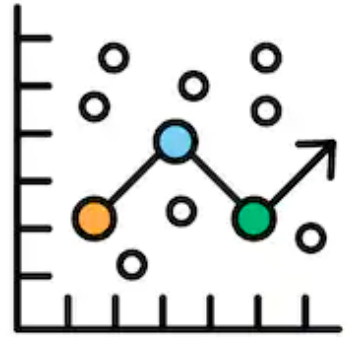
Structural Equation Model

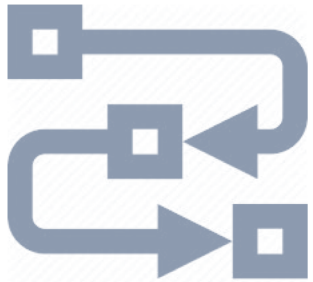
王　鹏

201910

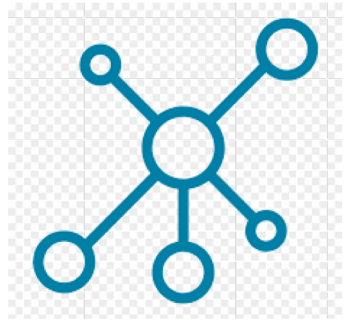
- **Structure** Equation Model (SEM)
- **Covariance** Structure Modelling (CSM)
- **Linear** Structure Relationship



