UNIVERSITY OF NOTRE DAME

Combining R with Psychology

-----An illustration with SEM

江歌
Ge Jiang
University Of Notre Dame
gjiang2@nd.edu

The 7th R conference, May 24-25th

Outline

Why Psychologists Need R

- The Argument of Science
- Qualitative vs Quantitative

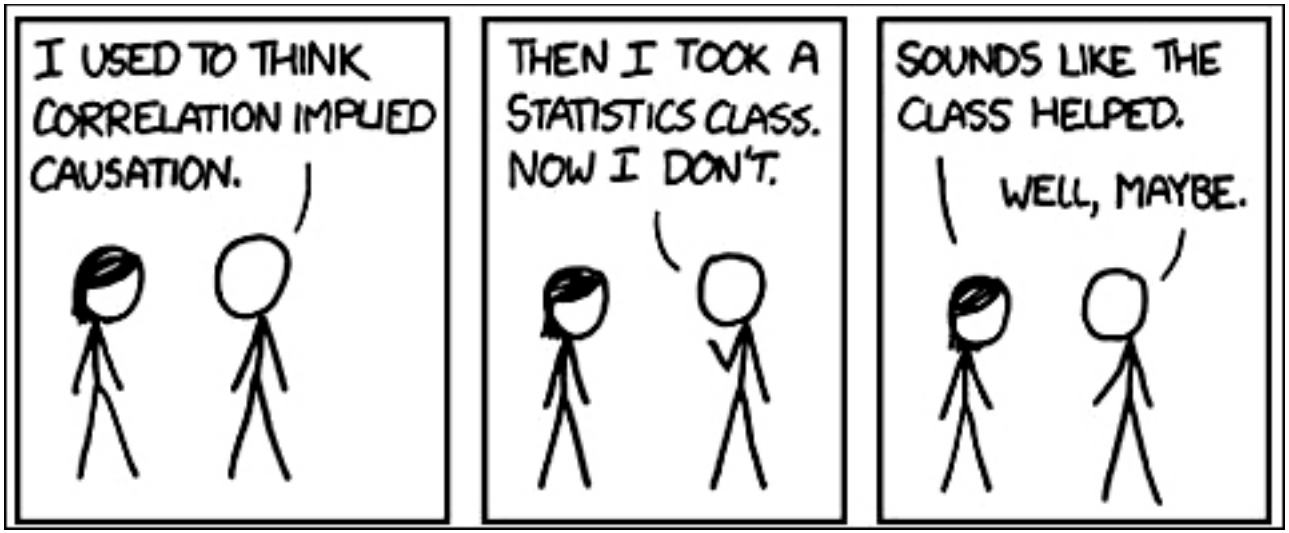
SEM

- Models
- R packages in SEM
- An illustration Example

Ongoing Project With R

- Test Statistics
- Simulation Design and Data
- Results

The Argument of Science



After we are able to:

- Quantify human or animal behavior with systematic and objective methods
- Test the hypotheses we put out
- Repeatability of psychological experiments

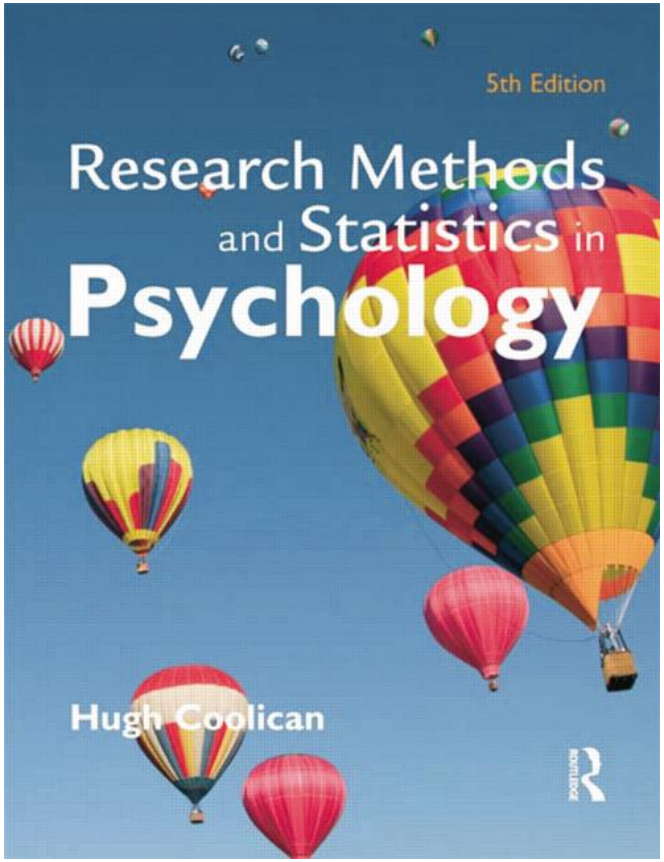
Qualitative vs Quantitative

Need:

- Not every psychological question can be solved with SPSS.

Must:

- Developing theory
- Latent Variable
- More complex models?



