

SHOPPING FOR EFFICIENT CONFIDENCE INTERVALS IN STRUCTURAL
EQUATION MODELS

Donna Mohr and Yong Xu

University of North Florida

Authors Note

Parts of this work were incorporated in Yong Xu's Masters Thesis while at the University of North Florida.

Correspondence concerning this article should be directed to

Donna Mohr
Department of Mathematics and Statistics
University of North Florida
4567 St. Johns Bluff Road, S
Jacksonville, FL 32256
Email: dmohr@unf.edu
phone: (904) 620-2884 FAX: (904) 620-2818

or

Yong Xu
Department of Statistics and Mathematics
Shandong Economic University
China
Email: Xuyong007@hotmail.com

Abstract

Most users of Structural Equation Models are aware that Wald-type standard errors for parameter estimates can vary remarkably depending on the arbitrary choice of how the scale is identified. When the focus is on *H₀: coefficient is 0*, tests based on a likelihood-ratio are invariant to the scale identification. However, a simple example shows that confidence intervals based on the likelihood ratio still display different relative precisions depending on the choice of scale identification. Is it legitimate for the researcher to ‘shop’ for a scale identification that gives the best relative precision for a certain parameter? A series of examples suggests that shopping has very little negative impact on coverage probabilities, and can yield substantially tighter confidence intervals.

Keywords: SEM, log-likelihood, latent variables

Introduction

When fitting a Structural Equation Model (SEM), the user has numerous seemingly arbitrary choices for how to fix the scaling. Typically, we set certain parameters, such as a coefficient that relates a manifest variable to its underlying latent variable, equal to 1. Users who have explored several alternative ways of fixing the scale are familiar with seeing the t -tests for the underlying latent regression coefficients vary in what is sometimes a dramatic fashion. Gonzalez and Griffin (2001) have given an excellent explanation of why these Wald-type tests are sensitive to the scale identification. As an alternative when testing hypotheses of the form $H_0: \text{coefficient} = 0$, they recommend Chi-squared tests based on the log-likelihood function because they are invariant with respect to the choice of scale identification.

In a growing number of applications in psychology, economics, and biology, researchers are interested in the actual values of the latent regression coefficients. Examples of this are found in the application of SEM to longitudinal data, as in work by McArdle and Hamagami (2001). In this paper, our focus is on the behavior of confidence intervals formed using the log-likelihood function, in the manner recommended by Neale and Miller (1997) and used in the computer package Mx.

We will first show by example that these confidence intervals have relative errors which are sensitive to the scale identification. Then we explore the effects of ‘shopping’, that is, of fitting many different scale identifications, and choosing the one which gives the best relative error. In the examples where we have tried this experiment, we have seen that there is little penalty (in the form of smaller than nominal coverage probabilities) and

potentially large savings in accuracy. This suggests that shopping is a legitimate strategy for finding scale specifications that carry information in an efficient manner.

Confidence Intervals based on the Profile Log-Likelihood

Neale and Miller (1997) provide a description of how confidence intervals are computed from profile log-likelihood functions. In the final paragraphs of that article, they comment that though these intervals are preserved by transformations of the parameter, they are not invariant with respect to the scale identification. Since this method of constructing confidence intervals is central to this paper, we will briefly review it in the context of an example.

Lattin, Carroll and Green (2003, Chapter 10) give several examples of SEM. We will focus on Problem 10.5, itself taken from Fredricks and Dossett (1983), which we will refer to as the FD example. The path diagram is shown in Figure 1. For the moment, let us assume that our emphasis is on estimating the latent regression coefficient γ_1 . There are at least 27 different ways to make the scale identification, even if we do not attempt to fix any of the specific variances for the manifest variables. For the moment, we will compare only two of these:

Specification 1: Set $\text{Var}(F1) = \text{Var}(D2) = \text{Var}(D3) = 1$

Specification 2: Set $\lambda_{x1} = \lambda_{y1} = \lambda_{y3} = 1$

Here, the λ are the coefficients connecting a manifest variable to an underlying latent variable.

A data set of 236 observations was simulated using this path model and parameters, shown in Table 1, like those estimated from the textbook's correlation matrix. The results

below are based on Maximum Likelihood Estimates (MLE) fit to the sample covariance matrix for this simulated data, shown in Table 2.

