Maximum Likelihood and Fit Statistics

Steven M. Boker

Department of Psychology
University of Virginia

Structural Equation Modeling
Psyc–8501–001

Overview

- How and why do we fit an SEM model?
- Maximum Likelihood for Covariance
- Maximum Likelihood for Covariance and Means
- Full Information Maximum Likelihood
- Assumptions
- Model Comparison Fit Statistics
  - Minus 2 Log Likelihood (-2LL or Chi Square or χ^2)
  - Root Mean Square Error of Approximation (RMSEA)
  - Akaike Information Criterion (AIC)
  - Bayes Information Criterion (BIC)

How and why do we fit an SEM model?

- Have a theory.
- Gather or find data.
- Construct several alternative models designed to test aspects of the theory.
  - Are these models the same as theories?
  - Models say how theories would account for these data given a particular set of statistical assumptions.
- Check whether the data violate the assumptions.
- Fit the alternative models.
- Compare fit statistics of the alternative models.
- Improve your understanding of the theory and perhaps make changes to the theory.

How and why do we fit an SEM model?

- Have a theory.
- Gather or find data.
- Construct several alternative models designed to test aspects of the theory.
  - Are these models the same as theories?
  - Models say how theories would account for these data given a particular set of statistical assumptions.
- Check whether the data violate the assumptions.
- Fit the alternative models.
- Compare fit statistics of the alternative models.
- Improve your understanding of the theory and perhaps make changes to the theory.

Fitting a Model

1. Data has an observed covariance (and possibly a vector of means).
2. The model has a specified structure and starting values.
3. Calculate the expected covariance matrix (and expected vector of means).
4. The difference between the observed and expected covariance matrices (and means) is calculated.
    - This difference is called the “function value”.
5. Specify new starting values so as to reduce the function value.
6. If the function value is still getting smaller, go back to step 3.
7. Report the parameter estimates, the function value, and fit statistics.

Fitting a Model

1. Data has an observed covariance (and possibly a vector of means).
2. The model has a specified structure and starting values.
3. Calculate the expected covariance matrix (and expected vector of means).
4. The difference between the observed and expected covariance matrices (and means) is calculated.
    - This difference is called the "function value".
5. Specify new starting values so as to reduce the function value.
6. If the function value is still getting smaller, go back to step 3.
7. Report the parameter estimates, the function value, and fit statistics.

Five Fit Functions

1. Unweighted Least Squares (ULS).
   - Scales for all variables must be the same.
2. Generalized Least Squares (GLS).
   - Assumes multivariate normal.
3. Maximum Likelihood (ML and FIML).
   - Assumes multivariate normal.
4. Asymptotic Distribution Free (ADF).
   - Does not make distributional assumptions, but sample size must be large.
5. Bayesian Estimation Methods.
   - Slow estimation, but best when there are unequal model priors.

Five Fit Functions

1. Unweighted Least Squares (ULS).
   - Scales for all variables must be the same.
2. Generalized Least Squares (GLS).
   - Assumes multivariate normal.
3. Maximum Likelihood (ML and FIML).
   - Assumes multivariate normal.
4. Asymptotic Distribution Free (ADF).
   - Does not make distributional assumptions, but sample size must be large.
5. Bayesian Estimation Methods.
   - Slow estimation, but best when there are unequal model priors.

Maximum Likelihood

- Suppose we have data.
- Suppose we have a specified model structure.
- For this model structure there are many model instances, each with different parameter values.
- For this model structure, we wish to select the parameter values that have the greatest likelihood to have produced the data.

Maximum Likelihood

- Suppose I wish to predict the outcome of a coin toss.
- I have three models:
  1. The probability of heads is $p(heads) = 1.00$.
  2. The probability of heads is $p(heads) = 0.50$.
  3. The probability of heads is $p(heads) = 0.00$.
- I flip the coin 20 times and it comes up heads 13 times.
- So, $p(heads) = 13/20 = 0.65$
- By inspection, it looks like Model 2 is the most likely model to have produced the data.

