

Announcing

STATA[®]

14

release

Available now at stata.com

Unicode

Bayesian analysis

Endogenous treatment effects

IRT (item response theory)

Panel and multilevel survival models

Small-sample inference for mixed models

Markov-switching regression

Survey for multilevel models

```
. describe
```

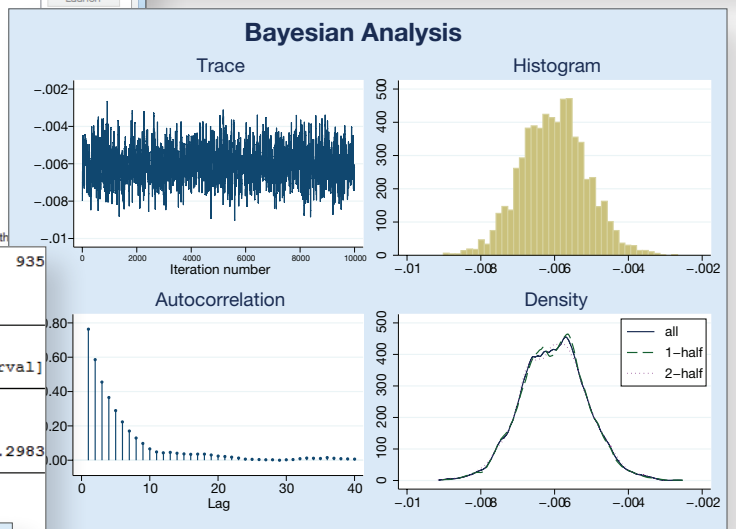
Contains data from auto_ja.dta

obs:	74	1978年自動車データ
vars:	12	24 Mar 2015 10:27
size:	3,182	(_dta has notes)

variable name	storage type	display format	value label	variable label
メーカー名	strl8	%-18s		メーカー名
価格	int	¥8.0gc		価格
走行距離	int	¥8.0g		走行距離(マイル)
修理層78	int	¥8.0g		修理層(1978年時点)
室内高	float	¥6.1f		天井の余裕(インチ)
トランク	int	¥8.0g		トランク容量(立方フィート)
車両重量	int	¥8.0gc		車両重量(ポンド)
全長	int	¥8.0g		全長(インチ)
最小回転半径	int	¥8.0g		最小回転半径(フィート)
	int	¥8.0g		変速比
	float	¥6.2f		
	byte	¥9.0g		生産国

Postestimation Selector

- Marginal analysis
 - Tests, contrasts, and comparisons of parameter estimates
 - Linear tests of parameter estimates
 - Nonlinear tests of parameter estimates
 - Contrasts
 - Contrasts of margins
 - Pairwise comparisons
 - Pairwise comparisons of margins
 - Linear expressions of parameter estimates
 - Nonlinear expressions of parameter estimates
 - Likelihood-ratio test comparing models
 - Seemingly unrelated regression by combining models
- Specification, diagnostic, and goodness-of-fit analysis
- Predictions
 - Probabilities, influence statistics, residuals, etc.
 - Nonlinear predictions of other predictions, parameters, and data; with

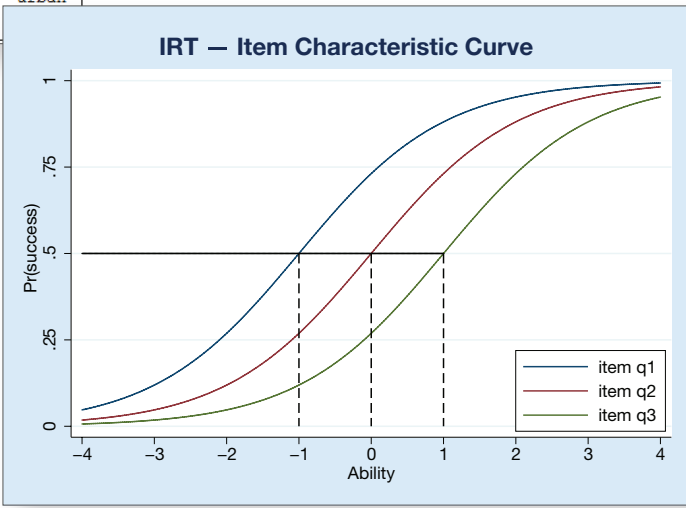


Endogenous treatment-effects estimation Number of obs = 935

Outcome model : exponential

Treatment model: probit

	wage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
ATE	urban	481.0465	29.72084	16.19	0.000	422.7948 539.2983
	(1 vs 0)					
POmean	urban					



- Regression models for fractional data
- Treatment effects adds survival models, balance, more
- Power analysis for survival methods, ANOVA, and contingency tables
- Censored Poisson model
- SEM adds Satorra-Bentler, survival models, more
- ICD-10 • Hurdle model • Structural break tests
- Stata in Spanish and Japanese

More inside!

Unicode

Did you see the output from **describe** on the first page? That's auto.dta in case you couldn't tell. You'd be excused for not knowing because it's in Japanese. All of which is our way of saying that Stata now supports Unicode, and it supports it everywhere. In variable names, in labels, in filenames, and in the string variables in your data.

Your use of Unicode may not be as extreme as our Japanese example. Realize that you can make tables and graphs labeled Übersetzung and Kofferraumvolumen (**Kubikfuß**). Just set the variable labels, whether they are named **übersetzung** and **kofferraumvolumen** or **gear_ratio** and **trunkspace** or even 変速比 and トランク.

Multilevel survival models

We model the effects of laparoscopic surgery and age on length of hospital stay (LOS) for adult patients with appendicitis. We believe that doctors affect the length of a patient's stay, so we include a random effect for doctor. We further believe that hospitals vary in their discharge procedures and thus also affect LOS, so we include a random effect for hospital. We will assume that hospitals nest doctors nest patients.

We will model LOS using survival-time analysis. We first **stset** our survival data by typing **stset los**.

In these data, we observe the LOS for all patients; there is no censoring. Rather oddly but not uncommonly, the "failure" is the happy event of departing the hospital.

