The DiffTest Trauma

(and how I'm recovering from it)

Paul Johnson[1]

[1]Center for Research Methods and Data Analysis

2019

KU

Outline

1. R User Point of view
2. Special Issues in SEM
3. What is CRMDA doing about it?

Outline

R's anova() function

- Is this model better than that model? *We are always asking that.*
- In R (R Core Team, 2017), one of the staples of regression analysis is the `anova()` function.
- `anova()` is a generic function comparing fitted models. There are versions for many kinds of models.
- In regression, for example

```
m1 <- lm(y ~ x1 + x2 + x3 + x4, data =
    fakedata)
m2 <- lm(y ~ x1 + x2, data = fakedata)
anova(m2, m1)
```

- The output will be an F test, indicating whether or not m1 "got worse" when x3 and x4 were omitted

Write that out as a hypothesis test

- Models are nested if the "smaller" one can be achieved by deleting features from the larger one. (or the larger one is achieved by adding features to the smaller one)
- the larger model is

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i + \beta_4 x4_i + \varepsilon_i$$

- the smaller model omits variables x3 and x4:

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \varepsilon_i$$

- The null hypothesis is $H_0 : \beta_3 = \beta_4 = 0$
- In ordinary least squares regression analysis, the F test is appropriate.

Here's what confuses most new R users

- The `anova()` function is a general purpose thing, applies to many model types, even ones that are not based on "sums of squares" or ANOVA concepts.
- `anova()` does provide an "F" test for regression, as ANOVA users expect
- `anova()` for other types of models will provide a likelihood ratio test or a score test or a ...
  - R design philosophy urges us to re-use function names were possible.

What about lavaan?

- `lavaan` implemented code for the `anova()` generic function.
- `anova()` re-organizes the information and passes it along to a lavaan specific model comparison function called `lavTestLRT()`. Users can access that directly if they don't want to use `anova()`

What is a Likelihood Ratio Test?

Basic Idea.

- Step 1. Run 1 model.
- Step 2. Add some features and run that again.
- These models are "nested"
  - may not delete elements while adding more.
  - may not change sample size while adding features
- Step 3. Compare the 2 models by comparing their likelihoods. Let
  - L_{max} be the value of the likelihood value at its maximum for the model with the most features.
  - L_0 be the value of the likelihood function in a "smaller" model.
- Let λ, (Greek "lambda"), be the "likelihood ratio" of L_0 to L_{max}:

$$\lambda = \frac{L_0}{L_{max}} = \frac{L_{smaller\,model}}{L_{bigger\,model}}$$

Intuitively,

What is a Likelihood Ratio Test? ...

  - if λ is near 1, then the models are about the same, so choose the smaller one.
  - if λ is very very small, then the smaller model is much worse, so keep the other.

- Here's the magic trick. If the sample size upon which these models are calculated is infinite, then

$-2\ln(\lambda)$ has a χ^2 distribution with k degrees of freedom

where k is the number of parameters removed in the smaller model.

- Sometimes called $-2LLR$ (LLR = log of likelihood ratio)
- Math reminder:
  The log of a ratio is the difference of the logs:

$$ln(a/b) = ln(a) - ln(b)$$

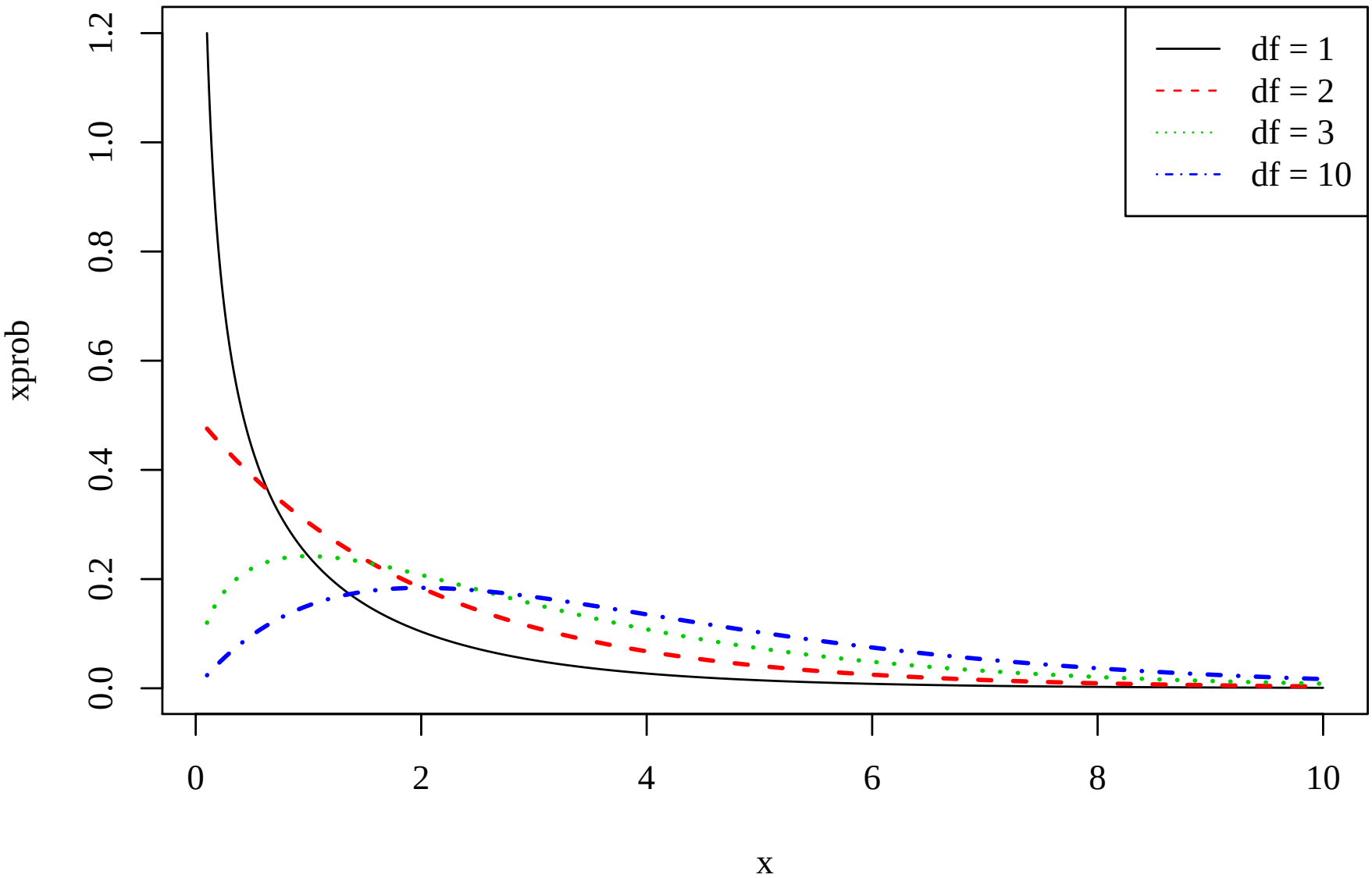
What is a Likelihood Ratio Test? ...

- Hence

$$
\begin{aligned}
-2ln(\lambda) = -2ln\left(\frac{L_0}{L_{max}}\right) &= -2\{ln(L_0) - ln(L_{max})\} \\
&= 2ln(L_{max}) - 2ln(L_0) \\
&= 2\{ln(L_{max}) - ln(L_0)\}
\end{aligned}
$$

- In order to claim that a difference is "statistically significantly different from 0 with probability 0.05", you hope that number is bigger than the critical value of the χ^2 statistic.

In Case you Can't visualize a Chi-square distribution



Critical values of a Chi-square distribution (α 0.05)

	df	critical.value
1	1.00	3.84
2	2.00	5.99
3	3.00	7.81
4	4.00	9.49
5	5.00	11.07
6	6.00	12.59
7	7.00	14.07
8	8.00	15.51
9	9.00	16.92
10	10.00	18.31

Roughly speaking, the -2LLR value needs to be larger than the degrees of freedom, by a scale of 2 to 4, in order to be "statistically significant".