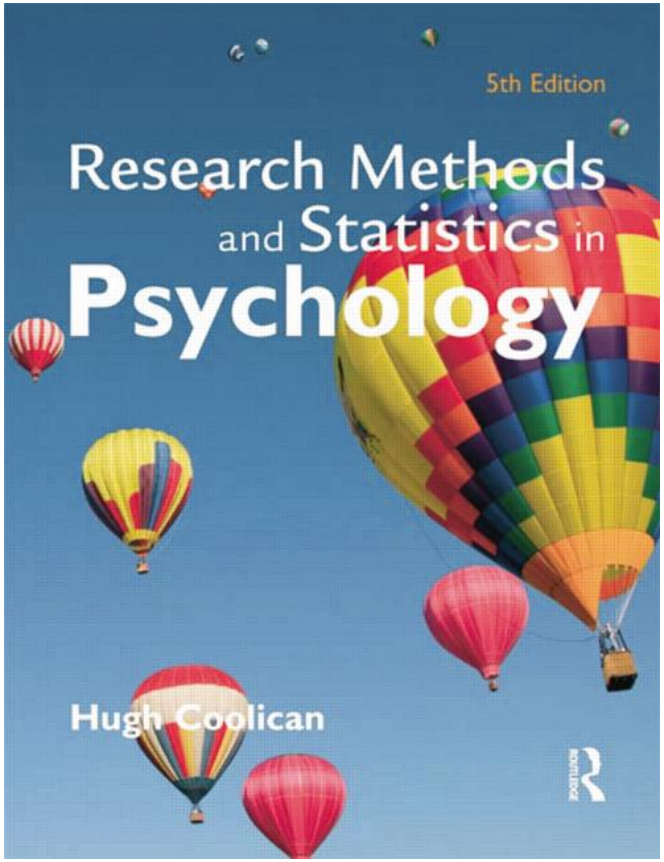
Qualitative vs Quantitative

Need:

- Not every psychological question can be solved with SPSS.

Must:

- Developing theory
- Latent Variable
- More complex models?



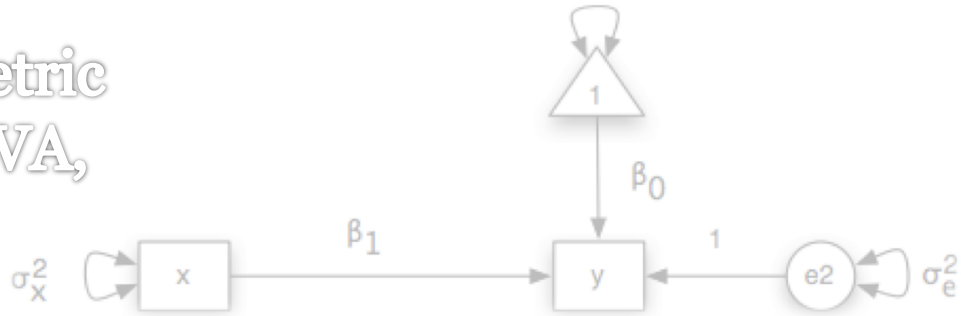
What is SEM?

Structural equation modelling (SEM) is a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions(Wright, 1921).

Why?

- Some complex ideas or hypotheses
- Integrated most parametric models, including ANOVA, Linear Regression, and Factor Analysis

Regression Model

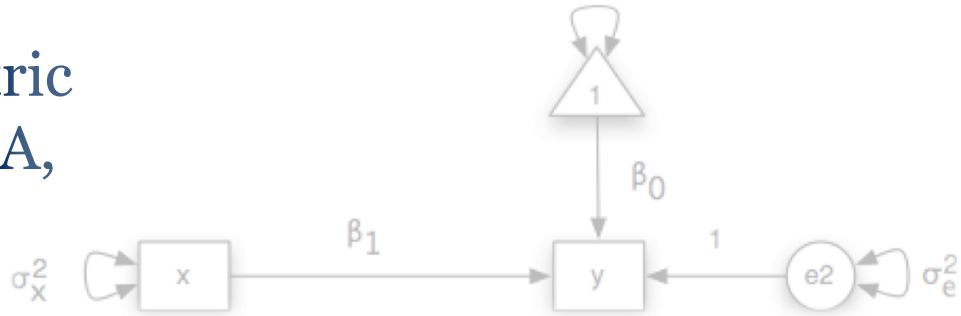


What is SEM?

Structural equation modelling (SEM) is a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions(Wright, 1921).

Why?

- Some complex ideas or hypotheses
- Integrated most parametric models, including ANOVA, Linear Regression, and Factor Analysis



What is SEM?

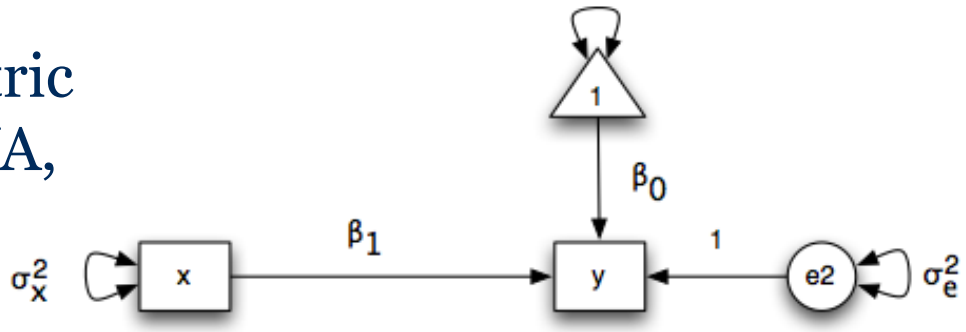
Structural equation modelling (SEM) is a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions(Wright, 1921).

Why?

- Some complex ideas or hypotheses
- Integrated most parametric models, including ANOVA, Linear Regression, and Factor Analysis

Regression Model

$$y = \beta_0 + \beta_1 x + e$$



R packages in SEM

Software:

