Psychology 454: Latent Variable Modeling
Review and final comments

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

November, 2016

Outline

Conceptual overview

- A theory of data

Correlation &Regression

- Regression
- Multivariate Regression and Partial Correlation

Goodness of fit measures

- Absolute and Relative Fit measures
  - Absolute fit indices
  - Incremental or relative fit indices
- Practical Advice

Measuring change and factorial invariance

- Factorial Invariance
- Measuring change

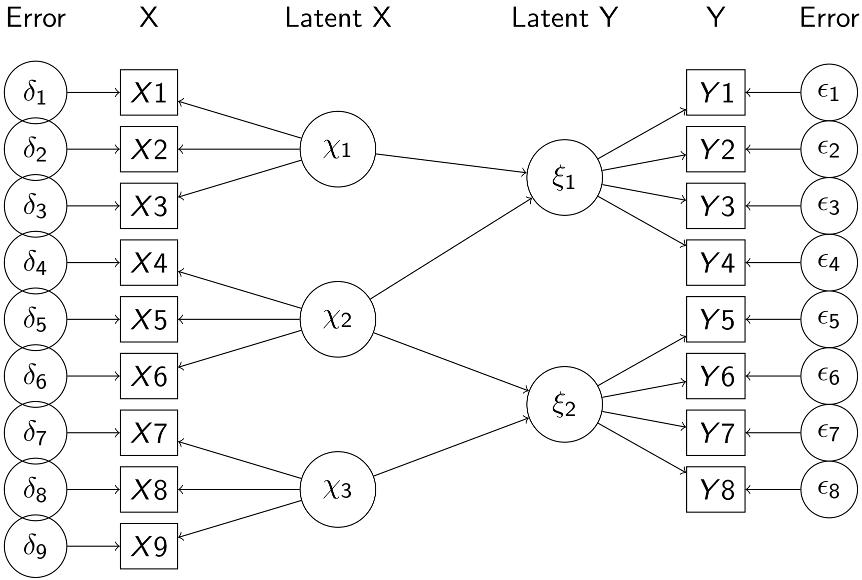
Regression without correcting for reliability

- Model misspecification
- Mediation

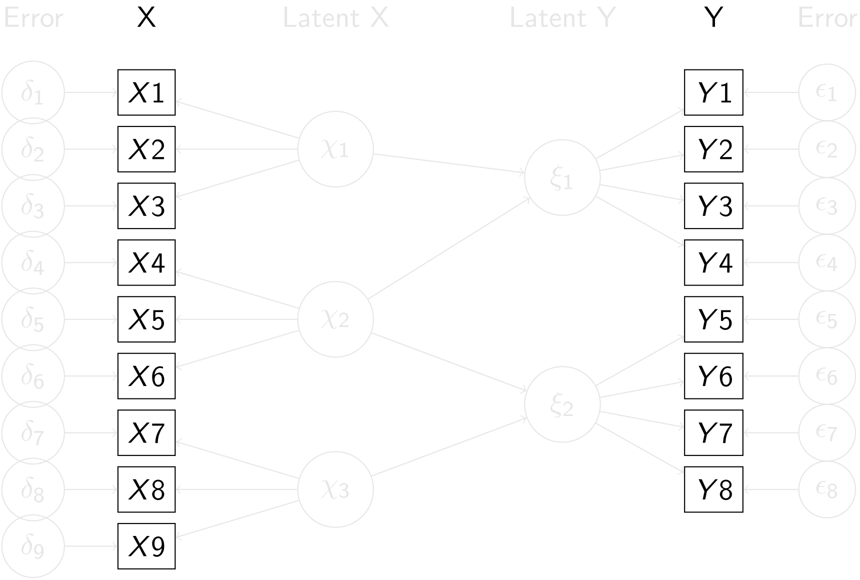
Latent model approach

Problems with SEM

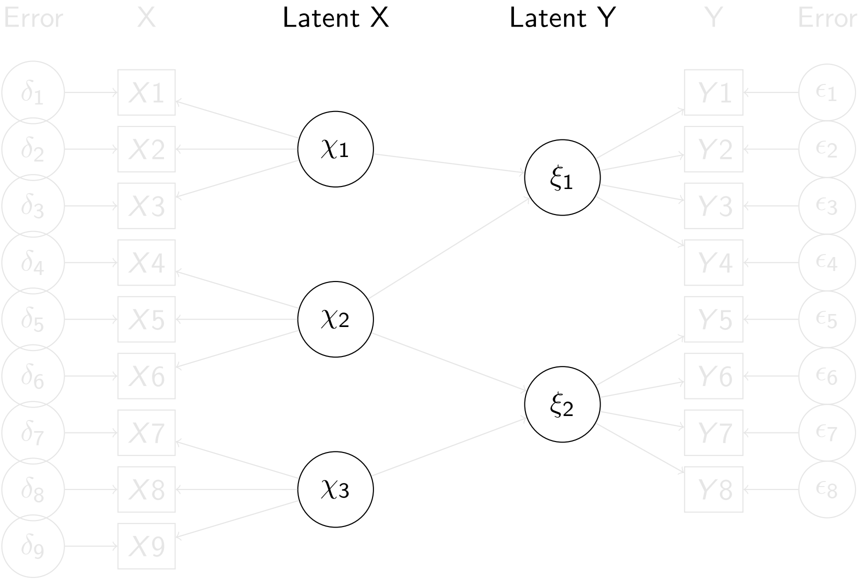
Latent Variable Modeling: A conceptual Syllabus



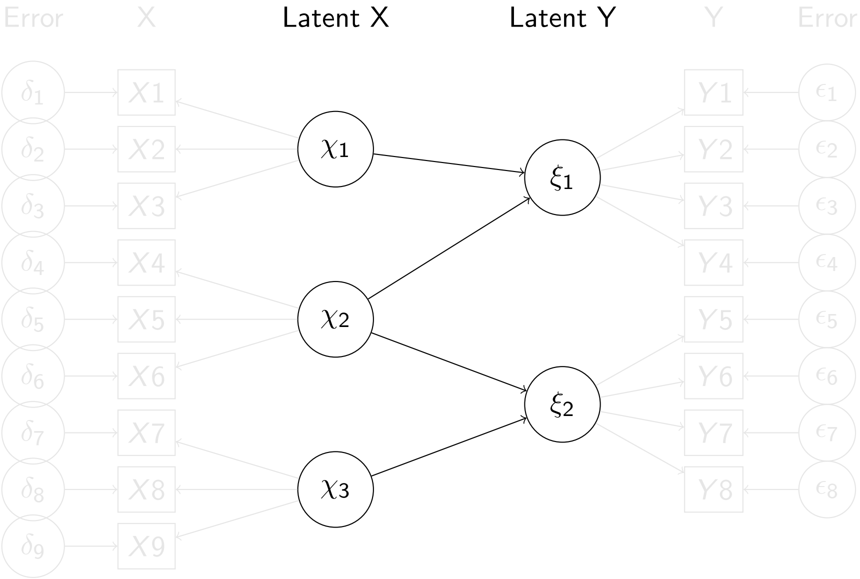
Observed Variables



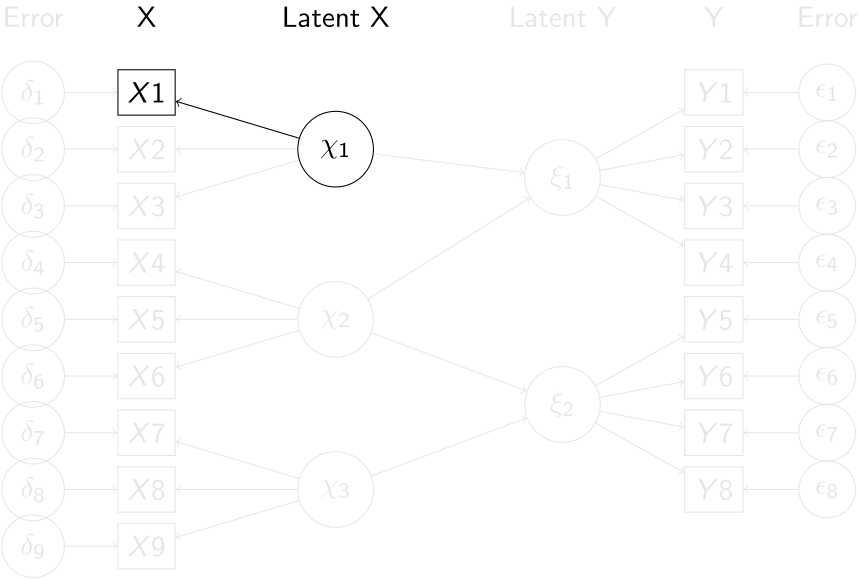
Latent Variables



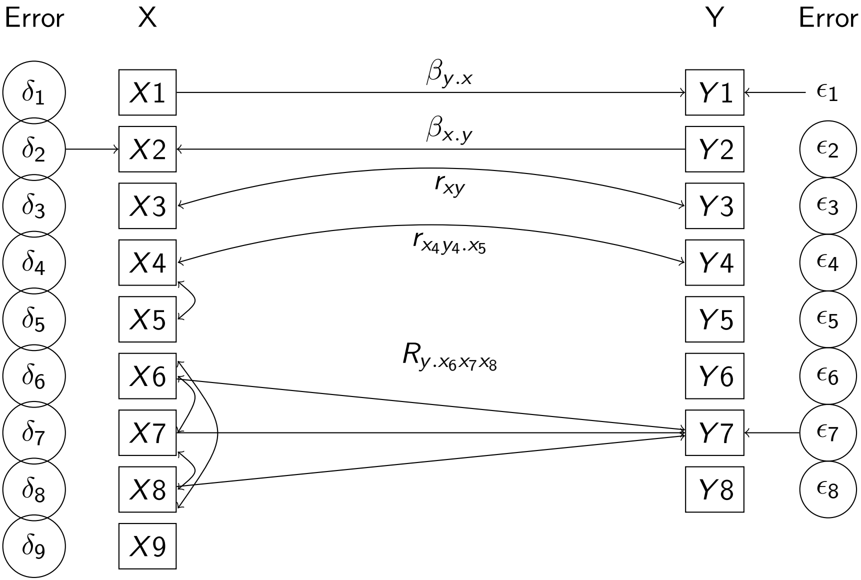
Theory


Error
X
Latent X
Latent Y
Y
Error
δ_1
δ_2
δ_3
δ_4
δ_5
δ_6
δ_7
δ_8
δ_9
$X1$
$X2$
$X3$
$X4$
$X5$
$X6$
$X7$
$X8$
$X9$
χ_1
χ_2
χ_3
ξ_1
ξ_2
$Y1$
$Y2$
$Y3$
$Y4$
$Y5$
$Y6$
$Y7$
$Y8$
ϵ_1
ϵ_2
ϵ_3
ϵ_4
ϵ_5
ϵ_6
ϵ_7
ϵ_8