The key quantity for inference is

$$F = n \times \{ \ln \|\Sigma\| - \text{tr}(S\Sigma^{-1}) - \ln \|S\| - p \} = -2 \{ \log \text{likelihood} - \text{constant} \}$$

where S and Σ are the sample and predicted covariance matrices (of dimension $p \times p$).

Under Specification 1, the MLE was $\hat{\gamma}_1 = .3221$ and $F = 5.1873$. To form the 95% confidence interval, we search either side of $\hat{\gamma}_1$ for two new values, $\hat{\gamma}_{1L}$ and $\hat{\gamma}_{1U}$ such that the F computed by fixing γ_1 at these values and re-maximizing the remaining parameters equal $9.0288 = 5.1873 + 3.8415$. The value 3.8415 is the 95th percentile from a Chi-Squared distribution with 1 df. Since we re-maximize the other parameters each time we try a new γ_1 , the resulting log-likelihoods are called profile log-likelihoods. Figure 2a shows the values of F created by varying γ_1 . The resulting confidence interval is $\gamma_1 \in (.134, .52)$

Figure 2b shows the construction of the confidence interval for γ_1 using Specification 2. Note that the MLE has changed due to the new choice of scale, so it is hard to directly compare the two graphs. To compare the two confidence intervals, we will focus on their relative error = $(\hat{\gamma}_{1U} - \hat{\gamma}_{1L}) / |\hat{\gamma}_1|$, which amounts to rescaling each interval so it is centered at 1.0. The results are shown in Figure 3. Now we clearly see that the interval produced by Specification 1 is far more accurate, in the sense of having much smaller relative error.

Shopping for a good scale identification

Given that the accuracy of the confidence interval is sensitive to the scale identification, this raises the possibility that the user can ‘shop’. That is, the user may try

many different scale identifications until finding one where the estimated confidence interval is as short (relative to the parameter estimate) as possible. Analogous to a multiple comparison problem, we might reasonably wonder whether such a strategy has an actual coverage probability less than the stated confidence level. We used a simulation based on the FD example to examine this possibility.

In the simulation, 2500 data sets were generated from the FD-model, using the parameter values given in Table 1. Each data set had 236 observations, as in the original example. The MLE were obtained, and then 95% confidence intervals were generated separately for γ_1 , γ_2 , and β using the profile log-likelihoods under 27 different scale specifications. All programs were in double-precision FORTRAN, using IMSL routines DUMING and DZBREN for the minimization tasks. For each of the scale specifications, we track the percentage of times the confidence intervals contained the true value of the corresponding parameter, and the mean relative error of the interval.

The coverage probabilities and mean relative errors are summarized in Table 3. Under every specification, the actual coverage probability was very close to the nominal 95%. However, the mean relative errors for parameter γ_1 varied tremendously, from .666 to .965. By contrast, the relative errors for γ_2 are far more stable. Note (from Table 1) that γ_2 is much closer to 0 than γ_1 , and it seems that in some neighborhood of 0 the relative errors are less sensitive to the scale identification.

The surprise comes in the behavior of the confidence interval chosen from the best identification, that is, by scanning across the calculated confidence intervals for all 27 specifications and reporting the specification with the shortest interval. These are summarized in Table 3, in the row labeled ¶. Of course, the mean relative error is slightly

smaller than the mean of any one method of specification. The surprise is that the actual coverage probability is still very close to the 95% level.

Shopping for several parameters simultaneously

Normally we have several parameters for which we need confidence intervals. We will suppose that we are interested in all three latent regression coefficients γ_1 , γ_2 , and β . Usually there will not be a single specification that happens to be best for all the parameters of interest. In that case, we will have to compromise. Denote the estimated relative error in the confidence interval for parameter j under specification s as $RE(j, s)$. Let $BestRE(j)$ be the smallest relative error found after scanning all the specifications. This will normally come from a different s for each of the parameters of interest. We will choose the single specification that minimizes

$$ORE(s) = \sum_j RE(j, s) / BestRE(j)$$

where the summation is over all parameters j of interest to the researcher.

The results of following this strategy in the FD simulations are shown at the bottom of Table 3 in the line labeled §. The confidence intervals based on this interval are almost as short as the ones chosen by optimizing each parameter separately. Unlike the intervals obtained by following separate strategies for each parameter, these are all comparable, in the sense they are all based on the same scale identification.

