Likelihood Ratio Tests for Multiply Imputed Datasets:

Introducing milrtest

Rose Anne Medeiros
rosem@ats.ucla.edu

Statistical Consulting Group
Academic Technology Services
University of California, Los Angeles

2008 Fall North American Stata Users Group meeting

Introduction

- Analyzing multiply imputed (MI) datasets typically involves estimating the desired model on each of the *m* imputed datasets.
- The final coefficient estimates are based on the mean of the parameter estimates across the *m* imputed datasets.
- The final estimates of the standard errors incorporate both the standard errors from the individual analyses, and the variance of the standard errors across the *m* imputed datasets.

- Estimates of the s.e. allow for hypothesis tests for individual coefficients, however, testing nested models is somewhat more difficult.
- Several variants of the Wald test exist (see Schafer 1997, and Li, Raghunathan & Rubin 1991).
- The classic likelihood ratio (LR) test cannot be implemented as is because the final estimates do not come directly from a single model, and hence it is unclear what the proper value of the likelihood is for a given model.
- A variant of the LR test is described by Meng and Rubin (1992).

In Stata

- In Stata M.I. datasets can be analyzed using the user-written package `mim` (Carlin, Calati & Royston 2008).
- `mim` includes the multiparameter (Wald) test from Li, Raghunathan and Rubin (1991).
- The program presented here, `milrtest`, adds to the available tests by implementing the LR test of Meng and Rubin (1992).

Review and Notation

A likelihood ratio test compares a full model (h_1) with a restricted model where some parameters are constrained to some value(h_0), often zero. The log likelihoods for the two models are compared to asses fit.

The likelihood ratio test statistic:

$$d' = 2(\ell\ell_1 - \ell\ell_0)$$

Coefficient estimates based on the *m* MI datasets (Little & Rubin 2002):

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

Setup

1. For each of the m imputed datasets:
   - Run the h_1 model.
   - Run the h_0 model.
   - Calculate d' (LR test).
2. From the m repetitions of the h_0 model, calculate $\bar{\theta}_0$.
3. From the m repetitions of the h_1 model, calculate $\bar{\theta}_1$.

4. For each of the m imputed datasets:
   - Calculate the likelihood for h_1 with the parameters constrained to $\bar{\theta}_1$.
   - Calculate the likelihood for h_0 with the parameters constrained to $\bar{\theta}_0$.
   - Calculate the likelihood ratio test d_L, using the above likelihoods.
5. Calculate the mean of d', $\bar{d}'_m$ (i.e. the LR test statistics from the unconstrained models).
6. Calculate the mean of d_L, $\bar{d}_L$ (i.e. the LR test statistic from the constrained models).
7. Calculate the test statistic and degrees of freedom.

The Test Statistic

$$D_L = \frac{\bar{d}_L}{k(1 + r_L)}$$

where:

$$k = df_1 - df_0$$

and

$$r_L = \frac{(m+1)}{k(m-1)}(\bar{d}'_M - \bar{d}_L)$$

combine D_L and r_L:

$$D_L = \frac{\bar{d}_L}{k + \frac{m+1}{m-1}(\bar{d}'_M - \bar{d}_L)}$$

Degrees of freedom

$D_L \sim F(k, w(r_L))$, where:

$$w(r_L) = \begin{cases} 4 + (\nu - 4)\{1 + (1 - 2\nu^{-1})r_L^{-1}\}^2 & \nu > 4 \\ \frac{1}{2}\nu(1 + \frac{1}{k})(1 + r_L^{-1})^2 & \text{otherwise.} \end{cases}$$

where:

$$\nu = k(m-1)$$

and

$$r_L = \frac{m+1}{k(m-1)}(\bar{d}'_M - \bar{d}_L)$$

Syntax

`milrtest` *`test_varlist`*

- *`test_varlist`* should contain the variables to be restricted in the null model.
- Must be run after a `mim` regression command. The model run should be the alternative (i.e. unrestricted) model.
- Currently only available after `regress`, `logit`, and `ologit`.
- `milrtest` inherits sample restrictions from `mim`.
- $m \geq 4$ required.

An Example

- Uses a subset of data from a study of college students' romantic relationships (n=2386).
- The percent of missing values on each variable ranges from less than 1% to 9%, with most variables missing around 8% to 9% of values.
- The variables engaged, married, and cohabiting are dummy variables for relationship status, dating is the reference group.