Regression analysis

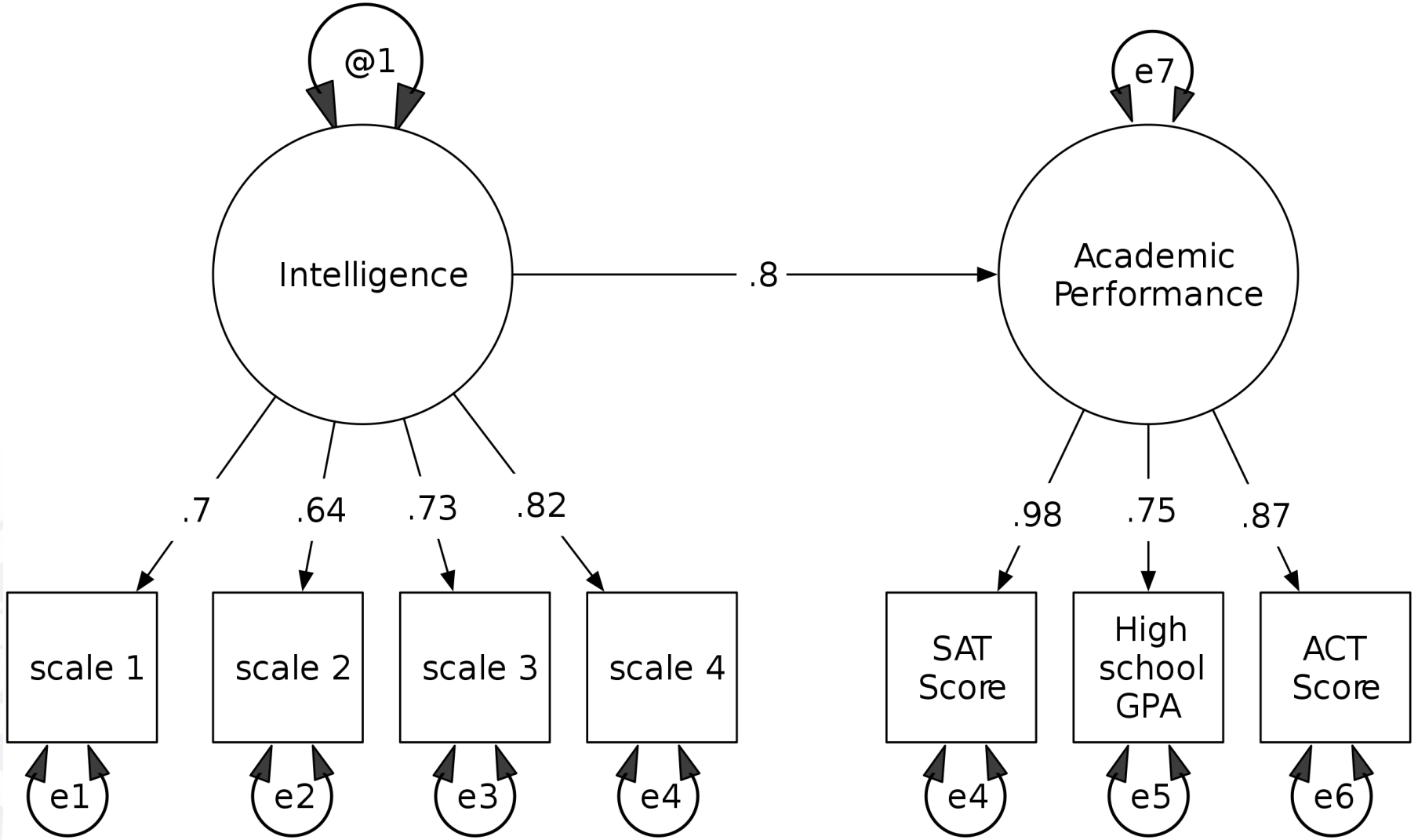


Path analysis



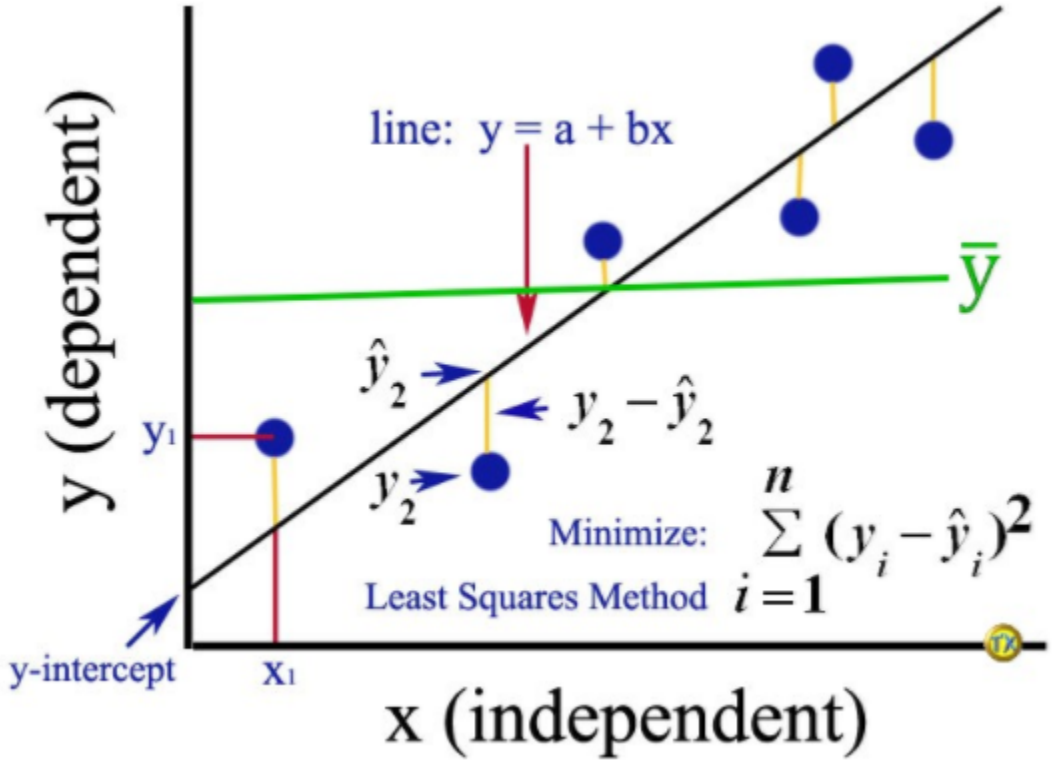
Confirmatory factor analysis

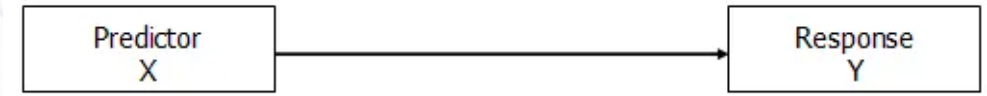
• • •



Regression analysis

回归分析（Regression Analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法





Step1：计算中心点（x和y的平均值）

Step2：线性拟合（最小二乘）

Step3：计算斜率b、截距a和决定系数R^2

$y_i \qquad \bar{y} \qquad \hat{y} \qquad a = \bar{y} - b\bar{x}$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{F-statistics} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

$SSE = \Sigma(y_i - \hat{y}_i)^2$	SSE = sum of squares due to error
$SST = \Sigma(y_i - \bar{y})^2$	SST = total sum of squares
$SSR = \Sigma(\hat{y}_i - \bar{y})^2$	SSR = sum of squares due to regression

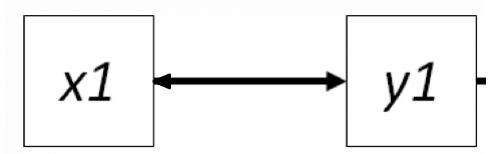
$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

$$R^2 = \frac{SSR}{SST}$$

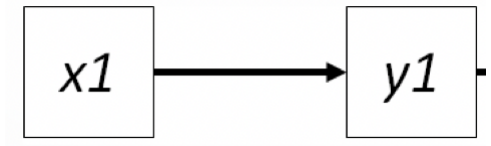
Novogene 诺禾致源

Providing advanced genomic solutions!

