

NEW

Announcing **STATA**[®]

13

Stata 13 ships June 24. Order now at stata.com

Highlights of What's New in Stata 13

Treatment effects

- Inverse probability weights (IPW)
- Regression adjustment
- Propensity-score matching
- Covariate matching
- Doubly robust methods
- Continuous, binary, and count outcomes

Multilevel/mixed models

- Negative binomial
- Ordered logistic
- Ordered probit
- Multinomial logistic
- GLM
- Hierarchical models

Long strings

- 2 billion characters
- Text strings
- Binary large objects (BLOBs)
- Import/Export/ODBC/SQL
- Work just like Stata strings

Generalized SEM

- Generalized linear responses: binary, count, ordered outcomes
- Multilevel/hierarchical models: nested and crossed models
- Random slopes and intercepts
- Fast

Power and sample size

- Means, proportions, variances, correlations
- Case-control and cohort studies
- Interactive control panel
- Tabular results
- Automatic graphs

Forecasting

- Time series and panels
- One to thousands of equations
- Identities
- Add factors
- Dynamic and static
- CIs via stochastic simulation
- Compare scenarios

Effect sizes

- Means
- ANOVA
- Linear regression
- Confidence intervals
- Cohen's d , Hedges's g , Glass's Δ , η^2 , ω^2

Panel data

- Ordered outcomes
- RE ordered probit
- RE ordered logistic
- Cluster-robust SEs to relax distributional assumptions and allow for correlated data

Project Manager

- Organize files (1-10,000)
- Multiple projects
- Filter on filenames
- Click to open
- Click to run

More statistics

- Ordered probit with selection
- Poisson with endogenous covariates
- Robust SEs for quantile regression
- ML estimation without programming
- Fractional-polynomial prefix

More documentation

- Treatment-effects manual
- Multilevel mixed-effects manual
- Power and sample-size manual
- 2,000 more pages
- 11,000+ total pages

And more

- Factor variables show labels
- Import delimited with preview
- Import from Haver Analytics
- Business calendars from data
- Java plugin API
- FTP and secure HTTP

Turn the page to learn more.



Long strings

- Each up to 2 billion characters in length (previous maximum was 244)
- All string functions work with them
- All of Stata works with them
- Read/write files into/from them
- Plain text and binary, including BLOBs—binary large objects

Nearly everybody has wanted strings longer than 244 characters, Stata's old limit. People asked for 500, 1,000, even 2,000 characters. So we settled on 2 billion.

These long strings work just like strings in Stata have always worked.

All of Stata understands them—string functions, commands, ODBC, **import**, **export**, **sort**, **by**, etc. Use them in analysis. Use them in data management. Use them in programming.

You can read entire files into them, even binary files such as Word documents and JPEG images.

See the video overview at stata.com/videos13/long-strings.

Learn everything about using long strings in Stata in the *User's Guide* at stata.com/manuals13/u12.4.

Multilevel mixed-effects GLM

New estimators

- Probit
- Complementary log–log
- Ordered logistic
- Ordered probit
- Negative binomial
- Multinomial logistic

New features

- Constraints on variance components
- Cluster-robust SEs to relax distributional assumptions and allow for correlated data
- Posterior mode and mean estimates of random effects

Models

- Nested models (hierarchical)
- Crossed models
- Random intercepts
- Random coefficients (slopes)

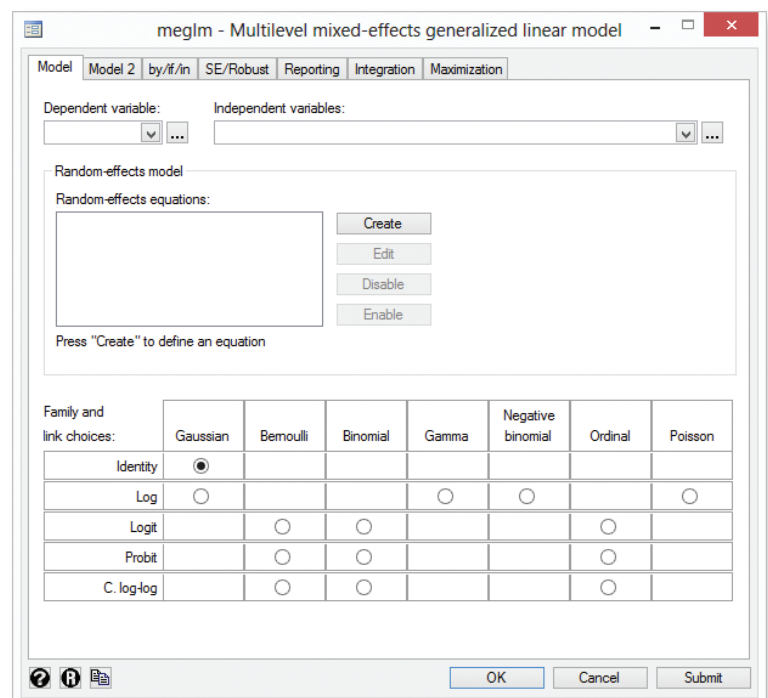
Stata now fits multilevel generalized linear models.

