simsem: SIMulated Structural Equation Modeling in R

Alexander M. Schoemann
Sunthud Pornprasertmanit
Patrick J. Miller

Motivation

- A comprehensive framework to simulate data, analyze data, combine results, and play with simulation results.

Monte Carlo Simulations

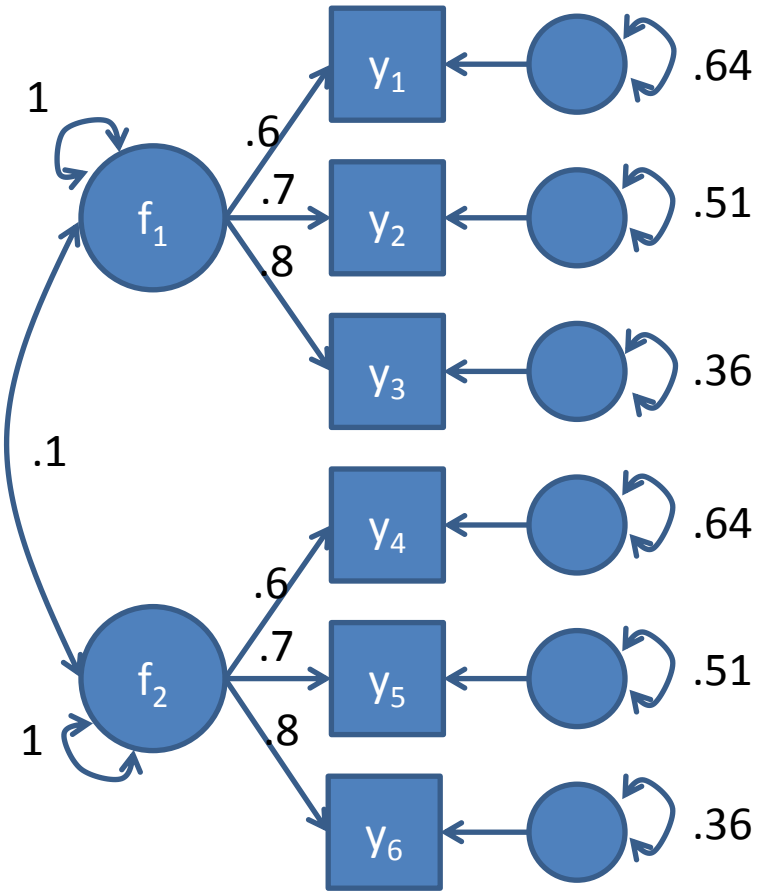
- Monte Carlo simulations are a popular tool for methodologists with many uses
  - Determine the accuracy of new methods
  - Compare different methods
  - Perform power analyses
  - Determine model fit in SEM

Monte Carlo Simulations

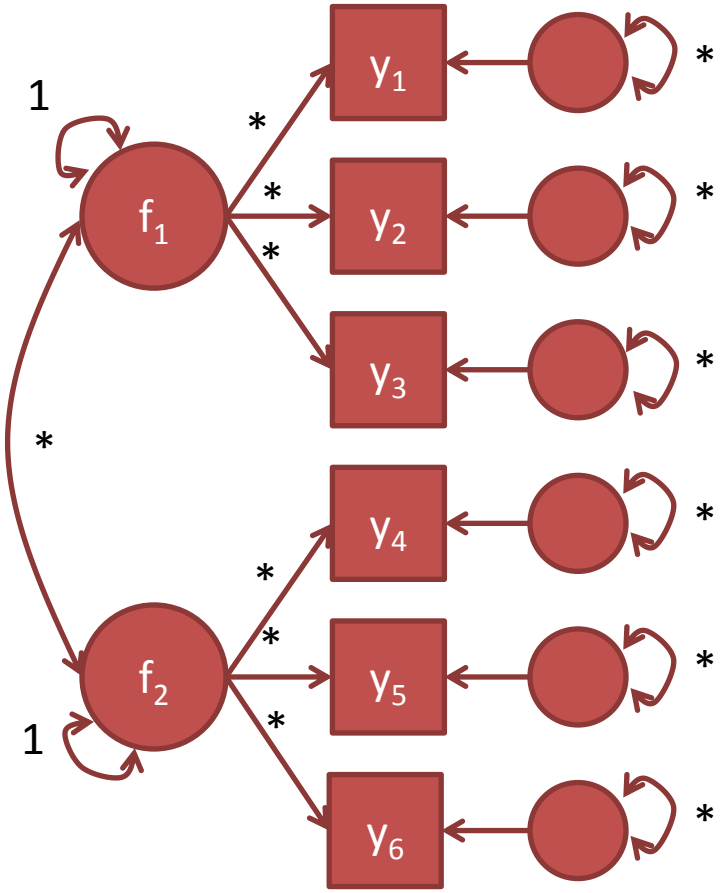
- General steps in a Monte Carlo Simulation
  1. Specify population parameters
  2. Create a sample of size N, based on population parameters
  3. Analyze sample data from step 2 with chosen statistical method(s).
  4. Repeat steps 2 and 3 for each of r replications.