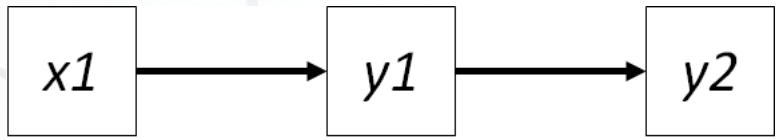
Path analysis



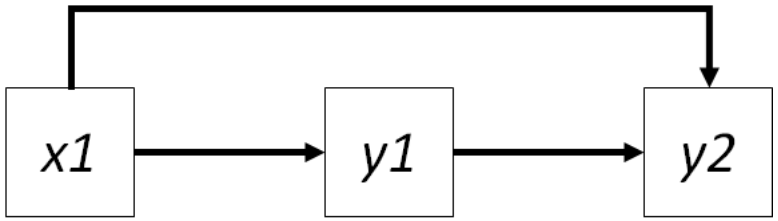
correlation coefficient (cor(x1,x2)



regression coefficient (β_{x1y1})



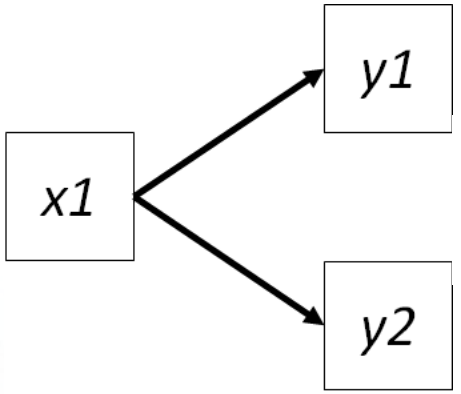
path coefficient (β_{x1y1} * β_{y1y2})



partial regression coefficient

$$b_{y2x1} = \frac{r_{x1y2} - (r_{x1y1} \times r_{y1y2})}{1 - r^2_{x1y1}}$$

$$Total\ effect = b_{y2x1} + b_{y2y1} * cor_{x1y1}$$

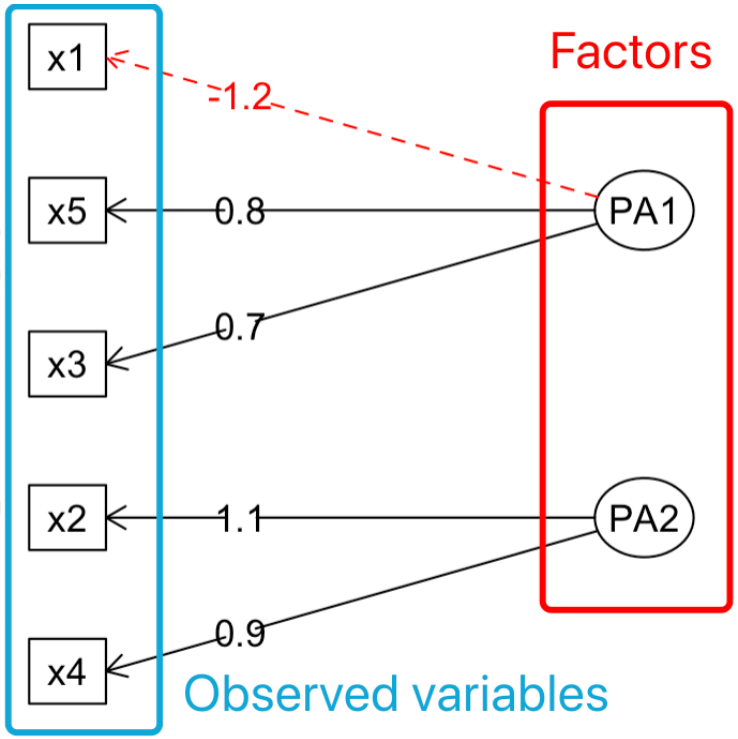


$$r_{y1y2 \bullet x1} = \frac{r_{y1y2} - (r_{x1y1} \times r_{x1y2})}{\sqrt{((1 - r^2_{x1y1})(1 - r^2_{x1y2}))}}$$