- R
- Rstudio
- WebSEM

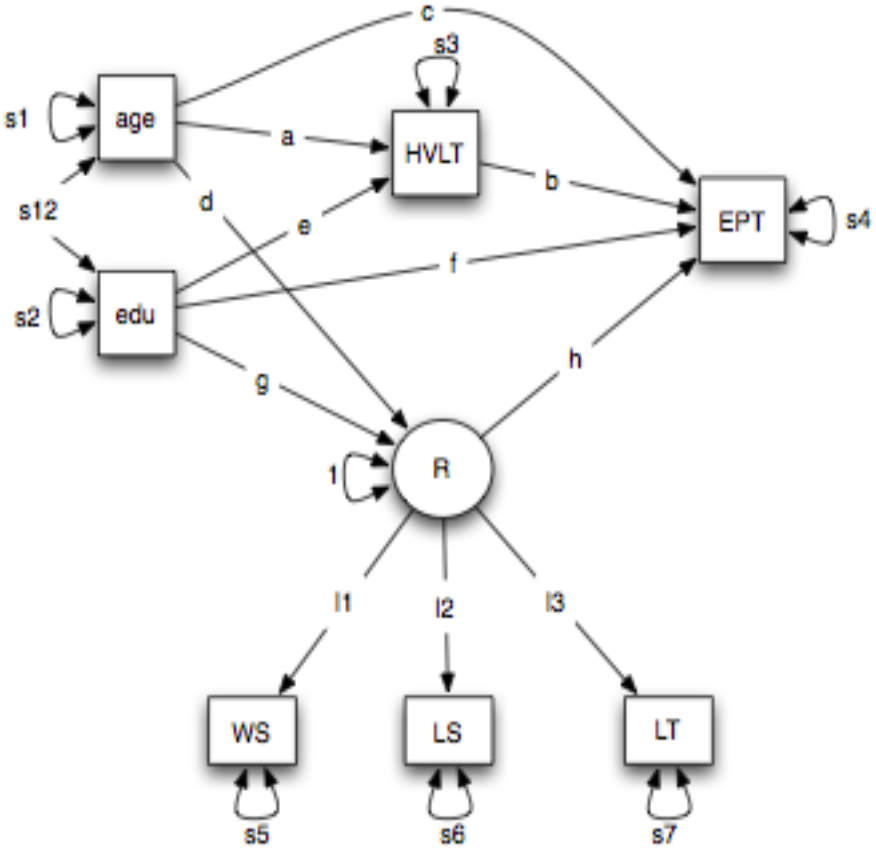
Packages for Testing Models:

- library(lavaan)
- library(sem)
- library(rsem)

Ancillary Packages

- library(RAMpath, Mi, Psych, bmem)

An illustration Example



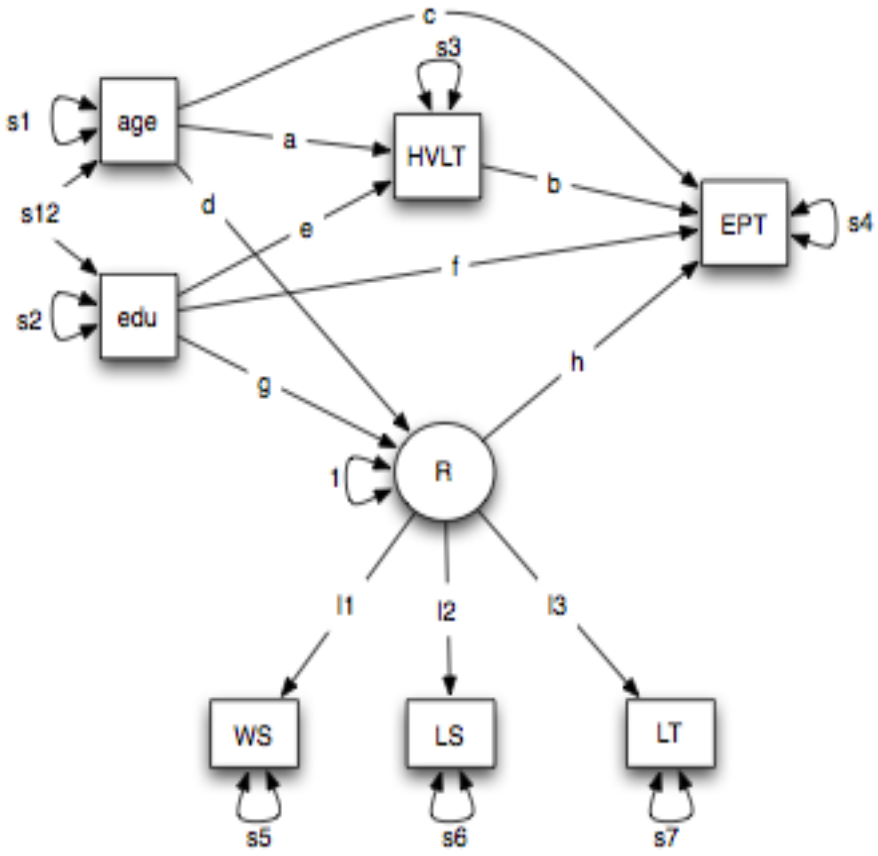
Data

- ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly) study.

Variables:

- HVLT: Hopkins Verbal Learning Test
- EPT: Everyday Problems Test
- WS: Word Series
- LS: Letter Series
- LT: Letter Sets
- R: Reasoning ability*

An illustration Example




```
library(lavaan)
model.lavaan<-'
R =~ ws + ls + lt   #=~ factor model
R ~ age + edu       #~ regression
hvltt ~ age +edu    # two mediators
ept ~ R + hvltt + age + edu
ab:=a*b             #defining parameter
'                   #mediation model
res.lavaan<-sem(model.lavaan,
data=dset, std.lv=T)  #std: give s.e.
summary(res.lavaan, fit=T)
mod.lavaan<-modindices(res.lavaan,
standard=F)   #model modification
```

An illustration Example

R Output:

- lavaan (0.5-16) converged normally after 36 iterations
- Number of observations 99

- Estimator ML
- Minimum Function Test Statistic 50.734
- Degrees of freedom 9
- P-value (Chi-square) 0.000

- Model test baseline model:
- Minimum Function Test Statistic 405.887
- Degrees of freedom 20
- P-value 0.000

- User model versus baseline model:
- Comparative Fit Index (CFI) 0.892
- Tucker-Lewis Index (TLI) 0.760

Loglikelihood and Information Criteria:

- Loglikelihood user model (H0) -1841.979
- Loglikelihood unrestricted model (H1) -1816.612

- Number of free parameters 16
- Akaike (AIC) 3715.958
- Bayesian (BIC) 3757.480
- Sample-size adjusted Bayesian (BIC) 3706.951

- Root Mean Square Error of Approximation:

- RMSEA 0.216
- 90 Percent Confidence Interval 0.161 0.276
- P-value RMSEA <= 0.05 0.000
- Standardized Root Mean Square Residual:

- SRMR 0.135

An illustration Example

R Output:

	Estimate	Std.err	Z-value	P(>|z|)
Latent variables:				
R =~				
ws	3.570	0.312	11.460	0.000
ls	3.942	0.343	11.479	0.000
lt	1.522	0.195	7.821	0.000
Regressions:				
R ~				
age	-0.065	0.020	-3.314	0.001
edu	0.207	0.042	4.925	0.000
hvltt ~				
age	-0.375	0.085	-4.434	0.000
edu	0.459	0.172	2.674	0.007
.....				