We will fit a Weibull model for length of stay:

```
. mestreg lap_surg age || hospital: ||
      doctor: , distribution(weibull)
```

Group Variable	No. of Groups	Observations per Group	Minimum	Average	Maximum
hospital	12	1	38.1	73	
doctor	185	1	2.5	8	

	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
lap_surg	7.025563	1.410613	9.71	0.000	4.739957 10.41329
age	.9716888	.0127353	-2.19	0.028	.9470459 .9969729
_cons	3.34e-21	7.60e-21	-20.70	0.000	3.84e-23 2.89e-19
/ln_p	2.369253	.0477084	49.66	0.000	2.275746 2.462759
hospital					
var(_cons)	.8650249	.436397			.3218143 2.325155
hospital>					
doctor					
var(_cons)	.7086671	.175355			.436333 1.150977

LR test vs. Weibull model: chi2(2) = 153.09 Prob > chi2 = 0.0000

Stata will report the results in either the accelerated failure-time or the proportional-hazards metric. These

Small-sample inference for linear mixed-effects models

Stata fits linear mixed-effects models and, until now, provided only large-sample inference based on normal and χ^2 distributions.

In small samples, the sampling distributions of test statistics are known to be t and F in simple cases, and those distributions can be good approximations in other cases. Stata 14 provides five methods for small-sample inference, including Satterthwaite and Kenward–Roger.

In addition to adjusting the confidence intervals and significance tests reported by Stata's **mixed** estimation command, small-sample statistics are also provided for subsequent estimation of linear combinations and linear hypothesis tests.

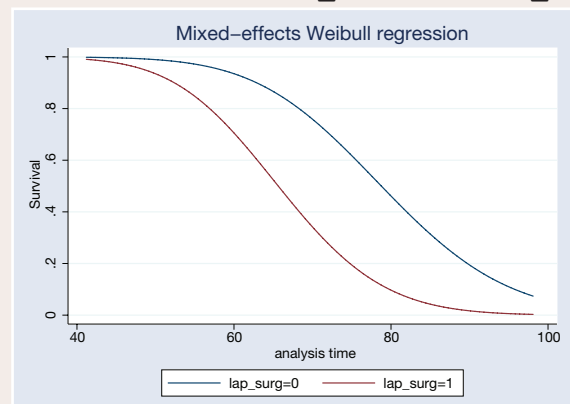
- Parametric survival models with multiple levels of random effects
- Random intercepts and random coefficients
- Crossed random effects
- Right-censoring
- Single- and multiple-record survival-time data
- Normal random effects rather than often less plausible gamma frailties
- Fits exponential, loglogistic, Weibull, lognormal, and gamma survival models
- Graphs of marginal survivor, cumulative hazard, and hazard functions
- Fully integrated with Stata's **st** and **me**

results are in the hazard metric. Laparoscopic surgery has a hazard ratio > 1 and so decreases LOS. Age has a hazard ratio < 1 and so increases LOS.

Far more importantly, we see a large variation across both doctor and hospital that might have contaminated our results had we not taken them into account.

We can plot the marginal survivor function for surgery method:

```
. stcurve, survival at1(lap_surg=0) at2(lap_surg=1)
```



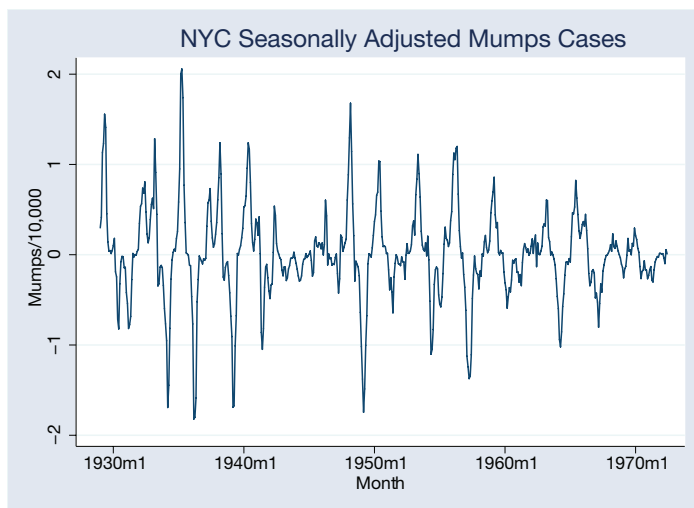
Markov-switching models

- Markov transition modeling
 - › Autoregressive model
 - › Dynamic regression model
- State-dependent variance parameters
- Tables of
 - › Transition probabilities
 - › Expected state durations
- Predictions
 - › Expected values of dependent variable
 - › Probabilities of being in a state
 - › Static (one-step)
 - › Dynamic (multistep)
 - › RMSEs of predictions

Markov switching is about time-series models in which the parameters change over time between regimes, and the switching is either abrupt or smooth. Smooth switching is achieved by autoregressively smoothing the transition. Abrupt switching is called dynamic. When the switching occurs is unknown, as are the number of switching points. The number of regimes is known.

Markov-switching models have been used to study asymmetric behavior of recessions and expansions; recessions happen fast, subsequent expansions, more slowly. They have been used for many other problems as well.

Say we have data on the incidence of mumps per 10,000 residents in New York City between 1928 and 1972.



There are periods of high and low volatility. The volatility is sometimes greater than at other times, and we are going to look at that. We are going to assume two regimes and fit a dynamic (abrupt-change) model.

```
. mswitch dr S12.mumpspsc,
  varswitch switch(LS12.mumpspsc, noconstant)
```

Note the two variance parameters **sigma1** and **sigma2**. They confirm our intuition of low- and high-variance regimes.

Note the two transition probability terms **p11** and **p21**. This is the Markov transition model. The full set of transition probabilities is as follows:

from/to state	1	2
1	0.76	1 - 0.76
2	0.15	1 - 0.15

The states are persistent. State 1 transits to state 1 with probability 0.76. State 2 transits to state 2 with probability 0.85 (1 - 0.15).

Survey for multilevel models

Stata 14 now provides survey-adjusted point estimates, standard errors, and tests for multilevel models. That includes adjustments for stratification, clustering, sampling weights, and finite-population corrections.

You can now use Stata's **svy:** prefix to fit multilevel mixed-effects models for continuous, binary, ordinal, count, and survival data models.

Sometimes, researchers analyze multistage survey data using single-level models. Stata 13 could do that. To properly adjust a multilevel model, however, we need to exploit the weights available at each stage of the survey. Stata now allows you to enter those weights.

You just survey set your data.