Multilevel data naturally divides into groups that have something in common—students attending the same school; or patients treated by the same doctor, with doctors themselves working at the same hospital. Stata's multilevel estimators support two-level data, three-level data, and higher-level data.

Generalized outcomes can be continuous, binary, count, ordered, or categorical.

See the video overview at stata.com/videos13/multilevel-mixed-effects.

Learn everything about multilevel modeling in Stata in the all-new 356-page manual at stata.com/manuals13/me.



Power and sample size

Solve for

- Power
- Sample size
- Minimum detectable effect
- Effect size

Comparisons of

- Means (t-tests)
- Proportions
- Variances
- Correlations

Designs

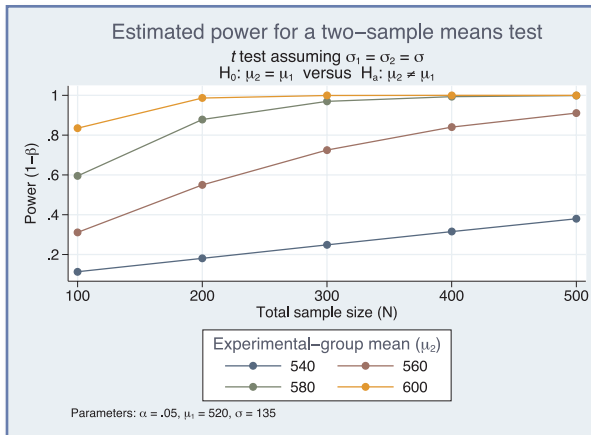
- Matched case-control studies
- Cohort studies
- Cross-sectional studies

Tables and graphs

- Automated
- Custom

Say we are planning an experiment to determine whether students who prepare for the SAT exam obtain higher math scores by taking classes rather than studying independently. The national average math score is 520 with a standard deviation of 135. We want to see the power obtained for sample sizes of 100 through 500 when scores increase by 20, 40, 60, and 80 points or, equivalently, when average scores increase to 540, 560, 580, and 600. We type

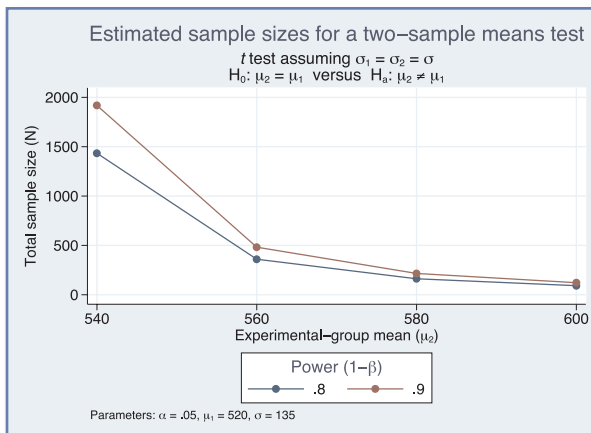
```
. power twomeans 520 (540 560 580 600) , n(100 200 300 400 500) sd(135) graph
```



We assumed above that those studying independently will obtain the national average of 520 and that the standard deviations will be 135 for both groups. We could relax either assumption with **power**.

We might be more interested in the sample size necessary to detect the increased scores at power levels of 0.8 and 0.9. In that case, we type

```
. power twomeans 520 (540 560 580 600) , power(0.8 0.9) sd(135) graph
```



In our example, we dealt with a comparison of means across two independent samples. **power** can also produce comparisons of proportions, variances, and correlations for one sample, two samples, or paired samples.

See the video overview at stata.com/videos13/power-and-sample-size.

Learn everything about power and sample size in Stata in the all-new 225-page manual at stata.com/manuals13/pss, and in particular, see the introduction at stata.com/manuals13/pss_intro.

Treatment effects

Estimators

- Inverse probability weights (IPW)
- Propensity-score matching
- Covariate matching
- Regression adjustment
- Doubly robust methods
 - › Augmented IPW
 - › IPW with regression adjustment

Features

- Multilevel and multivalued treatments
- Average treatment effects (ATEs)
- ATEs on the treated (ATETs)
- Potential-outcome means (POMs)
- Continuous, binary, and count outcomes

Endogenous treatment estimators

- Linear regression
- Poisson regression

A treatment is a new drug regimen, a surgical procedure, a training program, or even an ad campaign, intended to affect an outcome such as blood pressure, mobility, employment, or sales.

In the best of worlds, we would measure the difference in outcomes by designing an experiment that assigns subjects randomly to the treatment and control.

We can't always do that. When we need to make do with observational data—when the subjects themselves choose whether to be treated or the choice is otherwise nonrandom—we need a treatment-effects estimator.