Get Started

Data-Generating Model



Analysis Model



Get Started

```
library(lavaan)

popModel <- "
f1 =~ 0.6*y1 + 0.7*y2 + 0.8*y3
f2 =~ 0.6*y4 + 0.7*y5 + 0.8*y6
f1 ~~ 0.1*f2
f1 ~~ 1*f1
f2 ~~ 1*f2
y1 ~~ 0.64*y1
y2 ~~ 0.51*y2
y3 ~~ 0.36*y3
y4 ~~ 0.64*y4
y5 ~~ 0.51*y5
y6 ~~ 0.36*y6
"

data <- simulateData(popModel, sample.nobs = 200)
```

Get Started

```
analyzeModel <- "
f1 =~ y1 + y2 + y3
f2 =~ y4 + y5 + y6
"

data <- cfa(analyzeModel, data = data, std.lv = TRUE)

# Use simsem to simulate and analyze multiple data sets

library(simsem)

Output1 <- sim(1000, analyzeModel, n=200, generate=popModel,
   lavaanfun = "cfa", std.lv=TRUE)

summary(Output1)
```

```
> summary(Output1)
RESULT OBJECT
Model Type
[1] "lavaan"
========= Fit Indices Cutoffs ============
           Alpha
Fit Indices      0.1     0.05     0.01    0.001     Mean     SD
      chisq   13.705   15.625   20.026   25.550    8.095  4.074
      aic   3199.306 3217.672 3253.027 3292.959 3135.957 50.059
      bic   3242.185 3260.550 3295.905 3335.837 3178.835 50.059
      rmsea    0.060    0.069    0.087    0.105    0.019  0.026
      cfi      0.980    0.973    0.958    0.927    0.994  0.010
      tli      0.963    0.949    0.920    0.864    1.000  0.028
      srmr     0.047    0.051    0.058    0.069    0.033  0.010
```

========= Parameter Estimates and Standard Errors ============

	Estimate.Average	Estimate.SD	Average.SE	Power..Not.equal.0.	Std.Est	Std.Est.SD	Average.Param	Average.Bias	Coverage
f1=~y1	0.605	0.074	0.075	1.000	0.605	0.058	0.60	0.005	0.949
f1=~y2	0.696	0.076	0.076	1.000	0.699	0.058	0.70	-0.004	0.946
f1=~y3	0.796	0.076	0.078	1.000	0.799	0.057	0.80	-0.004	0.946
f2=~y4	0.598	0.074	0.075	1.000	0.600	0.059	0.60	-0.002	0.952
f2=~y5	0.699	0.077	0.076	1.000	0.699	0.059	0.70	-0.001	0.956
f2=~y6	0.801	0.080	0.078	1.000	0.801	0.060	0.80	0.001	0.938
y1~~y1	0.628	0.077	0.079	1.000	0.631	0.070	0.64	-0.012	0.943
y2~~y2	0.503	0.083	0.081	1.000	0.509	0.081	0.51	-0.007	0.938
y3~~y3	0.354	0.090	0.090	0.946	0.359	0.092	0.36	-0.006	0.959
y4~~y4	0.631	0.077	0.079	1.000	0.637	0.070	0.64	-0.009	0.950
y5~~y5	0.505	0.081	0.082	0.999	0.508	0.082	0.51	-0.005	0.953
y6~~y6	0.351	0.093	0.091	0.932	0.354	0.096	0.36	-0.009	0.960
f1~~f2	0.198	0.091	0.088	0.601	0.198	0.091	0.20	-0.002	0.943

```
========= Correlation between Fit Indices ============
       chisq    aic    bic  rmsea    cfi    tli   srmr
chisq  1.000  0.013  0.013  0.941 -0.909 -0.991  0.853
aic    0.013  1.000  1.000  0.011 -0.035 -0.011 -0.005
bic    0.013  1.000  1.000  0.011 -0.035 -0.011 -0.005
rmsea  0.941  0.011  0.011  1.000 -0.932 -0.931  0.788
cfi   -0.909 -0.035 -0.035 -0.932  1.000  0.909 -0.741
tli   -0.991 -0.011 -0.011 -0.931  0.909  1.000 -0.851
srmr   0.853 -0.005 -0.005  0.788 -0.741 -0.851  1.000
================== Replications =====================
Number of replications = 1000
Number of converged replications = 997
Number of nonconverged replications:
   1. Nonconvergent Results = 0
   2. Nonconvergent results from multiple imputation = 0
   3. At least one SE were negative or NA = 0
   4. At least one variance estimates were negative = 3
   5. At least one correlation estimates were greater than 1 or less than -1 = 0
```

