Small-sample inference for linear mixed-effects models (DDF adjustments)

Xiao Yang

Senior Statistician and Software Developer
StataCorp LP

2015 Stata Conference

Outline

- Motivation
- Currently supported methods
  - "Exact" methods
  - Approximate methods
  - Which one to use?
- Postestimation
  - Currently available commands
  - Small-sample adjustments for contrasts

The `mixed` command fits linear mixed-effects models. Mixed effects are fixed effects plus random effects. For example,

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + u_j + \epsilon_{ij},$$

where $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, s$.

In matrix notation, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$.

- β_0, β_1, ..., β_p are fixed effects
- u_j's are random effects
- $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma_e^2)$
- $\mathbf{X}$ and $\mathbf{Z}$ are design matrices

Researchers are often interested in making inferences about fixed effects.

- Large-sample approximation
  - sampling distributions of the test statistics are approximated by normal and χ^2
  - default in `mixed`
- In special cases, sampling distributions of the test statistics are known to be t or F distributions.
  - simple balanced split-plot design
  - simple balanced repeated measures
- In small samples, large-sample approximations may lead to anticonservative results.

Introducing `mixed, dfmethod()` ...

- In small samples, the null sampling distributions of test statistics for fixed effects are not known in general (except for special cases).
- Sampling distributions are approximated by t and F.
- Approximations differ in how respective denominator degrees of freedom (DDF) are computed.
- Five methods for calculating DDFs.
- New in Stata 14, need to specify the `dfmethod()` option.

Choosing between DDF methods is not an easy task!

- All DDF methods are only approximations (except in some rare cases).
- The choice of DDFs is highly dependent on the models, the data structure, the size of the dataset, and the balance of the dataset.
- No single method works for all possible models.

Example 1: Simple repeated-measures design

- From table 4.3 of Winer, Brown, and Michels[1].
- The reaction time for 5 subjects each tested with 4 drugs was recorded in the variable `score`. `drug` is the repeated-measures factor.

```
. tabdisp person drug, cellvar(score)
```

person	drug 1	drug 2	drug 3	drug 4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

[1] B. J. Winer, D. R. Brown, and K. M. Michels. *Statistical Principles in Experimental Design*. 3rd ed. New York, NY: McGraw-Hill, Inc., 1991.

- Use anova command:

```
. anova score person drug, repeated(drug)
...
Between-subjects error term:  person
                     Levels:  5          (4 df)
     Lowest b.s.e. variable:  person

Repeated variable: drug
                                          Huynh-Feldt epsilon        =  1.0789
                                          *Huynh-Feldt epsilon reset to 1.0000
                                          Greenhouse-Geisser epsilon =  0.6049
                                          Box´s conservative epsilon =  0.3333
```

Source	df	F	Prob > F Regular	H-F	G-G	Box
drug	3	24.76	0.0000	0.0000	0.0006	0.0076
Residual	12					

- Large-sample inference

```
. mixed score i.drug || person:, reml
...
Mixed-effects REML regression                   Number of obs      =        20
Group variable: person                          Number of groups   =         5
                                                Obs per group:
                                                              min =         4
                                                              avg =       4.0
                                                              max =         4
                                                Wald chi2(3)       =     74.28
Log restricted-likelihood = -49.640099          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |
          2  |        -.8   1.939072    -0.41   0.680    -4.600511    3.000511
          3  |      -10.8   1.939072    -5.57   0.000    -14.60051   -6.999489
          4  |        5.6   1.939072     2.89   0.004     1.799489    9.400511
             |
       _cons |       26.4   3.149604     8.38   0.000     20.22689    32.57311
------------------------------------------------------------------------------
...
```

- Small-sample inference

```
. mixed score i.drug || person:, reml  dfmethod(repeated)
...
Mixed-effects REML regression                   Number of obs     =         20
Group variable: person                          Number of groups  =          5
                                                Obs per group:
                                                              min =          4
                                                              avg =        4.0
                                                              max =          4
DF method: Repeated                             DF:           min =       4.00
                                                              avg =      10.00
                                                              max =      12.00
                                                F(3,     12.00)   =      24.76
Log restricted-likelihood = -49.640099          Prob > F          =     0.0000

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |
          2  |        -.8   1.939072    -0.41   0.687    -5.024874    3.424874
          3  |      -10.8   1.939072    -5.57   0.000    -15.02487   -6.575126
          4  |        5.6   1.939072     2.89   0.014     1.375126    9.824874
             |
       _cons |       26.4   3.149604     8.38   0.001      17.6553     35.1447
------------------------------------------------------------------------------
...
```