KU

LR tests, in practice

- Using SAS, SPSS
  - Run one model, save output, print it out.
  - Run another model, save output, print it out
- Use a calculator to calculate -2 LLR and look it up in the χ^2 table.
- Only feasible if output includes the numbers we need (which, we find out next, is often not true in SEM)

Outline

The χ^2 model test in SEM

- SEM output usually includes a statistic referred to as "model Chi-square"
  - the large model is the "saturated model", one which has the largest possible value of the likelihood (because it includes all possible parameters)
  - the smaller model is the likelihood for the user's chosen specification.
- User hopes that -2LLR is *not statistically significant*, as if to say "my model explains a sufficient amount of variance because it is close enough to the best possible model."
- Often used as a first piece of evidence that the model is "good enough".

Problems with model χ^2 test

1. The "model χ^2" test is only approximately distributed as a χ^2 statistic.
   1. The LLR test assumes the sample size is infinite (but it is generally not).
2. The model χ^2 test may be unreasonably discouraging, even minute differences are "statistically significant" with larger samples.

Published Indictments of the χ^2 test

- There are publications documenting the inaccuracy of the SEM χ^2-based LLR test, even for large samples.
  Satorra, A., and Bentler, P. M. (1994). "Corrections to test statistics and standard errors in covariance structure analysis," in *Latent Variables Analysis: Applications for Developmental Research*, eds A. von Eye, and C. C. Clogg (Thousand Oaks, CA: SAGE Publications), 399–419.
  Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In Heijmans, R.D.H., Pollock, D.S.G. & Satorra, A. (eds.), *Innovations in multivariate statistical analysis.* A Festschrift for Heinz Neudecker (pp.233-247). London: Kluwer Academic Publishers.
  Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507-514.
  Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75, 243-248.

KU

Published Indictments of the χ^2 test ...

- And, by the developers of Mplus
  Asparouhov, T. & Muthén, B. (2006). Robust Chi Square Difference Testing with mean and variance adjusted test statistics. Mplus Web Notes: No. 10. May 26, 2006.
  Asparouhov, T. & Muthén, B. (2010). Computing the strictly positive Satorra-Bentler chi-square test in Mplus.http://www.statmodel.com/examples/webnotes/SB5.pdf
- Hence, we have competing suggestions, all of which are backed up by simulations and theorems, which suggest they have the best way to re-scale and modify the "model χ^2 statistic", and probably also the model standard errors.
- From now on, we will refer to these things generically as "model χ^2 statistics", even though they are, in fact, not χ^2 distributed. Although some re-scaled values are close to χ^2 than others.

What do we know, for sure

- Even if the data is truly drawn from a multivariate-normal data generator, the "model χ^2 statistic" that is typically reported with ML estimates is not accurate
- As the data deviates from normality, the deviation of the "model χ^2 statistic" from a χ^2 distribution becomes more pronounced.
- Some re-scaling of the "model χ^2 statistic" is called for, and either the Satorra-Bentler suggestions or the Asparouhov & Muthen suggestions, are improvements.

KU

Caution about rescaled statistics

- If the software rescales "model χ^2 statistic", then the reported value cannot be compared across models.
- It is NOT allowed to subtract them to form a post-hoc LLR test.
- On Unscaled fit statistics, We can improvise $2\{ln(L_1) - ln(L_2)\}$ by taking

$$\text{reported model 1 } \chi^2 : 2ln(L_1) - 2ln(L_0)$$

and

$$\text{reported model 2 } : 2ln(L_2) - 2ln(L_0)$$

subtract one from the other:

$$2ln(L_1) - 2ln(L_0) - 2ln(L_2) + 2ln(L_0)$$

$$2ln(L_1) - \cancel{2ln(L_0)} - 2ln(L_2) + \cancel{2ln(L_0)}$$

- When the scaling has been applied, the "model χ^2 statistic" is no longer a simple sum, and the usual thing to do is always wrong.

How did Commercial Software Adapt?