Play with the output

- `summaryFit(Output1)`
- `summaryParam(Output1)`
- `getCutoff(Output1, alpha=0.05)`
- `plotCutoff(Output1)`
- `summaryConverge(Output1)`

- See `class?SimResult` for available functions

Varying Sample Sizes

```
Output2 <- sim(NULL, analyzeModel, n=100:1000, generate=popModel,
   lavaanfun = "cfa", std.lv=TRUE)
summary(Output2)
```

```
> summary(Output2)
RESULT OBJECT
Model Type
[1] "lavaan"
========= Fit Indices Cutoffs ============
     N  chisq       aic       bic rmsea   cfi   tli  srmr
1  100 15.705  1629.736  1668.456 0.073 0.973 0.950 0.054
2  325 15.905  5182.826  5228.549 0.061 0.979 0.961 0.045
3  550 16.106  8735.916  8788.642 0.049 0.985 0.973 0.036
4  775 16.307 12289.006 12348.734 0.037 0.991 0.984 0.028
5 1000 16.508 15842.096 15908.827 0.024 0.997 0.995 0.019
```

========== Parameter Estimates and Standard Errors ============

	Estimate.Average	Estimate.SD	Average.SE	Power..Not.equal.0.	Std.Est	Std.Est.SD	Average.Param	Average.Bias	Coverage	r_coef.n	r_se.n
f1=~y1	0.604	0.053	0.051	1.000	0.604	0.042	0.60	0.004	0.943	-0.023	-0.906
f1=~y2	0.698	0.056	0.052	1.000	0.699	0.043	0.70	-0.002	0.940	-0.002	-0.901
f1=~y3	0.798	0.056	0.053	1.000	0.800	0.042	0.80	-0.002	0.943	-0.032	-0.896
f2=~y4	0.598	0.053	0.051	1.000	0.598	0.043	0.60	-0.002	0.949	0.056	-0.905
f2=~y5	0.697	0.054	0.052	1.000	0.698	0.042	0.70	-0.003	0.958	0.071	-0.899
f2=~y6	0.801	0.055	0.053	1.000	0.803	0.041	0.80	0.001	0.947	0.018	-0.880
y1~~y1	0.633	0.056	0.054	1.000	0.634	0.050	0.64	-0.007	0.939	0.016	-0.882
y2~~y2	0.507	0.060	0.055	0.999	0.510	0.060	0.51	-0.003	0.952	-0.001	-0.861
y3~~y3	0.356	0.065	0.061	0.982	0.359	0.067	0.36	-0.004	0.951	0.076	-0.829
y4~~y4	0.637	0.057	0.054	1.000	0.640	0.051	0.64	-0.003	0.942	-0.024	-0.885
y5~~y5	0.509	0.059	0.055	1.000	0.511	0.059	0.51	-0.001	0.938	-0.027	-0.874
y6~~y6	0.352	0.064	0.062	0.974	0.354	0.067	0.36	-0.008	0.952	0.125	-0.797
f1~~f2	0.196	0.066	0.060	0.857	0.196	0.066	0.20	-0.004	0.939	0.048	-0.916

```
========= Correlation between Fit Indices ============
       chisq    aic    bic  rmsea    cfi    tli   srmr      n
chisq  1.000 -0.008 -0.008  0.866 -0.711 -0.808  0.556 -0.008
aic   -0.008  1.000  1.000 -0.218  0.289  0.023 -0.661  1.000
bic   -0.008  1.000  1.000 -0.218  0.289  0.023 -0.661  1.000
rmsea  0.866 -0.218 -0.218  1.000 -0.916 -0.877  0.701 -0.218
cfi   -0.711  0.289  0.289 -0.916  1.000  0.866 -0.707  0.290
tli   -0.808  0.023  0.023 -0.877  0.866  1.000 -0.561  0.023
srmr   0.556 -0.661 -0.661  0.701 -0.707 -0.561  1.000 -0.661
n     -0.008  1.000  1.000 -0.218  0.290  0.023 -0.661  1.000
================== Replications =====================
Number of replications = 901
Number of converged replications = 900
Number of nonconverged replications:
   1. Nonconvergent Results = 0
   2. Nonconvergent results from multiple imputation = 0
   3. At least one SE were negative or NA = 0
   4. At least one variance estimates were negative = 1
   5. At least one correlation estimates were greater than 1 or less than -1 = 0
NOTE: The sample size is varying.
```