- To display the DF value for each coefficient, just type

```
. mixed, dftable(pvalue)
```

score	Coef.	Std. Err.	DF	t	P>|t|
drug					
2	-.8	1.939072	12.0	-0.41	0.687
3	-10.8	1.939072	12.0	-5.57	0.000
4	5.6	1.939072	12.0	2.89	0.014
_cons	26.4	3.149604	4.0	8.38	0.001

```
. estat df
Degrees of freedom
```

	Repeated
score	
drug	
2	12
3	12
4	12
_cons	4

Example 2: Random-coefficient model for longitudinal data

- Simulated dataset from Kenward and Roger[2].
- 24 subjects, identified by **id**, split into 3 groups of 8. The subjects of each group are being observed on the same time points. The three sets of time points are chosen to be nonoverlapping: $(0, 1, 2)$, $(3, 4, 5)$, and $(6, 7, 8)$.

$$y_{ij} = \beta_0 + \beta_1 \, \text{time}_{ij} + u_j + \gamma_j \, \text{time}_{ij} + \epsilon_{ij}$$

- $\begin{bmatrix} u_j \\ \gamma_j \end{bmatrix} \sim N\left(\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$ and $\epsilon_{ij} \sim N(0, \sigma_e^2)$.
- Data are simulated from the model with $\beta_1 = 0$.

[2] M. G. Kenward and J. H. Roger. "Small sample inference for fixed effects from restricted maximum likelihood". In: *Biometrics* 53 (1997), pp. 983–997.

- Large-sample inference

```
. mixed y time || id: time, reml cov(unstructured)
...
Mixed-effects REML regression                   Number of obs      =        72
Group variable: id                              Number of groups   =        24

                                                Obs per group:
                                                              min =         3
                                                              avg =       3.0
                                                              max =         3

                                                Wald chi2(1)       =      4.34
Log restricted-likelihood = -109.39153          Prob > chi2        =    0.0372

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .2765987   .1327319     2.08   0.037     .0164489    .5367485
       _cons |   1.045034   .2504823     4.17   0.000     .5540973     1.53597
------------------------------------------------------------------------------
...
```

- The default large-sample inference for **time** suggests that the fixed time effect is significant at a 5% level (p-value $= 0.037$).

- Small-sample inference with the kroger method

```
. mixed y time || id: time, reml cov(unstructured)  dfmethod(kroger)
...
Mixed-effects REML regression                   Number of obs     =         72
Group variable: id                              Number of groups  =         24

                                                Obs per group:
                                                              min =          3
                                                              avg =        3.0
                                                              max =          3
DF method: Kenward-Roger                        DF:           min =      11.68
                                                              avg =      17.19
                                                              max =      22.69

                                                F(1,     22.69)   =       4.24
Log restricted-likelihood = -109.39153          Prob > F          =     0.0512

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .2765987     .13434     2.06   0.051    -.0015158    .5547132
       _cons |   1.045034   .2700712     3.87   0.002     .4548251    1.635242
------------------------------------------------------------------------------
...
```

- After adjusting for a small sample, we do not have sufficient evidence to reject the null hypothesis of no time effect at a 5% significance level.

`mixed, dfmethod()` provides five DDF methods.

Method	ML	REML
`residual`	YES	YES
`repeated`	YES	YES
`anova`	YES	YES
`satterthwaite`	NO	`eim`, `oim`
`kroger`	NO	`eim`, `oim`

`residual`

For "exact" methods, computing DF for each coefficient is based on the single hypothesis test $H_o : \beta_i = 0$, for $i = 1, 2, \ldots, p$.