A true multiple comparison problem arises as one considers the probability that all three of the true parameter values lie within their corresponding confidence intervals. In the simulation, this probability was 85.4%, indicating that Bonferroni-style adjustments may be appropriate for those concerned with the family-wise confidence level.

Conclusion

We have reproduced these results in several other examples. One (Example 10.6 from Lattin, Carroll, and Green) had two latent regression coefficients, and another (Example 10.2) had one coefficient. While examples do not yield universal truths, there is a consistent pattern in which the confidence intervals chosen *post hoc* for having the shortest relative length do a good job of maintaining the required confidence level.

Unfortunately, there seems to be no way to guess in advance which specification will yield the best results. In some simulations, the best results were found by fixing the variances of the disturbances in the underlying latent variables (as in the FD example). In others, one particular set of manifest variable coefficients seemed to be most advantageous. The driving factor seems to be the size of the variance: those variables (whether manifest or latent) with the largest variation are the ones that need to be anchored in some fashion. Naturally, this is not something one would know in advance of fitting the model.

Users of SEMs may have been unaware of the advantages of shopping, or they may have been ashamed to admit they shop, feeling that this was akin to fishing for the hypothesis that will (finally) yield a significant p-value. But the work of Gonzalez and Griffin (2001) shows us that various identifications carry information in more or less efficient ways for a particular subset of the parameter vector. Given that confidence levels seem to be maintained, the potential advantages in precision vindicate shopping as legitimate way of screening for an efficient identification.

References

Fredricks, A. J., & Dossett, D. L. (1983). Attitude-behavior relations: a comparison of the Fishbein-Ajzen and the Bentler-Speckart models. *Journal of Personality and Social Psychology, 42*, 201-212.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: every “one” matters. *Psychological Methods, 6*, 258-269.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Brooks/Cole—Thomson Learning.

McArdle, J. J., & Hamagami, F. (2001) Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayers (Eds.), *New methods for the analysis of change* (pp. 139-175). Washington, DC: American Psychological Association.

Neale, M. C., & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27*, 113-120.

Table 1. True Parameter Values used in Simulations from FD model. The λ represent the coefficients between the manifest variables and the underlying latent variable. The e are the error terms in the manifest variables $X1, \dots, Y4$.

1.	$\gamma_1 = .7008$		7.	$\lambda_{X1} \equiv 1$		13.	Var(e1) = .33411
2.	$\gamma_2 = .1390$		8.	$\lambda_{X2} = 1.0819$		14.	Var(e2) = .34584
3.	$\beta = .4119$		9.	$\lambda_{Y1} \equiv 1$		15.	Var(e3) = .55685
4.	Var(F1) = .42346		10.	$\lambda_{Y2} = .9912$		16.	Var(e4) = .57887
5.	Var(D1) = .42346		11.	$\lambda_{Y3} \equiv 1$		17.	Var(e5) = .50643
6.	Var(D2) = .28103		12.	$\lambda_{Y4} = .9748$		18.	Var(e6) = .42226

Table 2. Sample covariance matrix for a simulated data set from FD model

	X1	X2	Y1	Y2	Y3	Y4
X1	.8197	.4188	.2520	.1918	.1462	.1672
X2		.8371	.1147	.1584	.1053	.1013
Y1			1.0211	.7029	.2649	.2856
Y2				.9994	.1856	.2510
Y3					.9558	.3914
Y4						.9169

Table 3. Coverage probabilities and mean relative errors obtained from simulation.

Parameter identifications correspond to numbering in Table 1.

