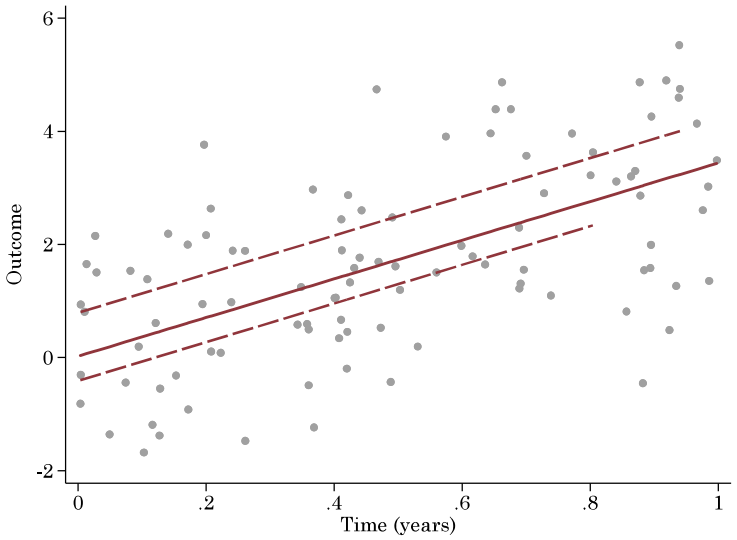
Stata 20 will provide correct inference on random effects

Matteo Bottai, Sc.D.
Karolinska Institutet

References for Details on the Theory

- Bottai. Confidence regions when the Fisher information is zero. Biometrika, 90(1): 73-84, 2003
- Bottai, Orsini. Confidence intervals for the variance component of random-effects linear models. The Stata Journal, 4(4): 429-435, 2004
- Ekvall, Bottai. Confidence regions near singular information and boundary points with applications to mixed models. Annals of Statistics, 50(3): 1806-1832, 2022
- Ekvall, Bottai. Uniform inference in linear mixed models. Under review

A Random Intercept Model



The National Longitudinal Survey

```
. webuse nlswork
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. qui xtreg ln_w grade age, mle i(id)
. ereturn display
```

ln_wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_wage						
grade	.0810052	.0019417	41.72	0.000	.0771996	.0848109
age	.0171891	.0003295	52.17	0.000	.0165433	.0178349
_cons	.1324334	.025892	5.11	0.000	.081686	.1831808
sigma_u						
_cons	.2961065	.0039158	75.62	0.000	.2884318	.3037813
sigma_e						
_cons	.3043363	.001397	217.86	0.000	.3015983	.3070743

The confidence interval for σ_u is incorrect.

The Problem

Coverage of "xtreg" and "xtvc' in 1,000 samples simulated with

$$y_{i,j} = u_i + e_{ij}$$
$$i = 1, \ldots, m \quad j = 1, 2, 3, 4$$
$$u \sim N(0, \sigma_u) \perp e \sim N(0, 1)$$

		Number of Subjects (m)		
		10	50	100
$\sigma_u = 0.01$	xtreg	0.081	0.139	0.139
	xtvc	0.905	0.931	0.929
$\sigma_u = 0.10$	xtreg	0.260	0.362	0.449
	xtvc	0.908	0.925	0.945
$\sigma_u = 1.00$	xtreg	0.978	0.957	0.956
	xtvc	0.952	0.959	0.950

Monte Carlo 95% confidence interval = ± 0.014

Some Details: Mixed Effects Models

The standard linear mixed model is

$$Y = X\beta + ZU + E$$

where $U \sim N(0, \Psi)$ and $E \sim N(0, \sigma_e I)$ are independent.

Some Details: Confidence Regions

We consider the confidence region

$$\mathbb{C}(\alpha) = \{\theta \in \mathbb{P} : T(\theta) \leq q_{1-\alpha}(\theta)\}$$

for a test statistic $T(\theta)$, parameter set $\mathbb{P}$, and cutoff $q_{1-\alpha}(\theta)$.

The region has uniform coverage if

$$\lim_{n \to \infty} \sup_{\theta \in \mathbb{P}} P[\theta \in \mathbb{C}(\alpha)] = 1 - \alpha$$

Some Details: Likelihood

We define the log-likelihood

$$l(\theta, y) = \sum_{i=1}^{n} \log f(y_i, \theta)$$

with derivatives $l'(\theta, y)$ and $l''(\theta, y)$, and maximizer

$$\widehat{\theta} = \arg\max_{\theta \in \mathbb{P}} l(\theta, y)$$

Some Details: Test Statistics

We consider three test statistics

$$
\begin{aligned}
T_{\text{score}}(\theta) &= \begin{cases} nl'(\theta, y)^2 i(\theta)^{-1} & \text{if } \theta \neq 0 \\ nl''(\theta, y)^2 i(0)^{-1} & \text{if } \theta = 0 \end{cases} \\
T_{\text{LR}}(\theta) &= 2[l(\widehat{\theta}, y) - l(\theta, y)] \\
T_{\text{Wald}}(\theta) &= -(\widehat{\theta} - \theta)^2 l''(\widehat{\theta}, y)
\end{aligned}
$$

where $i(\theta) = E_\theta[l''(\theta, y)]$ and $i(0) = E_0[l''(0, y)^2]$.

Only the score statistic $T_{\text{score}}(\theta)$ has uniform coverage.

Conclusions

Things get worse with multiple random effects.

- Variances and correlations may be on the boundary.
- The "mixed" confidence intervals are hopeless.

Luckily

- ✓ The theory is now available.
- ✓ Stata developers are eager to implement it in Stata 20.