- $v_{df} = n - rank(X)$ for all tests.
- `residual` **provides exact degrees of freedom only in the 'iid' case**.
- **For other mixed models, provides poor approximation**.
- Available for completeness.

repeated

- Partitions the residual degrees of freedom into the between-subject degrees of freedom and the within-subject degrees of freedom.
- Gives exact DF values for special balanced repeated-measures models with the spherical covariance structure.
- **Supported only with two-level models.**
- **Leads to poor approximations for more complex mixed-effects models or with unbalanced data.**

anova

- Checks if the fixed effect is contained in some random-effects equations.
- If contained in some random-effects equations, then DF equals the smallest number of levels among the level variables minus one.
- If not contained in any random-effects equation, then

$$v_{df} = n - rank(X, Z)$$

- **Gives an exact sampling distribution of the test statistics only when random effects are simple and balanced and the error terms are i.i.d.**
- **Leads to poor approximations for more complex mixed-effects models or with unbalanced data.**

Conclusion for "exact" methods

- `residual`, `repeated`, `anova`.
- Available for both `ML` and `REML`.
- Based on single-hypothesis tests.
- Available for multiple-hypotheses tests only if all corresponding single-hypothesis DFs are the same, $\nu_{ddf} = \nu_{df}$.
- If all corresponding single-hypothesis DFs are different, ν_{ddf} is not defined.

satterthwaite

- For a single-hypothesis test, Giesbrecht and Burns[3] developed a method of computing the DDF that is analogous to Satterthwaite's approximation of the degrees of freedom.
- For a multiple-hypotheses test, Fai and Cornelius[4] decomposed the contrast matrix using the spectral decomposition and repeatedly applied Giesbrecht and Burns's method to get the single-degree-of-freedom t test, then used the relationship between t and F to get the DDF.

[3] F. G. Giesbrecht and J. C. Burns. "Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results". In: *Biometrics* 41 (1985), pp. 477–486.

[4] A. H. Fai and P. L. Cornelius. "Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments". In: *Journal of Statistical Computation and Simulation* 54 (1996), pp. 363–378.

satterthwaite

- **Fai and Cornelius[5] prove that** `satterthwaite` **is good at approximating unbalanced split-plot designs.**
- **Schaalje, McBride, and Fellingham[6] recommend using the** `satterthwaite` **method only when the covariance structure of the data is compound symmetry and the sample size is moderately large.**

---

[5] A. H. Fai and P. L. Cornelius. "Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments". In: *Journal of Statistical Computation and Simulation* 54 (1996), pp. 363–378.

[6] G. B. Schaalje, J. B. McBride, and G. W. Fellingham. "Adequacy of approximations to distributions of test statistics in complex mixed linear models". In: *Journal of Agricultural, Biological, and Environmental Statistics* 7 (2002), pp. 512–524.

kroger

- Kenward and Roger[7] proposed the scaled F-test statistic,

$$F_{KR} = \frac{\lambda}{\ell}(\mathbf{C}'\widehat{\beta} - \mathbf{b})'(\mathbf{C}'\widehat{\mathbf{\Phi}}_A\mathbf{C})^{-1}(\mathbf{C}'\widehat{\beta} - \mathbf{b}) \sim F_{\ell,ddf_{kr}}$$

- **Accounts for the small-sample bias and the variability of the estimated random effects to obtain an adjusted estimator of the fixed-effects covariance matrix $\widehat{\mathbf{\Phi}}_A$.**
- Uses a Taylor expansion for $(\mathbf{C}'\widehat{\mathbf{\Phi}}_A\mathbf{C})^{-1}$ and matches moments of F_{KR} with those of the approximating F distribution to obtain ddf_{kr} and λ.

[7] M. G. Kenward and J. H. Roger. "Small sample inference for fixed effects from restricted maximum likelihood". In: *Biometrics* 53 (1997), pp. 983–997.

`kroger`

- `kroger` **yeilds to exact F distribution when the exact F distribution is available, and improves the approximation when the exact F distribution is not available.**
- Computing $\widehat{\mathbf{\Phi}}_A$ involves taking first and second derivatives of the covariance matrix of $\mathbf{y}$ w.r.t. each random component.
- $\widehat{\mathbf{\Phi}}_A$ is invariant under reparameterization if the covariance matrix of $\mathbf{y}$ can be written as a linear function of random components.
- The second derivatives require more computational resources and may not be numerically stable; therefore, they are ignored.

