Statistical Analysis Using Structural Equation Models

EPsy 8266

Christopher David Desjardins

Research Methodology Consulting Center

4/2/19

Early Disruptive Behavior, IQ, and Later School Achievement and Delinquent Behavior

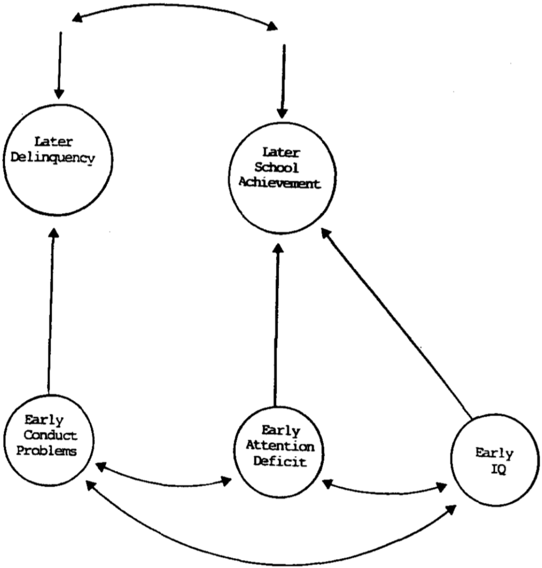
In this paper we develop a structural equation model to test the hypothesis that the **correlations between school achievement and delinquency are noncausal** and arise from the **common and correlated effects of early disruptive behaviors and early IQ** on both outcomes.

The models in Figs. 1a and 1b differed in their assumptions about the extent to which the **school achievement/delinquency correlation is predicted and explained by early behavior and IQ**.

The model in **Fig. 1a assumes that the common and correlated effects of the early explanatory variables are sufficient** to explain any correlation between school achievement and delinquency.

The model in **Fig. 1b assumes that even after the common and correlated effects of early behavior and IQ are taken into account significant correlations remain** between school performance and delinquency.

Theoretical model: 1b



Sample

709 of 1265 children used in this study.

A multiple-indicator approach was used to define latent constructs

"... the IQ and scholastic ability measure used indicators derived from split half measures of these tests."

Their SEM

1. Conduct problems, attention deficit and IQ at age eight are assumed to be correlated exogenous or independent variables.
2. Early conduct problems are related to later delinquency but are not directly related to later scholastic ability.
3. Early attention deficit and IQ are related to later scholastic ability but are not related to later delinquency.
4. When the antecedent effects of early behaviors and IQ are taken into account, scholastic ability and delinquent behaviors are permitted to be correlated.

Read in their data

```
library(lavaan)
lower.cor <- "
1
.329 1
.487 .277 1
.313 .710 .416 1
-.225 -.189 -.220 -.277 1
-.216 -.180 -.228 -.269 .883 1
.366 .322 .259 .346 -.217 -.222 1
.254 .279 .175 .261 -.158 -.146 .572 1
.162 .243 .142 .228 -.137 -.140 .438 .361 1
-.229 -.215 -.270 -.370 .696 .685 -.251 -.199 -.175 1
-.241 -.248 -.296 -.389 .688 .689 -.247 -.183 -.177 .908 1
"
s <- c(4.617, 4.982, 3.054, 3.388, 12.506, 12.049, 1.110, 1.609,
       .420, 7.657, 7.798)
cov.mat <- getCov(lower.cor, sds = s,
                  names = c("cp.mom", "cp.teach", "ad.mom", "ad.teach",
                            "iq.halfa", "iq.halfb", "del.mom", "del.self",
                            "del.cop", "ab.halfa", "ab.halfb"))
```

Fit their measurement model

```
meas.mod <- "
cp =~ cp.mom + cp.teach
ad =~ ad.mom + ad.teach
iq =~ iq.halfa + iq.halfb
del =~ del.mom + del.self + del.cop
ab =~ ab.halfa + ab.halfb

# reporter covariance
cp.mom ~~ ad.mom + del.mom
# ad.mom ~~ del.mom # reporter not included
cp.teach ~~ ad.teach
"
meas.fit <- cfa(meas.mod, sample.cov = cov.mat, sample.nobs = 709)
```

How does the measurement model fit?

```
fitmeasures(meas.fit, c("chisq", "df", "pvalue", "rmsea", "tli", "cfi"))

##  chisq     df pvalue  rmsea    tli    cfi
## 25.226 31.000  0.758  0.000  1.002  1.000

params <- parameterestimates(meas.fit, standardized = TRUE)
subset(params, lhs == "del" & rhs == "ab")

##    lhs op rhs    est    se      z pvalue ci.lower ci.upper std.lv
## 40 del ~~  ab -2.101 0.309 -6.793      0   -2.707   -1.495 -0.311
##    std.all std.nox
## 40  -0.311  -0.311
```