Unanalyzed (residual) correlations

Factor analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called **factors.**



$$X_i = a_1F_1 + a_2F_2 + \ldots + a_pF_p + U_i$$

$$PC_i = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

Exploratory factor analysis (EFA) VS Confirmatory factor analysis (CFA)

根据数据探索公因子 **根据公因子验证数据集**

Bartlett's test **变量之间两两相关，才可提取公因子** **KMO test**

≈

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} r_{ij \bullet 1,2 \cdots k}^2}$$

```
corr.dt <- psych::corr.test(select_dt)$r
msa <- psych::KMO(corr.dt)$MSA
cor.bar <- psych::cortest.bartlett(corr.dt, n = n)
```

Data Matrix => Covariance-variance matrix

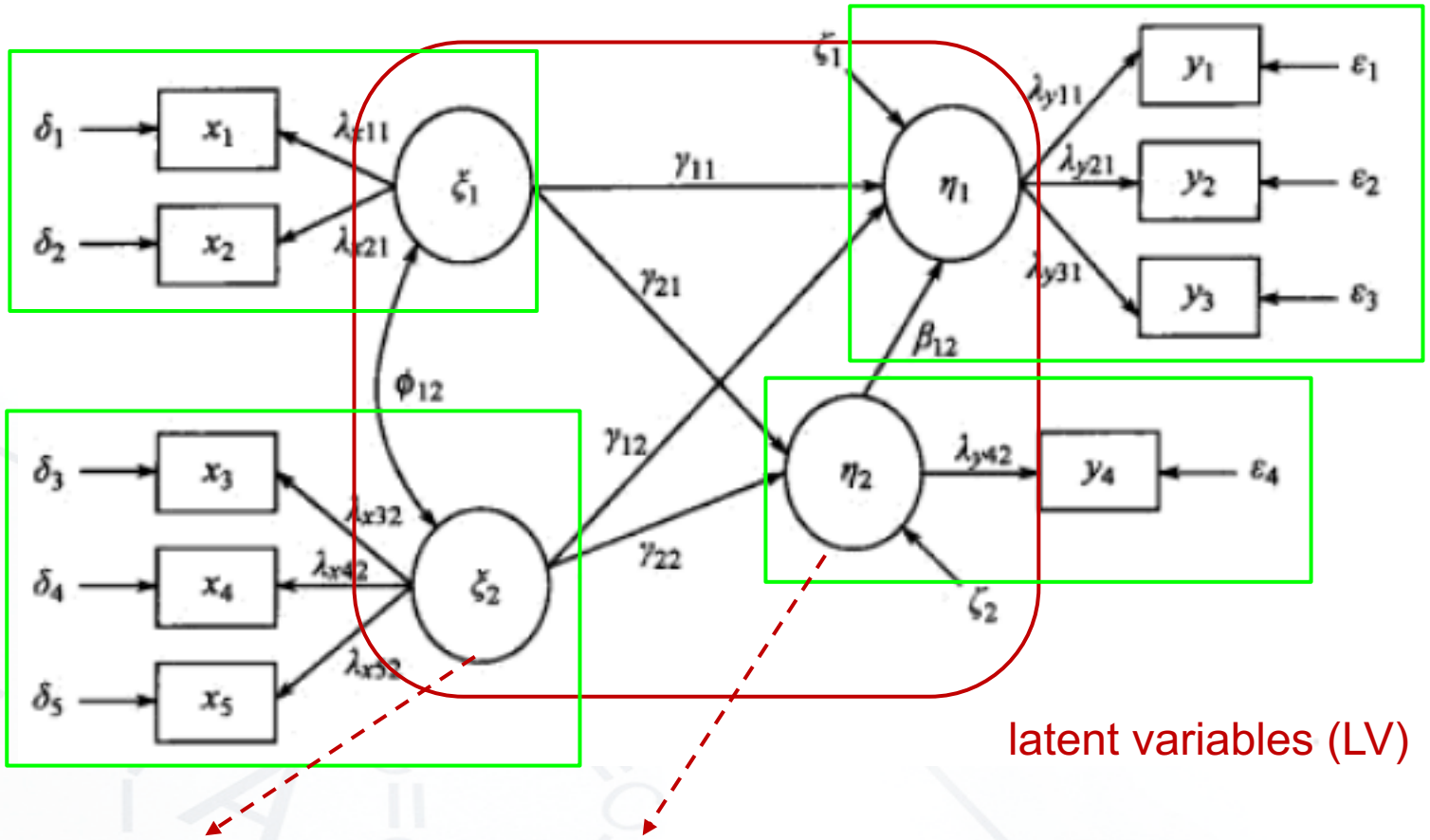
p<0.05: no sig diff

KMO>0.60: strong cor

Novogene 诺禾致源

Providing advanced genomic solutions!