Maximum Likelihood

- But now suppose I only had two models:
  1. The probability of heads is $p(heads) = 1.00$.
  2. The probability of heads is $p(heads) = 0.00$.
- I flip the coin 20 times and it comes up heads 13 times.
- So, $p(heads) = 13/20 = 0.65$ in the data.
- Now it looks like Model 1 is the most likely model to have produced the data.
- The model that is chosen as having maximum likelihood is dependent on which models are in the set!

Maximum Likelihood for Coin Toss

- The likelihood function is the joint probability of observing the data.
- For a coin toss with heads probability p, this is a Bernoulli distribution

$$f_p(x_i) = p^{x_i}(1-p)^{1-x_i}$$

- The likelihood $\mathcal{L}$ of the data given p is the product

$$\mathcal{L}(p) = \prod_{i=1}^{N} f_p(x_i)$$

$$\mathcal{L}(p) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i}$$

$$\mathcal{L}(p) = p^{\sum_{i=1}^{N} x_i}(1-p)^{N-\sum_{i=1}^{N} x_i}$$

Maximum Likelihood for Coin Toss

- Let's go back to our example and calculate the likelihood for each of the three models.
  1. The probability of heads is $p(heads) = 1.00$.
  $$\mathcal{L}(1) = 1^{13} \times 0^{7} = 0$$
  2. The probability of heads is $p(heads) = 0.50$.
  $$\mathcal{L}(.5) = .5^{13} \times .5^{7} = 0.00000095$$
  3. The probability of heads is $p(heads) = 0.00$.
  $$\mathcal{L}(0) = 0^{13} \times 1^{7} = 0$$
- This is usually expressed in terms of -2 times the natural log of the likelihood (written -2LL).
  $$-2\ln(\mathcal{L}(.5)) = 27.72$$

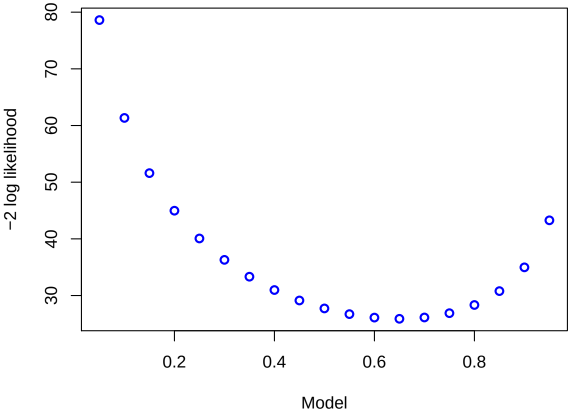
Maximum Likelihood for Coin Toss

- The closer the -2LL is to zero, the more likely are the data given the chosen model.
- Let's use R to calculate the -2LL for all the model probabilities from 0.5 to .95 stepping by .05.

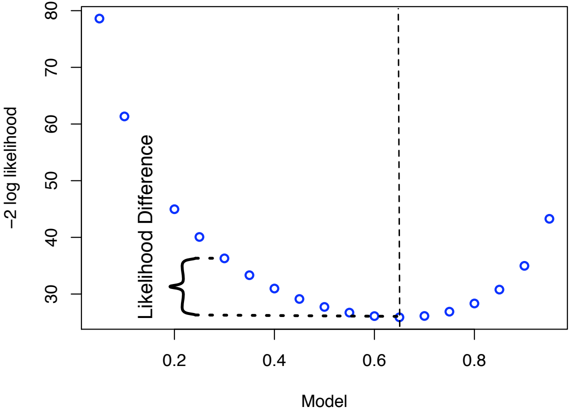
```
ll <- rep(NA, 19)
model <- rep(NA, 19)
i <- 1
for(p in seq(.05, .95, by=.05)) {
    model[i] <- p
    ll[i] <- -2 * ((13 * log(p)) + ((20 - 13) * log(1 - p)))
    i <- i + 1
}

pdf("LogLikelihoodCoin.pdf", height=5, width=6)
plot(model,ll, lwd=2, col="blue",
    xlab = "Model",
    ylab="-2 log likelihood")
dev.off()
```

