. In this new edition, Palmer and Sterne have substantially expanded the scope of the collection to cover in even more depth many contemporary advances that will help keep the reader up to date.

The second edition retains its original topic-specific sections devoted to the fundamentals of meta-analysis: fitting models, meta-regression, and graphical and analytic tools for detecting bias. It also retains a section devoted to advanced methods. Readers of the first edition will find new articles in these sections, in particular, ones that take advantage of major changes that occurred in Stata since the first edition, such as the introduction of the **gsem** command.

This edition also adds three new topic-specific sections for multivariate or multiple outcomes meta-analysis, individual participant data (IPD) meta-analysis, and network meta-analysis. The addition of these sections gives readers access to new commands that address recent methodological developments in the field.

The new edition adds 11 articles to the original collection of 16 articles. The articles cover topics ranging from standard and cumulative meta-analysis and forest plots to contour-enhanced funnel plots and nonparametric analysis of publication bias. In their articles, the authors present conceptual overviews of the techniques, thorough explanations, and detailed descriptions and syntax of new commands. They also provide examples using real-world data. In short, this collection is a complete introduction and reference for performing meta-analyses in Stata.

Read the table of contents and order online at **stata-press.com/books/meta-analysis-in-stata**.

New from the Stata Bookstore
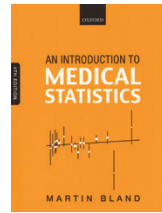
An Introduction to Medical Statistics, Fourth Edition

Author:	Martin Bland
Copyright:	2015
Price:	$44.75

This is a great statistics text for students in the medical sciences and an ideal self-study text for medical practitioners. Bland covers both introductory statistical concepts and advanced topics that would be taught in a graduate-level course on medical statistics.

Read the table of contents and order online at **stata.com/bookstore/introduction-to-medical-statistics**.

Italian Stata Users Group meeting

Florence, Italy   November 12–13, 2015
stata.com/meeting/italy15

Keep up with future Stata Users Group meetings at **stata.com/meeting**. Want to be notified when new information is posted? Sign up at **stata.com/alerts**.

Metrika Consulting celebrates 25 years with StataCorp

For over 30 years, StataCorp has been a leader in statistical software; for 25 of those years, Metrika Consulting has been our valued distributor in the Nordic and Baltic countries. In 1990, Metrika Consulting established a research-oriented consulting company with expertise in data analysis and became the first distributor of Stata software outside the United States. Today, it continues in its history of Stata sales, consulting, and first-class service. Thank you to Metrika Consulting for 25 years of service and many more to come.

NetCourses®

Learn how to exploit the full power of Stata with our affordable, convenient, web-based courses. We offer NetCourses for Stata users of all experience levels, from beginning to advanced.

Introduction to Stata

Learn how to use all of Stata's tools and become a sophisticated Stata user. You will understand the Stata environment, how to import and export data from different formats, how Stata's intuitive syntax works, data management in Stata, and more.

January 15–February 26, 2016 .... $95.00

Introduction to Statistical Graphics Using Stata

Learn how to communicate your data with Stata's powerful graphics features. Topics include using graphs to check model assumptions; formatting, saving, and exporting your graphs for publication; using the Graph Editor; creating custom graph schemes; creating complex graphs by layering and combining multiple graphs; using **margins** and **marginsplot**; and more.

January 15–February 26, 2016 .... $150.00

Introduction to Stata Programming

Become an expert in organizing your work in Stata. Make the most of Stata's scripting language to improve your workflow and create concretely reproducible analyses. Learn how to speed up your work and do more complete analyses.

January 15–February 26, 2016 .... $150.00

Advanced Stata Programming

Learn how to create and debug your own commands that are indistinguishable from the commands that ship with Stata.

January 15–March 4, 2016 .... $175.00

Introduction to Univariate Time Series with Stata

Learn univariate time-series analysis with an emphasis on the practical aspects most needed by practitioners and applied researchers.