SEM procedure

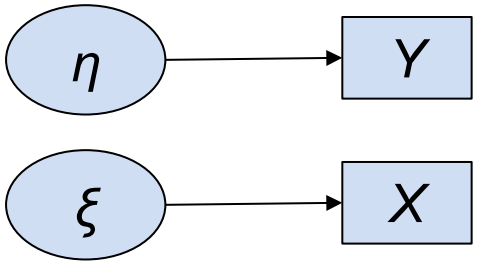


latent variables (LV)

Exogenous LV (ξ) Endogenous LV (η)

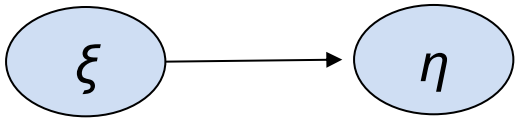
Measurement variables (MV)

Measurement model



$$Y = \Lambda_y \eta + \epsilon$$
$$X = \Lambda_x \xi + \delta$$

Structural model



$$\eta = B\eta + \Gamma\xi + \zeta$$

SEM procedure

Step1: Model specification and Identifiability

```
data(HolzingerSwineford1939)
### initial model format
initial_model_form <- ' visual  =~ x1 + x2 + x3;
                        textual =~ x4 + x5 + x6;
                        speed   =~ x7 + x8 + x9'
```

formula type	operator	mnemonic
latent variable definition	=~	is measured by
regression	~	is regressed on
(residual) (co)variance	~~	is correlated with
intercept	~ 1	intercept

```
fa_is_fit('visual  =~ x1 + x2 + x3', data = HolzingerSwineford1939, n = 301)
fa_is_fit('textual =~ x4 + x5 + x6', data = HolzingerSwineford1939, n = 301)
fa_is_fit('speed   =~ x7 + x8 + x9', data = HolzingerSwineford1939, n = 301)
```

Identifiability && Factor analysis

Knowns ➔ Unknowns

$$t \leq \frac{n(n+1)}{2}$$

df = Knowns - Unknows

```
> fa_is_fit('visual  =~ x1 + x2 + x3', data = HolzingerSwineford1939, n = 301)
      MSA    Chisq       pvalue df
1 0.630835 110.2237 9.821809e-24  3
> fa_is_fit('textual =~ x4 + x5 + x6', data = HolzingerSwineford1939, n = 301)
       MSA    Chisq        pvalue df
1 0.7460931 492.7478 1.779544e-106  3
> fa_is_fit('speed   =~ x7 + x8 + x9', data = HolzingerSwineford1939, n = 301)
       MSA    Chisq       pvalue df
1 0.6480035 155.1489 2.041369e-33  3
```

Novogene 诺禾致源

Providing advanced genomic solutions!

SEM procedure

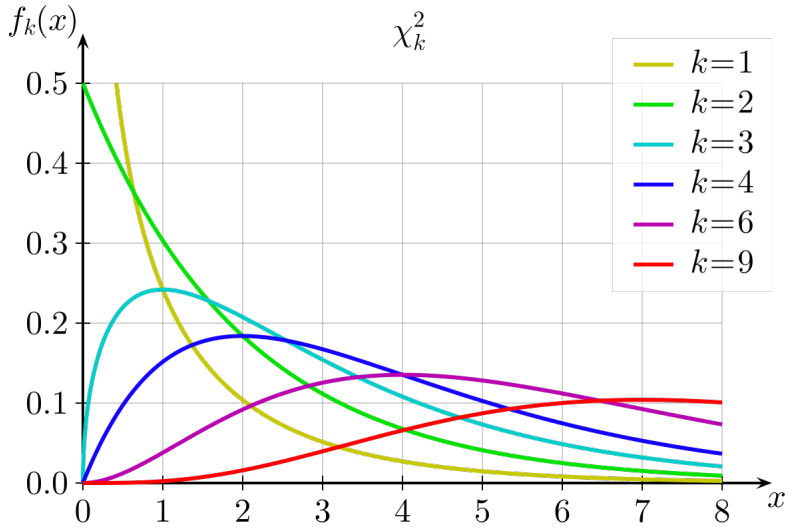
Step2: Estimation of free parameters

Also recall our formula for the maximum-likelihood fitting function:

$$F_{ML} = log|\hat{\Sigma}| + tr(S\hat{\Sigma}^{-1}) - log|S| - (p+q)$$

where Σ is the modeled covariance matrix, S is the observed covariance matrix, p is the number of endogenous variables, and q is the number of exogenous variables. tr is the trace of the matrix (sum of the diagonal) and the $^{-1}$ is the inverse of the matrix.