	Estimate	2.5%	97.5%
a	-0.375	-0.564	-0.179
c	-0.003	-0.209	0.156
d	-0.065	-0.106	-0.024
e	0.459	0.116	0.753
f	0.325	-0.048	0.673
g	0.206	0.121	0.297
b	0.348	0.138	0.560
h	2.355	1.398	3.521
a*b	-0.130	-0.244	-0.055
d*h	-0.153	-0.280	-0.054
e*b	0.160	0.040	0.354
g*h	0.485	0.259	0.832
a*b+d*h	-0.283	-0.424	-0.147
e*b+g*h	0.645	0.340	0.986

An illustration Example

R Output:

	Estimate	Std.err	Z-value	P(>|z|)
Latent variables:				
R =~				
ws	3.570	0.312	11.460	0.000
ls	3.942	0.343	11.479	0.000
lt	1.522	0.195	7.821	0.000
Regressions:				
R ~				
age	-0.065	0.020	-3.314	0.001
edu	0.207	0.042	4.925	0.000
hvltt ~				
age	-0.375	0.085	-4.434	0.000
edu	0.459	0.172	2.674	0.007
.....				

	Estimate	2.5%	97.5%
a	-0.375	-0.564	-0.179
c	-0.003	-0.209	0.156
d	-0.065	-0.106	-0.024
e	0.459	0.116	0.753
f	0.325	-0.048	0.673
g	0.206	0.121	0.297
b	0.348	0.138	0.560
h	2.355	1.398	3.521
a*b	-0.130	-0.244	-0.055
d*h	-0.153	-0.280	-0.054
e*b	0.160	0.040	0.354
g*h	0.485	0.259	0.832
a*b+d*h	-0.283	-0.424	-0.147
e*b+g*h	0.645	0.340	0.986

Mediation effect

An illustration Example

Model Modification

- > mod.lavaan
- lhs op rhs mi epc
- 1 R =~ ws 0.000 0.000
- 2 R =~ ls 0.000 0.000
- 3 R =~ lt 0.000 0.000
- 4 ws ~~ ws 0.000 0.000
- 5 ws ~~ ls 3.722 4.343
- 6 ws ~~ lt 0.135 0.226
- 7 ls ~~ ls 0.000 0.000
- 8 ls ~~ lt 2.997 -1.173
- 9 lt ~~ lt 0.000 0.000
- .....

```
library(lavaan)
model.new<-'
R =~ ws + ls + lt
R ~ age + edu
hvltt ~ age +edu
ept ~ R + hvltt + age + edu
ws~~ls          #added line
'

res.lavaan<-sem(model.new,
data=dset, std.lv=T)
summary(res.lavaan, fit=T)
```

An illustration Example

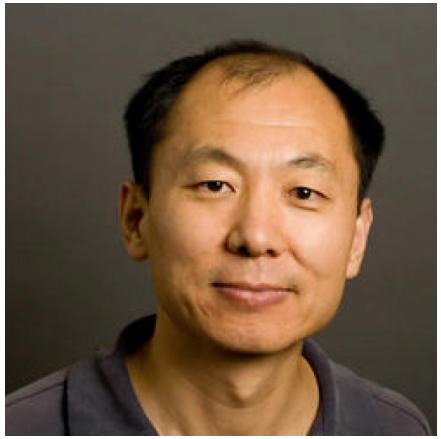
Model Modification

- > mod.lavaan
- lhs op rhs mi epc
- 1 R =~ ws 0.000 0.000
- 2 R =~ ls 0.000 0.000
- 3 R =~ lt 0.000 0.000
- 4 ws ~~ ws 0.000 0.000
- 5 ws ~~ ls 3.722 4.343
- 6 ws ~~ lt 0.135 0.226
- 7 ls ~~ ls 0.000 0.000
- 8 ls ~~ lt 2.997 -1.173
- 9 lt ~~ lt 0.000 0.000
- .....

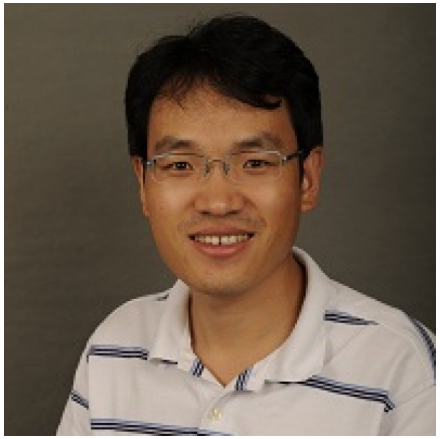
```
library(lavaan)
model.new<-'
R =~ ws + ls + lt
R ~ age + edu
hvltt ~ age +edu
ept ~ R + hvltt + age + edu
ws~~ls          #added line
'
res.lavaan<-sem(model.new,
data=dset, std.lv=T)
summary(res.lavaan, fit=T)
```

Library(rsem)

rsem is a public package that can be downloaded from CRAN which is specified for SEM, especially for doing **robust SEM** and handling **missing data**.