After setting your data, you can fit single-level or multilevel models. If you fit a single-level model, Stata automatically produces the single-level weights it needs from the multistage weights.

Bayesian analysis

- Bayesian estimation
 - › Continuous, binary, ordered, and count outcomes
 - › Univariate, multivariate, and multiple-equation models
 - › Linear models, nonlinear models, and generalized nonlinear models
 - › 10 likelihood models, including univariate and multivariate
- normal, logit, probit, ordered, Poisson ...
 - › 18 prior distributions, including normal, lognormal, multivariate normal, gamma, beta, Wishart ...
 - › Specialized priors, such as flat, Jeffreys, and Zellner's g
 - › User-defined likelihoods and priors
 - › Or write your own programs to
- calculate likelihood function and choose built-in priors
 - › Or write your own programs to calculate posterior density directly
- MCMC methods
 - › Adaptive Metropolis–Hastings (MH)
 - › Adaptive MH with Gibbs updates
 - › Full Gibbs sampling for certain likelihood and prior

Your Bayesian analysis can be as simple or as complicated as your research problem. Here's an overview.

First, fit the model. If we wanted to estimate the mean cholesterol level of children aged 5–10 whose parents have high cholesterol and if we wanted to use a normal model for cholesterol levels with noninformative priors for the parameters—flat prior for the mean and Jeffreys prior for the variance—we would type

```
. bayesmh chol, likelihood(normal({var}))  
  prior({chol:_cons}, flat)  
  prior({var}, jeffreys)
```

Point estimates, credible intervals, etc., are reported.

If we instead wanted to assume an informative normal prior centered at 190 mg/dL with a variance of 100, all based on previous studies, we would type

```
. bayesmh chol, likelihood(normal({var}))  
  prior({chol:_cons}, normal(190,100))  
  prior({var}, jeffreys)
```

Either way, convergence of MCMC and the distributions of the parameters can be explored using

```
. bayesgraph diagnostics {chol:_cons} {var}
```

We may be interested in estimating the probability that the mean cholesterol level is greater than 200 based on

the current sample.

```
. bayestest interval {chol:_cons}, lower(200)
```

Change point analysis

As an example, let's look at the British coal mining disaster dataset (1851–1962). Variable **count** records the number of disasters involving 10 or more deaths. There was a fairly abrupt decrease in the rate of disasters around 1887–1895. Let's estimate the date when the rate of disasters changed.

We will fit the model

```
count ~ Poisson( $\mu_1$ ), if year < cp  
count ~ Poisson( $\mu_2$ ), if year >= cp
```

cp —the change point—is the main parameter of interest. We are doing what's called a change-point analysis.

We will use noninformative priors for the parameters: flat priors for the means and a uniform on [1851, 1962] for the change point.

We will model the mean of the Poisson distribution as a mixture of μ_1 and μ_2 using a nonlinear specification.

As an aside, we will use the **noglmtransform** option so that cp is modeled as calendar year instead of $\ln(\text{calendar year})$.

What is Bayesian analysis?

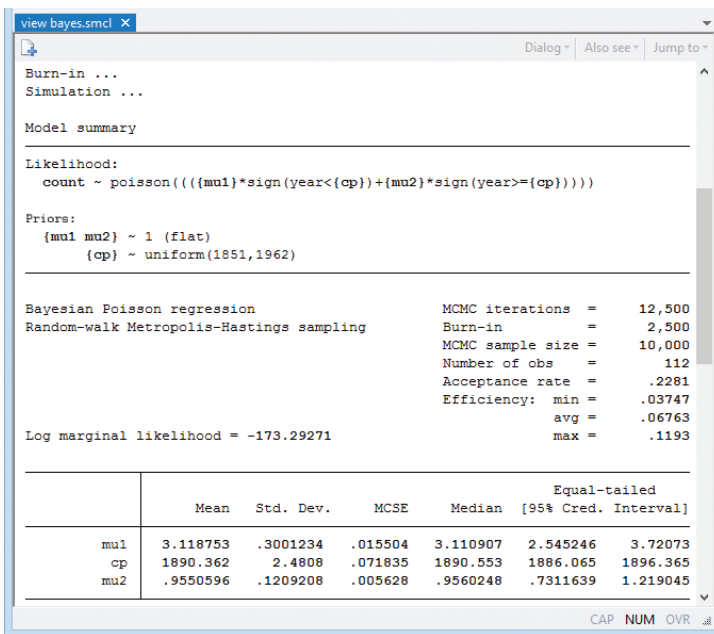
Bayesian analysis is a statistical analysis that answers research questions about unknown parameters using probability statements. For example, what is the probability that the average male height is between 70 and 80 inches or that the average female height is between 60 and 70 inches? Or, what is the probability that people in a particular state vote Republican or vote Democrat? Or, what is the probability that a person accused of a crime is guilty?

Such probabilistic statements are natural to Bayesian analysis because of the underlying assumption that all parameters are random quantities. In Bayesian analysis, a parameter is summarized by an entire distribution of values instead of the one fixed value used in classical frequentist analysis. The estimation of this distribution, the posterior distribution of a parameter of interest, is at the heart of Bayesian analysis.

- combinations
- › Graphical tools to check MCMC convergence visually
- › Explore MCMC efficiency by computing effective sample sizes, autocorrelation times, and efficiencies
- Bayesian summaries
 - › Posterior means and SDs
 - › Monte Carlo standard errors (MCSEs)
- › Credible intervals (CrIs)
- › Compute any of the above for parameters or functions of parameters
- Hypothesis testing
 - › Interval-hypothesis testing for parameters or functions of parameters
 - › Model-based hypothesis testing by computing model posterior probabilities
- Model comparison
 - › Bayesian information criteria such as deviance information criterion
 - › Bayes factors
- Save your MCMC and estimation results for future use

To fit the model, we type

```
. bayesmh count = ({mu1}*sign(year < {cp}) +
                  {mu2}*sign(year >= {cp})),
  likelihood(poisson, noglmtransform)
  prior({mu1 mu2}, flat)
  prior({cp}, uniform(1851,1962))
  initial({mu1 mu2} 1 {cp} 1906)
```



The posterior mean estimate of the change point is 1,890.362.