Say we have a training program in which participants enroll voluntarily. In the raw data, participants do poorly relative to nonparticipants, even after the training. Even so, the training program might have improved their outcomes—say, hourly wages—over what they would have been. To determine this, we need a treatment-effects estimator.

There are different treatment-effect estimators for different situations. Let us tell you about them.

When we know the determinants of participation, the appropriate estimators include IPW and propensity-score matching. We might type

```
. teffects ipw      (wage) (trained x1 x2)
. teffects psmatch (wage) (trained x1 x2)
```

When we instead know the determinants of outcome, the appropriate estimators include regression adjustment and covariate matching. We might type

```
. teffects ra      (wage x1 x3) (trained)
. teffects nnmatch (wage x1 x3) (trained)
```

When we know both, we can use the doubly robust estimators—augmented IPW and IPW with regression adjustment.

We might type

```
. teffects aipw   (wage x1 x3) (trained x1 x2)
. teffects ipwra  (wage x1 x3) (trained x1 x2)
```

We only need to be right about one of the adjustments—**wage** needs to be a function of **x1** and **x3** or **trained** needs to be a function of **x1** and **x2**. That is a feature of the doubly robust methods.

To obtain the doubly robust IPW regression-adjusted results, we type

```
. teffects ipwra (wage x1 x3) (trained x1 x2)
```

```
Iteration 0: EE criterion = 2.523e-16
Iteration 1: EE criterion = 4.680e-30
```

```
Treatment-effects estimation      Number of obs      =      1000
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: logit
```

wage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
ATE trained (1 vs 0)	.4646677	.080218	5.79	0.000	.3074432	.6218921
POmean trained 0	8.275523	.0628417	131.69	0.000	8.152356	8.398691

The output reveals that the average treatment effect (ATE)—the effect we would have observed had the entire population been treated—is 0.46, meaning 46 cents more in the hourly wage. (The baseline wage is \$8.28. This is the wage had no one been treated).

If we instead want the average treatment effect for the treated (ATET), we add the **atet** option:

```
. teffects ipwra (wage x1 x3) (trained x1 x2), atet
```

```
Iteration 0: EE criterion = 2.523e-16
Iteration 1: EE criterion = 3.211e-30
```

```
Treatment-effects estimation      Number of obs      =      1000
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: logit
```

wage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
ATET trained (1 vs 0)	.8212578	.0948967	8.65	0.000	.6352636	1.007252
POmean trained 0	7.974128	.0886919	89.91	0.000	7.800295	8.14796

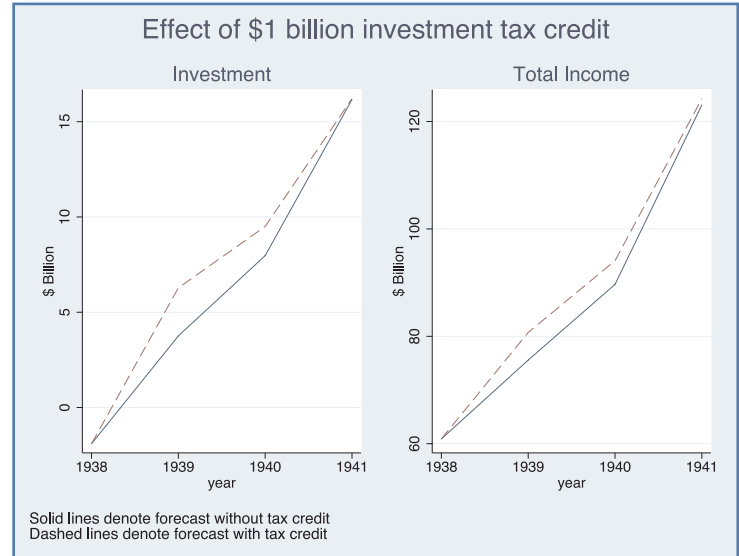
The ATET is 82 cents more per hour over the baseline among the treated, which is \$7.97.

For more information on treatment effects, see the video overview at [stata.com/videos13/treatment-effects](https://www.stata.com/videos13/treatment-effects).

Learn everything about treatment effects in Stata in the all-new 158-page manual at [stata.com/manuals13/te](https://www.stata.com/manuals13/te).

Forecasting

- Time series and panel datasets
- Multiple estimation results
 - › OLS, VARs, VECs, ARIMA, 3SLS, more
 - › Estimated with Stata or
 - › Obtained from outside sources
 - › One equation or thousands
- Identities
- Add factors and other adjustments
- Dynamic or static (1 step ahead) forecasts
- Solve simultaneous systems
- Confidence intervals via stochastic simulation
- Compare forecasts of alternative scenarios



Let us show you. We will forecast a simple seven-equation model produced by one estimation command, but we could forecast a model with thousands of equations obtained from thousands of estimation commands.