Ke-Hai Yuan



Johnny Zhang

Library(rsem)

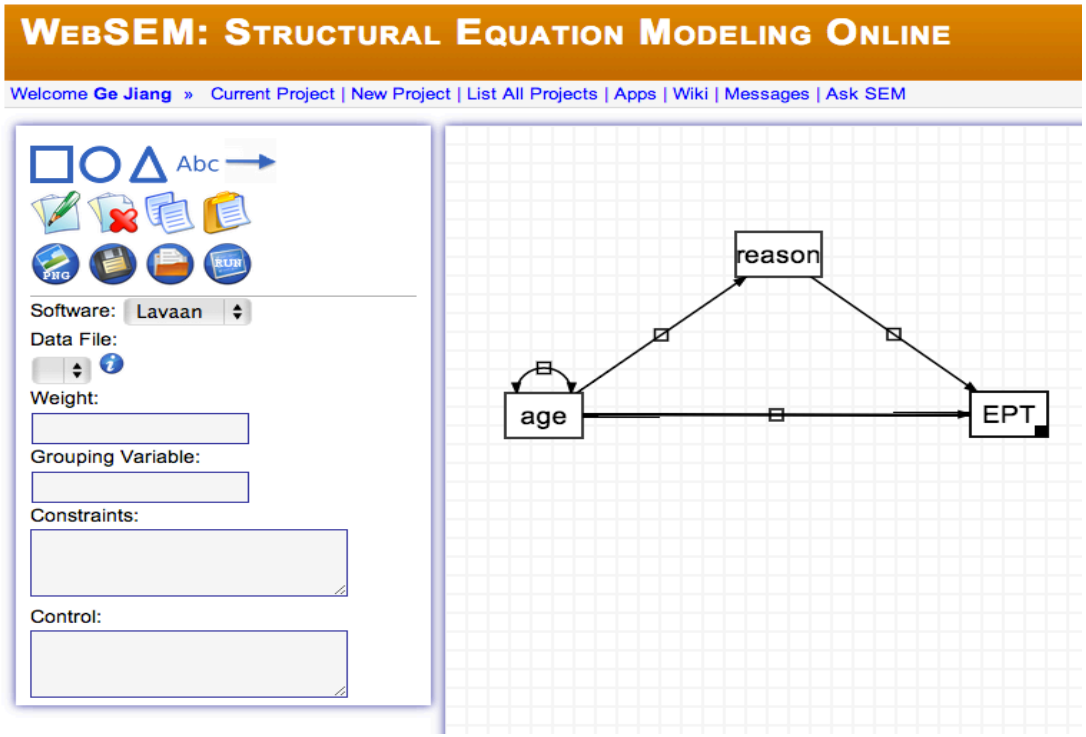
rsem is a public package that can be downloaded from CRAN which is specified for SEM, especially for doing **robust SEM** and handling **missing data**.

Help Pages

rsem-package	Robust Structural Equation Modeling with Missing Data and Auxiliary
mardiamv25	Simulated data
mardiamv25_contaminated	Simulated data
rsem	The main function for robust SEM analysis
rsem.Ascov	Sandwich-type covariance matrix
rsem.DP	Generate a duplication matrix
rsem.emmusig	Robust mean and covariance matrix using Huber-type weight
rsem.fit	Calculate robust test statistics
rsem.gname	Internal function

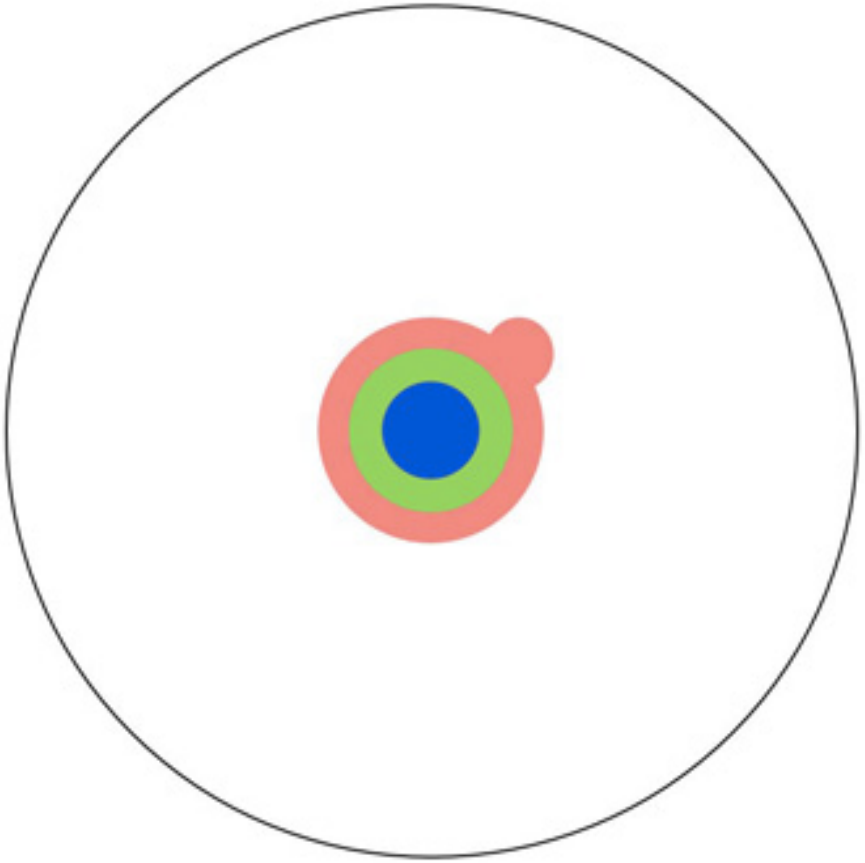
WebSEM

- WebSEM is a online website that is specified for using SEM with R cores.
- https://websem.psychstat.org/ (Registration required)