January 15–March 4, 2016 .... $295.00

Introduction to Panel Data Using Stata

Become an expert in the analysis and implementation of linear, nonlinear, and dynamic panel-data estimators using Stata. Geared for researchers and practitioners in all fields, this course focuses on the interpretation of panel-data estimates and the assumptions underlying the models that give rise to them.

January 22–March 4, 2016 .... $295.00

Introduction to Survival Analysis Using Stata

Learn how to effectively analyze survival data using Stata. We cover censoring, truncation, hazard rates, and survival functions. Discover how to set the survival-time characteristics of your dataset just once and apply any of Stata's many estimators and statistics to those data.

January 15–March 4, 2016 .... $295.00

stata.com/netcourse

The dates above don't work for you? No problem! NetCourseNow allows you to set the time and work at your own pace as well. It also gives you a personal instructor to guide you through the course.

stata.com/netcourse/ncnow

Public training schedule

Learn Stata from StataCorp's experts. These two-day courses take place in Washington, DC, and are ideal for researchers and individuals who want to learn more or gain a deeper understanding of Stata.

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

January 11–12, 2016 • March 8–9, 2016 • April 26–27, 2016

Aimed at both new Stata users and those who wish to learn techniques for efficient day-to-day use of Stata, this course teaches you to use Stata in a reproducible manner, making collaborative changes and follow-up analyses much simpler.

Time-Series Analysis Using Stata

February 8–9, 2016

Learn how to efficiently use Stata for time-series analysis. Topics include data management, estimation of univariate and multivariate stationary and nonstationary time series, model selection, hypothesis testing, and interpretation.

Structural Equation Modeling Using Stata

March 10–11, 2016

Learn how to illustrate, specify, and estimate structural equation models in Stata using Stata's SEM Builder and the **sem** command. The course introduces several types of models, including path analysis, confirmatory factor analysis, full structural equation models, and latent growth curves.

Handling Missing Data Using Multiple Imputation

April 21–22, 2016

Learn all aspects of multiple-imputation (MI) analysis, including creation of MI data using the multivariate normal and chained-equations (or fully conditional specification) imputation methods, manipulation of MI data, and analysis of MI data.

stata.com/public-training

Stata 14.1

Stata 14.1 is available now and is a no-cost update to Stata 14. Just type **update query** and follow the instructions.

Here are some of the treats in Stata 14.1.

Bayesian analysis. The new Bayesian analysis features introduced in Stata 14 now have dramatically improved performance and capabilities when working with latent variables and with panel and multilevel models. And, there is more in Bayesian analysis. See **help whatsnew** after updating.

Multiple imputation. You can now add your own imputation methods to **mi impute**—Stata's main engine for handling multiple imputation of missing data. It's easy. After updating, type **help mi impute usermethod** for documentation and examples.

IRT. Those who use IRT (item response theory) will be glad to see the new **diflogistic** command that performs the logistic regression test for uniform and nonuniform differential item functioning (DIF). The Mantel–Haenszel test has improvements too. Type **help difmh** after updating.

Mixed models. Small-sample denominator-degree-of-freedom adjustments such as Kenward–Roger and Satterthwaite can now be performed when estimating contrasts via the **contrast** command after estimating linear mixed models.

MLEs of expressions. You can now predict expected means after **mlexp**—maximum likelihood estimation (MLE) of any expression you specify. This also implies that **margins** can be used after **mlexp** to estimate marginal and conditional effects, estimated marginal means, and predictive margins. **mlexp** also now supports linear constraints.

ICD-10. The ICD-10 facilities now work with all codes from 2003 to 2015.

Propensity-score matching. Propensity-score matching via **teffects psmatch** is now much faster for large datasets.

More CIs. You can now get confidence intervals (CIs) for variances and standard deviations from the **ci** commands.

Exporting results to Excel®. The syntax for exporting estimation and other results to Excel spreadsheets has been dramatically simplified. You can also easily choose and export estimation and other results using the intuitive and functional dialog box.

High DPI support. Under Windows, high DPI monitors are now better supported. A host of icon sizings keep your icons looking good when working at higher resolutions.

So… be sure to update your Stata 14 to Stata 14.1. If you're not yet on Stata 14, what are you waiting for?