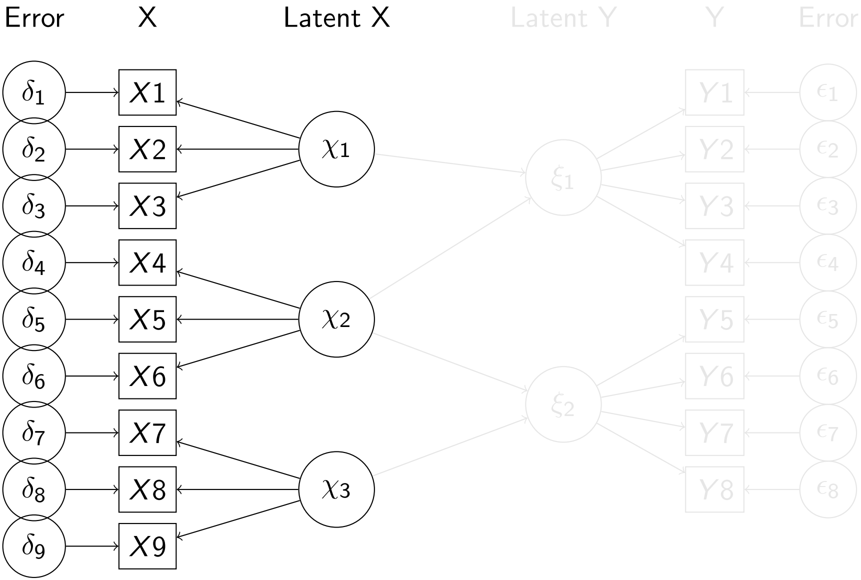
A theory of data and fundamentals of scaling



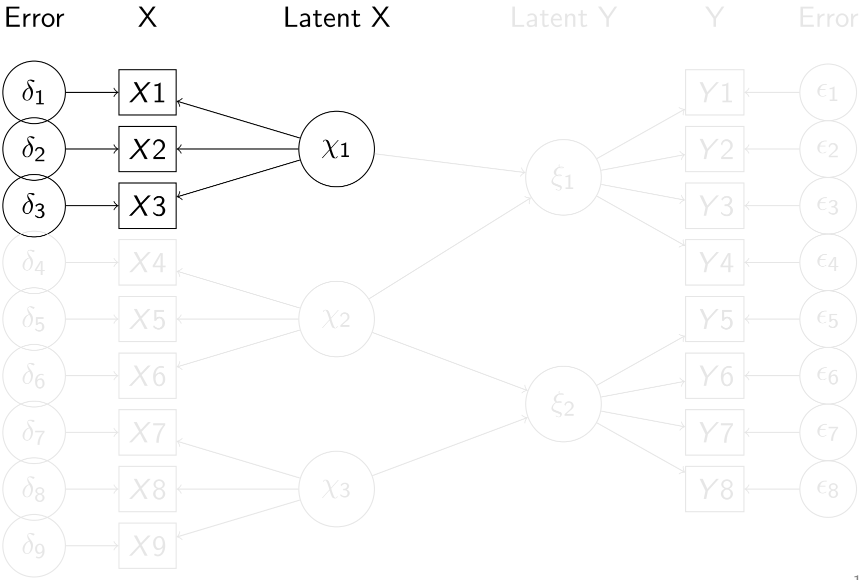
Correlation, Regression, Partial Correlation, Multiple Regression



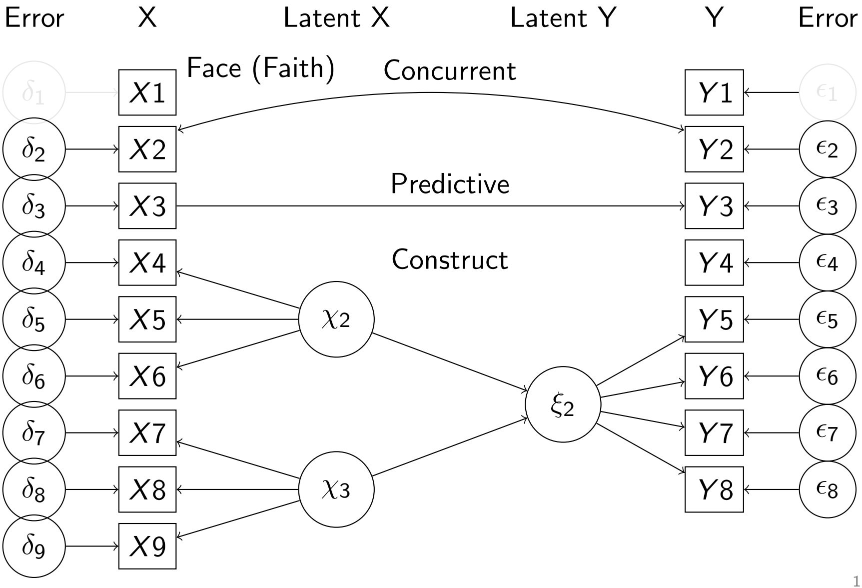
Measurement: A latent variable approach.



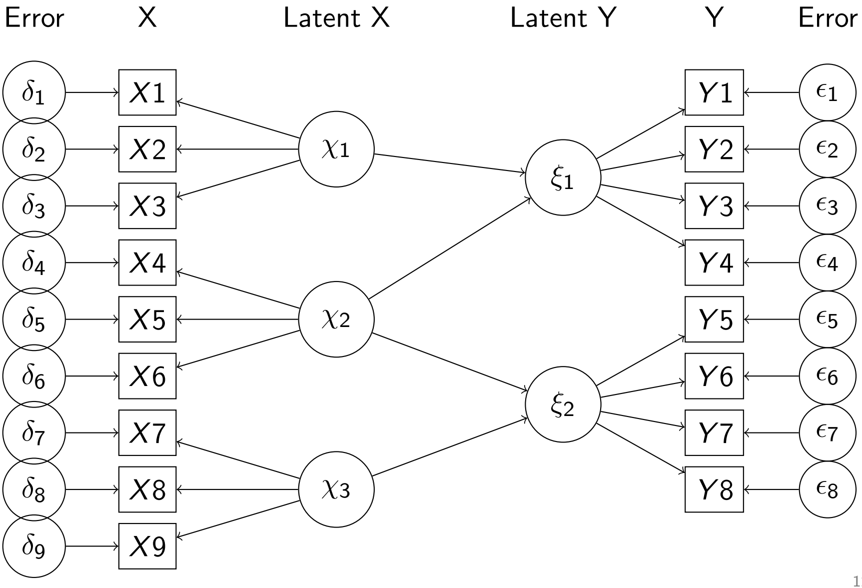
Reliability: How well does a test reflect one latent trait?



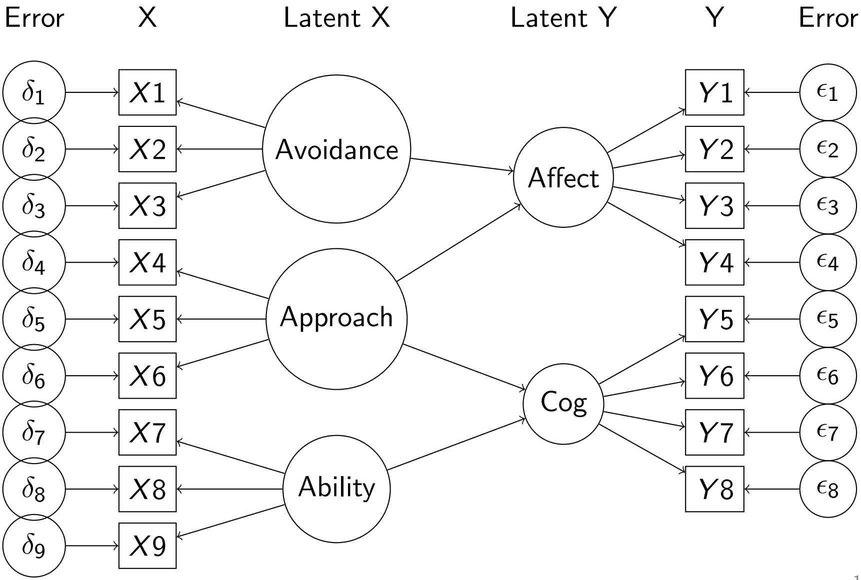
Face, Concurrent, Predictive, Consruct



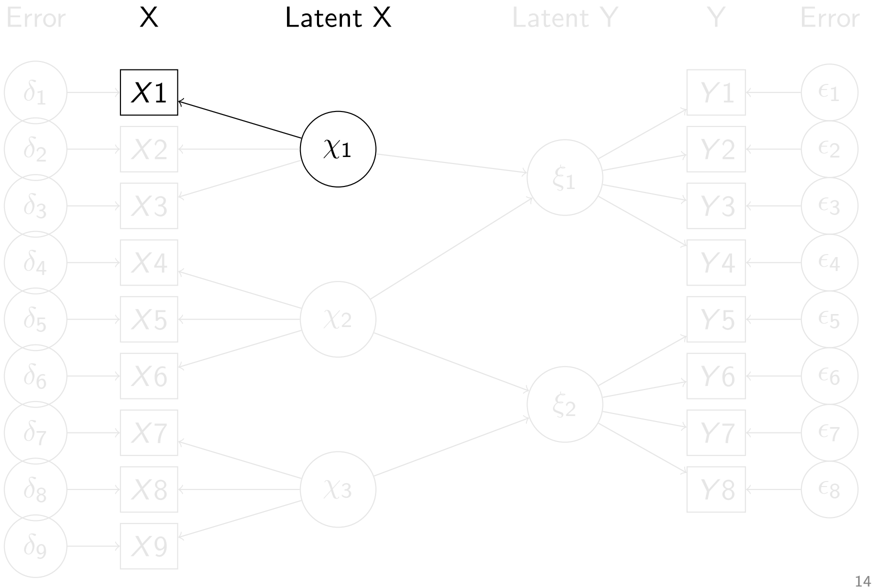
Psychometric Theory: Data, Measurement, Theory



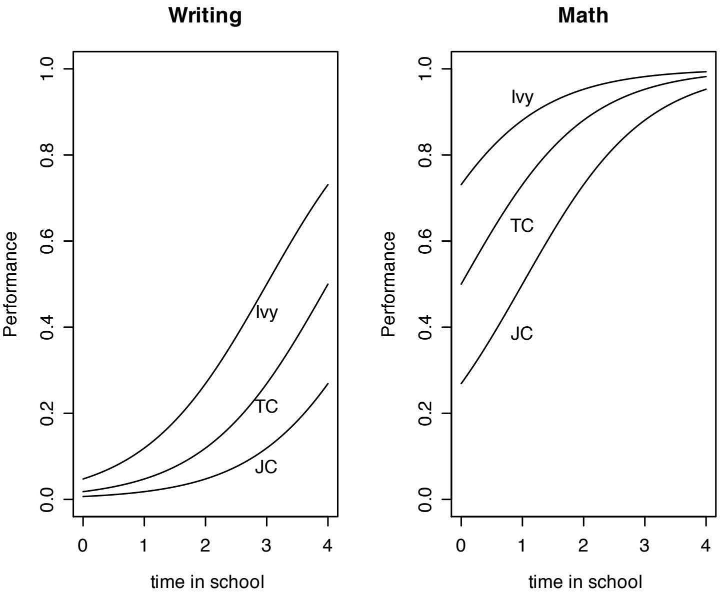
Psychometric Theory: Data, Measurement, Theory



A theory of data and fundamentals of scaling



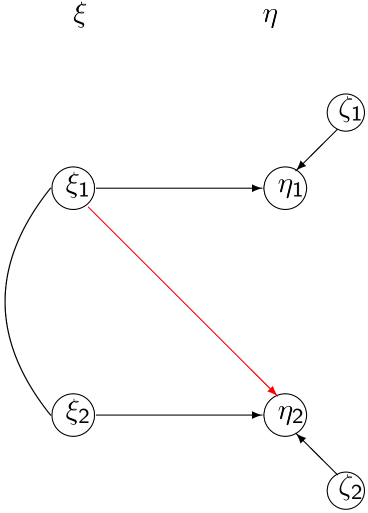
Effect of teaching, effect of students, or just scaling?