The models:

h_1: `reg distress rc01 rc02 age engaged married cohabiting`

h_0: `reg distress rc01 rc02 age`

```
 mim: reg distress rc01 rc02 age engaged married cohabiting

Multiple-imputation estimates (regress)                  Imputations =       5
Linear regression                                        Minimum obs =    2385
                                                         Minimum dof =   108.8

------------------------------------------------------------------------------
    distress |      Coef.  Std. Err.      t    P>|t|    [95% Conf. Int.]   MI.df
-------------+----------------------------------------------------------------
        rc01 |  -1.38278   .139585    -9.91   0.000   -1.65679 -1.10878   781.4
        rc02 |  -1.16774    .13375    -8.73   0.000   -1.43086 -.904618   326.0
         age |   .065342   .019917     3.28   0.001    .026014  .104669   163.4
     engaged |  -.470156    .29352    -1.60   0.111    -1.0504  .110085   141.8
     married |  -.142893   .337372    -0.42   0.673   -.811571  .525784   108.8
  cohabiting |   .656153   .536409     1.22   0.222   -.396464  1.70877  1000.0
       _cons |   21.2969   .569379    37.40   0.000    20.1755  22.4184   247.2
------------------------------------------------------------------------------
```

```
milrtest engaged married cohabiting

Test statistic: F( 3,  415.116) =    1.557
Prob > F    0.1993

quietly: mim: reg distress rc01 rc02 age engaged married cohabiting
mim: testparm engaged married cohabiting

 ( 1)  engaged = 0
 ( 2)  married = 0
 ( 3)  cohabiting = 0

       F(  3, 431.9) =    1.56
            Prob > F =    0.1990
```

A cautionary tale

Using the naive approach and averaging the likelihood ratio tests across the *m* imputed datasets:

$\chi^2 = 5.5718, df = 3$

$p \leq .1344$

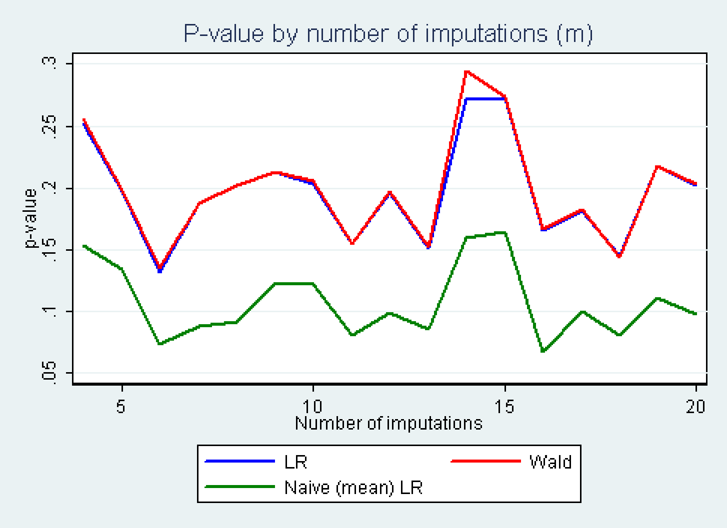
Which is far lower than the $p \leq 0.2$ obtained from both the Wald and the LR tests.

A comparison

The version of the Wald test implemented in `mim` is known to be unstable at low values of *m*. So the question is, how does the LR test implemented here compare?

Using the same data:

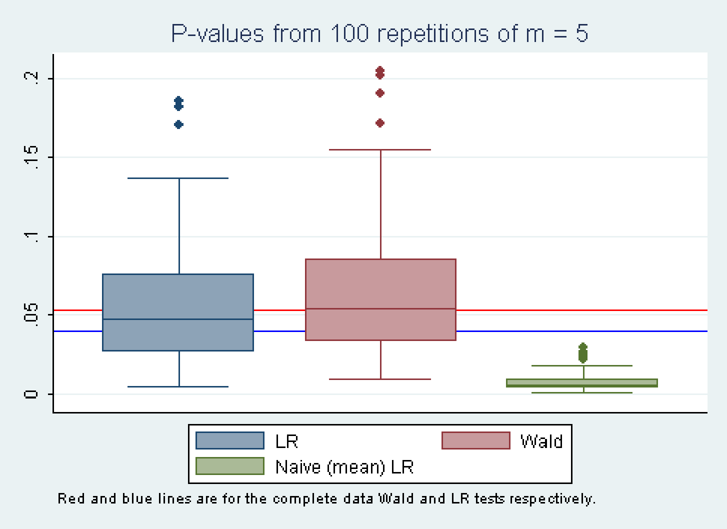
- MI datasets were created with $4 \leq m \leq 20$.
- The alternative (versus null) model above was tested using the LR and Wald tests with each of the 17 datasets.

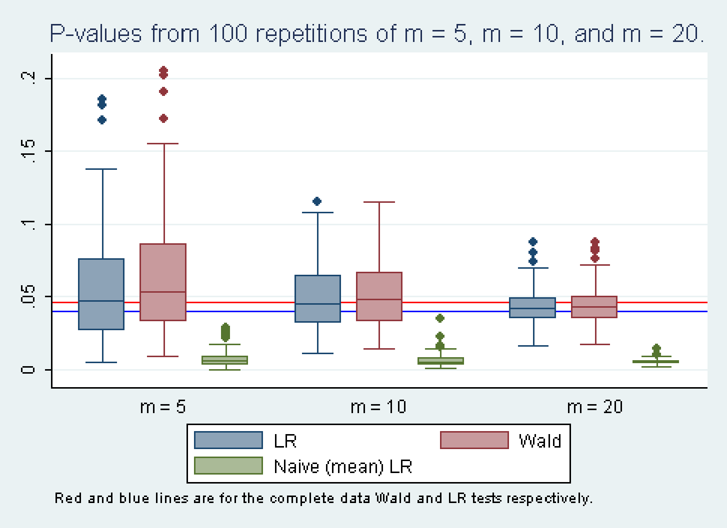

P-value by number of imputations (m)
p-value
.05
.1
.15
.2
.25
.3
5
10
15
20
Number of imputations
LR
Wald
Naive (mean) LR