Maximum Likelihood for Coin Toss



Maximum Likelihood for Coin Toss



Maximum Likelihood

- We would *like* to select the most likely model given the data.

$$\max p(Model|Data)$$

- But what we are *actually* choosing is the maximum probability of the data given the model.

$$\max p(Data|Model)$$

- By Bayes theorem we know that

$$p(Model|Data) = \frac{p(Data|Model)p(Model)}{p(Data)}$$

- But we know that $p(Data)$ is constant.
- So, if every model in our set is equally probable, maximum likelihood would choose the most likely model.

Maximum Likelihood

- Suppose we have two models:
  1. A person is President Obama.
  2. A person is not President Obama.
- Suppose in the population of the U.S. we find that $p(BrownHair) = .7$ and $p(Male) = .5$
- The probability of A and B occurring is the probability of A occurring times the probability of B occurring given that A has already occurred. $p(A \cap B) = p(A)p(B|A)$
- I now have two models:
  1. The probability of brown hair and male if a person is President Obama is $p(BrownHair \cap Male) = 1.0 \times 1.0 = 1.0$.
  2. The probability of brown hair and male if a person is not President Obama is $p(BrownHair \cap Male) = 0.7 \times 0.5 = 0.35$.
- If we select a person that is a male and has brown hair, then given these models, by maximum likelihood I will select the model that says that person is President Obama!

Maximum Likelihood Base Rate Fallacy

- What went wrong with our logic?
- Let's look again at the two models
  1. A person is President Obama.
  2. A person is not President Obama.
- In the 2000 U.S. Census, the population of the U.S. was 281,421,906.
- There is only one President Obama.
- So, the prior probability of each model is
  1. $p(ObamaModel) = 1/281421906 = 0.00000000355$
  2. $p(NotObamaModel) = 281421905/281421906 = 0.99999999645$
- When the prior probability of models is different from one another, then Bayesian estimation techniques maximize

$$p(Data|Model)p(Model)$$

and so we would always choose the NotObamaModel since $p(ObamaModel)$ is so low.

Maximum Likelihood Base Rate Fallacy

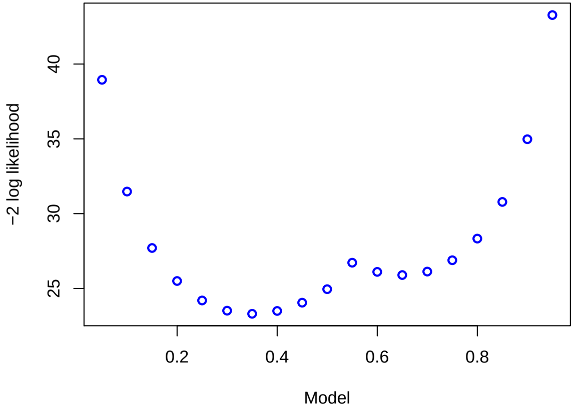
- A more subtle version of this fallacy involves when you pick comparison models that are "straw man" models to compare to your preferred model.
- If you know that there is almost no chance that the comparison model is correct, then you have set up a comparison with a base rate fallacy.
- If you want to believe the model comparison statistics, then you must be willing to say that each model in the comparison has an equal prior probability.
- Model comparison statistics are not valid if you already know the answer!
- One way to know the answer is to look at your data before you build your theory and your models.
  - In that case, the model comparison statistics are not informative: They don't add any more evidence than you already had.