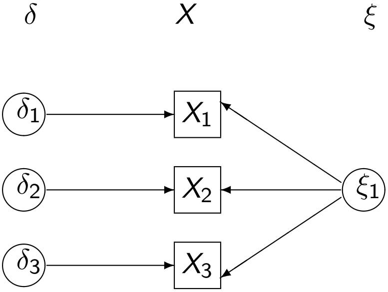
The best scale is the one that works best

1. Money is linear but negatively accelerated with utility.
2. Perceived intensity is a log function of physical intensity.
3. Probabilty of being correct is a logistic or cumulative normal function of ability.
4. Energy used to heat a house is linear function of outdoor temperature.
5. Time to fall a particular distance varies as the square root of the distance.
6. Gravitational attraction varies as $1/distance^2$
7. Hull speed of sailboat varies as square root of length of boat.
8. Sound intensity in db is log(observed/reference)
9. pH of solutions is -log(concentration of hydrogen ions)

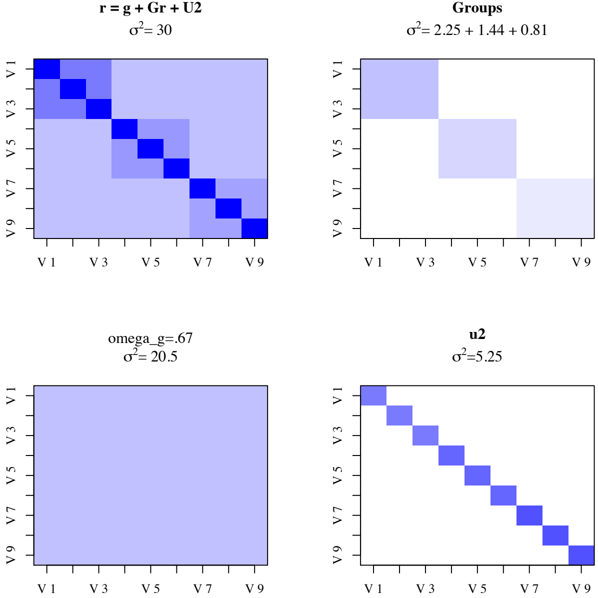
Theory: A regression model of latent variables



A measurement model for X Reliability



A graphic representation of a general factor + 3 group factors



Factor analytic definition of a general factor

Let F = [g + G] be a column wise concatenation of a general factor and a set of group factors:

Variable	g	Group 1	Group 2	Group 3	...
	g	F1	F2	F3	
V1	g1	a1	0	0	
V2	g2	a2	0	0	
V3	g3	a3	0	0	
V4	g4	0	b4	0	
V5	g5	0	b5	0	
V6	g6	0	b6	0	
V7	g7	0	0	c7	
V8	g8	0	0	c8	
V9	g9	0	0	c9	

Variance decomposition of R

$R = FF' + U^2$ where $F = [g + G]$

For orthogonal F and G, the correlation matrix is a function of the general loadings as well as the group loadings:

$$R = gg' + GG' + U^2$$

The amount of variance attributable to the general factor is just ω_g (McDonald, 1999) where

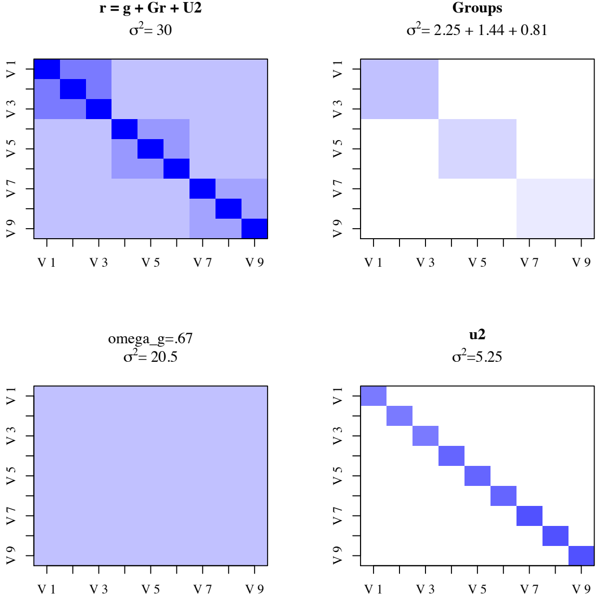
$$\omega_g = \frac{1'gg'1}{1'R1}$$

The total amount of reliable variance (that which is attributable to general + groups) is ω_t

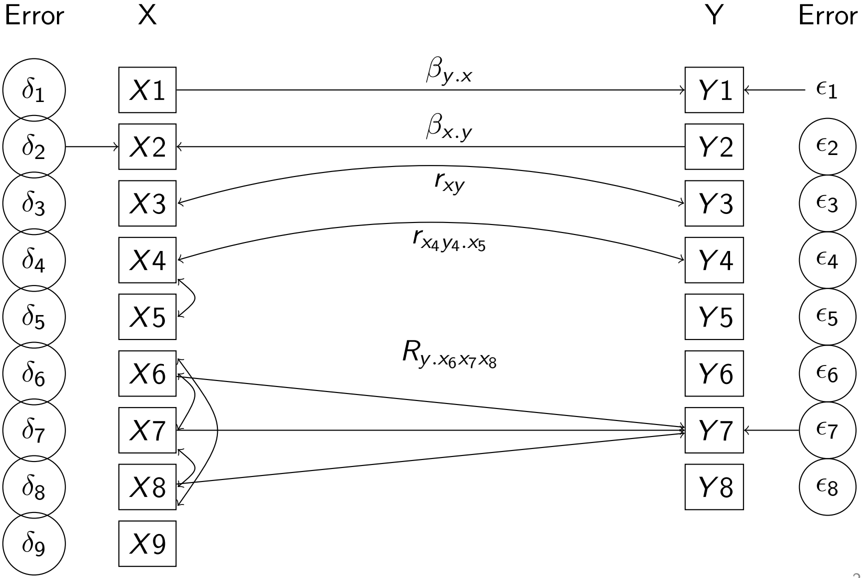
$$\omega_t = \frac{1'gg'1 + 1'GG'1}{1'R1}$$

The problem then is how to find ω_g.

A graphic representation of the problem



Correlation, Regression, Partial Correlation, Multiple Regression



Bivariate Regression

δ X Y ϵ

$$X \xrightarrow{\beta_{y.x}} Y \leftarrow \epsilon$$

$$\hat{y} = \beta_{y.x}x + \epsilon$$

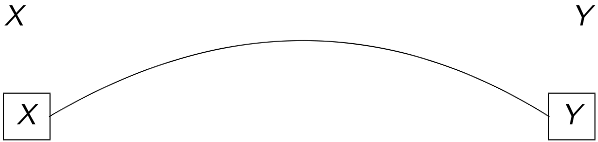
$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\delta \rightarrow X \xleftarrow{\beta_{x.y}} Y$$

$$\hat{x} = \beta_{x.y}y + \delta$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$$

Bivariate Correlation

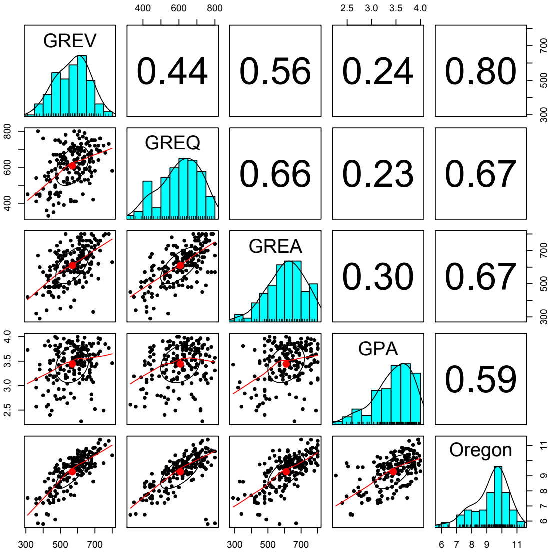


$\hat{x} = \beta_{x.y} y + \delta$ $\hat{y} = \beta_{y.x} x + \epsilon$

$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$ $\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$

$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma^{2x} \sigma_y^2}}$$

Scatter Plot Matrix showing correlation and LOESS regression

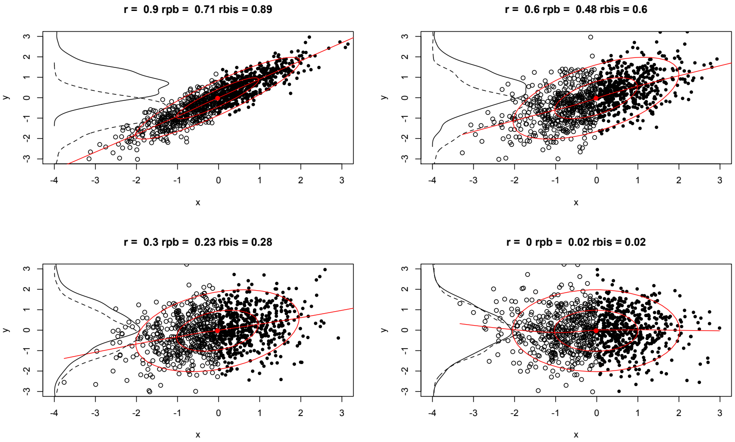


Alternative versions of the correlation coefficient

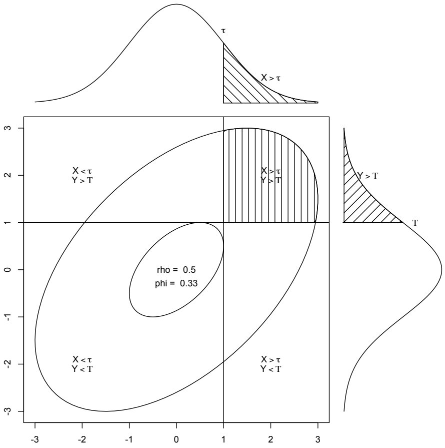
Table: A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

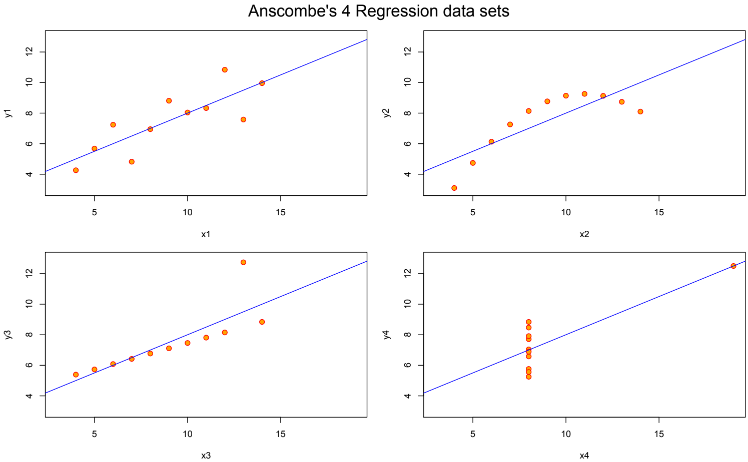
The biserial correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation



Cautions about correlations: Anscombe data set

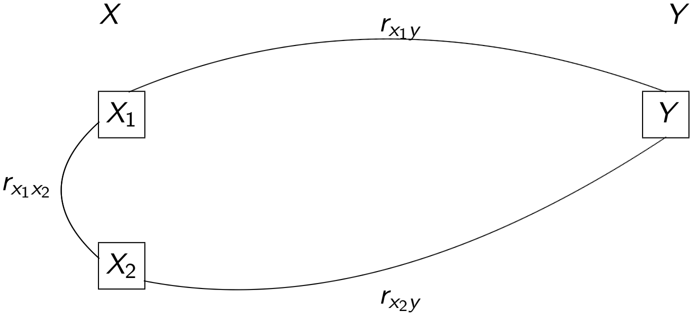


The ubiquitous correlation coefficient

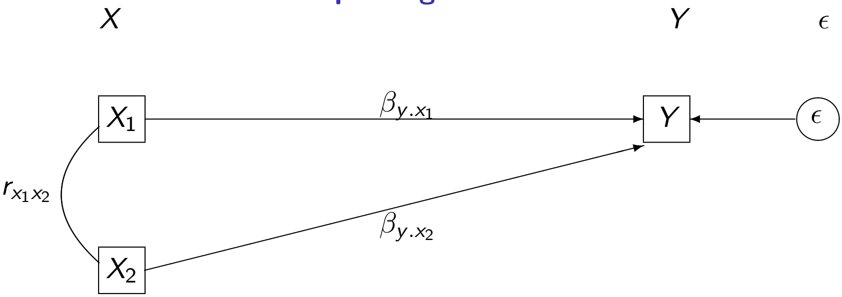
Table: Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{C_{xy}}{\sigma_x^2}$	$r = b_{y.x}\frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r\frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = \frac{d\sqrt{df}}{2}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = \frac{d^2 df}{4}$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

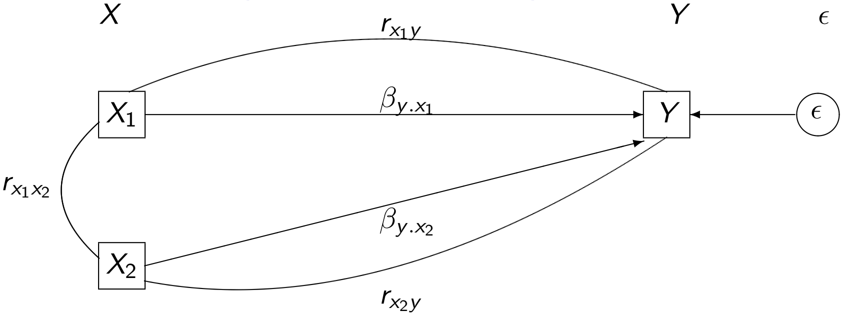
Multiple correlations



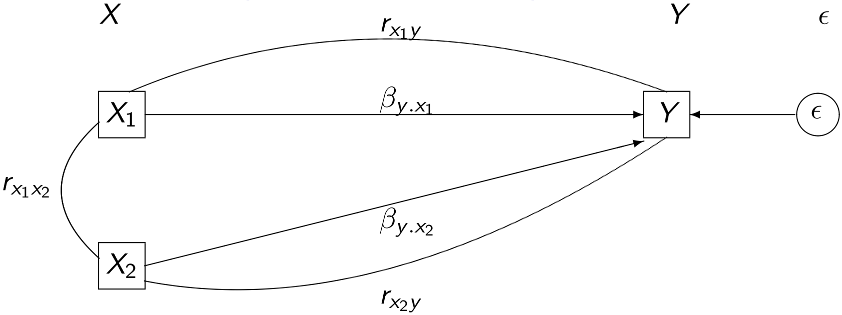
Multiple Regression



Multiple Regression: decomposing correlations



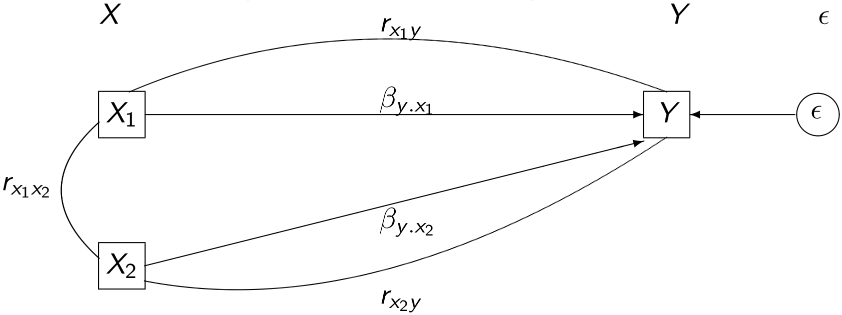
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect}$$

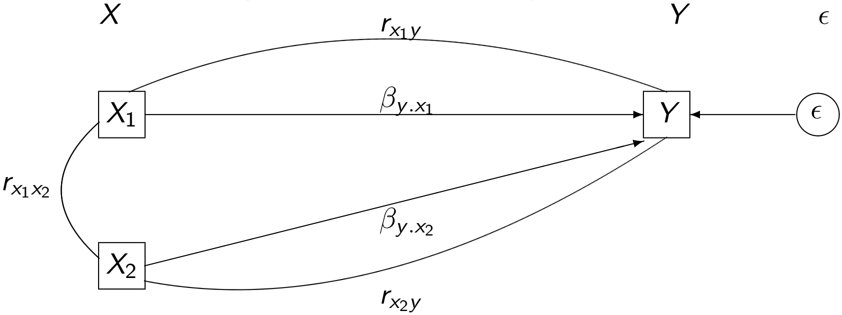
Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{\color{blue}{direct}} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{\color{red}{indirect}} \qquad \beta_{y.x_1} = \frac{\color{blue}{r_{x_1 y}} \color{red}{- r_{x_1 x_2} r_{x_2 y}}}{1 - \color{red}{r^2_{x_1 x_2}}}$$

$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{\color{blue}{direct}} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{\color{red}{indirect}} \qquad \beta_{y.x_2} = \frac{\color{blue}{r_{x_2 y}} \color{red}{- r_{x_1 x_2} r_{x_1 y}}}{1 - \color{red}{r^2_{x_1 x_2}}}$$

Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{\textit{direct}} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{\textit{indirect}}$$

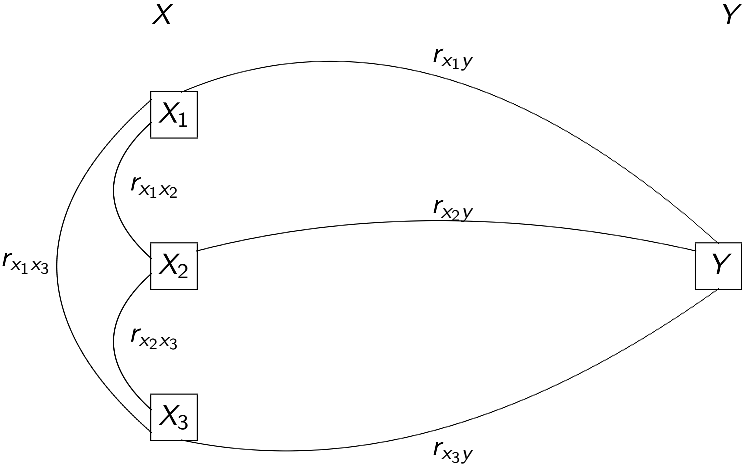
$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{\textit{direct}} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{\textit{indirect}}$$

$$\beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r_{x_1x_2}^2}$$

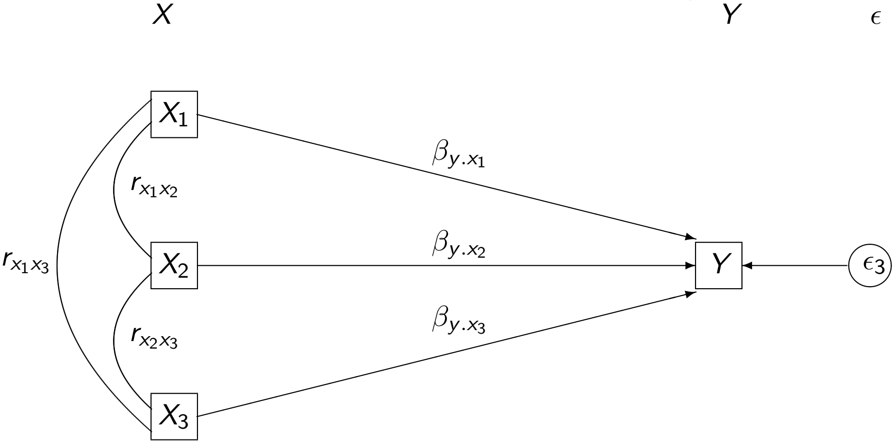
$$\beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r_{x_1x_2}^2}$$

$$R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2}$$

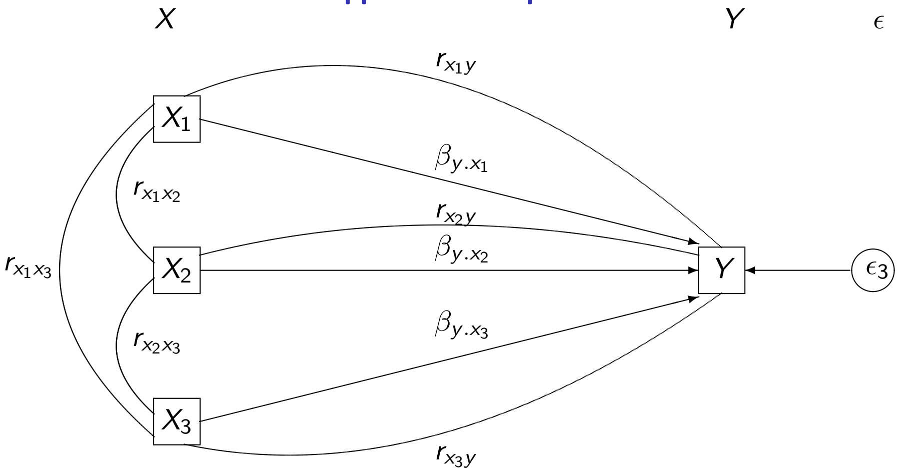
What happens with 3 predictors? The correlations



What happens with 3 predictors? β weights



What happens with 3 predictors?