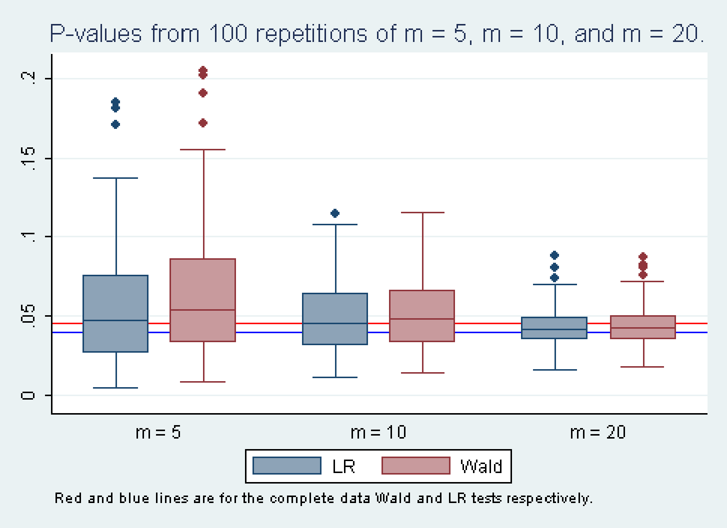
A more in-depth comparison

Using data from the study described above:

- Started with a subset of those cases with complete data on the necessary variables (n=2150).
- Compared the null and alternative models above using the standard LR and Wald tests.
- Created a single dataset with data missing completely at random. Percent missing for each variable ranged from less than 1% to about 30%, with a mean of about 15% missing.
- Imputed the missing values 100 times with $m = 5$, $m = 10$ and $m = 20$.
- Compared the null and alternative models from above using the `milrtest` and `mim: testparm`, saving the results.


P-values from 100 repetitions of m = 5
0
.05
.1
.15
.2
LR
Wald
Naive (mean) LR
Red and blue lines are for the complete data Wald and LR tests respectively.


P-values from 100 repetitions of m = 5, m = 10, and m = 20.
0
.05
.1
.15
.2
m = 5
m = 10
m = 20
LR
Wald
Naive (mean) LR
Red and blue lines are for the complete data Wald and LR tests respectively.


P-values from 100 repetitions of m = 5, m = 10, and m = 20.
0
.05
.1
.15
.2
m = 5
m = 10
m = 20
LR
Wald
Red and blue lines are for the complete data Wald and LR tests respectively.

Returned Arguments

scalars:

r(d_m)	Mean of likelihood ratio chi-squares for h1 vs h0 in unconstrained models
r(d_L)	Mean of likelihood ratio chi-squares for h1 vs h0 in constrained models
r(p)	p value of final statistic
r(df_d)	denominator degrees of freedom
r(df_n)	numerator degrees of freedom
r(test_stat)	F statistic
r(m)	number of imputed datasets used in estimation
r(h0_c_*m*)	LL of constrained model under h0
r(h1_c_*m*)	LL of constrained model under h1
r(h0_uc_*m*)	LL of unconstrained model under h0
r(h1_uc_*m*)	LL of unconstrained model under h1

macros:

r(cmd)	Name of the estimation command
r(h0_model)	Model under the null hypothesis
r(h1_model)	Model under the alternative hypothesis

matrices:

r(h0_coefs)	Coefficient estimates for null model
r(h1_coefs)	Coefficient estimates for alternative model

Programming notes

- The likelihoods for the constrained models are calculated using Mata.
- Currently these Mata functions are embedded in the appropriate .ado file.

`milrtest` can be downloaded from the ATS website, http://www.ats.ucla.edu/stat/stata/ado/analysis/milrtest.pkg or located using `findit milrtest`

References

Allison, P. D. (2001) Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Carlin, J. B., J. C. Calati, & P. Royson (2008) A new framework for managing and analyzing multiply imputed data in Stata. The Stata Journal 8(1): 49-67.

Carlin, J.B., N. Li, P. Greenwood, & C. Coffey (2003) Tools for analyzing multiple imputed datasets. The Stata Journal 3(3): 226-244.

Li, K.H., T.E. Raghunathan, & D.B. Rubin (1991) Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. Journal of the American Statistical Association 86(416): 1065-1073.

Little, R. J. A., & D. B. Rubin (2002) Statistical analysis with missing data. Hoboken, N.J: Wiley.

Meng, X., & D. B. Rubin. (1992) Performing likelihood ratio tests with multiply-imputed data sets. Biometrika 79: 103-111.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability 72. New York, NY: Chapman & Hall/CRC.