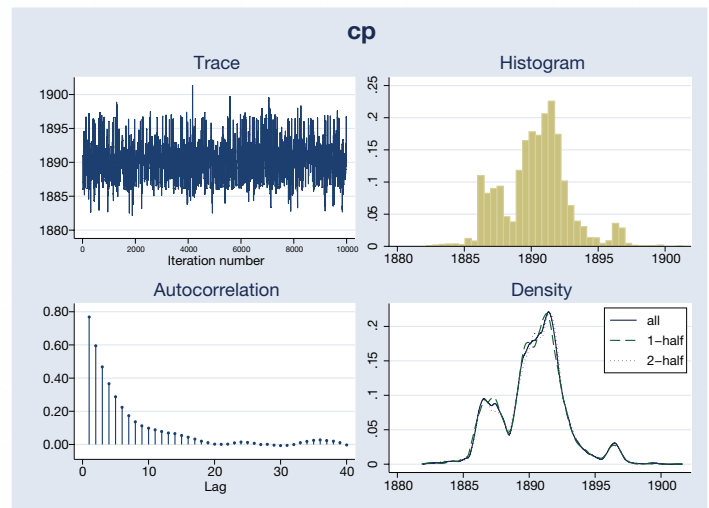
The standard error of the posterior mean estimate (MCSE) is 0.07. The MCSE is about the accuracy of our simulation results. We would like it to be zero, but that would take an infinite number of MCMC iterations. We used 10,000 and have results accurate to about one decimal place. That's good enough, but if we wanted more accuracy, we could increase the MCMC sample size.

The corresponding 95% CrI of **cp** is [1886, 1896]. The probability that the change point is between 1886 and 1896 is about 0.95.

Next

The interpretation of our change-point results is valid only if MCMC converged. We can explore convergence visually for **cp**.

```
. bayesgraph diagnostics {cp}
```

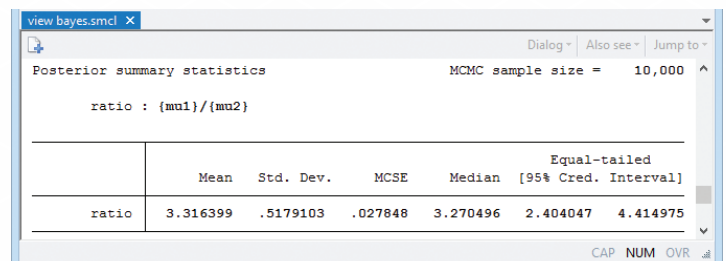


The graphical diagnostics look reasonable. The marginal posterior distribution of the change point has the main peak at about 1890 and two smaller bumps around the years 1886 and 1896, which correspond to local peaks in the number of disasters.

Change-point analysis—follow on

We might be interested in estimating the ratio between the two means. If we were, it would be easy to get:

```
. bayesstats summary (ratio: {mu1}/{mu2})
```

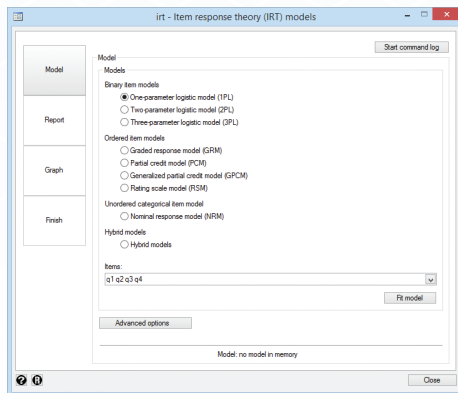


IRT (item response theory)

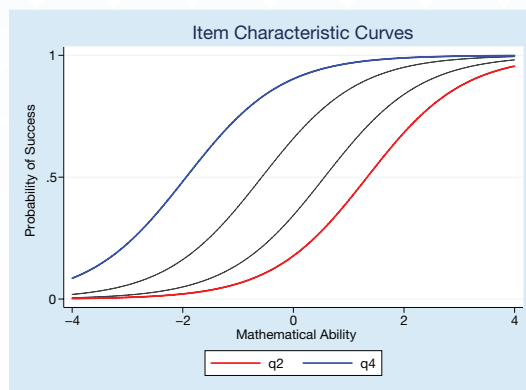
- Binary response models
 - › One-parameter logistic (1PL)
 - › Two-parameter logistic (2PL)
 - › Three-parameter logistic (3PL)
- Ordinal response models
 - › Rating scale
 - › Graded response
 - › Partial credit
- Categorical response model
 - › Nominal response
- Hybrid models with differing response types
- Graphs
 - › Item characteristic curve
 - › Test characteristic curve
 - › Item information function
 - › Test information function
- Control panel to guide you through the analysis

What's this about?

IRT stands for “item response theory”. IRT models explore the relationship between a latent (unobserved) trait and items that measure aspects of the trait. This often arises in standardized testing, where the items are a set of questions and the trait, an unobserved ability.



trait—by graphing the item characteristic curves (ICCs) using **irtgraph icc**.



We made the easiest question blue and the hardest one red. The probability of succeeding on the easiest is higher than it is for the hardest. In this case, that's true for every level of ability. We fit a 1PL model. 2PL and 3PL would not have prevented curves from crossing each other.

Let's see it work

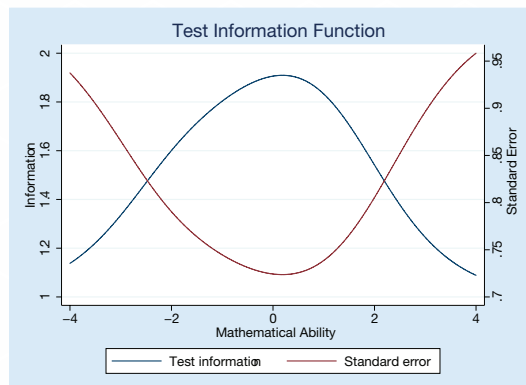
We have a test designed to assess mathematical ability based on four questions (aka, items) that are scored incorrect (0) or correct (1). We fit a one-parameter logistic model by typing **irt 1pl q1-q4**. Or we fit our model from IRT's Control Panel (shown above).