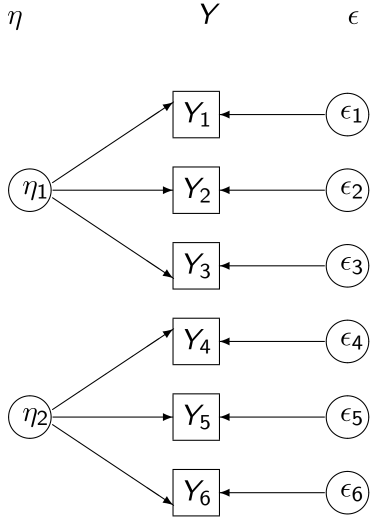
$$r_{x_1 y} = \underbrace{\beta_{y.x_1}}_{\textit{direct}} + \underbrace{r_{x_1 x_2}\beta_{y.x_1} + r_{x_1 x_3}\beta_{y.x_3}}_{\textit{indirect}} \quad r_{x_2 y} = \dots \quad r_{x_3 y} = \dots$$

The math gets tedious

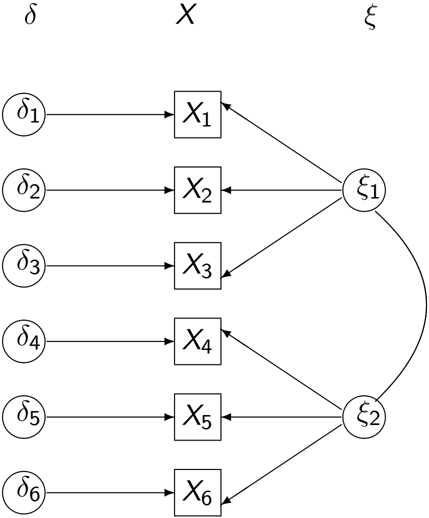
Multiple regression and matrix algebra

- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
  - Each equation is expressed as a $r_{x_i y}$ in terms of direct and indirect effects.
  - Direct effect is $\beta_{y.x_i}$
  - Indirect effect is $\sum_{j \neq i} beta_{y.x_j} r_{x_j y}$
- How to solve these equations?
- Tediously, or just use matrix algebra

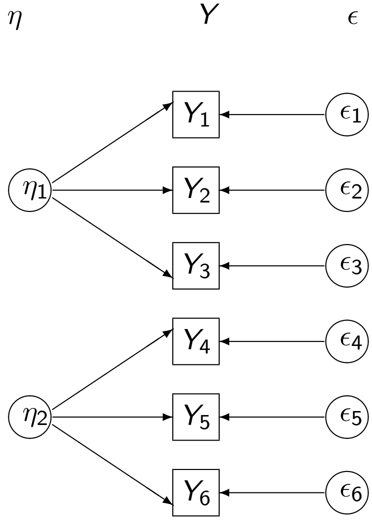
A measurement model for Y - uncorrelated factors



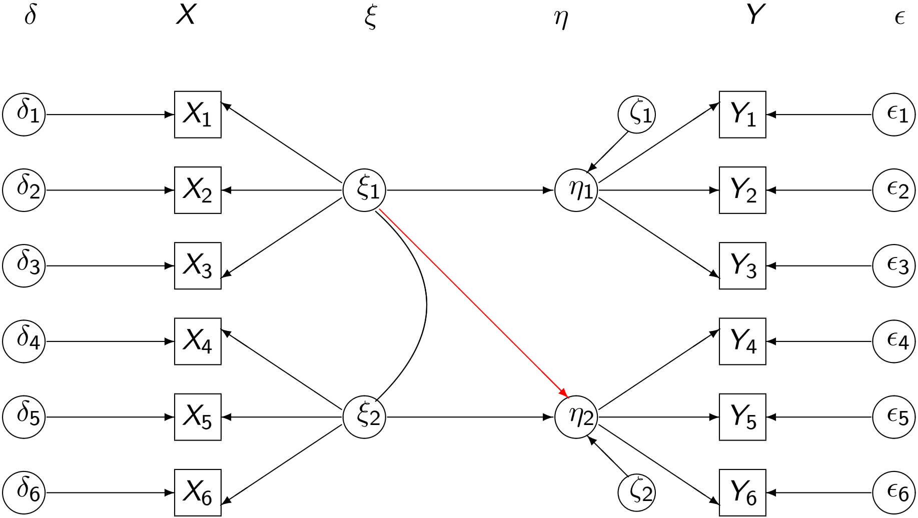
A measurement model for X – Correlated factors



A measurement model for Y - uncorrelated factors



A complete structural model



Exploratory versus confirmatory

- The real power of confirmatory analysis is the ability to test paths and models.
- EFA fits $\mathbf{R} = \mathbf{FF'} + \mathbf{U^2}$
  - for k variables and nf factors
  - The number of parameters to estimate is $k * nf - (nf * (nf - 1)/2)$
  - The number of observed correlations is $k * (k - 1)/2$
  - degrees of freedom $= k * (k - 1)/2 - k * nf - (nf * (nf - 1)/2)$
  - all parameters are allowed to vary
- Confirmatory analysis fixes some parameters to zero, doesn't estimate others
  - This makes a more constrained model, with more degrees of freedom
- Compare the EFA and CFA of the Thurstone data sets (12 versus 24 df)
- Fixing parameters to specified values is also possible

A number of tests of fit taken from Marsh et al. (2005)

1. Marsh, Hau & Grayson (2005) lists 40 different proposed measures of goodness of fit
2. Measures of absolute fit
   - I_o = index of fit for original or baseline model
   - I_t = index of fit for target or "true" model
3. Measures of incremental fit Type I
   - $\frac{|I_t - I_o|}{Max(I_o, I_t)}$ which is either
     - $\frac{I_o - I_t}{I_o}$
     - or $\frac{I_t - I_o}{I_t}$
4. Measures of incremental fit Type II
   - $\frac{|I_t - I_o|}{E(I_t - I_o)}$ which is either
     - $\frac{I_o - I_t}{I_o - E(I_t)}$
     - or $\frac{I_t - I_o}{E(I_t) - I_o}$

Fit functions from Jöreskog

1. Ordinary least squares $F = \frac{1}{2}tr(S - \Sigma)^2$
   - The squared difference between the observed (S) and model (Σ) covariance matrices
   - tr means trace of the sum of the diagonal values of the matrix of squared deviations
2. Generalized least squares $F = \frac{1}{2}tr(I - S^{-1}\Sigma)^2$
   - I is the identity matrix
   - if the model = data, then $S^{-1}\Sigma$ should be I
   - weight the fit by the inverse of the observed covariances
3. Maximum Likelihood $F = log|\Sigma| + tr(S\Sigma^{-1}) - log|S| - p$
   - weight the fit by the inverse of the modeled covariance
   - p is the number of variables
   - tr (I) = p, and thus the ML should be 0 if the model fits the data

Fit-function based indices

1. Fit Function Minimum fit function (FF)
   - $FF = \frac{\chi^2}{(N-1)}$
2. Likelihood ratio $LHR = e^{\frac{-1}{2}FF}$
3. χ^2 (minimum fit function chi square)
   - $\chi^2 = tr(\Sigma^{-1}S - I) - log|\Sigma^{-1}S| = (N-1)FF$
4. $p(\chi^2)$ probability of observing a χ^2 this larger or larger given that the model fits
5. $\frac{\chi^2}{df}$ has expected value of 1

Non-centrality based indices

1. Non-centrality parameter
   - $NCP = \chi^2 - df$
2. Dk: Rescaled noncentrality paramter (McDonald & Marsh, 1990)
   - $Dk = FF - \frac{df}{N-1} = \frac{\chi^2}{(N-1)} - \frac{df}{N-1} = \frac{\chi^2 - df}{N-1}$
3. PDF (population discrepancy function = DK normed to be non-negative)
   - $PDF = max(\frac{\chi^2 - df}{N-1}, 0)$
4. Mc: Measure of centrality (CENTRA, MacDonald Fit Index (MFI)
   - $Mc = e^{\frac{-(\chi^2 - df)}{2(N-1)}}$

Error of approximation indices

How large are the residuals, estimated several different ways

1. RMSEA (root mean square error of approximation)
    - $RMSEA = \sqrt{PDF/df} = \sqrt{\frac{max(\frac{\chi^2 - df}{N-1}, 0)}{df}}$
    - based upon the non-central χ^2 distribution to find the error of fit
2. MSEA (mean square error of approximation – unnormed version of RMSEA)
    - $MSEA = \frac{Dk}{df} = \frac{\chi^2 - df}{(N-1)df}$
3. RMSEAP (root mean square error of approximation of close fit)
    - RMSEA $< .05$
4. RMR Root mean square residual (or, if S and Σ are standardized, the SRMR). Just
    - square root of the average squared residual
    - $\sqrt{\frac{2\sum(S-\Sigma)^2}{p*(p+1)}}$

Information indices

Compare the information of a model with the number of parameters used for the model. These allow for comparisons of different models with different degrees of freedom.

1. AIC (Akaike Information Criterion for a model penalizes for using up df)
   - $AIC = \chi^2 + p * (p + 1) - 2df = \chi^2 + 2K$
   - where $K = \frac{p*(p+1)}{2} - df$
2. Baysian Information Criterion
   - $-2Log(L) + plog(N) = \chi^2 - Klog(N(.5(p(p + 1))$

Goodness of fit indices

1. GFI from LISREL
   - $GFI = 1 - \frac{tr(\Sigma^{-1}S - I)^2}{tr(\Sigma^{-1}S)^2}$
2. Adjusted Goodness of Fit (Lisrel)
   - $AGFI = 1 - \frac{p(p+1)}{2df}(1 - GFI)$
3. Unbiased GFI (from Steiger)
   - $\text{GFI} = \frac{p}{2\frac{(\chi^2 - df)}{(N-1)} + p}$

Comparing solutions to solutions

1. Incremental fit indices without correction for model complexity
   - RNI (relative non-centrality) McDonald and Marsh
   - CFI Comparative fit index (normed version of RNI) Bentler
   - Normed Fit index (Bentler and Bonett)
2. Incremental fit indices correcting for model complexity
   - Tucker - Lewis Index
   - Normed Tucker Lewis
   - Incremental Fit index
   - Relative Fit Index
3. Parsimony indices

Incremental fit indices without correction for model complexity

1. RNI (relative non-centrality) McDonald and Marsh
   - $RNI = 1 - \frac{Dk_t}{Dk_n}$
   - where $DK = \frac{\chi^2 - df}{N-1}$ for either the null or the tested model
2. CFI Comparative fit index (normed version of RNI) Bentler
   - Just norm the RNI to be greater than 0.
   - $CFI = 1 - \frac{MAX(NCP_t, 0)}{MAX(NCP_n, 0)}$
3. Normed Fit index (Bentler and Bonett)

Fitting functions from Loehlin

1. Let S be the "strung out" data matrix
2. Let Σ be the "strung out" model matrix
3. $Fit = (S - \Sigma)' W^{-1} (S - \Sigma)$
4. Where W =
   - Ordinary Least Squares $W = I$
   - Generalized Least Squares $W = SS'$
   - Maximum likelihood $W = \Sigma\Sigma'$

Practical advice

1. Taken from Kenny
   - http://davidkenny.net/cf/fit.htm
2. Bentler and Bonnet Normed Fit Index
   - $\frac{\chi^2_{Null} - \chi^2_{Model}}{\chi^2_{Null}}$
   - Between .90 and .95 is acceptable
   - $>$ .95 is "good"
3. RMSEA
   - if $\chi^2 < df$, then RMSEA $= 0$
   - "good" models have $RMSEA < .05$
   - "poor" models have$RMSEA > .10$
4. p of close fit
   - Null hypothesis is that RMSEA is .05
   - test if $RMSEA > .05$
   - Claim good fit if $p(RMSEA > .05) < .05$

Considering rules of thumb and fit

1. Fit functions have distributions and thus are susceptible to problems of type I and type II error.
    - Compare the fits for correct model as well as those for a simpler incorrect (?) model
2. Should we just use chi square and reject models that don't fit, or should we reason about why they don't fit.
3. All models are wrong, some are useful.
4. "The epistemological basis of statistics has moved away from being a set of procedures, applied mechanistically, and moved toward building and evaluating statistical and scientific models. " (Rodgers, 2010)

What does it mean if the model does not fit

1. Model is wrong
2. Measurement is wrong
3. Structure is wrong
4. Assumptions are wrong
5. at least one of above, but which one?