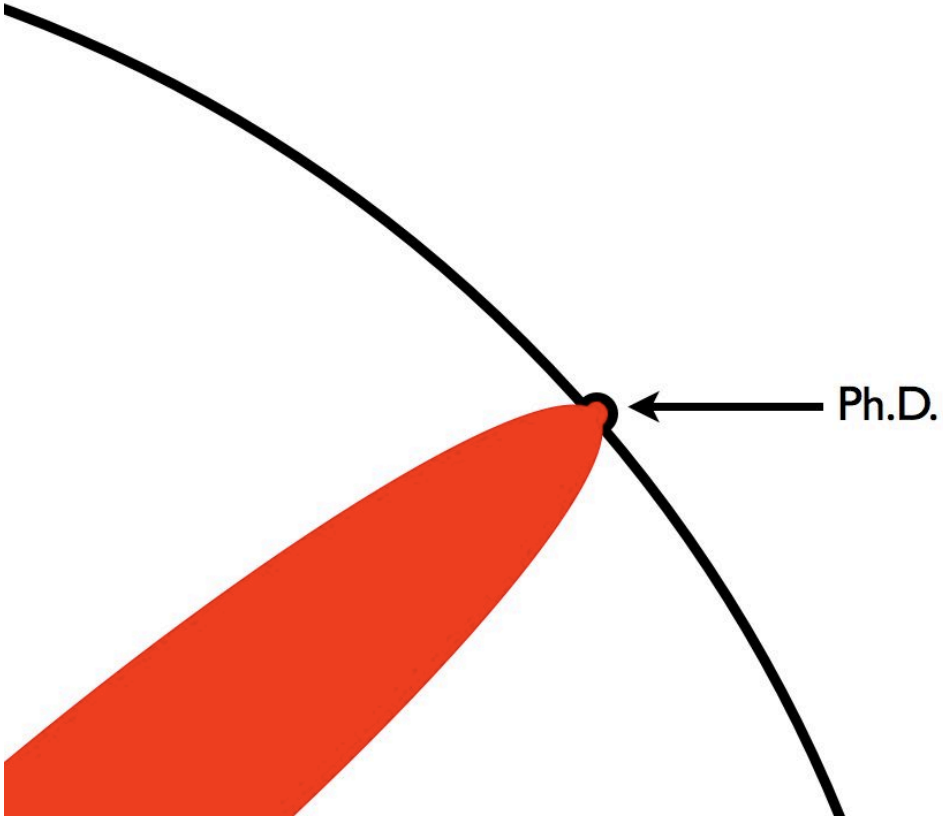
Ongoing Project

This is your knowledge circle:



Ongoing Project

This is what Ph.D does.....



Robust Test Statistics

Do you still remember the p-value we saw from R output?

```
lavaan (0.5-16) converged normally after  36 iterations

  Number of observations                            99

  Estimator                                         ML
  Minimum Function Test Statistic                50.734
  Degrees of freedom                                 9
  P-value (Chi-square)                           0.000
```

This p-value is crucial in determining the model fit.

H_0 : **The model closely fit the data.**

Robust Test Statistics

Assumptions:

- Multivariate normality assumption
- Large sample size

Chi-square statistic:

$$T_{ML} = n[tr(S\Sigma^{-1}) - log|S\Sigma^{-1}| - p]$$

asymptotically follows χ^2_{df} distribution.

In the case when both assumptions are violated…

Robust Test Statistics

Satorra-Benter Scaled Statistics:

$$T_{RML} = \tau^{-1} T_{ML}$$
$$\tau = tr(U\Gamma)/df$$

which is a scaling constant that corrects TML so that the mean of the sampling distribution of TML will be closer to the expected mean under the correct model.

Taking sample size into account:

$$c = tr(U\Gamma)/rank(U\Gamma)$$
$$T_3 = c^{-1} T_{ML}$$

$$m = \frac{1}{2}(\tau + c)$$
$$T_4 = m^{-1} T_{ML}$$

Simulation Design

3-Factor Model

$$X = \Lambda f + \varepsilon$$

Mean and Covariance Structure

$$E(X) = \Lambda\mu_f + \mu_\varepsilon \quad Cov(X) = \Sigma = \Lambda\Phi\Lambda' + \Psi$$

$$\Lambda = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1.0 & 03 & 0.4 \\ 0.3 & 1.0 & 0.5 \\ 0.4 & 0.5 & 1.0 \end{pmatrix}$$

where $\lambda' = (.70, .70, .75, .80, .80)$ and Ψ is a diagonal matrix chosen to make the diagonal elements in Σ all to be 1.

Rep=500

N=(50 , 55 , 60 , 65 , 70 , 75 , 80 , 85 , 90)

R codes:

```
lambda=matrix(rep(0,45),15)
lambda[1:5,1]=lambda[6:10,2]=lambda[11:15,3]=c(0.7,0.7,0.75,0.8,0.8)

Phir=matrix(c(1,0.3,0.4,0.3,1,0.5,0.4,0.5,1),3)

Psir=diag(rep(0,15),15)
diag(Psir)=1-diag(lambda%*%Phir%*%t(lambda)) ##fixing the variance of Xs to be 1
```

Simulation Design

f and ε variates are assumed to follow different combinations of distributions:

1. $X=\Lambda f+\varepsilon, f\sim N(0,\Phi), \varepsilon \sim N(0,\Psi)$
2. $X=(\Lambda f+\varepsilon)/r, f\sim N(0,\Phi), \varepsilon \sim N(0,\Psi), r\sim\sqrt{\chi^2_5/3}$
3. $X=(\Lambda f+\varepsilon)/r, f\sim N(0,\Phi), \varepsilon \sim Exp(0,\Psi), r\sim\sqrt{\chi^2_5/3}$
4. $X=(\Lambda f+\varepsilon)/r, f\sim Exp(0,\Phi), \varepsilon \sim N(0,\Psi), r\sim\sqrt{\chi^2_5/3}$
5. $X=(\Lambda f+\varepsilon)/r, f\sim Exp(0,\Phi), \varepsilon \sim Exp(0,\Psi), r\sim\sqrt{\chi^2_5/3}$

The rescaling constant is chosen because $E(\chi^2_5/3)=1$.