Conclusion for Approximate Methods

- `satterthwaite` and `kroger` are only available under `REML`.
- You can choose to use either `oim` or `eim` in the computation of `satterthwaite` or `kroger`; `eim` is the default.
- For a single-hypothesis test, DFs are the same between `satterthwaite` and `kroger`, but tests statistics and therefore tests are not necessarily identical.
- Suitable for complex covariance structures and unbalanced data.
- Computationally intensive.

Which is the best dfmethod()?

- **[Spilke et al]** [7]: Assessed the performance of five DDF methods on RCB, split-plot, strip plots with missing data under REML. Prefered `kroger`.
- **[Alnosaier]** [1]: Assessed the performance of `satterthwaite` and `kroger` for PBIB, BIBD, and RCB with missing data designs through simulation. Prefered `kroger` method.
- **[Schaalje et al]** [6]: Assessed the performance of `kroger` and `satterthwaite` for split-plot and repeated measures designs; both methods are affected by covariance structure complexity, sample size, and imbalance. Prefered `kroger` method.
- **[Gregory]** [4]: Compared four DDF methods in the unbalanced two-way factorial design. Found no significant differences between those methods.

Which one is the best `dfmethod()`?

- Prefer the `kroger` method when the sample size is small, the covariance structure is complicated, and/or the data is unbalanced.
- Even the `kroger` method sometimes produces inflated Type I error rates (e.g., AR(1) error covariance structure).
- The approximation methods can be computationally intensive.
- More research needs to be done to determine which method is the best for different mixed-effects models.

Postestimation

Stata provides additional postestmation commands and options for small-sample inference after `mixed`:

- `estat df`
- `test, small`
- `testparm, small`
- `lincom, small`
- `contrast, small` (forthcoming)

Example 3: Unbalanced split-plot design

- There are 30 observations, 8 subjects, whole plot factor *a* (2 levels), sub-plot factor *b* (4 levels), unbalanced

```
. tabdisp b s, cellvar(y) by(a) concise
```

a and b	s 1	2	3	4	5	6	7	8
1								
1	3	6	3	3				
2	4	5	4	3				
3	7		7	6				
4	7	8	9	8				
2								
1					1	2	2	2
2					2	3	4	3
3					5	6	5	6
4					10		9	11

estat df

estat df is a convenient tool to calculate and compare the DF's for different methods.

- Fit the model based on large-sample inference

```
. mixed y a##b || s:, reml
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
2.a	-2	.6288677	-3.18	0.001	-3.232558	-.767442
b						
2	.25	.5359916	0.47	0.641	-.8005243	1.300524
3	3.108222	.5862035	5.30	0.000	1.959284	4.25716
4	4.25	.5359916	7.93	0.000	3.199476	5.300524
a#b						
2 2	1	.7580066	1.32	0.187	-.4856656	2.485666
2 3	.6417778	.7943057	0.81	0.419	-.9150328	2.198588
2 4	4.044205	.7943057	5.09	0.000	2.487395	5.601016
_cons	3.75	.4446766	8.43	0.000	2.87845	4.62155

estat df

- Compare different DF methods using the method() option

```
. estat df, method(residual repeated anova satterthwaite kroger)
Degrees of freedom
```

	Residual	Repeated	ANOVA	Satterthwaite	Kenward-Roger
y					
a					
1	(empty)				
2	22	6	16	18.29179	18.29179
b					
1	(empty)				
2	22	16	16	16.01983	16.01983
3	22	16	16	16.66069	16.66069
4	22	16	16	16.01983	16.01983
a#b					
1 1	(empty)				
...	(empty)				
2 1	(empty)				
2 2	22	16	16	16.01983	16.01983
2 3	22	16	16	16.36871	16.36871
2 4	22	16	16	16.36871	16.36871

estat df

- Post the desired DF (`kroger` in our example) using the `post` option.

```
. estat df, method(kroger) post
```

- It is the same as refitting the model using the `dfmethod()` option in `mixed`.