Specification & Respecification

1. Is the measurement model consistent
    - revise it
        - evaluate loadings
        - evaluate error variances
        - more or fewer factors
        - correlated errors?
2. Structural model:
    - adjust paths
    - drop paths
    - add paths
3. Equivalent models
    - What models are equivalent
    - Do they make equally good sense

Factorial invariance: Does a test measure the same thing in different groups?

1. Across groups
   - Different schools
   - Different groups (e.g., ethnicity, age, gender)
2. Across time
   - Is todays' measure the same as next year's measure?
3. Types of invariance
   - Configural: Are the arrows the same
   - Weak invariance: Are the loadings the same across groups
   - Strong invariance: Loadings and intercepts are equal across groups
   - Super strong: Loadigs, intercepts and means are equal across groups

Measuring structure at two (or more) time points

1. Is the structure the same
   - Structural Invariance (is the graph the same)
   - Measurement invariance (are the loadings the same)
   - Strong measurement invariance (are the item intercepts the same?)
   - Measuring change
2. Do the means change (is there growth)
   - This is the means of the latent trait, not the means of the items
3. Do the latent traits correlate across two or more occasions?
   - Just two occasions, can not separate trait from state effects
   - With > 2 occasions, can examine trait and state effects
4. Compare several different simulations

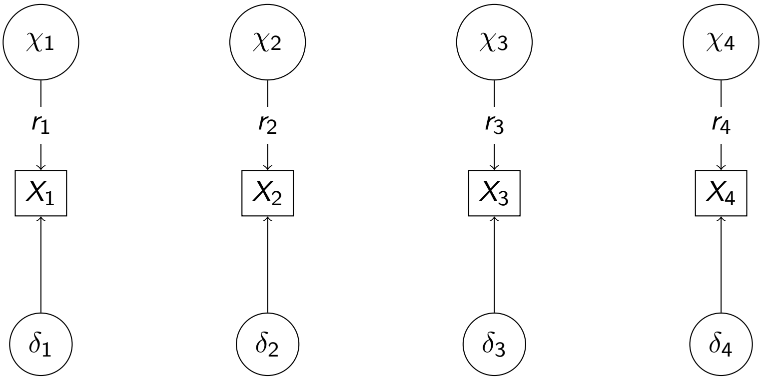
Observed scores over 4 time points

Do they differ in means? Do they measure the same thing?

$\boxed{X_1}$ $\boxed{X_2}$ $\boxed{X_3}$ $\boxed{X_4}$

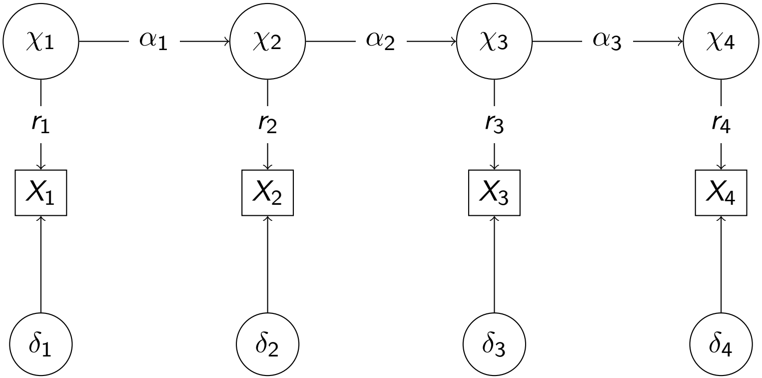
Modeled 4 time points

Are the measures measures of the same construct? Are the measures invariant?

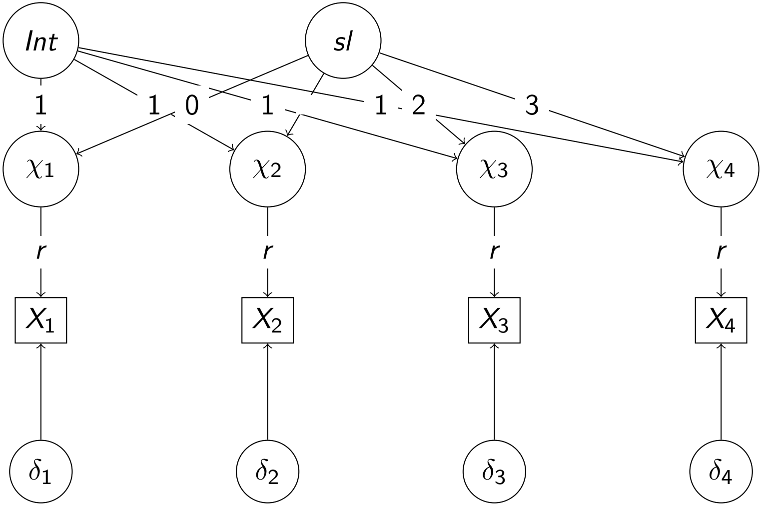


Model state change over 4 time points: simplex but with too many parameters

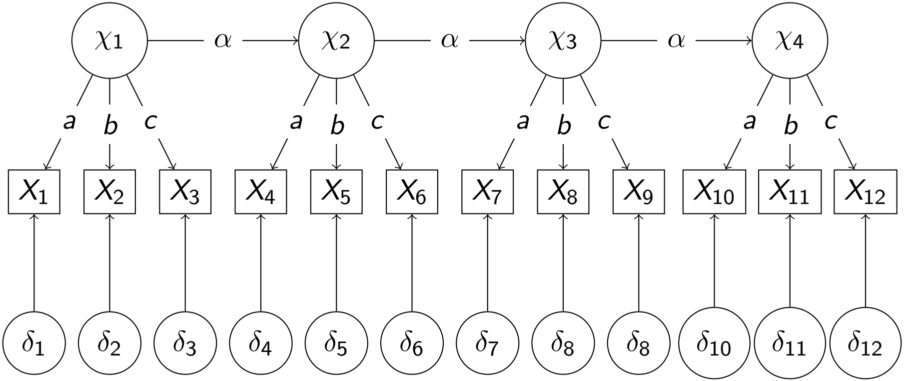
$\sqrt{r_i}$ is reliability, α_i is autocorrelation.



Model change over 4 time points: traits are stable



Model change over 4 time points: states with measurement



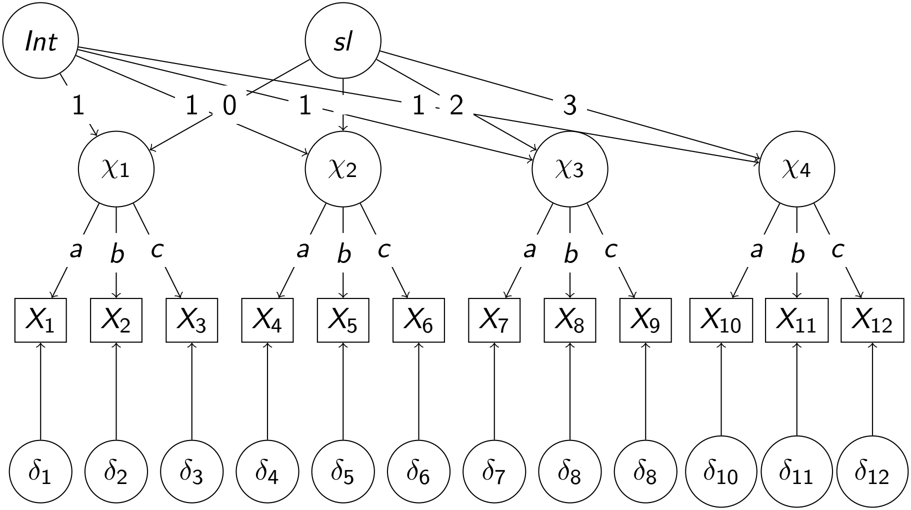
State growth

R code

```
models <- 'L1 =~ a*x1 +b *x2 + c* x3
           L2 =~ a* x4 + b * x5 + c* x6
           L3 =~ a*x7 + b*x8 +c*x9
           L4 =~ a*x10 + b*x11 + c*x12

           L2 ~ alpha* L1
           L3 ~ alpha * L2
           L4 ~ alpha *  L3'

fits <- growth(models,data=simp4$observed)
summary(fits)
```

```
lavaan (0.5-19) converged normally after  32 iterations

  Number of observations                          1000

  Estimator                                         ML
  Minimum Function Test Statistic               84.661
  Degrees of freedom                                67
  P-value (Chi-square)                           0.071

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  Z-value  P(>|z|)
```

Model change over 4 time points: traits with measurement



Model misspecification: failure to include variables

A classic problem in statements of causal structure is the failure to include appropriate variables. Such model misspecification is the bane of using correlations to infer anything about causality, for there is always the lurking third variable that could explain the relationship.

1. In an attempt to demonstrate this effect, consider the correlation between three variables at time 1 as predictors of an important outcome at time 2.
2. The measured variables at time 1 are
   - Yellow Fingers
   - Yellow Teeth
   - Bad Breath.
3. The outcome variable is probability of Lung Cancer (rescored with a logistic transformation to be a continuous variable ranging from -3 to 3.) [1]

[1] As I hope is obvious, this is an artificial example. It was inspired, in part, by the webpage on causal and statistical reasoning at Carnegie Mellon University (www.cmu.edu/CSR/index.html

For the purposes of this demonstration, we create an artificial correlation matrix of these four variables by defining a latent variable, θ, with factor loadings theta. The product of $\theta\theta^T$ is the observed correlation matrix:

R code

```
theta <- matrix(c(0.8, 0.7, 0.6, 0.5), nrow = 4)
observed <- theta %*% t(theta)
diag(observed) <- 1
rownames(observed) <- colnames(observed) <- c("breath", "teeth",
 "fingers", "cancer")
observed
```

```
        breath teeth fingers cancer
breath    1.00  0.56    0.48   0.40
teeth     0.56  1.00    0.42   0.35
fingers   0.48  0.42    1.00   0.30
cancer    0.40  0.35    0.30   1.00
```

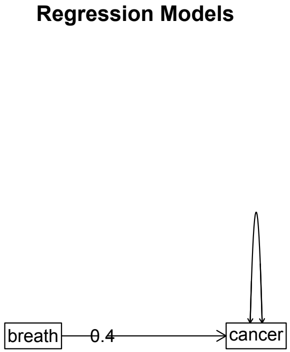
Misspecified linear regression: A series of models

R code

```
setCor(y="cancer",x="breath",data=observed)
```

```
Call: setCor(y = "cancer", x = "breath", data = observed)

Multiple Regression from matrix input

Beta weights
       cancer
breath    0.4

Multiple R
       cancer
cancer    0.4
multiple R2
       cancer
cancer   0.16

 Unweighted multiple R
cancer
   0.4
 Unweighted multiple R2
cancer
  0.16
>
```