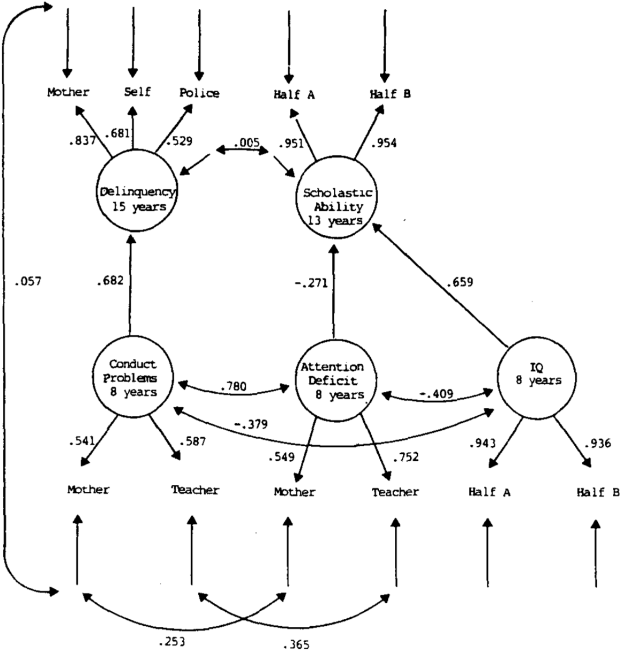
Fit the SR model

```
sr.mod <- "
# measurement models
cp =~ cp.mom + cp.teach
ad =~ ad.mom + ad.teach
iq =~ iq.halfa + iq.halfb
del =~ del.mom + del.self + del.cop
ab =~ ab.halfa + ab.halfb

# structural paths
del ~ cp
ab ~ ad + iq

# covariances among factors
del ~~ ab
cp ~~ ad + iq
ad ~~ iq

# reporter covariance
cp.mom ~~ ad.mom + del.mom
# ad.mom ~~ del.mom # not in their model why?
cp.teach ~~ ad.teach
"
sr.fit <- cfa(sr.mod, sample.cov = cov.mat, sample.nobs = 709)
```

The value of the overall goodness of fit index (Joreskog & Sorbom, 1989) was .99, suggesting a well-fitting model,

and the value of the log likelihood chi square associated with the fitted model showed no significant departures in the fit of the model from the data (Log likelihood chi square = 27.6; df = 34; p > .60).

```
anova(meas.fit, sr.fit)

## Chi Square Difference Test
##
##          Df   AIC   BIC  Chisq Chisq diff Df diff Pr(>Chisq)
## meas.fit 30 37977 38142 24.772
## sr.fit   34 37972 38118 27.029     2.2569       4     0.6886

fitmeasures(sr.fit, c("chisq", "df", "pvalue", "gfi"))

##  chisq     df pvalue    gfi
## 27.029 34.000  0.796  0.993
```


Mother
Self
Police
Half A
Half B
.837
.681
.529
.005
.951
.954
Delinquency 15 years
Scholastic Ability 13 years
.057
.682
-.271
.659
Conduct Problems 8 years
Attention Deficit 8 years
IQ 8 years
.780
-.409
-.379
.541
.587
.549
.752
.943
.936
Mother
Teacher
Mother
Teacher
Half A
Half B
.253
.365

```
summary(sr.fit, standardized = TRUE, fit.measures = TRUE)

##   Estimator                                         ML
##   Model Fit Test Statistic                      27.029
##   Degrees of freedom                                34
##   P-value (Chi-square)                           0.796
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             4503.399
##   Degrees of freedom                                55
##   P-value                                        0.000
##
##   Comparative Fit Index (CFI)                    1.000
##   Tucker-Lewis Index (TLI)                       1.003
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.000
##   90 Percent Confidence Interval          0.000  0.018
##   P-value RMSEA <= 0.05                          1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.014
##
```

```
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   cp =~
##     cp.mom            1.000                               2.492    0.541
##     cp.teach          1.172    0.137    8.539    0.000    2.920    0.586
##   ad =~
##     ad.mom            1.000                               1.674    0.548
##     ad.teach          1.520    0.155    9.808    0.000    2.545    0.752
##   iq =~
##     iq.halfa          1.000                              11.787    0.943
##     iq.halfb          0.956    0.025   38.902    0.000   11.272    0.936
##   del =~
##     del.mom           1.000                               0.928    0.837
##     del.self          1.179    0.086   13.749    0.000    1.095    0.681
##     del.cop           0.239    0.020   11.765    0.000    0.222    0.529
##   ab =~
##     ab.halfa          1.000                               7.280    0.951
##     ab.halfb          1.022    0.022   45.891    0.000    7.437    0.954
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   del ~
##     cp                0.254    0.028    9.029    0.000    0.681    0.681
##   ab ~
##     ad               -1.178    0.177   -6.650    0.000   -0.271   -0.271
##     iq                0.407    0.021   19.850    0.000    0.659    0.659
```

```
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  .cp.mom ~~
##    .ad.mom            3.563    0.486    7.329    0.000    3.563    0.360
##    .del.mom           0.293    0.139    2.114    0.034    0.293    0.125
##  .cp.teach ~~
##    .ad.teach          6.162    0.685    8.996    0.000    6.162    0.685
##  .del ~~
##    .ab                0.024    0.182    0.134    0.894    0.008    0.008
##   cp ~~
##     ad                3.256    0.490    6.651    0.000    0.780    0.780
##     iq              -11.206    1.735   -6.458    0.000   -0.381   -0.381
##   ad ~~
##     iq               -8.081    1.163   -6.946    0.000   -0.410   -0.410
```