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Discrim	1.151231	.087828	13.11	0.000	.9790909 1.32337	
q1	Diff	-.5720887	.0852481	-6.71	0.000	-.7391719 -.4050054
q2	Diff	1.333696	.1134982	11.75	0.000	1.111244 1.556149
q3	Diff	.5615614	.084883	6.62	0.000	.3951938 .727929
q4	Diff	-1.939248	.1466532	-13.22	0.000	-2.226683 -1.651813

Coefficients labeled **Diff** report difficulty; question 4's coefficient is -1.94 —it's the easiest—and **q2** at 1.33 is the most difficult.

We can visualize the relationship between questions and mathematical ability—between the items and latent

irtgraph tif graphs the test information function.



This graph combines all the questions and shows where on the scale of mathematical ability we get the most from our test in terms of information. We wish the curves were flatter.

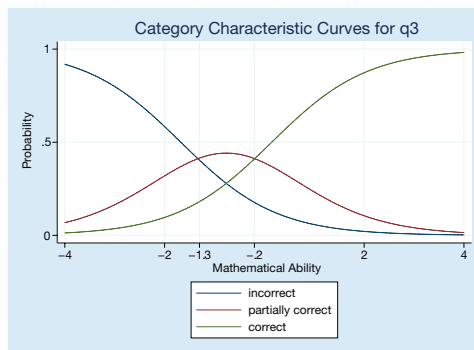
Stata can analyze ordinal and categorical responses, too. Here's another four-item test in which responses are graded. Each problem is scored 0 (incorrect), 1 (partially correct), or 2 (correct).

With these ordinal data, we will fit a graded response

model (we could instead fit a partial-credit model or a rating-scale model). We type **irt grm q1-q4**.

		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
q1	Discrim	2.461675	.2182542	11.28	0.000	2.033905 2.889445
	>=1	-1.854518	.0755534	-24.55	0.000	-2.0026 -1.706436
	=2	-1.086664	.0458764	-23.69	0.000	-1.17658 -.9967479
q2	Discrim	1.70363	.1209653	14.08	0.000	1.466543 1.940718
	>=1	-1.295414	.0615988	-21.03	0.000	-1.416145 -1.174682
	=2	-.5417255	.0378556	-14.31	0.000	-.6159211 -.4675298
q3	Discrim	1.041941	.0763572	13.65	0.000	.8922838 1.191599
	>=1	-1.673948	.1052791	-15.90	0.000	-1.880291 -1.467605
	=2	-.1471413	.0437065	3.37	0.001	.0614781 .2328044
q4	Discrim	.7866272	.0619737	12.69	0.000	.665161 .9080935
	>=1	-.8857031	.0800935	-11.06	0.000	-1.042684 -.7287227
	=2	1.380453	.1060525	13.02	0.000	1.172594 1.588312

Here's the category characteristic curve showing how question 3 relates to mathematical ability. We use **irtgraph icc**.



Respondents with mathematical ability levels below -1.3 are most likely to answer **q3** with a completely incorrect answer, those with levels between -1.3 and -0.2 are most likely to give a partially correct answer, and those with ability levels above -0.2 are most likely to give a completely correct answer. Question 3 focuses on the lower levels of mathematical ability.

From the test characteristic curve produced by **irtgraph tcc**, we see how the expected total test score relates to mathematical ability levels.

If we had the space, we'd show you the test characteristic curve. You would see that out of a possible 8 points on the test, a person with above-average mathematical ability would be expected to score above 5.

Not interested in standardized testing?

IRT models can be used to measure many types of latent traits. For example,

- attitudes
- personality traits
- health outcomes
- quality of life

Use IRT for analyzing any unobservable characteristic for which binary or categorical measurements are observed.

Panel-data survival models

- Random effects and random coefficients
- Right-censoring
- Exponential, loglogistic, Weibull, lognormal, and gamma survival models
- Proportional-hazards metric
- Accelerated failure-time metric
- Single- and multiple-record survival-time data

Stata fits survival models. In survival models relevant here, survival time is modeled using a parametric distribution, and right-censoring is allowed.

Stata fits panel-data models. In panel-data models relevant here, the data occur in groups of observations that share something in common that is modeled as unobserved random effects.

In Stata 14, we put the two models together.

We model the time to infection after catheter insertion. We have multiple observations on each patient.

xtstreg is fully integrated with Stata's **xt** and **st** features, so first, we must **stset** our survival data,

```
. stset time, failure(infect)
```

and we must **xtset** our panel data (same data),

```
. xtset patient
```

We fit a panel-data Weibull survival model of time to next infection on **age** and **female**. We type

```
. xtstreg age female, distribution(weibull)
```

	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.007329	.0137828	0.53	0.594	.9806742 1.034708
female	.1910581	.0999004	-3.17	0.002	.0685629 .5324042
_cons	.0073346	.0072307	-4.99	0.000	.0010623 .0506427
/ln_p	.222825	.1386296	1.61	0.108	-.0488841 .494534
/sigma2_u	.8234584	.4812598			.2619194 2.588902

LR test vs. Weibull model: $\chi^2(01) = 9.40$ Prob >= $\chi^2(01) = 0.0011$

The results look just as if **streg** had reported them, but with the addition of panel-data features and an estimated **/sigma2_u**, which is the estimated variance of the random effect.

We could fit a similar model using **streg** with shared frailties, but **streg** assumes the frailties follow a gamma distribution. **xtstreg** makes the often more plausible assumption that random effects are normally distributed, meaning frailties are lognormal.

Stata can also fit survival models with both random intercepts and random coefficients.

Treatment effects for survival models

- Exponential, loglogistic, Weibull, lognormal, survival distributions
- Right-censoring
- Integrated with **st**
- Methods
 - › Inverse-probability weighting (IPW)
 - › Survival regression adjustment (RA)
 - › Weighted regression adjustment (WRA)
- › Inverse-probability weighted regression adjustment (IPWRA)
- Multilevel and multivalued treatments
- Average treatment effect (ATE)
- Potential-outcome means (POMs)
- ATE among the treated (ATET)
- Diagnostics for balancing and overlap

Stata's treatment-effects estimators now support parametric survival-time models.

We want to measure the effect of (continued) smoking on time to second heart attack among women aged 45–55. Not all women, obviously, are observed to have a second heart attack, but we'll assume that many of these women do have second heart attacks (whether observed or not).

We are going to show you three models. In the first, we model time to second heart attack. In the second, we instead model treatment. In the third, we model both. Obviously, results depend on the model being correct.