In our simple example, we fit a classic model of the U.S. economy to pre-World War II data by using 3SLS:

```
. reg3 (c p L.p w) (i p L.p L.k) (wp y L.y yr), endog(w p y) exog(t wg g)
. estimates store klein
```

Now we tell **forecast** about our estimation results and the model's identities:

```
. forecast create kleinmodel
Forecast model kleinmodel started.

. forecast estimates klein
Added estimation results from reg3.
Forecast model kleinmodel now contains 3 endogenous variables.

. forecast identity y = c + i + g
Forecast model kleinmodel now contains 4 endogenous variables.

. forecast identity p = y - t - wp
Forecast model kleinmodel now contains 5 endogenous variables.

. forecast identity k = L.k + i
Forecast model kleinmodel now contains 6 endogenous variables.

. forecast identity w = wg + wp
Forecast model kleinmodel now contains 7 endogenous variables.

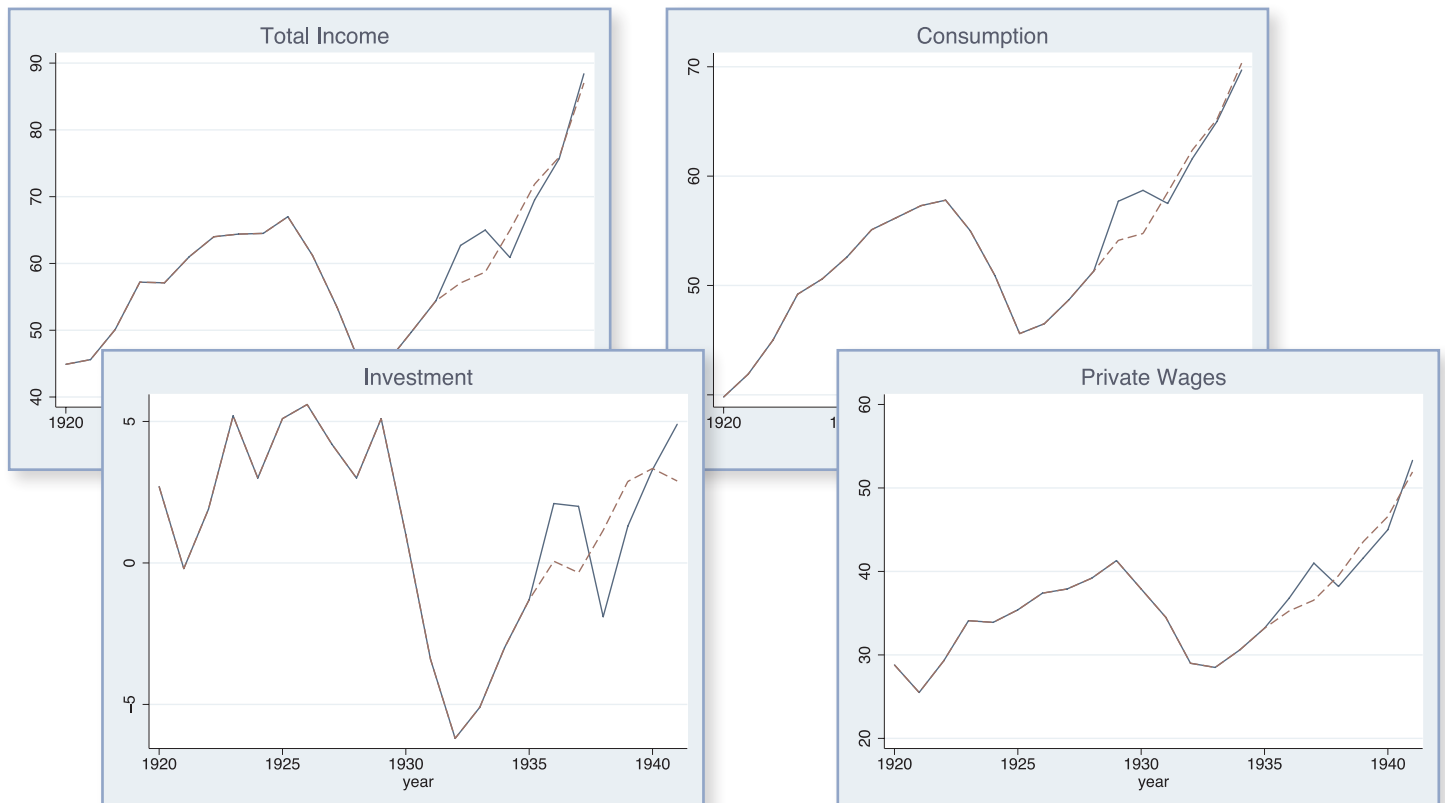
. forecast exogenous wg g t yr
Forecast model kleinmodel now contains 4 declared exogenous variables.
```

Even better, we could have entered the previous model into **forecast**'s Control Panel:

The image shows two overlapping Stata windows. The background window is titled "forecast - Econometric model forecasting" and displays the "Estimate model" control panel. On the left is a vertical sidebar with buttons for "Create", "Estimate Model", "Build", "Solve", and "Utilities". The main area is divided into "Step 1" and "Step 2". Step 1, "Choose an estimation command and press 'Go':", lists several options, with "Three-stage estimation for systems of simultaneous equations" selected and highlighted in blue. A "Go ->" button is to the right. Step 2, "Save estimation results to memory:", has radio buttons for "Save estimation results to memory:" (selected) and "Save estimation results to disk:". Below this is a text field for "Result set name" and a checkbox for "Append results to existing file". At the bottom, it says "Model = No model in memory".