MPLUS

- For models that use a scaled "model χ^2 statistic" MPlus offers the following procedure.
  1. Fit one model. Insert a stanza with the command "SAVEDATA", as in

```
SAVEDATA:
  DIFFTEST = model1.dif;
```

  the output file can be any legal file name.
  2. Then run a second model, and ask for a diff test compared against the saved model file

```
ANALYSIS: DIFFTEST IS model1.dif
```

Take a Look at a diff file

It is a single column of un-labeled output, with various values, vectors, and matrices thrown together.

```
 1.768739314948761E-004
             1
            22
            19
   1.00000000000000
  0.000000000000000E+000
  0.000000000000000E+000
  0.000000000000000E+000
  0.000000000000000E+000
  0.000000000000000E+000
...[snip]
 -1.455231515017422E-002
 -8.481935382156272E-002
 -0.158546226444418
  2.890432611036050E-002
  8.653350827758647E-003
 -4.346341800339478E-002
 -9.327387467894621E-002
 -1.992458362913379E-002
 -9.790624138486069E-002
 -0.103097018690428
 -0.206034705941789
 -4.600490626847783E-003
 -4.661276672608555E-002
 -6.300500112246536E-002
 -0.121049791982158
  5.383824055426895E-003
```

Take a Look at a diff file ...

```
-3.467974760189810E-002
  2.94053552225483
```

Bentler and Satorra's counter proposal

EQS

- Satorra and Bentler created another approximate scaled statistic that can be calculated using standard output, pencil, paper, and a calculator.
- The last version of EQS implemented that version of their test.
- In 2017, we obtained the newest version of EQS and quickly discovered that the raw data was not being imported correctly. We reported the bug and stopped using the software.

Outline

How did I bump into this problem?

- Mplus is an expensive program, I don't want to pay for it. But
- KU students have told me they must have it! There's a thing called DIFFTEST and journal reviewers require it. They cannot do without it. And the formula is confidential! [1]
- In Summer 2017, we tried to figure out
  1. what the Mplus DIFFTEST is doing,
  2. if we can implement an equivalent comparison in R, probably with lavaan models

[1] Well, wait.... Its partially confidential. The DIFFTEST formula is not spelled out with perfect clarity in the Mplus guides. But there are hints that may be sufficient in the eyes of the Mplus team.

KU

What did we find out?

- Project 1. Learn what is in the SAVEDATA result file and figure out what's in there, and then verify the DIFF Test Results from Mplus.
- Project 2. Generate equivalent test results from a function we would create with R code

Project 1: Mission Accomplished

View online: Replicating the Mplus DIFFTEST Procedure: An R Function to Reproduce Nested Model Comparisons

Check writeup folder in this workshop folder.

- That guide explains that our goal is to recover a test statistic

$$T_3 = a \cdot T.chisq + b$$

- where we need to create a b and $T.chisq$

$$T.chisq = \{T_{large} - T_{small}\} \cdot N \cdot 2$$

- and T_{large} and T_{small} are the *unscaled* final fits from the 2 models being compared. These unscaled values are not reported in the usual output, they instead must be harvested from the *.dif file that was saved for each model.
- The calculation of a and b required a lot of guessing about which values in the diff file can be assigned to which matrix, but, in the end, the results match.

Project 2: Develop a Free-standing Replication

- Project 1 showed that if we have the required matrices, we can replicate the Mplus DIFFTEST.
- Project 2 would have us generate the same matrices and re-generate matching test statistics.
  - So far, we have exact replications for
    - 1 group models
    - 2 (or multi) group models with equal sample sizes within groups
- We do not replicate Mplus results for multi-group models when case numbers differ between groups.
  - At last inspection, it appears the multi-group variance/covariance matrix in Mplus is adjusted in a way for which we are unable to account.

KU

Where does all of this leave you?

- If your journal reviewers insist on the Mplus T_3 scaled chi-square difference test, for the moment it is necessary to use Mplus to conduct the test.
- If your journal reviewers will accept either the Satorra (2000) proposal, or the Satorra-Bentler proposals from 2001 or 2010, these can be supplied without Mplus. They were implemented in the lavaan source code base and they are now documented for the `anova()` method in `lavaan`.
- We believe that Satorra (2000) is closest to correct among these suggestions.

References

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Session

```
sessionInfo()
```

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 19.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/atlas/libblas.so.3.10.3
LAPACK: /usr/lib/x86_64-linux-gnu/atlas/liblapack.so.3.10.3

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
     LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8
     LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8
    LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] xtable_1.8-4
```

Session ...

```
20
loaded via a namespace (and not attached):
[1] compiler_3.6.0 tools_3.6.0
```