In the event that $\Sigma = S$, then the first two terms would equal 0, and similarly for the second two terms. Thus a model where $F_{ML} = 0$ implies perfect fit because the observed covariance matrix has been *exactly* reproduced.




```
#bootstrap can efficiently test the significance of coefficiencts
fit <- sem(model = initial_model_form, data = HolzingerSwineford1939, se = 'bootstrap')
########
res <- summary(fit, fit.measures = T, standardized=T, rsq = T)
```

lhs	op	rhs	est	se	z	pvalue	std.lv	std.all
textual	=~	x6	0.9261462	0.06229082	14.868103	0.000000e+00	0.9166000	0.8380101
speed	=~	x7	1.0000000	0.00000000	*NA*	*NA*	0.6194737	0.5695147
speed	=~	x8	1.1799508	0.14484958	8.146042	4.440892e-16	0.7309485	0.7230444
speed	=~	x9	1.0815302	0.37177385	2.909108	3.624620e-03	0.6699795	0.6650092
x1	~~	x1	0.5490540	0.17221493	3.188190	1.431663e-03	0.5490540	0.4042006
x2	~~	x2	1.1338390	0.11561331	9.807167	0.000000e+00	1.1338390	0.8205622
x3	~~	x3	0.8443240	0.10439437	8.087831	6.661338e-16	0.8443240	0.6622852

est: estimation; **se**: standard error; **z**: Wald statistic; **pvalue**;

std.lv：only the latent variables are standardized

std.all：both latent and observed variables are standardized

SEM procedure

Step3: Assessment of model and model fit

p < 0.05, Σ和S之间有显著性差异（拟合效果不好）

•*Root-mean squared error of approximation* (RMSEA): this statistic penalizes models based on sample size. An acceptable value is generally <0.10 and a good value is anything <0.8.

•*Comparative fit index* (CFI): this statistic considers the deviation from a 'null' model. In most cases, the null estimates all variances but sets the covariances to 0. A value >0.9 is considered good.

•*Standardized root-mean squared residual* (SRMR): the standardized difference between the observed and predicted correlations. A value <0.08 is considered good.

```
  Akaike (AIC)                                   7517.490
  Bayesian (BIC)                                 7595.339
  Sample-size adjusted Bayesian (BIC)            7528.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.092
  90 Percent confidence interval - lower         0.071
  90 Percent confidence interval - upper         0.114
  P-value RMSEA <= 0.05                          0.001
```

Akaike Information Criterion (AIC)

Bayesian Information Criterion (BIC)

```
fitmeasures(fit)
```

npar	fmin	chisq	df	pvalue
21.000	0.142	85.306	24.000	0.000
baseline.chisq	baseline.df	baseline.pvalue	cfi	tli
918.852	36.000	0.000	0.931	0.896
logl	unrestricted.logl	aic	bic	ntotal
-3737.745	-3695.092	7517.490	7595.339	301.000
bic2	rmsea	rmsea.ci.lower	rmsea.ci.upper	rmsea.pvalue
7528.739	0.092	0.071	0.114	0.001
srmr				
0.065				

SEM procedure

Step4: Model modification

```
modindices(fit)
```

Modification indices： expected ***decrease*** in the model χ2

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
70	x5	~~	x7	1.232749336	-0.049451850	-0.049451850	-0.082796345	-0.082796345
71	x5	~~	x8	0.346612461	0.023016631	0.023016631	0.049337267	0.049337267
72	x5	~~	x9	0.998870619	0.039621251	0.039621251	0.078827548	0.078827548
73	x6	~~	x7	0.258682051	-0.019661782	-0.019661782	-0.036846352	-0.036846352
74	x6	~~	x8	0.274879372	0.017771665	0.017771665	0.042638750	0.042638750
75	x6	~~	x9	0.097078722	-0.010715488	-0.010715488	-0.023861887	-0.023861887
76	x7	~~	x8	34.145089360	0.536443970	0.536443970	0.859150977	0.859150977
77	x7	~~	x9	5.182955177	-0.186706891	-0.186706891	-0.277537687	-0.277537687
78	x8	~~	x9	14.946391738	-0.423095918	-0.423095918	-0.805202612	-0.805202612

epc: expected parameter change

sepc.lv: only standardizing the latent variables;

sepc.all: standardizing all variables;

sepc.nox: standardizing all but exogenous observed variables

SEM procedure

Step4: Model modification

```
fit.res <- fitted_model_search(model_formula = initial_model_form, data=HolzingerSwineford1939)
```