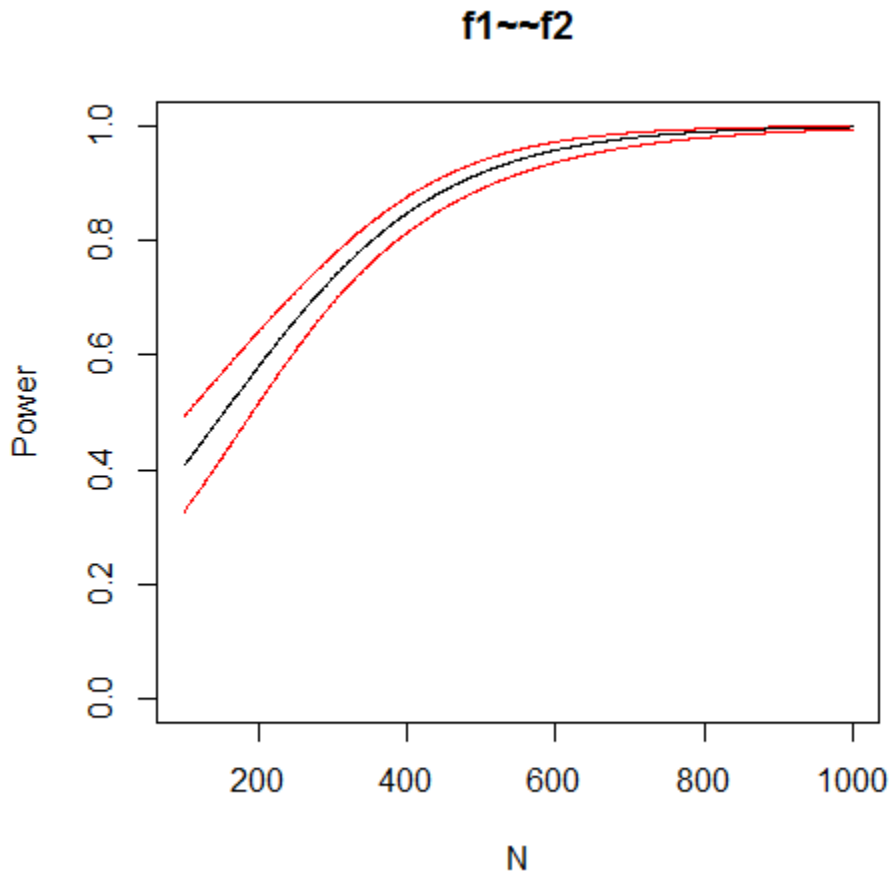
Varying Sample Sizes

```
powTable2 <- getPower(Output2)
findPower(powTable2, "N", 0.80)
```

```
> findPower(powTable2, "N", 0.80)
f1=~y1 f1=~y2 f1=~y3 f2=~y4 f2=~y5 f2=~y6 y1~~y1 y2~~y2 y3~~y3 y4~~y4 y5~~y5 y6~~y6 f1~~f2
   Inf    Inf    Inf    Inf    Inf    Inf    Inf    100    100    Inf    Inf    132    354
```

