Handling Missing Data in Research: A Practical Guide

full information maximum likelihood (FIML, direct-ml)

Waylon Howard

Webinar, November 20, 2024

ML estimation

- ML identifies the population parameters that are most likely given the observed data
- A likelihood (or log-likelihood) function is used to quantify how well the proposed parameters explain the observed data.
- ML requires a population distribution (normal)

ML estimation 😏

A density function gives the shape of the normal curve

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\frac{(Y_i-\mu)^2}{\sigma^2}}$$

L_i (the likelihood) gives the relative probability that Y_i came from a normal distribution with a particular mean and variance.

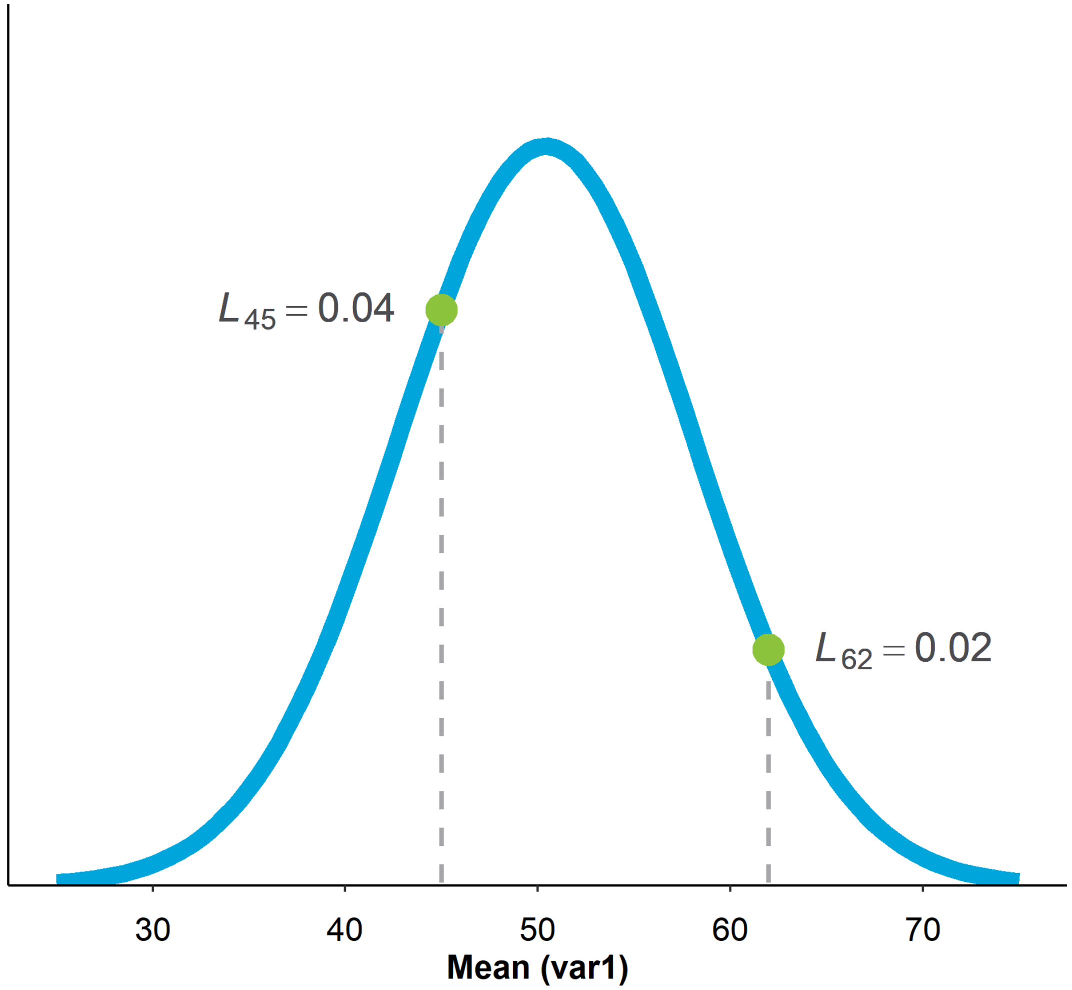
ML estimation

Applying the density function gives the relative probability (L_i) of each score from this normal distribution.

complete data:

var1 ($\mu = 50.42, \sigma = 7.65$)

ID	var1	L_i
1	36.6	0.010201
2	41.8	0.027624
3	42.6	0.030908
4	43.1	0.032971
5	43.4	0.034205
6	44.2	0.037444
7	44.9	0.040166
8	46.3	0.045074
9	48.6	0.050658
10	49.0	0.051223
11	50.0	**0.052038**
12	51.6	0.051508
13	54.6	0.044915
14	54.8	0.044264
15	55.7	0.041102
16	57.2	0.035227
17	57.6	0.033589
18	60.3	0.022677
19	60.9	0.020433
20	65.3	0.007888


$L_{45} = 0.04$
$L_{62} = 0.02$
30
40
50
60
70
Mean (var1)

Maximum Likelihood

Multiple each (L_i) to get sample likelihood.

0.0000000000000000000000000000163666415977258

Fit of this data to μ = 50.42, σ = 7.65

To avoid small numbers, we take the log of the likelihood.