Before we can start, we must **stset** our survival data. We type **stset atime, failure(fail)**. Variable **atime** records the time of second heart attack or censoring, and variable **fail** records whether the event was a second heart attack.

Here's our first model: Time to second heart attack is modeled as Weibull using age, exercise, quality of diet, and education. We type

```
. stteffects ra (age exercise diet education) (smoke)
```

	_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
ATE	smoke (Smoker vs Nonsmoker)	-1.956657	.3331787	-5.87	0.000	-2.609676 -1.303639
	Pomean smoke Nonsmoker	4.243974	.2620538	16.20	0.000	3.730358 4.75759

Here's our second model. We model continued smoking (and the censoring mechanism) as being determined by age, exercise, diet, and education. We

fit the model by typing

```
. stteffects ipw (smoke age exercise diet education) (age exercise diet education)
```

	_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
ATE	smoke (Smoker vs Nonsmoker)	-2.187297	.6319837	-3.46	0.001	-3.425962 -.9486314
	Pomean smoke Nonsmoker	4.225331	.517501	8.16	0.000	3.211047 5.239614

And in our final model, we assume that both survival time and continued smoking are determined by age, exercise, diet, and education.

```
. stteffects ipwra (age exercise diet education) (smoke age exercise education) (age exercise diet education)
```

	_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
ATE	smoke (Smoker vs Nonsmoker)	-2.285057	.7318456	-3.12	0.002	-3.719448 -.8506656
	Pomean smoke Nonsmoker	4.385841	.6427521	6.82	0.000	3.12607 5.645612

Now, compare results. They are all in agreement!

Endogenous treatment effects

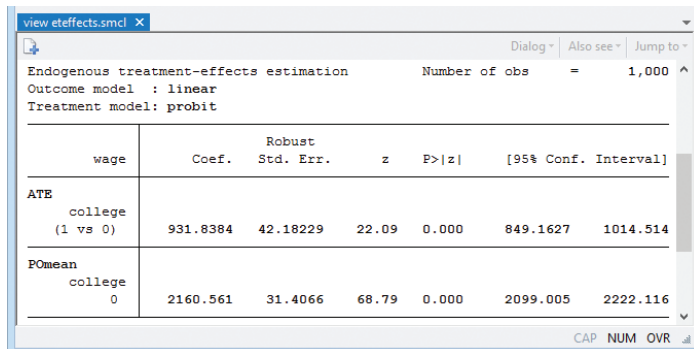
- Endogenous treatments
- Control function estimator
- Continuous, fractional, binary, and count outcomes
- Average treatment effect (ATE)
- ATE among the treated (ATET)
- Potential-outcome means (POMs)

Treatment-effects estimators extract experimental-style causal effects from observational data.

New in Stata 14 is dealing with endogeneity, which is to say, when the same unobserved variable(s) affected both treatment and outcome.

We want to measure the effect of a college degree on wages. College is our treatment and wages, our outcome. We fit a model of outcome on treatment. We worry that unobserved ability will affect both wages and college attainment. To eliminate the confounding effect of unobserved ability, we model college attainment. We type

```
. eteffects (wage tenure c.age##c.age)
      (college c.age##c.age i.pcollege)
```



	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
ATE					
college (1 vs 0)	931.8384	42.18229	22.09	0.000	849.1627 1014.514
Pomean					
college 0	2160.561	31.4066	68.79	0.000	2099.005 2222.116

We modeled wages as determined by job tenure and age, and college attainment, by age and number of parents who attended college. The treatment model was probit; the outcome model, linear.

The estimated ATE is \$931.84 per month for college attainment. The potential-outcome mean is the expected wage if no one attended college. It's \$2,160.56 per month.

If there is endogeneity and we had not accounted for it, we would have obtained incorrect estimates. In this case, we are using simulated data, and we can tell you that the true ATE was 924. If we were to estimate ATE ignoring the endogeneity, Stata would report an ATE of \$1,514.

Balance diagnostics for treatment effects

Treatment effects extract experiment-style causal effects from observational data. A key requirement is that our treatment-effects model explicitly or implicitly reweights the data such that the model-adjusted distribution of the covariates is comparable across treatment groups. Balance diagnostics check this.

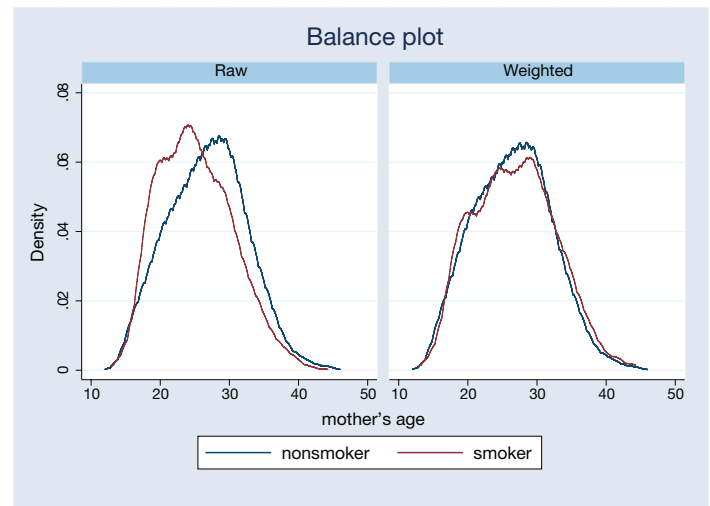
Four diagnostics and tests are provided.

One reports, for each covariate, the model-adjusted difference in means in the treatment groups and the ratio of variances, providing a useful diagnostic.

Another graphs the model-adjusted estimated pdfs of covariates; these pdfs can be examined visually to verify that they are approximately equal.

Another does the same but uses box plots rather than smoothed pdfs.

And finally, an overidentification test is provided. It statistically tests whether the model-adjusted means of the covariates are the same between groups.



Find your postestimation

Did you see the Postestimation Selector on the front page? You have got to try it. Bring up this little window, and, as you fit models, Stata will show the postestimation statistics, tests, and predictions that you could use right now. Fit a linear regression and one list appears. Fit a logistic regression and another list appears. There's overlap in the lists, of course, but each is tailored to the estimator you used and options you specified. It's useful for teaching, but it's even better in the hands of research professionals. Stata has so many postestimation features that too often researchers didn't realize Stata had one they needed.