Varying Sample Sizes

```
plotPower(Output2, powerParam = "f1~~f2")
```



Missing Data

```
missModel <- miss(pmMCAR = 0.1, numImps = 10)
Output3 <- sim(NULL, analyzeModel, n=100:1000, generate=popModel,
   miss=missModel, lavaanfun = "cfa", std.lv=TRUE)
summary(Output3)
```

```
> summary(Output3)
RESULT OBJECT
Model Type
[1] "lavaan"
========= Fit Indices Cutoffs ============
     N  chisq       aic       bic rmsea   cfi   tli  srmr
1  100 15.343  1495.381  1556.060 0.070 0.968 0.941 0.053
2  325 15.587  4726.438  4797.084 0.059 0.976 0.954 0.044
3  550 15.830  7957.496  8038.109 0.048 0.983 0.968 0.035
4  775 16.074 11188.554 11279.133 0.037 0.990 0.981 0.027
5 1000 16.317 14419.612 14520.158 0.026 0.997 0.994 0.018
========= Parameter Estimates and Standard Errors ============
       Estimate.Average Estimate.SD Average.SE Power..Not.equal.0. Std.Est Std.Est.SD Average.FMI1 SD.FMI1 r_coef.n r_se.n
f1=~y1            0.604       0.059      0.055                1.000   0.604      0.047        0.139   0.025   -0.018 -0.901
f1=~y2            0.697       0.060      0.056                1.000   0.699      0.048        0.151   0.029   -0.006 -0.892
f1=~y3            0.799       0.060      0.058                1.000   0.800      0.046        0.166   0.032   -0.039 -0.875
f2=~y4            0.599       0.057      0.055                1.000   0.600      0.046        0.140   0.028    0.022 -0.897
f2=~y5            0.698       0.062      0.056                1.000   0.699      0.048        0.152   0.030    0.040 -0.891
f2=~y6            0.801       0.061      0.058                1.000   0.802      0.047        0.167   0.033    0.013 -0.868
y1~~y1            0.632       0.062      0.058                1.000   0.633      0.057        0.156   0.023    0.052 -0.874
y2~~y2            0.506       0.067      0.061                0.996   0.510      0.067        0.191   0.032    0.013 -0.843
y3~~y3            0.355       0.072      0.069                0.961   0.358      0.074        0.217   0.035    0.079 -0.789
y4~~y4            0.635       0.062      0.059                1.000   0.638      0.056        0.157   0.024    0.020 -0.873
y5~~y5            0.507       0.065      0.061                0.999   0.510      0.067        0.191   0.032    0.013 -0.850
y6~~y6            0.351       0.075      0.070                0.961   0.354      0.077        0.219   0.036    0.116 -0.776
f1~~f2            0.196       0.068      0.063                0.832   0.196      0.068        0.096   0.030    0.049 -0.906
y1~1              0.003       0.054      0.050                0.059   0.003      0.054        0.076   0.012    0.045 -0.899
y2~1              0.000       0.052      0.050                0.052   0.000      0.053        0.070   0.012    0.029 -0.900
y3~1              0.002       0.054      0.050                0.052   0.002      0.054        0.066   0.011    0.014 -0.898
y4~1             -0.002       0.051      0.050                0.047  -0.002      0.052        0.077   0.012    0.017 -0.903
y5~1             -0.002       0.052      0.050                0.052  -0.002      0.053        0.071   0.012    0.072 -0.901
y6~1             -0.001       0.052      0.050                0.051  -0.001      0.052        0.066   0.011    0.007 -0.899
```