Model after modification：

```
visual  =~ x1 + x2 + x3; textual =~ x4 + x5 + x6; speed   =~ x7 + x8 + x9;
x7 ~~ x8; x3 ~~ x5; x4 ~~ x7; x1 ~~ x4
```

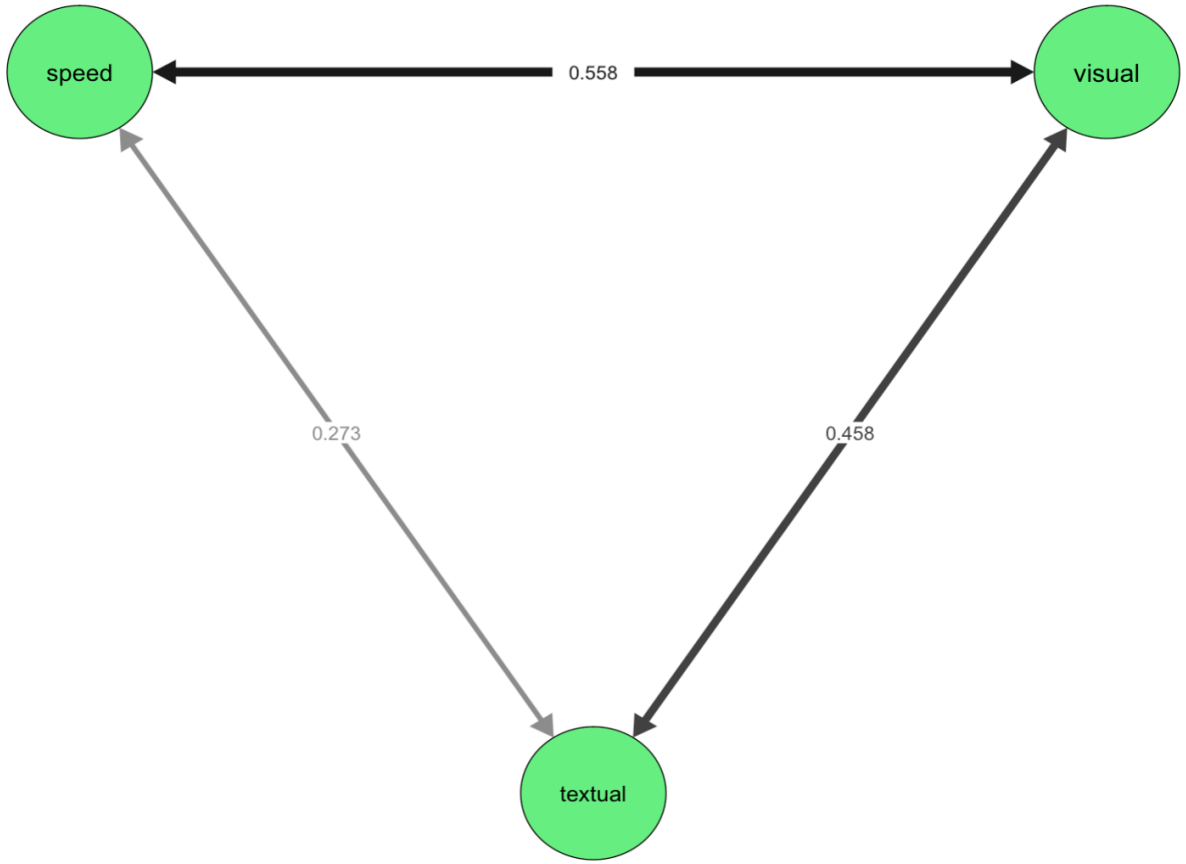
```
fit <- sem(model = fit.res$model, data = HolzingerSwineford1939, se = 'bootstrap' )
########
res <- summary(fit, fit.measures = T, standardized=T, rsq = T)
```