New in SEM (structural equation modeling)

- Survival models (parametric)
 - Latent predictors
 - Mediation models and more
 - Unobserved components
 - Multilevel survival models—random intercepts and random coefficients
 - Survival outcomes with other outcomes
 - Right-censoring
 - Left-truncation
 - Exponential, loglogistic, Weibull, lognormal, and gamma survival distributions
- Generalized models now support survey data
 - Adjusted point estimates, SEs, and tests
 - Sampling weights
 - Sampling weights at each stage of survey (multilevel models)
 - Clustered sampling
 - Stratified sampling and poststratification
 - Finite population corrections
 - Linearized, bootstrap, jackknife, or BRR standard errors

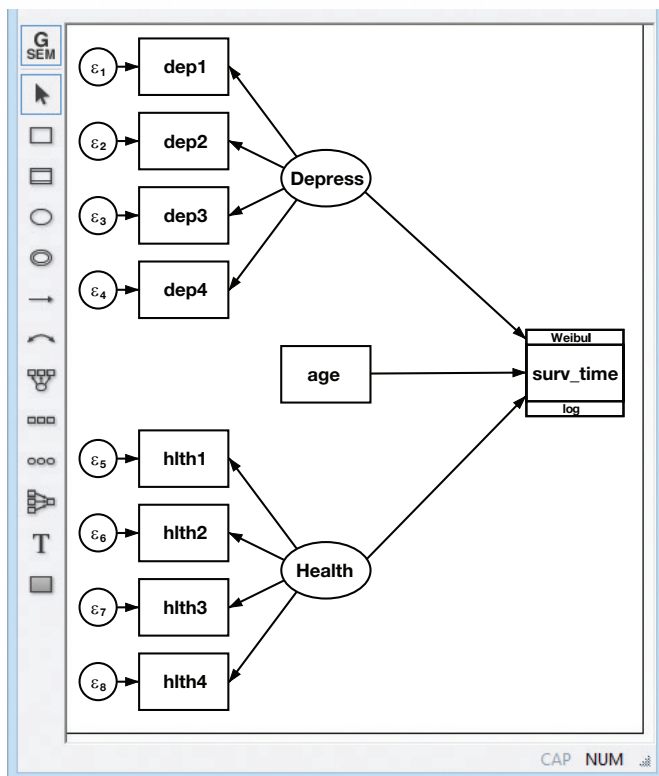
Example 1: Survival model

Let's do a survival model combined with CFA (confirmatory factor analysis). CFAs model the level of a latent trait using observable measurements.

We analyze survival times of nursing home residents. We have censored data; thankfully, not all the residents have died yet.

- We posit that survival times are determined by age, depression level, and overall health.
- We have four variables that each measure aspects of depression (our first latent trait).
- We have four variables that each measure aspects of health (our second latent trait).

We can create our model using Stata's SEM Builder:



Or we can go directly to typing a command:

```
. gsem (surv_time <- x Dep Health,
       family(weibull, fail(death)))
      (Depress -> dep1 dep2 dep3 dep4)
      (Health -> hlth1 hlth2 hlth3 hlth4)
```

Either way, we get the same output:

		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Generalized structural equation model						
Response		: surv_time				
Family		: Weibull				
Form		: proportional hazards				
Link		: log				
Log likelihood		= -5917.3602				
(1) [var(Dep)]_cons = 1						
(2) [var(Health)]_cons = 1						
Number of obs = 500						
No. of failures = 399						
Time at risk = 362.30429						
surv_time <-	x	-.9669588	.0919301	-10.52	0.000	-1.147138 - .7867791
	Dep	-1.417553	.1336268	-10.61	0.000	-1.679457 -1.155649
	Health	1.80293	.1614184	11.17	0.000	1.486556 2.119304
	_cons	1.560356	.1459481	10.69	0.000	1.274303 1.846409
dep1 <-	Dep	1.031049	.0409969	25.15	0.000	.9506962 1.111401
	_cons	.0914346	.0510285	1.79	0.073	-.0085794 .1914486
dep2 <-	Dep	.5164082	.0380608	13.57	0.000	.4418105 .5910059
	_cons	.0439688	.039877	1.10	0.270	-.0341866 .1221262

SEM produces a lot of output; we've selected just a portion of it.

By the way, another way to think about the observed variables measuring depression and health is that each measures depression (health) with error. Combining the multiple measures allows us to wash away the errors-in-variables bias.

Example 2: Survey data

We want to fit a CFA model for students' attitudes toward math using five ordinal measurements, **att1–att5**. That's easy enough:

```
. gsem (MathAtt -> att1 att2 att3 att4 att5), oprobit
```

However, our data were the result of multiple-stage cluster sampling. Schools were sampled, and then

- Satorra–Bentler scaled χ^2
 - › Adjustment for nonnormal data
 - › All relevant goodness-of-fit statistics adjusted
 - › Robust standard errors and postestimation tests

What is SEM?

SEM handles one or more latent (unobserved) variables. (They can be exogenous or endogenous.)

SEM handles one or more observed endogenous variables (and the structural relationships among them).

SEM handles multilevel random effects and random coefficients.

SEMs can be linear or generalized linear, meaning probit, logit, Poisson, and others.

students were sampled from the chosen schools. SEM's new survey features allow us to specify the primary sampling unit and the sampling weight. We just survey set the data:

```
. svyset school [pweight=finalweight]
```

If we put **svy:** in front of the same simple SEM command that we would have typed with random (i.i.d.) data, **gsem** now produces survey-adjusted results. Just type

```
. svy: gsem (MathAtt -> att1 att2 att3 att4 att5), oprobit
```

By the way, we could have specified this entire model, including the survey aspects of the sample, from Stata's SEM Builder.

view sem.smcl X

```
. svy: gsem (MathAtt -> att1 att2 att3 att4 att5), oprobit
(running gsem on estimation sample)

Survey: Generalized structural equation model

Number of strata = 1          Number of obs = 200
Number of PSUs  = 20         Population size = 2,852
                                   Design df = 19

( 1) [att1]MathAtt = 1
```

	Linearized		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
att1 <- MathAtt	1 (constrained)					
att2 <- MathAtt	.2771797	.1553982	1.78	0.090	-.0480726	.6024319
att3 <- MathAtt	-1.371444	.4569308	-3.00	0.007	-2.327811	-.4150767
att4 <- MathAtt	-.6330853	.2979832	-2.12	0.047	-1.256771	-.0093992
att5 <- MathAtt	.2698887	.0826312	3.51	0.002	.1169396	.4628377
att1 /cut1	-.6870853	.1207852	-5.69	0.000	-.9398916	-.4342791
/cut2	-.2281198	.1088803	-2.09	0.047	-.4487277	.0078176

New in power and sample size

- Contingency tables
 - › Stratified 2x2 tables (Cochran–Mantel–Haenszel)
 - › 1:M matched case–control studies
 - › Trend in Jx2 tables (Cochran–Armitage)
- Survival analysis
 - › 2-sample log-rank test
 - › 2-sample exponential test
 - › Cox PH regression
- Multiple values of parameters
- Automatic and custom tables and graphs

With Stata's **power**, you can compute power, sample size, and effect size. Enter any two and get the third.