Maximum Likelihood for Continuous Variables

- So far, we have only looked at discrete outcomes (heads/tails, Obama/NotObama).
- What happens when we move to continuous outcomes?
- Now we have a **very** large number of possible outcomes and possible model parameters.
- One way to deal with this is to use means, variances and covariances.
- We calculate the likelihood of the data given a model–predicted covariance structure.
- There is no way to calculate the likelihood of all of them.
- So, we use search strategies to find a model with minimum -2LL.
- But at the heart of it we are still doing the same thing:

$$\max p(Data|Model)$$

Maximum Likelihood Local Minimum



Maximum Likelihood for Univariate Normal

- Suppose we have a data matrix composed of independent rows where each row was drawn from the same distribution.
- Then the likelihood $\mathcal{L}$ of the data given a set of parameters Θ is the product of the univariate probability densities

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f_\theta(x_i)$$

- This calculation can be simplified by taking the log of each side so that the products turn into sums

$$\ln(\mathcal{L}(\theta)) = \sum_{i=1}^{N} \ln(f_\theta(x_i))$$

Maximum Likelihood for Univariate Normal

- The normal distribution with mean=mu and variance σ^2 has a probability density function

$$f_\theta(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Then for a sample matrix with N independent identically distributed rows we have

$$\mathcal{L}(\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{\sum_{i=1}^N (x_i-\bar{x})^2 + N(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

- Taking the log doesn't change the maximum of this function, so an easier form to calculate is

$$\ln(\mathcal{L}(\mu,\sigma^2)) = N\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{\sum_{i=1}^N (x_i-\bar{x})^2 + N(\bar{x}-\mu)^2}{2\sigma^2}$$

Maximum Likelihood for Covariance Matrices

- This same logic can be extended to multivariate normal data.
- Suppose we have a data set with N rows and p variables with an observed covariance matrix $\mathbf{S}$.
- Suppose we have a model that results in an expected covariance matrix $\mathbf{\Sigma}$.
- Then the maximum likelihood function value can be defined as

$$-2 \ln \mathcal{L} = (N - 1)(\ln |\mathbf{\Sigma}| - \ln |\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - p)$$

Maximum Likelihood for Covariance and Means

- This same logic can be extended to multivariate normal data when you have a model for the means.
- Suppose we have a data set with N rows and p variables with an observed covariance matrix $\mathbf{S}$ and vector of means $\mathbf{x}$.
- Suppose we have a model that results in an expected covariance matrix $\mathbf{\Sigma}$ and expected mean vector μ.
- Then the maximum likelihood function value can be defined as

$$\begin{aligned}-2\ln\mathcal{L} \quad &= (N-1) \quad \Big(\ln|\mathbf{\Sigma}| - \ln|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - p + \\ &\qquad\qquad\quad \frac{N}{N-1}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) + 1\Big)\end{aligned}$$

Full Information Maximum Likelihood

- This logic can also be applied to create a likelihood calculation for each row in the data matrix.
- The advantage here is that if some variable is missing for one row in the data matrix, the model expectation can be used for the data that does exist.
- Thus, for data that are missing at random, each row of the data contributes to the misfit to the extent that they do have data.
- In order to use FIML you must have a model for the means as well as the covariances.

Full Information Maximum Likelihood

- If person i is observed on k variables in a vector $\mathbf{x}_i$, the maximum likelihood function value looks like this.

$$-2\ln\mathcal{L} = \sum_{i=1}^{N}\left(-k\ln(2\pi) + \ln|\mathbf{\Sigma}_i| + (\mathbf{x}_i - \mu_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}_i - \mu_i)\right)$$

- Note that the expected covariance matrix $\mathbf{\Sigma}$ and expected mean vector μ have only the rows and columns that exist in the k variables that exist for person i.