The foreground window is titled "reg3 - Three-stage estimation for systems of simultaneous equations". It has tabs for "Model", "Model 2", "Est. method", "by/if/in", "Weights", "df adj.", "Reporting", and "Optimization". The "Model" tab is active. It shows "Multiple equations: (equation 1 is required)" with a list of "Equation 1" through "Equation 8". To the right, "Dependent variables for equation 1:" is set to "c" and "Independent variables for equation 1:" is set to "p L p w". There is a checkbox for "Suppress equation 1 constant term". Under "Options", "Label for equation 1:" is empty, and the "Three-stage estimator" radio button is selected. A "Constraints:" field is empty with a "Manage..." button. At the bottom are "OK", "Cancel", and "Submit" buttons.

We could graph the results:



See the video overview at stata.com/videos13/forecast.

Learn everything about **forecast** by reading the manual at stata.com/manuals13/forecast.

Generalized SEM

Generalized linear responses

- Continuous—linear, log-gamma
- Binary—probit, logit, complementary log-log
- Count—Poisson, negative binomial
- Categorical—multinomial logit
- Ordered—ordered logit, ordered probit

Multilevel/hierarchical data

- Nested: 2 levels, 3 levels, more levels
- Crossed
- Latent variables at different levels
- Random intercepts
- Random slopes (paths)
- Mixed models

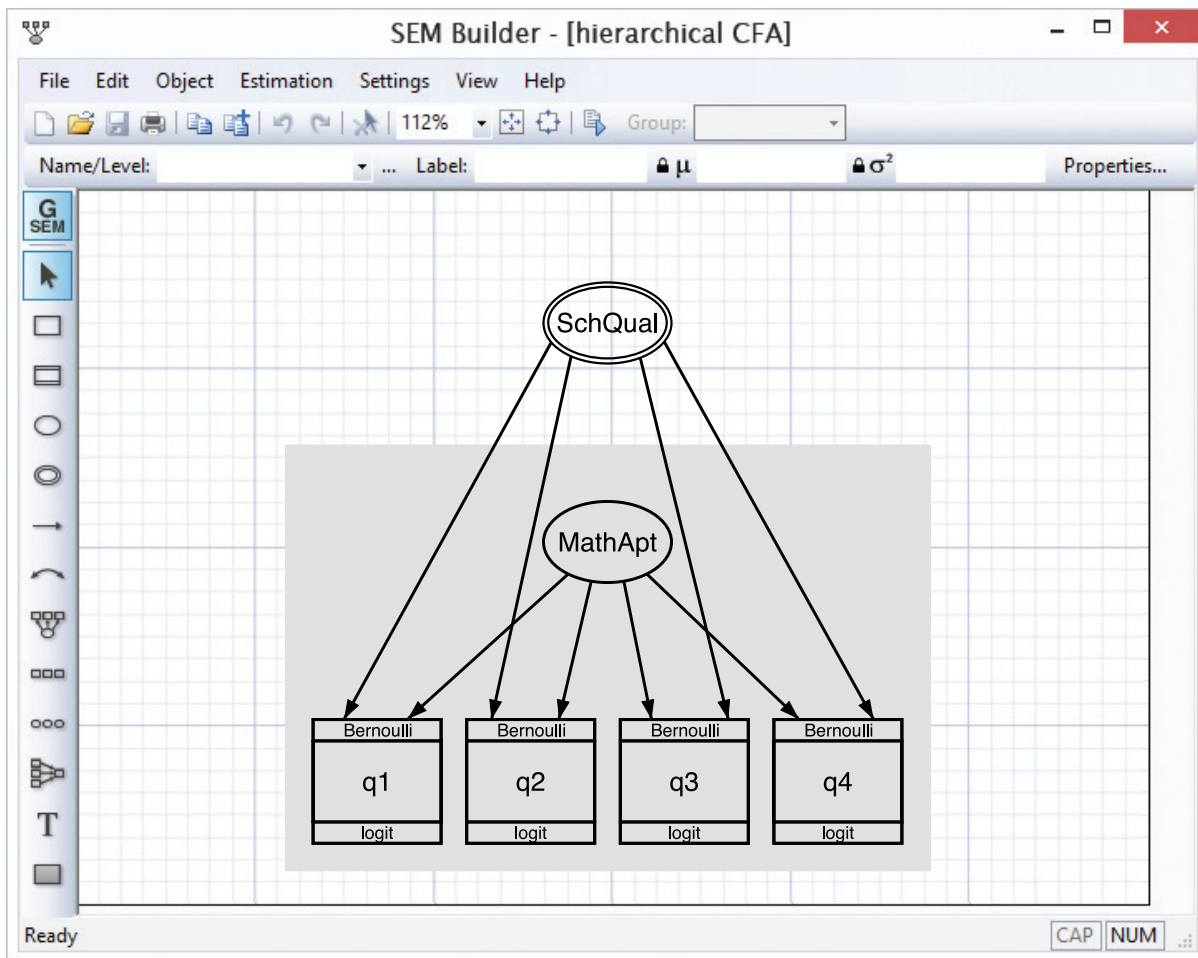
Path Builder supports GLM and multilevel models