R codes:

```
Phir12=egvec(Phir)%*
%diag(sqrt(egval(Phir)))%*
%t(egvec(Phir))
#square root of a matrix

for (i in 1:n){
    z1=matrix(rexp(3,rate=1)-1)
#standard exponential
    z2=matrix(rexp(15,rate=1))-1
    ch=sqrt(rchisq(1,5)/3)
#r constant
    f=Phir12%*%z1
    e=Psir12%*%z2
    x=(lambda%*%f+e)/ch
    data=cbind(data,y)
  }
```

How to do this in R

Packages:

- Mass, lavaan, rsem, mvtnorm

Steps:

- 1. Simulate data from the population model
- 2. Get Tml from Lavaan/rsem or calculate with Newton-Raphson Method
- 3. Get scaling constant from Lavaan/rsem robust statistics output or calculate relavant matrices
- 4. Calculate the standard errors and rejection rates of the four test statistics.

Results

f and ε variates are assumed to follow different combinations of distributions:

1. $X = \Lambda f + \varepsilon, f \sim N(0, \Phi), \varepsilon \sim N(0, \Psi)$
2. $X = (\Lambda f + \varepsilon)/r, f \sim N(0, \Phi), \varepsilon \sim N(0, \Psi), r \sim \sqrt{\chi^2_5/3}$
3. $X = (\Lambda f + \varepsilon)/r, f \sim N(0, \Phi), \varepsilon \sim Exp(0, \Psi), r \sim \sqrt{\chi^2_5/3}$
4. $X = (\Lambda f + \varepsilon)/r, f \sim Exp(0, \Phi), \varepsilon \sim N(0, \Psi), r \sim \sqrt{\chi^2_5/3}$
5. $X = (\Lambda f + \varepsilon)/r, f \sim Exp(0, \Phi), \varepsilon \sim Exp(0, \Psi), r \sim \sqrt{\chi^2_5/3}$

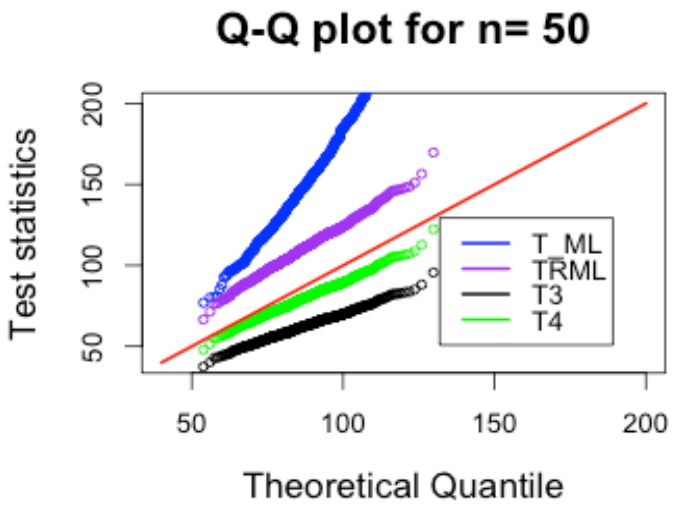
The rescaling constant is chosen because $E(\chi^2_5/3) = 1$.