1 predictor



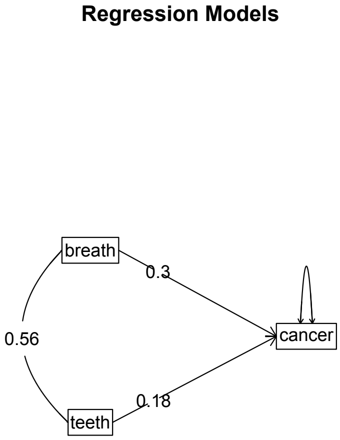
Misspecified linear regression: A series of models

R code

```
setCor(y="cancer",x=c("breath","teeth"),data=observed)
```

```
Call: setCor(y = "cancer", x = c("breath", "teeth"), data = observed)

Multiple Regression from matrix input

Beta weights
       cancer
breath   0.30
teeth    0.18

Multiple R
       cancer
cancer   0.43
multiple R2
       cancer
cancer   0.18

 Unweighted multiple R
cancer
  0.42
 Unweighted multiple R2
cancer
  0.18
```

2 predictors



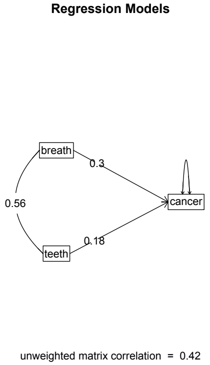
Misspecified linear regression: A series of models

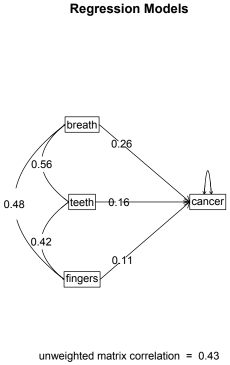
R code

```
setCor(y="cancer",x=c("breath","teeth","fingers"),data=observed)
```

```
Call: setCor(y = "cancer", x = c("breath", "teeth", "fingers"), data = observed)

Multiple Regression from matrix input

Beta weights
        cancer
breath    0.26
teeth     0.16
fingers   0.11

Multiple R
       cancer
cancer   0.44
multiple R2
       cancer
cancer   0.19

 Unweighted multiple R
cancer
  0.43
 Unweighted multiple R2
cancer
  0.19
```

2 and 3 predictors





Add in a perfectly causal variable

R code

```
theta <- matrix(c(1, 0.8, 0.7, 0.6, 0.5), nrow = 5)
observed <- theta %*% t(theta)
diag(observed) <- 1
rownames(observed) <- colnames(observed) <- c("smoking", "breath",
   "teeth", "fingers", "cancer")
 observed
```

```
observed
        smoking breath teeth fingers cancer
smoking     1.0   0.80  0.70    0.60   0.50
breath      0.8   1.00  0.56    0.48   0.40
teeth       0.7   0.56  1.00    0.42   0.35
fingers     0.6   0.48  0.42    1.00   0.30
cancer      0.5   0.40  0.35    0.30   1.00
```

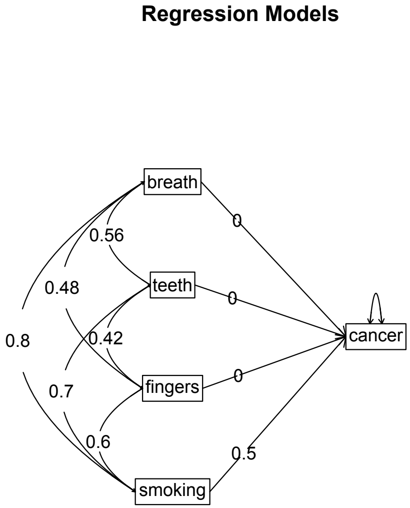
Try the regression with all 4 predictors

R code

```
setCor(y="cancer",x=c("breath","teeth","fingers","smoking"),data=observed)
```

```
Call: setCor(y = "cancer", x = c("breath", "teeth", "fingers", "smoking"),
    data = observed)

Multiple Regression from matrix input

Beta weights
        cancer
breath     0.0
teeth      0.0
fingers    0.0
smoking    0.5

Multiple R
       cancer
cancer    0.5
multiple R2
       cancer
cancer   0.25

 Unweighted multiple R
cancer
  0.46
 Unweighted multiple R2
cancer
  0.22
```

A perfect predictor



But what if the causal variable is not a perfect measure?

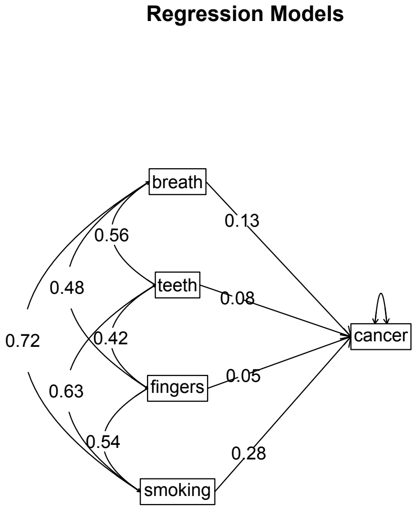
Add in some error to smoking

R code

```
theta <- matrix(c(.9, 0.8, 0.7, 0.6, 0.5), nrow = 5)
observed <- theta %*% t(theta)
diag(observed) <- 1
rownames(observed) <- colnames(observed) <- c("smoking", "breath",
   "teeth", "fingers", "cancer")
 observed
```

	smoking	breath	teeth	fingers	cancer
smoking	1.00	0.72	0.63	0.54	0.45
breath	0.72	1.00	0.56	0.48	0.40
teeth	0.63	0.56	1.00	0.42	0.35
fingers	0.54	0.48	0.42	1.00	0.30
cancer	0.45	0.40	0.35	0.30	1.00

An imperfect predictor



Improper specification of a mediation model will produce incorrect results

1. Although mediation models are popular ways of organizing regression models, they are not guaranteed to be corrrect.
2. Even with time lags built into the model, the relationships are not necessarily correct.
3. Consider our smoking - cancer example

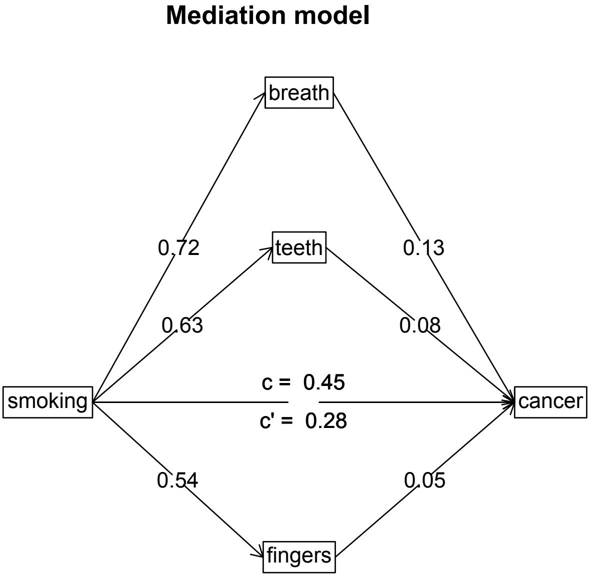
Does hygiene mediate smoking - cancer relationship?

R code

```
mediate(y="cancer",m=c("breath","teeth","fingers"),x="smoking",data=observed)
```

```
Mediation analysis
Call: mediate(y = "cancer", x = "smoking", m = c("breath", "teeth",
    "fingers"), data = observed)

The DV (Y) was  cancer . The IV (X) was  smoking . The mediating variable(s) =  breath teeth fingers .

Total Direct effect(c) of  smoking  on  cancer  =  0.45   S.E. =  0.03  t direct =  15.92
                              with probability =  0
Direct effect (c') of  smoking  on  cancer  removing  breath teeth fingers  =  0.28
                         S.E. =  0.05  t direct =  6.08   with probability =  1.8e-09
Indirect effect (ab) of  smoking  on  cancer  through  breath teeth fingers   =  0.17
Mean bootstrapped indirect effect =  0.15  with standard error =  0.03
                                                  Lower CI =  0.08    Upper CI =  0.22
Summary of a, b, and ab estimates and ab confidence intervals
           a cancer   ab mean.ab ci.ablower ci.abupper
breath  0.72   0.13 0.09    0.05       0.00       0.11
teeth   0.63   0.08 0.05    0.04       0.00       0.09
fingers 0.54   0.05 0.03    0.06       0.02       0.09

ratio of indirect to total effect=   0.39
ratio of indirect to direct effect=  0.63
```

Smoking effect on cancer is partially mediated by hygiene



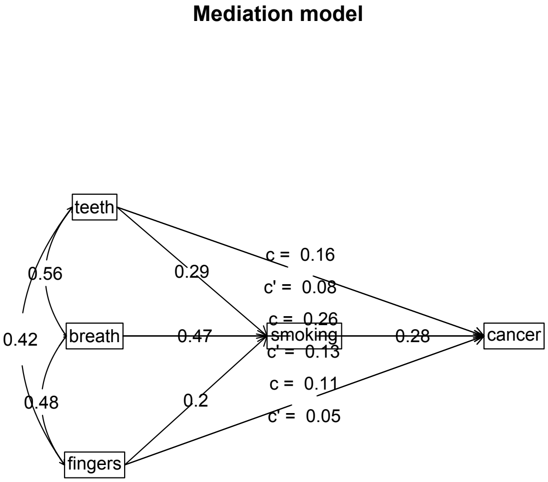
Try it the other way: smoking is the mediator

R code

```
mediate(y="cancer",x=c("breath","teeth","fingers"),m="smoking",data=observed)
```

```
Mediation analysis
Call: mediate(y = "cancer", x = c("breath", "teeth", "fingers"), m = "smoking",
    data = observed)

The DV (Y) was  cancer . The IV (X) was  breath teeth fingers . The mediating variable(s) =  smoking .

Total Direct effect(c) of  breath  on  cancer  =  0.26   S.E. =  0.03  t direct =  8.62   with p =  0
Direct effect (c') of  breath  on  cancer  removing  smoking  =  0.13   S.E. =  0.04  t direct =  3.12
Indirect effect (ab) of  breath  on  cancer  through  smoking   =  0.13
Mean bootstrapped indirect effect =  0.13  with standard error =  0.02  Lower CI =  0.08    Upper CI =

Total Direct effect(c) of  teeth  on  cancer  =  0.16   S.E. =  0.03  t direct =  5.19   p =  2.6e-07
Direct effect (c') of  teeth  on  cancer  removing  smoking  =  0.08   S.E. =  0.04  t direct =  2.17
Indirect effect (ab) of  teeth  on  cancer  through  smoking   =  0.08
Mean bootstrapped indirect effect =  0.07  with standard error =  0.01  Lower CI =  0.05    Upper CI =

Total Direct effect(c) of  fingers  on  cancer  =  0.11   S.E. =  0.03  t direct =  3.5   p=  0.00048
Direct effect (c') of  fingers  on  cancer  removing  smoking  =  0.05   S.E. =  0.03  t direct =  1.62
Indirect effect (ab) of  fingers  on  cancer  through  smoking   =  0.05
Mean bootstrapped indirect effect =  0.06  with standard error =  0.01  Lower CI =  0.04    Upper CI =
Summary of a, b, and ab estimates and ab confidence intervals
           a cancer   ab mean.ab ci.ablower ci.abupper
breath  0.47   0.28 0.13    0.13       0.08       0.17
teeth   0.29   0.28 0.08    0.07       0.05       0.10
fingers 0.20   0.28 0.05    0.06       0.04       0.08

ratio of indirect to total effect=   0.5 0.5 0.5
ratio of indirect to direct effect=  0.99 0.61 0.42
```