New models


- CFA with binary, count, or ordinal measurements
- Multilevel CFA
- Multilevel mediation
- Item-response theory (IRT)
- Latent growth curves with repeated measurements of binary, count, or ordinal responses
- Selection models
- Endogenous treatment effects
- Any multilevel SEM with generalized linear responses

Say we have a test designed to assess mathematical performance. The data record a set of binary variables measuring whether individual answers were correct. The test was administered to students at various schools.

We postulate that performance on the questions is determined by unobserved (latent) mathematical aptitude and by school quality representing latent characteristics of the school:



In the diagram, the values of the latent variable **SchQual** are constant within school and vary across schools.

We can fit the model from the path diagram by pressing .

Results will appear on the diagram.

Or we can skip the diagram and type the equivalent command,

```
. gsem (      MathApt  -> q1 q2 q3 q4)
      (SchQual[school] -> q1 q2 q3 q4), logit
```

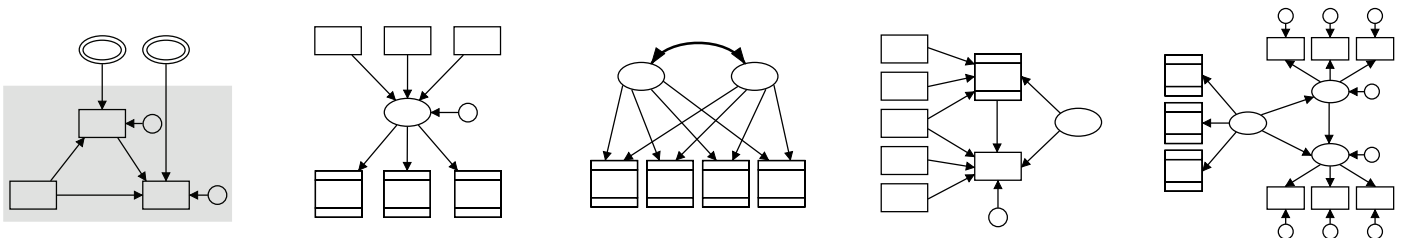
Either way, we get the same results:

Generalized structural equation model		Number of obs		=		2500	
Log likelihood = -6506.1646							
(1) [q1]SchQual[school] = 1							
(2) [q2]MathApt = 1							
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
q1 <-	SchQual[school]	1 (constrained)					
	MathApt	1.437913	.1824425	7.88	0.000	1.080333	1.795494
	_cons	.0459474	.1074647	0.43	0.669	-.1646795	.2565743
q2 <-	SchQual[school]	.1522361	.0823577	1.85	0.065	-.0091821	.3136543
	MathApt	1 (constrained)					
	_cons	-.377969	.0518194	-7.29	0.000	-.4795332	-.2764047
q3 <-	SchQual[school]	.5194866	.0965557	5.38	0.000	.3302408	.7087324
	MathApt	.8650544	.1098663	7.87	0.000	.6497204	1.080388
	_cons	.026989	.0667393	0.40	0.686	-.1038175	.1577955
q4 <-	SchQual[school]	.6085149	.119537	5.09	0.000	.3742266	.8428032
	MathApt	1.721957	.2466729	6.98	0.000	1.238487	2.205427
	_cons	-.3225736	.0845656	-3.81	0.000	-.4883191	-.1568281
var(SchQual[school])		.4167718	.1222884			.2344987	.7407238
var(MathApt)		1.004914	.1764607			.7122945	1.417744

Math aptitude has a larger variance and loadings than school quality. Thus, math aptitude is more important than school, although school is still important.

See the video overview at [stata.com/videos13/gsem](https://www.stata.com/videos13/gsem).

See the introduction to generalized SEM at [stata.com/manuals13/semintro1](https://www.stata.com/manuals13/semintro1), or see the entire SEM manual at [stata.com/manuals13/sem](https://www.stata.com/manuals13/sem). In the manual we added 18 new worked examples illustrating the models you can fit with the new features.