SEM Statistical Assumptions

- *Linearity*. For each unit change in the independent variable, there is some proportional change in the dependent variable.

$$y_i = b_0 + b_1 x_i + e_i$$

- The expected value of x is the mean of x

$$\mathcal{E}(x) = M_x = \bar{x} = \mu_x$$

- The expected value of the error is 0

$$\mathcal{E}(e) = M_e = \bar{e} = \mu_e = 0$$

SEM Statistical Assumptions

- The expected value of the square of the deviation scores is the variance
$$\mathcal{E}((x-\bar{x})(x-\bar{x})') = V_x$$
- The expected value of the square of the error is the variance of the error (*Homoscedasticity*)
$$\mathcal{E}((e)(e)') = V_e$$
- The independent variable x and the error e have an expected covariance of 0 (predictor variables and the residuals are uncorrelated).
$$\mathcal{E}((x-\bar{x})(e)')) = 0$$
- The error is approximately normally distributed with a mean of 0 and normal variance.
$$e \approx \mathrm{N}(0, V_e)$$

SEM Statistical Assumptions

- *Multivariate Normality*. The multivariate extension of the normality assumption in regression.
- *Sample Size*. Widely varies depending on the number of *indicators* (measured variables), the number of latent variables and the strength of *loadings* (coefficients). Simple models may be estimated with fewer than 100 cases, more complex models may require 500 or more cases.
- *No Multicollinearity*. This can result in problems like singular matrices when two variables are correlated too highly ($> .9$ or so).

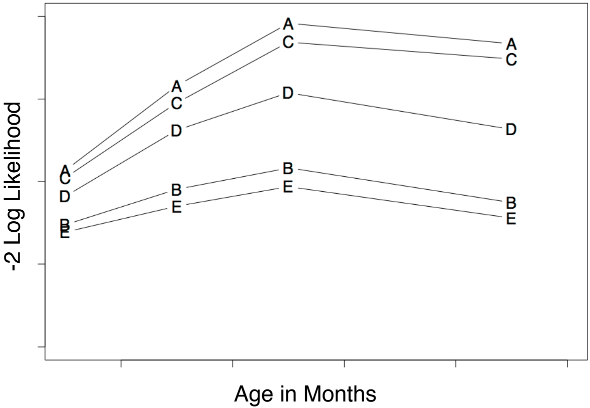
Model Comparison Fit Statistics

- Model Comparison Fit Statistics
  - Minus 2 Log Likelihood (-2LL or Chi Square or χ^2)
  - Root Mean Square Error of Approximation (RMSEA)
  - Akaike Information Criterion (AIC)
  - Bayes Information Criterion (BIC)

Minus 2 Log Likelihood

- This can be used to perform a *Likelihood Ratio Difference Test* when you have nested models.
- The difference between two nested models' Minus 2 Log Likelihoods is distributed as Chi Square with p degrees of freedom where p is the difference between the degrees of freedom in the two models.
- Nested models are ones where you only constrained parameters or you only released constraints on parameters.
- I tend to use the -2LL in a slightly different way.
- I test several models including one that has many constraints (a Null Comparison Model) up to a model that has released all of the constraints that are interesting (a Nearly Saturated Model). I look carefully at places where there are large jumps in the -2LL.

Minus 2 Log Likelihood



Root Mean Square Error of Approximation

- RMSEA was proposed as a goodness of fit index by Steiger & Lind in 1980.
- RMSEA attempts to make a fit statistic that is relatively independent of sample size.
- It is calculated as

$$RMSEA = \sqrt{\frac{(-2\ln\mathcal{L} - df)}{N\ df}}$$

- One way to think about RMSEA is as an index of being "close".
- Rules of thumb have grown to be < 0.05 is "good" fit and > 0.10 is "poor" fit.
- Rules of thumb are great but one person can only count up to two with their thumbs.

AIC and BIC

- The AIC and BIC attempt to compare non–nested models.
- There is still some controversy about their use.
- Akaike's Information Criterion is defined as

$$AIC = -2\ln\mathcal{L} + 2(p+1)$$

  where p is the number of free parameters in the model
- The Bayes Information Criterion is defined as

$$BIC = -2\ln\mathcal{L} + 2(p+1)\ln(N)$$

- As you can see, BIC attempts to account for the sample size as well as the complexity of the model.
- Again, lower AIC and BIC are better.
- AIC and BIC are also used with FIML.

Next Lecture

- Data Screening.
- Graphical Diagnostics.