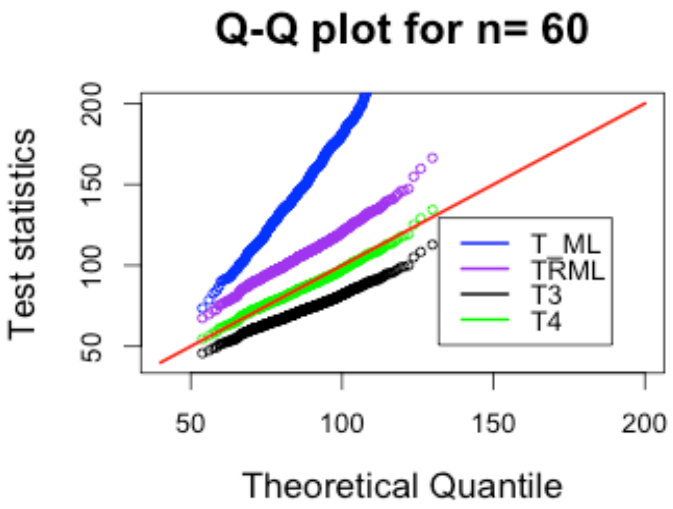
Results

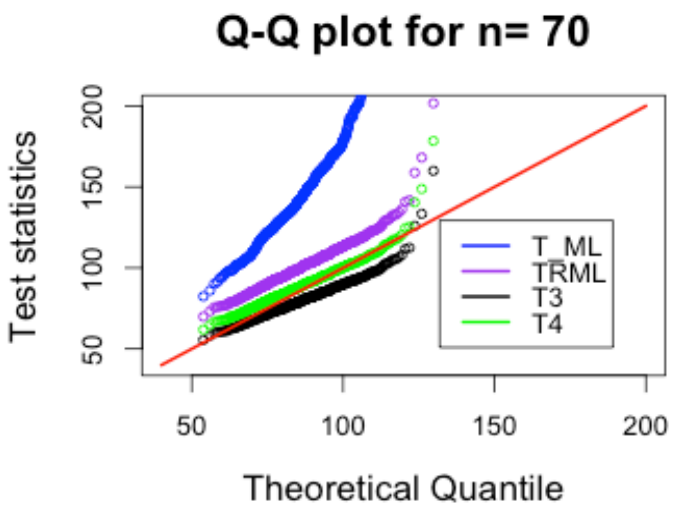
Exponential-Exponential

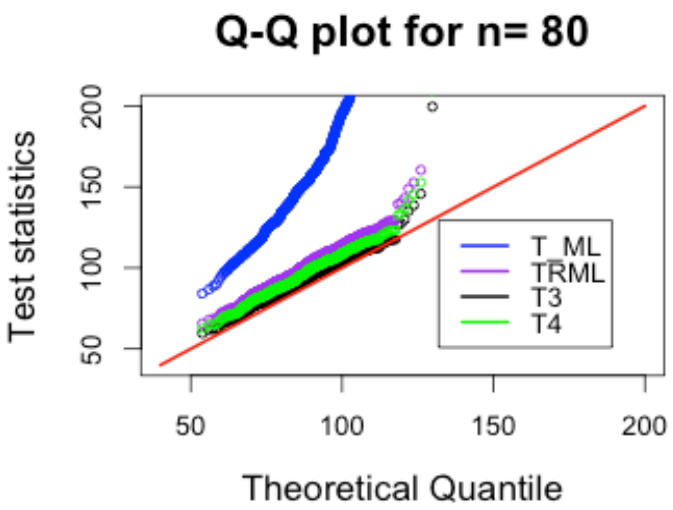
T	50	55	60	65	70	75	80	85	90
Rej	448	462	452	453	452	468	464	468	467
Tml	150.98	151.14	151.21	151.26	150.9	156.4	157.56	155.78	160.87
Sd	35.81	32.91	36.18	38.7	38	45.85	45.04	41.04	46.39
Rej	251	193	188	158	115	118	131	86	94
Trml	110.34	106.55	106.61	103.19	101.47	101.96	100.77	98.93	98.82
Sd	14.95	13.39	14.79	14.28	13.51	15.26	14.97	12.53	13.02
Rej	0	0	1	3	5	15	36	62	94
T3	62.14	66.14	72.3	75.91	80.48	86.73	91.5	95.52	98.82
Sd	8.42	8.31	10.03	10.51	10.71	12.98	13.6	12.09	13.02
Rej	2	4	17	17	20	47	72	74	94
T4	79.51	81.61	86.16	87.47	89.77	93.73	95.91	97.19	98.82
Sd	10.77	10.26	11.95	12.11	11.95	14.03	14.25	12.31	13.02

QQ-plot


Q-Q plot for n= 50
Test statistics
Theoretical Quantile
T_ML
TRML
T3
T4
50
100
150
200


Q-Q plot for n= 60
Test statistics
Theoretical Quantile
T_ML
TRML
T3
T4
50
100
150
200


Q-Q plot for n= 70
Test statistics
Theoretical Quantile
T_ML
TRML
T3
T4
50
100
150
200


Q-Q plot for n= 80
Test statistics
Theoretical Quantile
T_ML
TRML
T3
T4
50
100
150
200

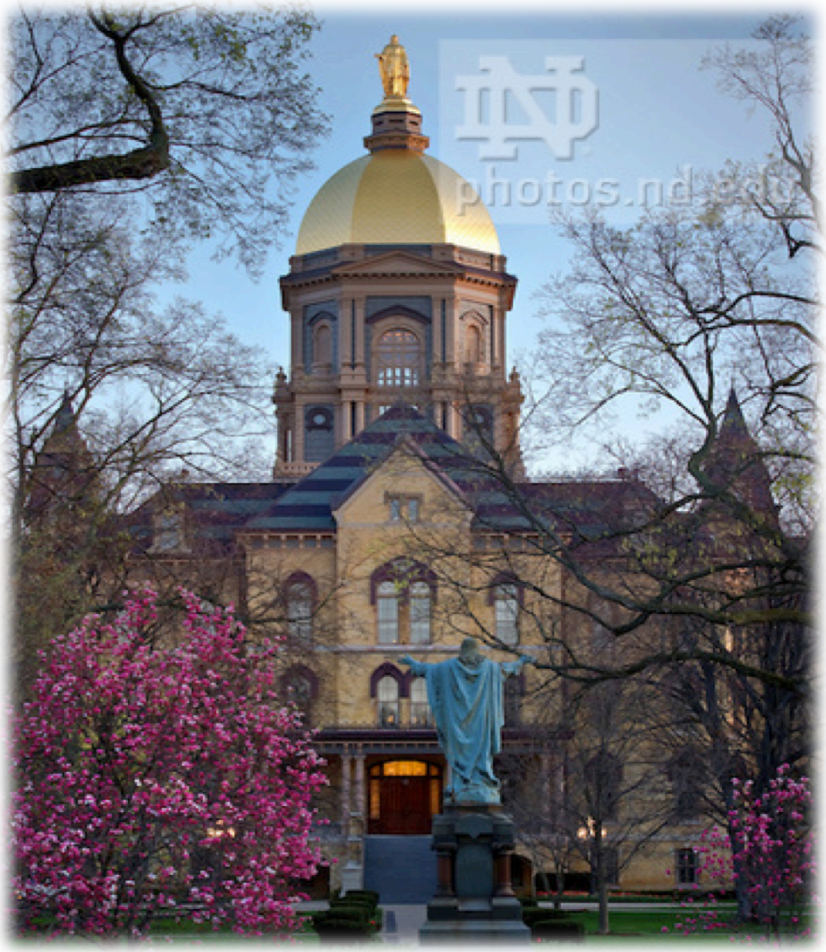
Conclusion

- This is an ongoing research project and the manuscript is also in progress.
- Although there are a lot of statistical softwares (eg., SAS, EQS etc.), R and R packages are pretty beneficial in solving lots of problems in psychology, especially in quantitative psychology.
- Modern SEM methods represent a confluence of work in many disciplines, including biostatistics, econometrics, psychometrics, and social statistics. The general synthesis of these various traditions dates to the late 1960s and early 1970s and will continue to develop in the next decade.

Reference

- Bentler, P. M., Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34*(2), 181-197.
- Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling, 7*(3), 356-410.
- Satorra, A., Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.
- Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling, 17*(2), 241-264.
- Yuan, K. H., Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica, 9*(3), 831-853.

Notre Dame



Thank You!