Model, with ID of parameters=1	γ_1		γ_2		β	
	Coverage Prob. (%)	Mean Rel Error	Coverage Prob. (%)	Mean Rel Error	Coverage Prob. (%)	Mean Rel Error
1. (7,9,11)	94.5	.806	94.3	11.81	94.9	1.07
2. (7,10,11)	94.9	.809	94.3	11.81	94.7	1.07
3. (8,9,12)	95.0	.837	94.7	12.02	94.5	1.09
4. (8,9,11)	95.0	.837	94.2	11.96	94.9	1.07
5. (4,9,11)	94.5	.666	94.2	11.53	94.9	1.07
6. (4,5,6)	95.0	.795	94.7	11.86	94.3	1.21
7. (4,5,11)	95.0	.795	94.2	11.53	94.7	1.05
8. (4,5,12)	95.0	.795	94.9	11.59	94.6	1.06
9. (4,6,9)	94.5	.666	94.7	11.85	94.8	1.23
10. (4,9,12)	94.5	.666	94.9	11.59	94.5	1.09
11. (4,6,10)	95.0	.669	94.7	11.85	94.7	1.23
12. (4,10,11)	95.0	.669	94.2	11.53	94.7	1.07
13. (4,10,12)	95.0	.669	94.9	11.59	94.5	1.08
14. (5,6,7)	94.8	.919	94.7	12.12	94.3	1.20
15. (5,7,11)	94.8	.919	94.3	11.81	94.7	1.05
16. (5,7,12)	94.8	.919	95.0	11.87	94.6	1.06
17. (6,7,9)	94.5	.806	94.7	12.12	94.8	1.23
18. (7,9,12)	94.5	.806	95.0	11.87	94.5	1.09
19. (6,7,10)	94.9	.809	94.7	12.12	94.7	1.23
20. (7,10,12)	94.9	.809	95.0	11.87	94.5	1.08
21. (5,6,8)	95.1	.965	94.7	12.28	94.3	1.21
22. (5,8,11)	95.1	.965	94.2	11.96	94.7	1.05
23. (5,8,12)	95.1	.965	94.7	12.02	94.6	1.06
24. (6,8,9)	95.0	.837	94.7	12.28	94.8	1.23
25. (6,8,10)	95.0	.840	94.7	12.28	94.7	1.23
26. (8,10,11)	95.0	.840	94.2	11.96	94.7	1.07
27. (8,10,12)	95.0	.840	94.7	12.02	94.5	1.08
¶ shortest interval	94.7	.646	94.6	11.44	94.1	1.00
§ compromise specification	94.5	.648	94.5	11.48	94.5	1.03

Figure 1. Path Diagram for FD data. Latent variables are F1, F2, and F3. D2 and D3 are disturbance terms in the endogenous latent variables.

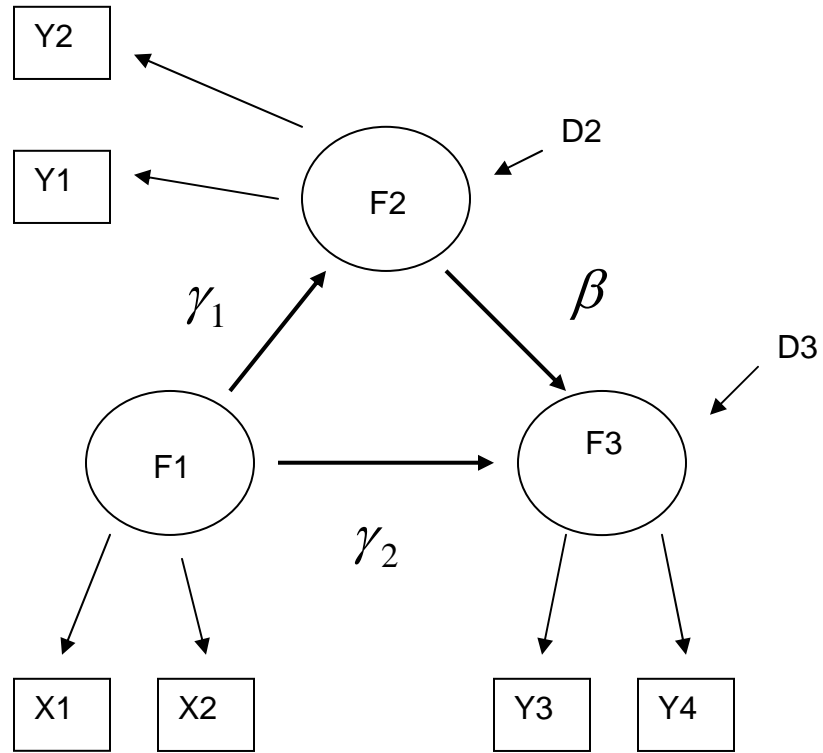


Figure 2. (A) Confidence interval for γ_1 constructed from profile log-likelihood using Specification 1, and (B) using Specification 2.