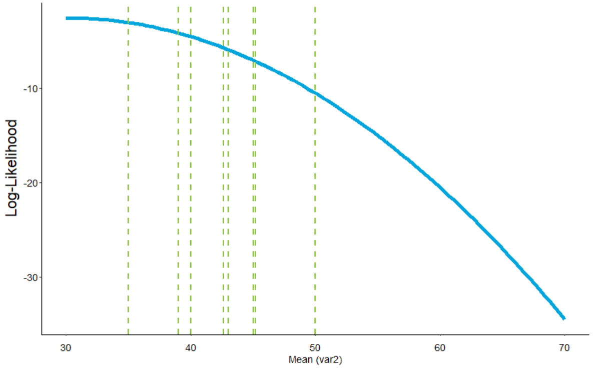
Add each $logL_i$ to get sample loglikelihood.

-68.58

ID	var1	L_i	$logL_i$
1	36.6	0.010201	-4.585258
2	41.8	0.027624	-3.589055
3	42.6	0.030908	-3.476754
4	43.1	0.032971	-3.412113
5	43.4	0.034205	-3.375376
6	44.2	0.037444	-3.284921
7	44.9	0.040166	-3.214733
8	46.3	0.045074	-3.099445
9	48.6	0.050658	-2.982664
10	49.0	0.051223	-2.97157
11	50.0	0.052038	**-2.955783**
12	51.6	0.051508	-2.966023
13	54.6	0.044915	-3.102986
14	54.8	0.044264	-3.117579
15	55.7	0.041102	-3.191692
16	57.2	0.035227	-3.345936
17	57.6	0.033589	-3.393553
18	60.3	0.022677	-3.786393
19	60.9	0.020433	-3.890587
20	65.3	0.007888	-4.842415

ID	var1	μ = 30	μ = 40	μ = 50	μ = 60	μ = 70
1	36.6	-3.326	-3.053	-4.487	-7.627	-12.474
2	41.8	-4.142	-2.982	-3.528	-5.781	-9.74
3	42.6	-4.309	-3.012	-3.422	-5.538	-9.361
4	43.1	-4.419	-3.036	-3.361	-5.391	-9.129
5	43.4	-4.487	-3.053	-3.326	-5.306	-8.992
6	44.2	-4.675	-3.105	-3.241	-5.085	-8.634
7	44.9	-4.849	-3.159	-3.176	-4.9	-8.33
8	46.3	-5.222	-3.293	-3.071	-4.556	-7.747
9	48.6	-5.906	-3.585	-2.971	-4.063	-6.862
10	49	-6.035	-3.645	-2.963	-3.987	-6.718
11	50	-6.368	-3.808	-2.954	-3.808	-6.368
12	51.6	-6.936	-4.103	-2.976	-3.556	-5.843
13	54.6	-8.118	-4.773	-3.135	-3.203	-4.978
14	54.8	-8.203	-4.823	-3.151	-3.185	-4.926
15	55.7	-8.591	-5.058	-3.231	-3.112	-4.699
16	57.2	-9.268	-5.479	-3.397	-3.021	-4.352
17	57.6	-9.455	-5.598	-3.447	-3.003	-4.266
18	60.3	-10.789	-6.471	-3.86	-2.955	-3.757
19	60.9	-11.102	-6.682	-3.968	-2.961	-3.661
20	65.3	-13.588	-8.416	-4.952	-3.194	-3.143
		-139.79	**-87.13**	**-68.62**	**-84.23**	**-133.98**

Possible population means for var2



Audition different parameters to quantify how well the proposed values explain the observed data.

Green dotted lines represent observed values for var2.

Note. Listwise var2 *M* = 41.98.

ID	var2	μ = 30	μ = 40	μ = 50	μ = 60	μ = 70
1	40	-5.252	-2.341	-5.252	-13.984	-28.538
2	40	-5.252	-2.341	-5.252	-13.984	-28.538
3	35	-3.068	-3.068	-8.89	-20.533	-37.998
4	43	-7.26	-2.603	-3.767	-10.753	-23.561
5	42.6	-6.962	-2.538	-3.935	-11.154	-24.194
6	39	-4.698	-2.37	-5.863	-15.177	-30.314
7	45	-8.89	-3.068	-3.068	-8.89	-20.533
8	45.2	-9.066	-3.128	-3.011	-8.717	-20.243
9	50	-13.984	-5.252	-2.341	-5.252	-13.984
10	40	-5.252	-2.341	-5.252	-13.984	-28.538
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
		-69.68	-29.05	-46.63	-122.43	-256.44

Reporting 📄 📄 📄

Step 3. Report methods and software.

> To address the issue of incomplete data, we used full information maximum likelihood (FIML) in R using the lavaan package (version 0.6-18). To account for non-normality in the data, we applied the robust standard error estimator by specifying the robust = TRUE option in the lavaan function. Additionally, we followed Graham's (2003) approach for incorporating auxiliary variables into the analysis. Specifically, we included five auxiliary variables: age, sex, race, education, and income to handle potential bias introduced by missing data.

Extended Reporting. Provide supplemental materials for details and code.

Myths 🤥

- FIML "fills in" missing data
- Only for SEM models
- Must have MAR to use MI/FIML
- Guaranteed better than MI
- Exogenous variables are always included
- No need with large samples

Any questions?