Effect sizes

Comparison of means

- Cohen's d
- Hedges's g
- Glass's Δ
 - › Point/biserial correlation
 - › Estimated from data or published summary statistics

Variance explained by regression and ANOVA

- Eta-squared and partial eta-squared (η^2)
- Omega-squared and partial omega-squared (ω^2)
- Estimated from data
- Overall statistics from data or published summary statistics

All with confidence intervals

Effect sizes provide results the way many researchers want to see them and the way many journals require.

Cohen's d , Hedges's g , and Glass's Δ help to judge practical as opposed to statistical significance.

Eta-squared and omega-squared measure variance explained in ANOVA and regression models, both overall and term-by-term.

See the video overview at stata.com/videos13/effect-sizes.

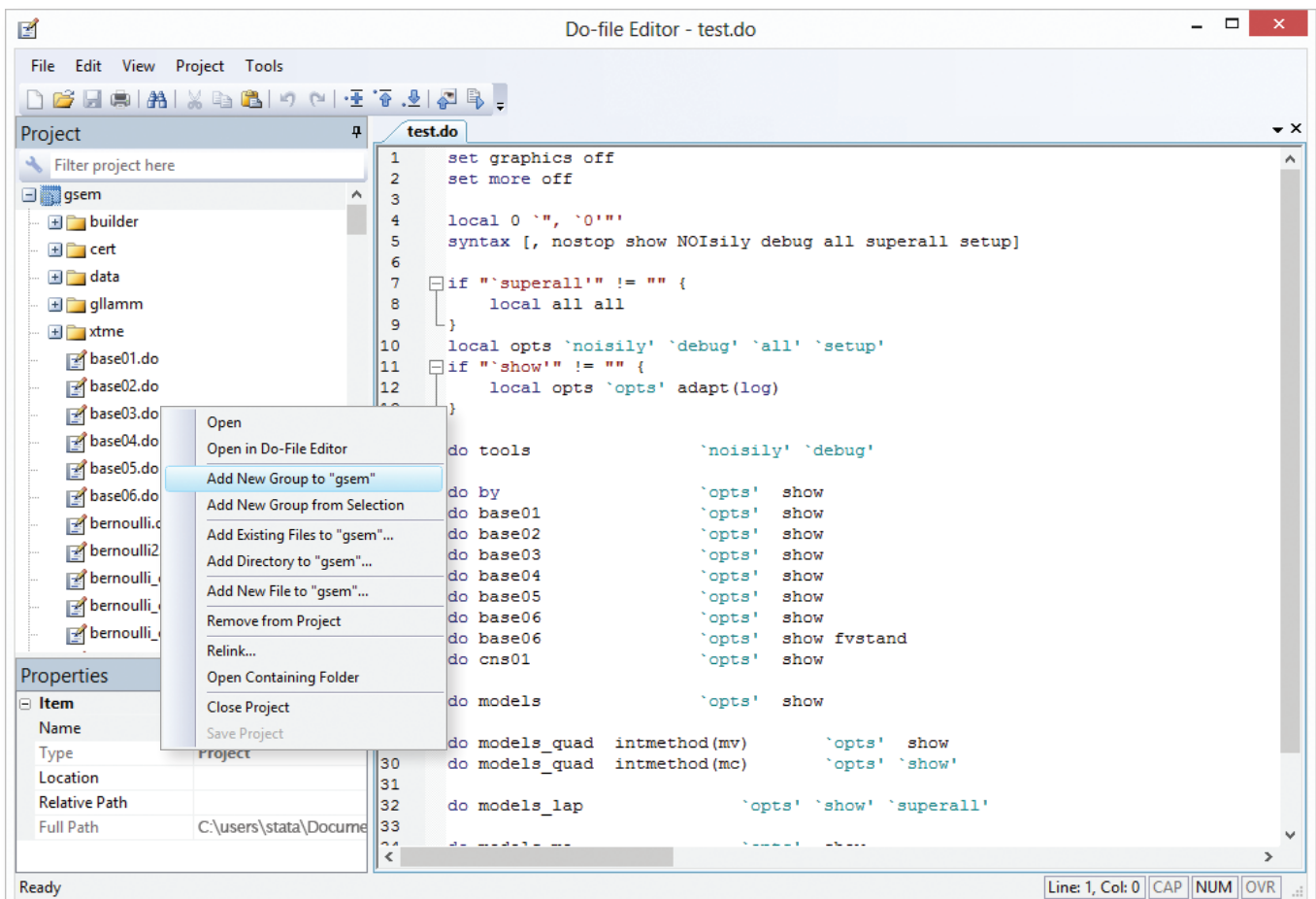
Learn everything about effect sizes in Stata in the manual entry at stata.com/manuals13/esize.

Project Manager

- Organize do-files, ado-files, datasets, raw files, etc.
- Manage hundreds, even thousands, of files per project
- Manage multiple projects
- Create groups in the project to categorize files
- Portable projects
- Relative or absolute paths

See all the files at a glance. Filter on filenames. Click to open, edit, or run.