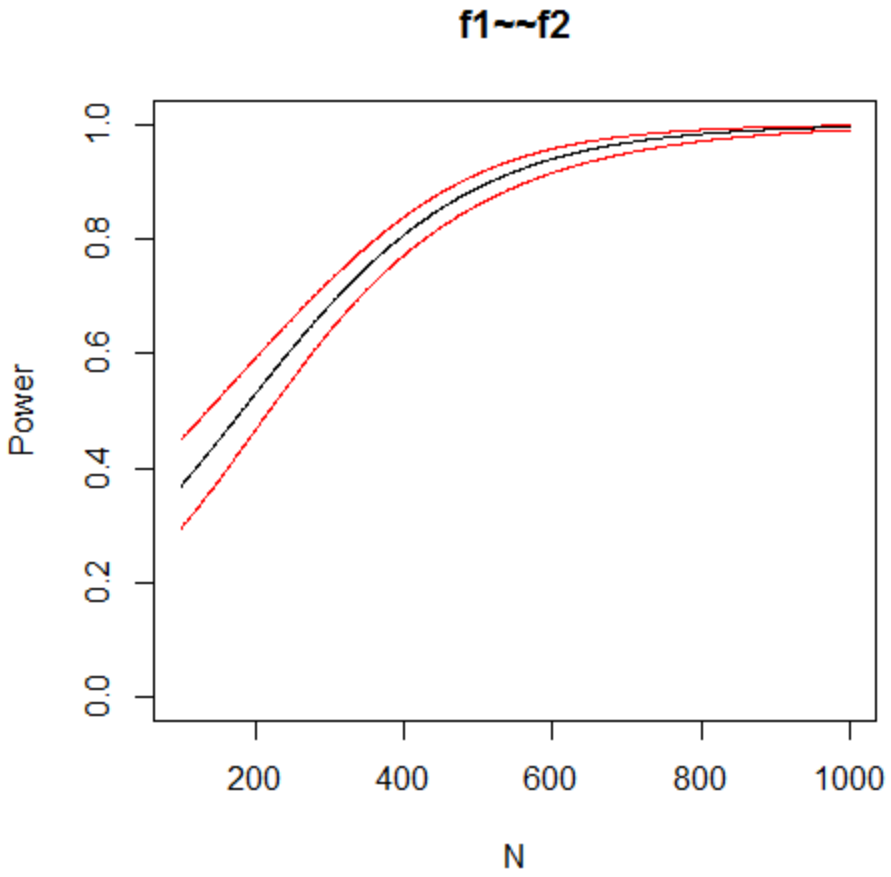
Missing Data

```
powTable3 <- getPower(Output3)
findPower(powTable3, "N", 0.80)
```

```
> findPower(powTable3, "N", 0.80)
f1=~y1 f1=~y2 f1=~y3 f2=~y4 f2=~y5 f2=~y6 y1~~y1 y2~~y2 y3~~y3 y4~~y4 y5~~y5 y6~~y6 f1~~f2
   Inf    Inf    Inf    Inf    Inf    Inf    Inf    107    167    Inf    100    164    394
```

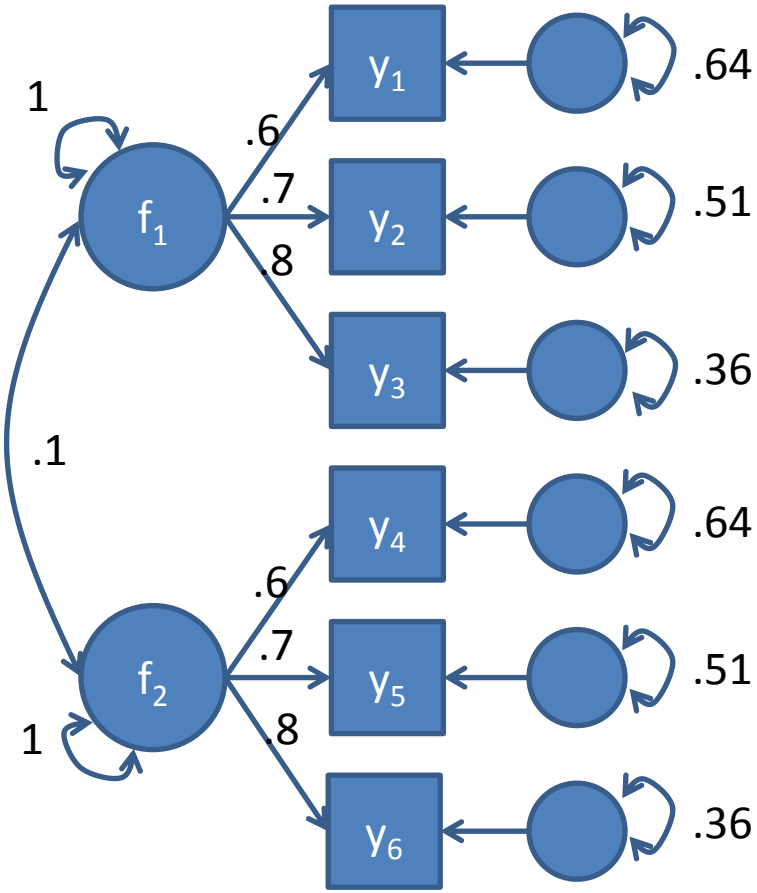
Missing Data

```
plotPower(Output3, powerParam = "f1~~f2")
```