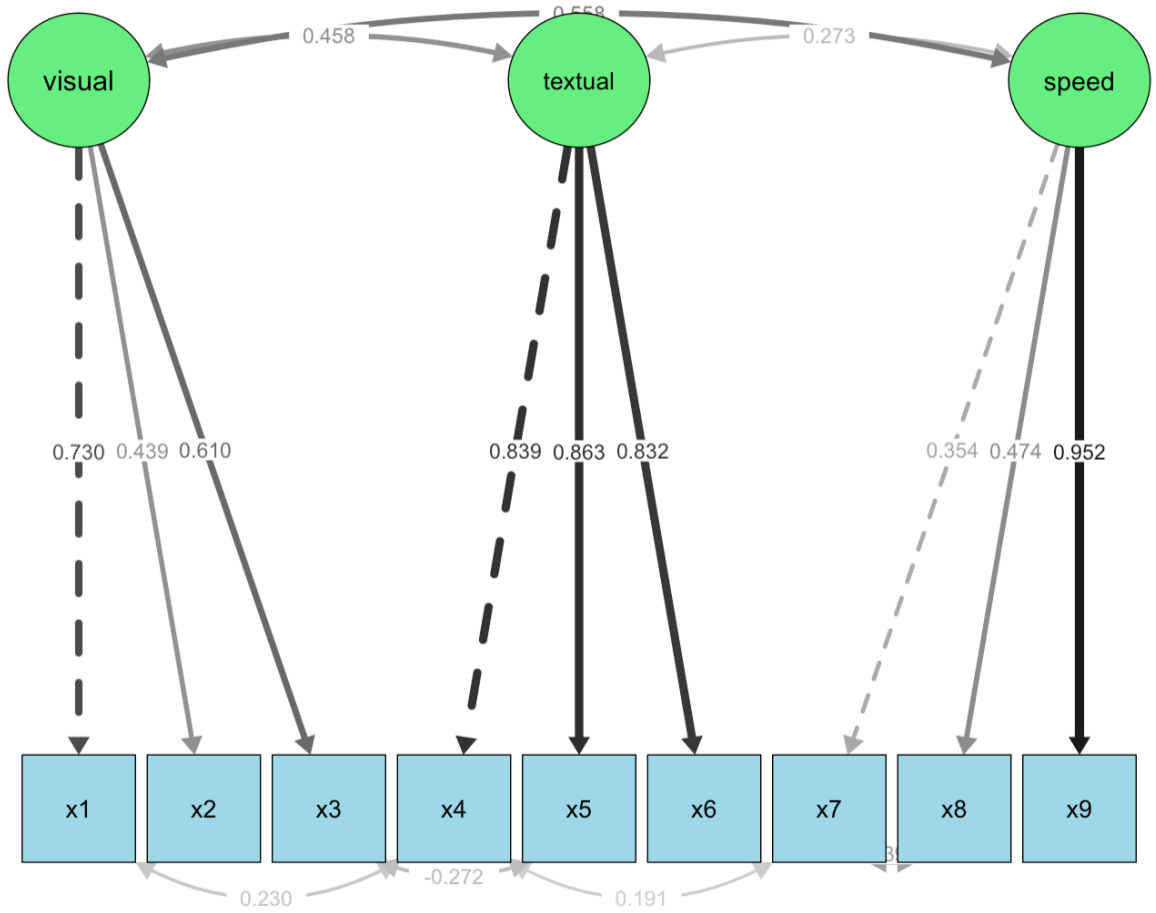
```
Model Test User Model:

  Test statistic                28.039
  Degrees of freedom                20
  P-value (Chi-square)           0.108
```

npar	fmin	chisq	df	pvalue
25.000	0.047	28.039	20.000	0.108
baseline.chisq	baseline.df	baseline.pvalue	cfi	tli
918.852	36.000	0.000	0.991	0.984
logl	unrestricted.logl	aic	bic	ntotal
-3709.112	-3695.092	7468.224	7560.901	301.000
bic2	rmsea	rmsea.ci.lower	rmsea.ci.upper	rmsea.pvalue
7481.615	0.037	0.000	0.066	0.745
srmr				
0.039				

SEM procedure

Step5: Model interpretation





Novogene 诺禾致源

Providing advanced genomic solutions!

SEM sample size

N/P > 10; N/t > 5

N: 样本量；P：观测指标的数目；t: 估计参数

indicators loading at .50). In comparison, the 10 cases per variable rule-of-thumb would have led to sample size recommendations ranging from 40 to 240, respectively. Furthermore, rather than increasing linearly

Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety

https://en.wikipedia.org/wiki/Structural_equation_modeling

https://jslefche.github.io/sem_book/index.html

http://lavaan.ugent.be/tutorial/index.html

/TJPROJ1/MICRO/wangpeng/R_and_D/SEM/sem.R

Novogene 诺禾致源

Providing advanced genomic solutions!

Thanks for your attention!

更多关注，敬请留意：www.novogene.cn