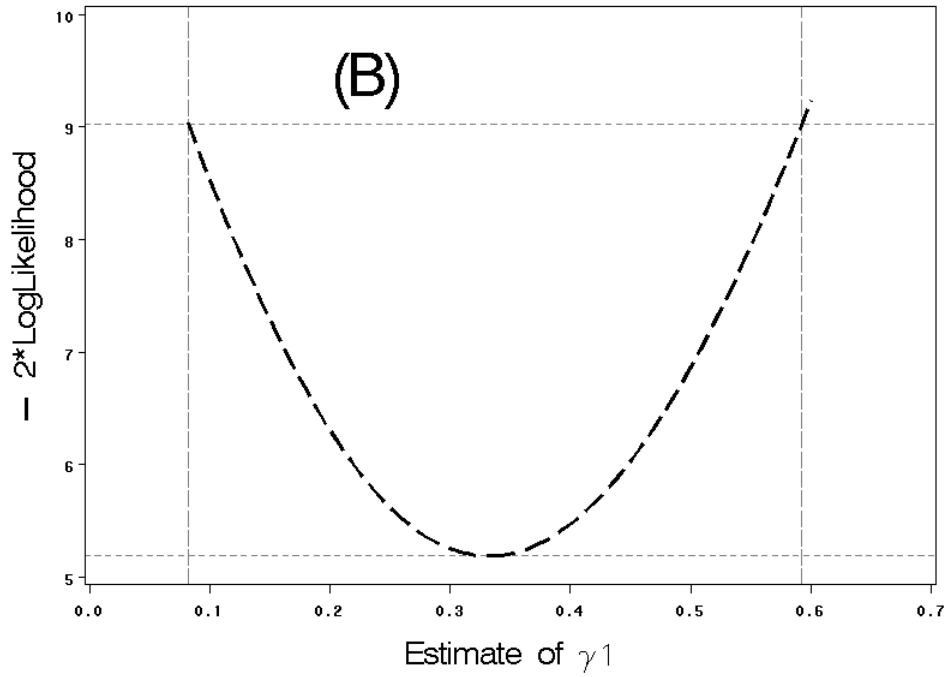
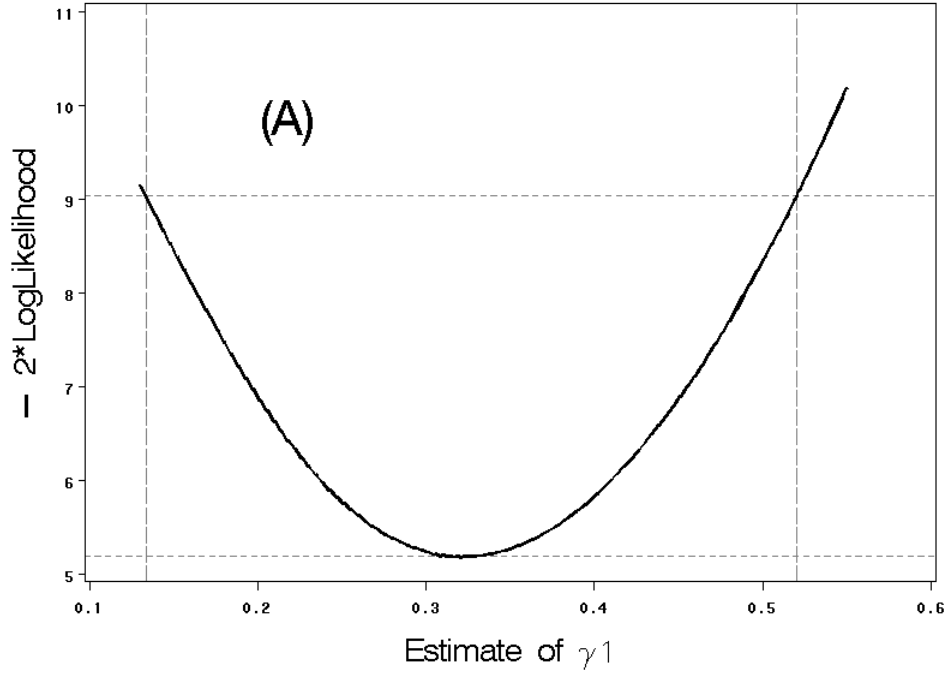


Figure 3. Confidence intervals for γ_1 rescaled relative to $\hat{\gamma}_1$.