```
. mixed y a##b || s:, reml dfmethod(kroger)
```

test, small

- Obtain the large-sample inference as usual.

```
. test 2.a
 ( 1)  [y]2.a = 0
           chi2(  1) =   10.11
         Prob > chi2 =    0.0015
```

- Use the small option to get small-sample adjustment.

```
. test 2.a, small
 ( 1)  [y]2.a = 0
       F(  1, 18.29) =   10.11
            Prob > F =    0.0051
```

testparm, small

- testparm also provides both tests, with and without small-sample adjustment.

```
. testparm a#b
 ( 1)  [y]2.a#2.b = 0
 ( 2)  [y]2.a#3.b = 0
 ( 3)  [y]2.a#4.b = 0
           chi2(  3) =   29.35
         Prob > chi2 =    0.0000
. testparm a#b, small
 ( 1)  [y]2.a#2.b = 0
 ( 2)  [y]2.a#3.b = 0
 ( 3)  [y]2.a#4.b = 0
       F(  3, 16.35) =    9.66
            Prob > F =    0.0007
```

lincom, small

- lincom also provides two sets of results.

```
. lincom 2.a + 2.a#4.b
 ( 1)  [y]2.a + [y]2.a#4.b = 0
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	2.044205	.6721771	3.04	0.002	.7267621	3.361648

```
. lincom 2.a + 2.a#4.b, small
 ( 1)  [y]2.a + [y]2.a#4.b = 0
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	2.044205	.6764554	3.02	0.007	.6311736	3.457237

contrast, small

- Suppose that we want to test the effect of factor *a*.
- The effect of factor *a* includes the main effect of *a* and the interaction effects that contain *a*.
- Currently, `contrast` only provides large-sample inference.

```
. contrast a

Contrasts of marginal linear predictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
y            |
           a |          1        1.79     0.1810
-------------------------------------------------
```

- **It is not the test for *2.a* from the coefficient table!**

contrast, small

We need to manually compute the small-sample inference.

- Write your own contrast

$$H_o : 2.a + \frac{1}{4} \times 2.a\#2.b + \frac{1}{4} \times 2.a\#3.b + \frac{1}{4} \times 2.a\#4.b = 0$$

- Use test, small

```
. test ([y]2.a + 0.25*[y]2.a#2.b+0.25*[y]2.a#3.b+0.25*[y]2.a#4.b = 0), small
 ( 1)  [y]2.a + .25*[y]2.a#2.b + .25*[y]2.a#3.b + .25*[y]2.a#4.b = 0
       F(  1,   5.95) =    1.78
            Prob > F =    0.2307
```

- contrast, small forthcoming

Thank you!

References

[1] W. S. Alnosaier. "Kenward-Roger Approximate F Test for Fixed Effects in Mixed linear models". PhD thesis. Oregon State University, 2007.

[2] A. H. Fai and P. L. Cornelius. "Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments". In: *Journal of Statistical Computation and Simulation* 54 (1996), pp. 363–378.

[3] F. G. Giesbrecht and J. C. Burns. "Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results". In: *Biometrics* 41 (1985), pp. 477–486.

References (cont.)

[4] K. B. Gregory. "A Comparison of Denominator Degrees of Freedom Approximation Methods in the Unbalanced Two-way Factorial Mixed Model". MA thesis. Texas A & M University, 2011.

[5] M. G. Kenward and J. H. Roger. "Small sample inference for fixed effects from restricted maximum likelihood". In: *Biometrics* 53 (1997), pp. 983–997.

[6] G. B. Schaalje, J. B. McBride, and G. W. Fellingham. "Adequacy of approximations to distributions of test statistics in complex mixed linear models". In: *Journal of Agricultural, Biological, and Environmental Statistics* 7 (2002), pp. 512–524.

References (cont.)

[7] J. Spilke, H. Piepho, and X. Hu. "A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data". In: *Journal of Agricultural, Biological and Environmental Statistics* 10 (2005), pp. 374–389.

[8] B. J. Winer, D. R. Brown, and K. M. Michels. *Statistical Principles in Experimental Design*. 3rd ed. New York, NY: McGraw-Hill, Inc., 1991.