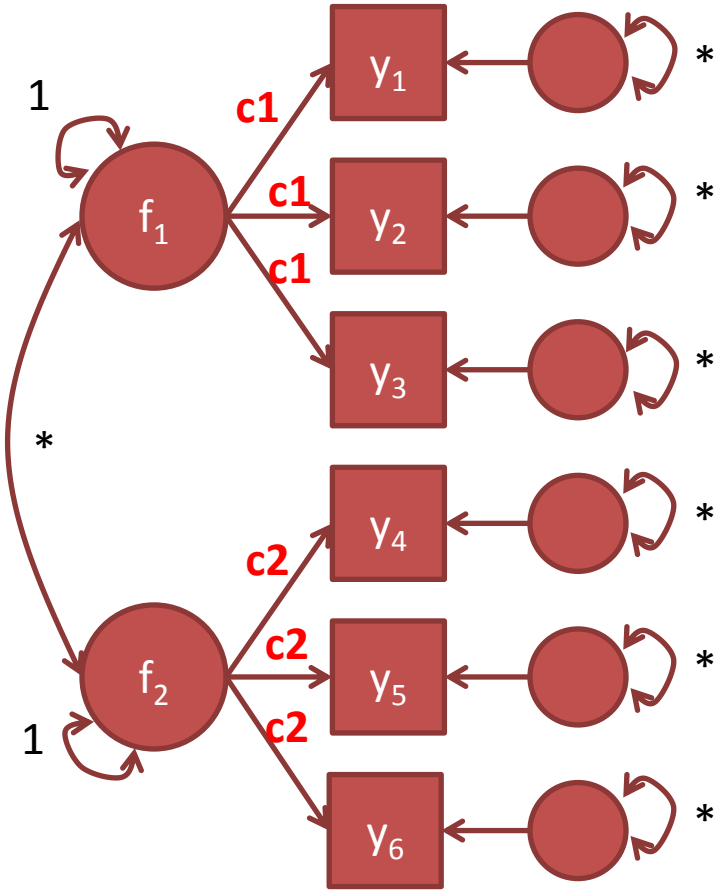


Power in Rejecting Misspecified Model

Data-Generating Model



Analysis Model



Power in Rejecting Misspecified Model

```
wrongModel <- "
f1 =~ c1*y1 + c1*y2 + c1*y3
f2 =~ c2*y4 + c2*y5 + c2*y6
"

Output4 <- sim(NULL, wrongModel, n=100:1000, generate=popModel,
   lavaanfun = "cfa", std.lv=TRUE)
summary(Output4)
```

```
> summary(Output4)
RESULT OBJECT
Model Type
[1] "lavaan"
========= Fit Indices Cutoffs ============
     N  chisq       aic       bic rmsea   cfi   tli  srmr
1  100 28.447  1630.049  1656.919 0.091 0.930 0.913 0.089
2  325 40.215  5191.432  5222.869 0.086 0.937 0.921 0.081
3  550 51.983  8752.815  8788.819 0.081 0.944 0.930 0.074
4  775 63.750 12314.198 12354.768 0.076 0.951 0.939 0.066
5 1000 75.518 15875.581 15920.718 0.070 0.958 0.947 0.059
========= Parameter Estimates and Standard Errors ============
       Estimate.Average Estimate.SD Average.SE Power..Not.equal.0. Std.Est Std.Est.SD r_coef.n r_se.n
c1                0.704       0.036      0.033                1.00   0.677      0.024   -0.026 -0.905
c2                0.703       0.034      0.032                1.00   0.676      0.025    0.075 -0.907
y1~~y1            0.585       0.054      0.051                1.00   0.541      0.033   -0.003 -0.877
y2~~y2            0.496       0.048      0.046                1.00   0.500      0.037   -0.020 -0.886
y3~~y3            0.443       0.045      0.043                1.00   0.471      0.040    0.053 -0.884
y4~~y4            0.587       0.054      0.051                1.00   0.542      0.034   -0.040 -0.880
y5~~y5            0.498       0.048      0.046                1.00   0.501      0.036   -0.037 -0.885
y6~~y6            0.442       0.044      0.042                1.00   0.472      0.039    0.107 -0.880
f1~~f2            0.199       0.068      0.062                0.85   0.199      0.068    0.052 -0.913
========= Correlation between Fit Indices ============
       chisq    aic    bic  rmsea    cfi    tli   srmr      n
chisq  1.000  0.699  0.699  0.652 -0.597 -0.605  0.215  0.698
aic    0.699  1.000  1.000  0.111  0.011 -0.016 -0.428  1.000
bic    0.699  1.000  1.000  0.111  0.011 -0.016 -0.428  1.000
rmsea  0.652  0.111  0.111  1.000 -0.944 -0.958  0.711  0.109
cfi   -0.597  0.011  0.011 -0.944  1.000  0.992 -0.789  0.013
tli   -0.605 -0.016 -0.016 -0.958  0.992  1.000 -0.767 -0.014
srmr   0.215 -0.428 -0.428  0.711 -0.789 -0.767  1.000 -0.430
n      0.698  1.000  1.000  0.109  0.013 -0.014 -0.430  1.000
================== Replications =====================
Number of replications = 901
Number of converged replications = 901
Number of nonconverged replications:
   1. Nonconvergent Results = 0
   2. Nonconvergent results from multiple imputation = 0
   3. At least one SE were negative or NA = 0
   4. At least one variance estimates were negative = 0
   5. At least one correlation estimates were greater than 1 or less than -1 = 0
NOTE: The sample size is varying.
NOTE: The data generation model is not the same as the analysis model. See the summary of the population
```