See the video at stata.com/videos13/project-manager, but you've got to try it.



Panel data—new estimators and extensions

New estimators

- Random-effects ordered probit
- Random-effects ordered logit

Cluster-robust standard errors

- Relax distributional assumptions
- Allow for correlated data
- Available on new estimators
- Also available on probit, logit, complementary log-log, and Poisson

It is difficult to say panel data without saying random effects. Panel data is repeated observations on individuals. Random effects are individual-level effects that are unrelated to everything else in the model.

Say we have data on 4,708 employees of a large multinational corporation. We have repeated observations on these employees over the years. On average, we have six years of data. For some employees, we have fifteen years.

Our data include professional status (1, 2, 3, or 4), age, education, and years of job experience.

We fit this model (shown to the right).

We find that professional status increases with education and experience. We also find that individuals have a large permanent component (σ^2_u , the variance of the random effect, is both large and significant).

See the video overview at stata.com/videos13/panel-data.

Learn more about the new estimators and extensions in the manual entries at stata.com/manuals13/xtoprobit and stata.com/manuals13/xtologit.

```

xtoprobit status educ age age2 experience
Random-effects ordered probit regression      Number of obs   =   28508
Group variable: idcode                       Number of groups =   4708

Random effects u_i ~ Gaussian                Obs per group:  min =    1
                                                avg =    6.1
                                                max =   15

Integration method: mvaghermite              Integration points =   12

Log likelihood = -26032.326                   Wald chi2(4)     =   5676.07
                                                Prob > chi2      =    0.0000
  
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
educ	.333993	.0091215	36.62	0.000	.3161152 .3518707
age	.2129696	.0123513	17.24	0.000	.1887616 .2371777
age2	-.004221	.0002029	-20.80	0.000	-.0046188 -.0038232
experience	.1918314	.004661	41.16	0.000	.1826959 .2009669
/cut1	7.625209	.2156879	35.35	0.000	7.202468 8.047949
/cut2	8.498753	.217015	39.16	0.000	8.073411 8.924094
/cut3	10.07729	.219697	45.87	0.000	9.646696 10.50789
/sigma2_u	1.498732	.0506867			1.402609 1.601443

LR test vs. oprobit regression: $\chi^2(01) = 9467.03$ Prob>= $\chi^2 = 0.0000$

And lots more

We should mention the improved help-file searching, and Poisson with endogenous covariates, and that Stata now supports FTP and secure HTTP, and fast PDF manual navigation, and seven noncentral- t and noncentral- F functions, and Java API, and ordered probit with Heckman-style sample selection, and the new way of estimating ML models without writing an evaluator program, and the new fractional-polynomial prefix command, and intraclass correlations, and that quantile regression can now produce robust estimates of standard errors, and that factor variables now support value labels for labeling output, and the new way to import data from Haver Analytics, and automatic business-calendar creation, and the new import commands that make reading data really easy, and how you can create Word and Excel files from Stata, and how you can solve arbitrary nonlinear systems, and a lot of other things.

And Stata 13 contains the most requested feature of all: You can now type **cls** to clear the Results Window.

Learn more about everything at stata.com/stata13.

Stat/Transfer Version 12

- Supports new Stata 13 datasets
- Fully supports Stata long strings and BLOBs

Stat/Transfer 12 also adds support for Excel 2013, gretl, JMP Version 10, JMP Compressed Files, SPSS 21 gsav Compressed Files, SAS 64 Bit Catalogs, and SYSTAT 13.

It also adds cross-platform licensing. One activation code can be used on Windows, Mac, or Linux.

Find out more at stata.com/products/stat-transfer.



stata.com/giftshop



StataCorp
 4905 Lakeway Drive
 College Station, TX 77845-4512
 USA

Return service requested.

Contact us

979-696-4600
service@stata.com

979-696-4601 (fax)
stata.com

Please include your Stata serial number with all correspondence.

Find a Stata distributor near you
stata.com/worldwide



 Copyright 2013 by StataCorp LP. Stata is a registered trademark of StataCorp LP.

STATA[®] CONFERENCE

Join us in the Big Easy and learn about Stata 13!

Mix and mingle with Stata users and developers. After a day of learning and networking, join us Thursday evening for a relaxing, authentic creole dinner at the renowned Tujague's. While in town, why not take in a jazz band, tour the French Quarter, or explore the world-famous Bourbon Street?

Make New Orleans your destination!

The conference includes, in addition to user contributions, presentations by StataCorp developers on new Stata 13 features. Meet with the developers who wrote Stata 13.

Don't miss this opportunity to exchange ideas with Stata users and developers.

See the program and register online at
stata.com/new-orleans13.

Dates: July 18–19, 2013
 Venue: Hyatt French Quarter New Orleans
 800 Iberville Street
 New Orleans, Louisiana
 Cost: \$195 regular; \$75 student
 \$45 dinner at Tujague's (optional)



Stata 13 ships June 24. Order now at stata.com