Among other new features, **power** now provides PSS for matched case–control studies.

For instance, consider cancer among smokers and nonsmokers. How many case–control pairs do we need to achieve 80% power of detecting a 1.7 odds ratio with a 5% significance test if we used a 2-sided association test? If we knew from previous studies that

the probability of exposure (smoking) for controls was roughly 0.22, we would type

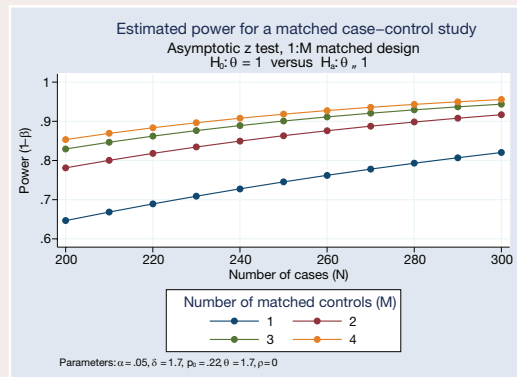
```
. power mcc .22, oratio(1.7)
```

and learn that we need 285 cases and 285 controls.

1:M matching is often used to reduce the required number of cases because cases are often more difficult to obtain than controls. It is thus useful to evaluate designs with different values of M.

We could plot power curves for designs with 1:1, 1:2, 1:3, and 1:4 matching by typing

```
. power mcc 0.22, oratio(1.7) n(200(10)300) m(1 2 3 4) graph
```



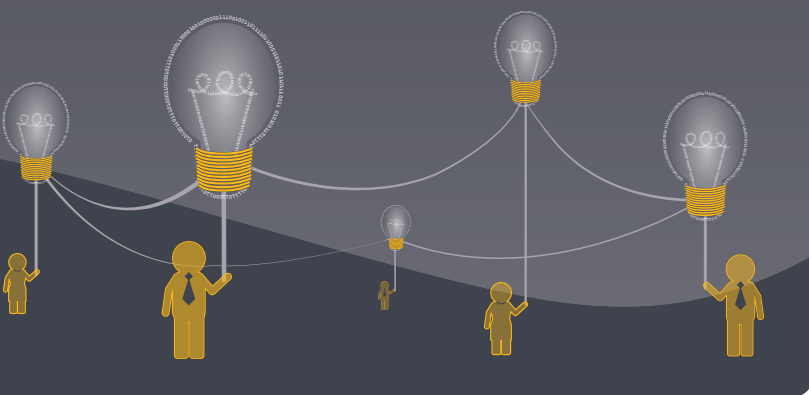
STATA CONFERENCE

July 30–31, 2015

Hyatt Regency Columbus

Join us in Columbus, Ohio, for two days of networking and Stata exploration. Don't miss this opportunity to hear all about What's New in Stata 14. Then, why not take advantage of your surroundings? Just like Stata, Ohio's capital city offers a little something for everyone.

stata.com/columbus2015



Margins gets better

One of Stata's neatest commands, **margins**, is difficult to explain. With **margins**, you can do what-if analyses. What would have been observed if everyone in the data was male? Female? What would have happened if the men in the data had their same characteristics but were relabeled women, and the women had their same characteristics but were relabeled men? If you can think of a counterfactual, potential outcome, comparison, or contrast, **margins** can do it.

margins is used after fitting a model. **margins** combines the fitted results, the data, and a little that you type to produce estimates of marginal effects, marginal means, predictive margins, population-averaged effects, and least-squares means and presents the estimates in tables or graphs.

margins now automatically produces multiple results for the multiple outcomes for estimators like multinomial logistic, ordered logistic, and multivariate regression. We wish we had space to show you some graphs.

margins now handles multilevel models and SEM by integrating over unobserved (latent) variables such as the random effect.

Stat/Transfer 13

Now with support for Stata 14, including Unicode and more than 2 billion observations.

Stat/Transfer 13 also adds support for Eviews and Genstat.

Other new options include

- Preserve numeric widths
- Control over SAS date and time formats
- Blank columns can be optionally transferred from worksheets
- Value labels can be written to Excel
- Editable schema for ASCII-delimited files—reorder, rename, reformat, assign labels to variables, and more

And more new features

We wish we had the space to tell you more about the following. Visit stata.com/stata14. You can read more, and even read the manual entries and worked examples.

Tests for structural break in time series let you test after estimation for a break at known or unknown dates.

Hurdle model estimation allows you to model censored and uncensored outcomes in separate equations; uncensored outcomes are observed when a hurdle is cleared.

Censored Poisson regression lets you model count processes with values that are not observed below a threshold, not observed above a threshold, or both simultaneously.

Sampling weights allowed with treatment effects is highly requested, so now Stata allows it.

Integrate better with Excel® by using Stata to insert graphs, formulas, formatted text, and more.

More than 2 billion observations are allowed by the multiprocessor version of Stata, Stata/MP. You are limited only by memory.

Stata's interface can now be in Spanish or Japanese, including all menus, dialogs, and other interface elements.

ICD-10 diagnosis codes are now understood by Stata just as ICD-9 codes are.

The 64-bit Mersenne Twister now provides Stata's pseudorandom numbers, and PRNGs for more distributions are available.

Quick starts have been added to the manual to give you a quick overview or refresher for common syntaxes.

Fractional regression lets you model variables that are fractions, proportions, or rates.

Upgrade now at stata.com/stata14.

Or just explore even more about the new features.