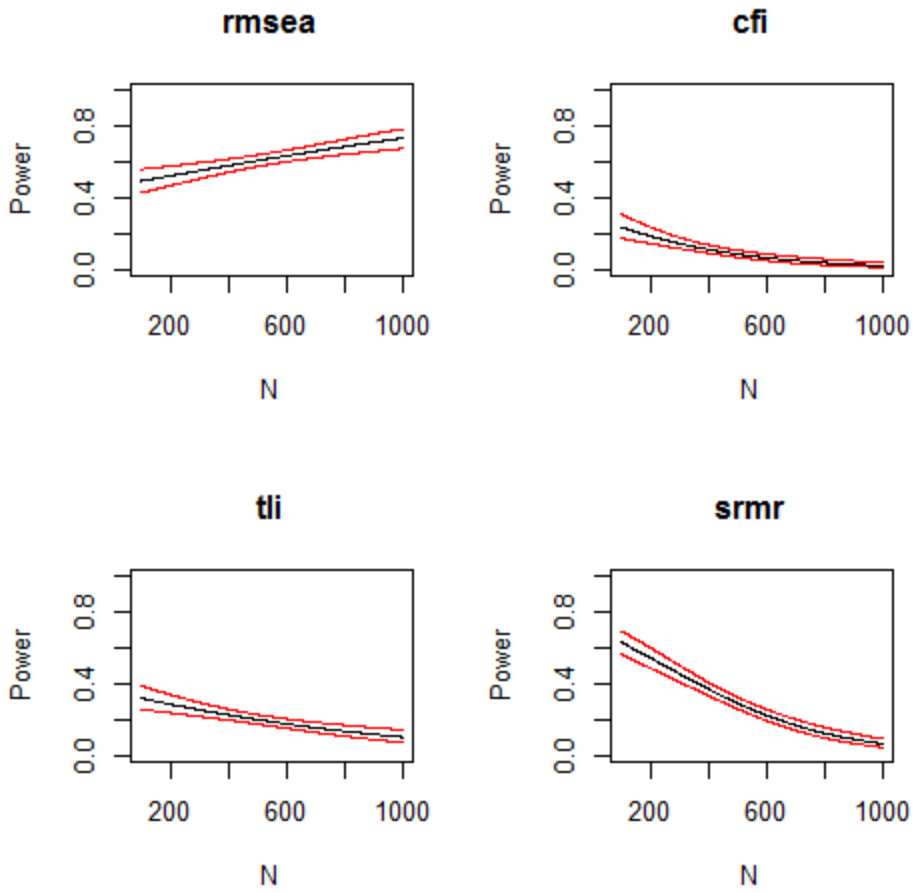
Power in Rejecting Misspecified Model

```
cutoff <- c(RMSEA = 0.05, CFI = 0.95, TLI = 0.95, SRMR = 0.06)
getPowerFit(Output4, cutoff=cutoff, nVal=200)
```

```
> getPowerFit(Output4, cutoff=cutoff, nVal=200)
       cfi        tli      rmsea       srmr
0.1715043 0.2687876 0.5389233 0.4957211
```

Power in Rejecting Misspecified Model

```
plotPowerFit(Output4, cutoff=cutoff)
```



Other Features

- Generate/Analyze Data with lavaan, OpenMx, or user-specified functions
- Can request additional output for each data analysis and save as a list
- Other types of missing data, such as MAR, dropout, or planned missing data
  - Missing data can be handled with FIML or multiple imputation through Amelia or mice

Other Features

- Nonnormal distributions
- Nested or nonnested model comparisons
- Use fixed values of covariates to generate data
- Parallel processing
- And many more…

Questions?

- Thanks to:
  - Paul Johnson
  - Todd Little
  - Yves Rosseel
  - And all the contributors to simsem development

simsem: simsem.org

example code available at: simsem.org