Smoking partially mediates hygiene effects



Latent variable modeling can correct for unreliability of measurement

1. Part of the problem is regression is measures are imperfect
2. We can not see that in the raw correlations
3. Can either estimate reliabilities (somehow) or can form latent variables

Try lavaan

R code

```
library(lavaan)
 model1 <- 'theta =~  breath + teeth + fingers
        '
fit1 <-  cfa(model=model1,sample.cov=observed,sample.nobs=1000,std.lv=TRUE)
summary(fit1)
lavaan.diagram(fit1)
```

```
avaan (0.5-18) converged normally after  14 iterations

  Number of observations                          1000

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0
  Minimum Function Value               0.0000000000000

Parameter estimates:
  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  theta =~
    breath            0.800    0.035   22.813    0.000
    teeth             0.700    0.034   20.413    0.000
    fingers           0.600    0.034   17.855    0.000
Variances:
    breath            0.360    0.041
    teeth             0.509    0.037
    fingers           0.639    0.036
```

Add in the Cancer variable

R code

```
model2 <- 'theta =~  breath + teeth + fingers
           cancer ~ theta
        '
fit2 <-  cfa(model=model2,sample.cov=observed,sample.nobs=1000,std.lv=TRUE)
summary(fit2)
lavaan.diagram(fit2)
```

```
lavaan (0.5-18) converged normally after  15 iterations
  Number of observations                          1000
  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 2
  P-value (Chi-square)                           1.000
Parameter estimates:
  Information                                 Expected
  Standard Errors                             Standard
                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  theta =~
    breath            0.800    0.032   24.874    0.000
    teeth             0.700    0.032   21.645    0.000
    fingers           0.600    0.033   18.319    0.000
Regressions:
  cancer ~
    theta             0.500    0.034   14.897    0.000
```

Add in smoking

R code

```
model3 <- 'theta =~ smoking +  breath + teeth + fingers
           cancer ~ theta
        '
fit3 <-  cfa(model=model3,sample.cov=observed,sample.nobs=1000,std.lv=TRUE)
summary(fit3)
lavaan.diagram(fit3)
```

```
lavaan (0.5-18) converged normally after  16 iterations

  Number of observations                          1000

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 5
  P-value (Chi-square)                           1.000
Parameter estimates:
  Information                                 Expected
  Standard Errors                             Standard
                   Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  theta =~
    smoking           0.900    0.026   34.113    0.000
    breath            0.800    0.028   28.838    0.000
    teeth             0.700    0.029   24.109    0.000
    fingers           0.600    0.030   19.824    0.000
Regressions:
  cancer ~
    theta             0.500    0.031   15.975    0.000
```

44 ways to fool yourself with SEM

Adapted from Rex Kline; Principals and Practice of Structural Equation Modeling, 2005

1. Specification
2. Data
3. Analysis and Respecification
4. Interpretation

Specification errors

1. Specifying the model after the data are collected.
    - Particularly a problem when using archival data.
2. Are key variables omitted?
3. Is the model identifiable?
4. Omitting causes that are correlated with other variables in the structural model.
5. Failure to have sufficient number of indicators of latent variables.
    - "Two might be fine, three is better, four is best, anything more is gravy" (Kenny, 1979)
6. Failure to give careful consideration to directionality.
    - Path techniques are good for estimating strengths if we know the underlying model, but are not good for determining the model (Meehl and Walker, 2002)

Specification errors (continued)

7. Specifying feedback loops ("non recursive models") as a way to mask uncertainty
8. Overfit the model, ignoring parsimony
9. Add disturbances ("measurement error correlations" aka "correlated residuals") with substantive reason
10. Specifying indicators that are multivocal without substantive reason

Data Errors

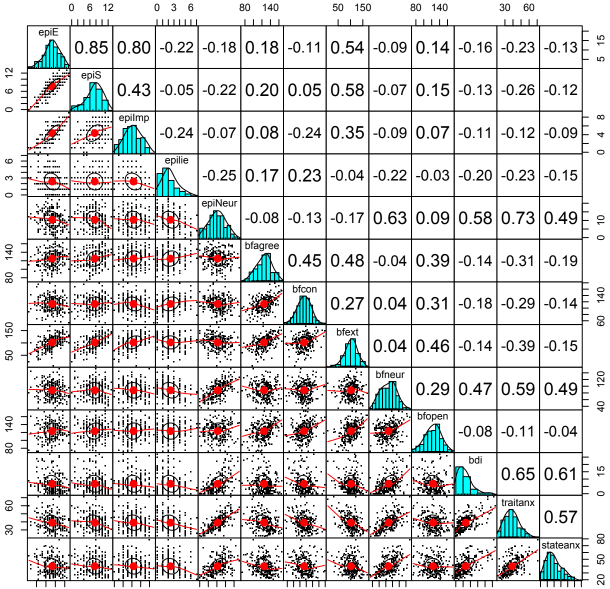
1. Failure to check the accuracy of data input or coding
    - Missing data codes (use a clear missing value)
    - Misytyped, mis-scanned data matrices
    - Improperly reversed items
        - Let the computer do it for you
        - Why reverse an item when a negative sign will do it for you?
2. Ignoring the pattern of missing data, is it random or systematic.
3. Failure to examine distributional characteristics
    - Weird data -> weird results
4. Failure to screen for outliers
    - Outliers due to mistakes
    - Outliers due to systematic differences

Describe the data

```
> describe(epi.bfi)

 pairs.panels(epi.bfi,pch=".",gap=0)   #mind the gap
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
epiE	1	231	13.33	4.14	14	13.49	4.45	1	22	21	-0.33	-0.06	0.27
epiS	2	231	7.58	2.69	8	7.77	2.97	0	13	13	-0.57	-0.02	0.18
epiImp	3	231	4.37	1.88	4	4.36	1.48	0	9	9	0.06	-0.62	0.12
epilie	4	231	2.38	1.50	2	2.27	1.48	0	7	7	0.66	0.24	0.10
epiNeur	5	231	10.41	4.90	10	10.39	4.45	0	23	23	0.06	-0.50	0.32
bfagree	6	231	125.00	18.14	126	125.26	17.79	74	167	93	-0.21	-0.27	1.19
bfcon	7	231	113.25	21.88	114	113.42	22.24	53	178	125	-0.02	0.23	1.44
bfext	8	231	102.18	26.45	104	102.99	22.24	8	168	160	-0.41	0.51	1.74
bfneur	9	231	87.97	23.34	90	87.70	23.72	34	152	118	0.07	-0.55	1.54
bfopen	10	231	123.43	20.51	125	123.78	20.76	73	173	100	-0.16	-0.16	1.35
bdi	11	231	6.78	5.78	6	5.97	4.45	0	27	27	1.29	1.50	0.38
traitanx	12	231	39.01	9.52	38	38.36	8.90	22	71	49	0.67	0.47	0.63
stateanx	13	231	39.85	11.48	38	38.92	10.38	21	79	58	0.72	-0.01	0.76

Graphic descriptions using SPLOMs



Data errors (continued)

5. Assuming all relationships are linear without checking
   - graphical techniques are helpful for non-linearities
   - Simple graphical techniques do not help for interactions
6. Ignoring lack of independence among observations
   - Nesting of subjects within pairs, within classrooms, with managers
   - Can we model the nesting?

Errors of analysis and respecification

1. Failure to check the accuracy of computer syntax
   - Direction of effects
   - Error specifications
   - Omitted paths
2. Respecifying the model based entirely on statistical criteria
   - Just because it does not fit does not mean it should be fixed
3. Failure to check for admissible solutions
   - Are some of the paths impossible?
   - Do some of the variables have negative variances?
4. Reporting only standardized estimates
   - These are sample based estimates and reflect variances (errorful) and covariances (supposedly error free)
5. Analyzing a correlation matrix when the covariance matrix is more appropriate
   - Anything that has growth or change component must be done with covariances

Errors of Analysis and respecification (continued)

6. Analyzing a data set with extremely high correlations
   - solution will either be unstable or will not work if variables are too "colinear"
7. Not enough subjects for complexity of the data
   - This is ambiguous – what is enough?
   - Remember, the standard error of a correlation reflects sample size $se_r = \frac{1}{\sqrt{(1-r^2)(n-2)}}$
   - And thus, the t value associated with any correlation is $\frac{r}{\sqrt{(1-r^2)(n-2)}}$

Errors of Analysis and respecification (continued)

8. Setting scales of latent variables inappropriately.
    - particularly a problem with multiple group comparisons
9. Ignoring the start values or giving bad ones.
    - Supplying reasonable start values helps a great deal
10. Do different start values lead to different solutions?
11. Failure to recognize empirical underidentification
    - for some data sets, the model is underidentified even though there are enough parameters
    - Failure to separate measurement from structural portion of model
        - Use the two or four step procedure

Errors of Analysis and respecification (continued)

12. Estimating means and intercepts without showing measurement invariance
13. Analyzing parcels without checking if parcels are in fact factorially homogeneous.
    - Factorial Homogeneous Item Domains (FHID)
    - Homogenous Item Composites (HIC)
    - (but consider contradictory advice on parcels)

Errors of Interpretation

1. Looking only at indexes of overall fit
   - need to examine the residuals to see where there is misfit, even though overall model is fine
2. Interpreting good fit as meaning model is "proved".
   - consider alternative models
   - better able to reject alternatives
3. Interpreting good fit as meaning that the endogenous variables are strongly predicted.
   - What is predicted is the covariance of the variables, not the variables
   - Are the residual covariances small, not whether the error variance is small
4. Relying solely on statistical criterion in model evaluation
   - What can the model not explain
   - What are alternative models
   - What constraints does the model imply

Errors of interpretation (continued)

5. Relying too much on statistical tests
   - significance of particular path coefficients does not imply effect size or importance
   - Overall statistical fit (χ^2) is sensitive to model misfit as f(N)
6. Misinterpreting the standardized solution in multiple group problems
7. Failure to consider equivalent models
   - Why is this model better than equivalent models?
8. Failure to consider non-equivalent models
   - Why is this model better than other, non-equivalent models?
9. Reifying the latent variables
   - Latent variables are just models of observed data
   - "Factors are fictions"
10. Believing that naming a factor means it is understood

Errors of interpretation (continued)

11. Believing that a strong analytical method like SEM can overcome poor theory or poor design.
12. Failure to report enough so that you can be replicated
13. Interpreting estimates of large effects as evidence for "causality"

Final Comments

1. Theory First
   - What are the alternative theories?
   - Are there specific differences in the theories that are testable?
2. Measurement Model
   - Comparison of measurement models?
   - How many latent variables? How do you know?
   - Measurement Invariance?
3. Structural Model
   - Comparison of multiple models?
   - What happens if the arrows are reversed?
4. Theory Last
   - What do we know now that we did not know before?
   - What do we have shown is not correct?

Conclusion

1. Latent variable models are a powerful theoretical aid but do not replace theory.
2. Nor do latent modeling algorithms replace the need for good scale development.
3. Latent variable models are a supplement to the conventional regression models of observed scores.
4. Other latent models (not fully considered) include
   - Item Response Theory
   - Latent Class Analysis
   - Latent Growth Curve analysis

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics* chapter 10, (pp. 275–340). New York: Routledge.

McDonald, R. & Marsh, H. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*(2), 247–255.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1 – 12.