Why did they omit that one reporter covariance?

```r
sr.mod2 <- "
# measurement models
cp =~ cp.mom + cp.teach
ad =~ ad.mom + ad.teach
iq =~ iq.halfa + iq.halfb
del =~ del.mom + del.self + del.cop
ab =~ ab.halfa + ab.halfb

# reporter covariance
cp.mom ~~ ad.mom + del.mom
ad.mom ~~ del.mom # not in their model why?
cp.teach ~~ ad.teach

# structural paths
del ~ cp
ab ~ ad + iq

del ~~ ab
cp ~~ ad + iq
ad ~~ iq
"
sr.fit2 <- cfa(sr.mod2, sample.cov = cov.mat, sample.nobs = 709)
fitmeasures(sr.fit2, c("chisq", "df", "pvalue", "gfi"))

##  chisq     df pvalue    gfi
## 26.525 33.000  0.780  0.993
```

```
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   del ~
##     cp                0.252    0.028    8.970    0.000    0.675    0.675
##   ab ~
##     ad               -1.189    0.179   -6.635    0.000   -0.270   -0.270
##     iq                0.407    0.021   19.860    0.000    0.659    0.659
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  .cp.mom ~~
##    .ad.mom            3.655    0.506    7.220    0.000    3.655    0.366
##    .del.mom           0.345    0.156    2.204    0.027    0.345    0.145
##  .ad.mom ~~
##    .del.mom           0.066    0.093    0.709    0.478    0.066    0.042
##  .cp.teach ~~
##    .ad.teach          6.081    0.702    8.667    0.000    6.081    0.684
```

Correlation residuals

```
lavResiduals(sr.fit, type = "cor")$cov

##           cp.mom cp.tch ad.mom ad.tch iq.half iq.hlfb del.mm dl.slf del.cp ab.half ab.hl
## cp.mom     0.000
## cp.teach   0.012  0.000
## ad.mom     0.002  0.026  0.000
## ad.teach  -0.005  0.000  0.004  0.000
## iq.halfa  -0.030  0.022 -0.008  0.013  0.000
## iq.halfb  -0.023  0.029 -0.018  0.019  0.000   0.000
## del.mom    0.000 -0.012  0.015  0.011 -0.012  -0.018   0.000
## del.self   0.003  0.007 -0.023 -0.011  0.009   0.020   0.002  0.000
## del.cop   -0.033  0.032 -0.012  0.017 -0.007  -0.011  -0.005  0.001  0.000
## ab.halfa   0.009  0.043  0.012  0.017  0.005  -0.001  -0.003  0.003 -0.018  0.000
## ab.halfb  -0.002  0.011 -0.013 -0.001 -0.005   0.001   0.002  0.020 -0.020  0.000   0.00

# How many are greater than .1?
#  - divided by 2 bc this is a full matrix not just a tri
sum(abs(lavResiduals(sr.fit, type = "cor")$cov) > .1)  / 2

## [1] 0
```

Standardized residuals

```
lavResiduals(sr.fit, type = "cor")$cov.z

##          cp.mom cp.tch ad.mom ad.tch iq.half iq.hlfb del.mm dl.slf del.cp ab.half ab.hl
## cp.mom    0.000
## cp.teach  0.829  0.000
## ad.mom    0.387  1.248  0.000
## ad.teach -0.320  0.120  0.465  0.000
## iq.halfa -1.277  1.069 -0.358  0.954  0.000
## iq.halfb -0.944  1.410 -0.759  1.318 -1.111   0.000
## del.mom   0.006 -0.876  0.675  0.759 -0.530  -0.814   0.000
## del.self  0.135  0.324 -0.854 -0.514  0.330   0.719   0.600  0.000
## del.cop  -1.221  1.162 -0.391  0.605 -0.236  -0.361  -0.712  0.056  0.000
## ab.halfa  0.402  2.047  0.666  1.482  1.732  -0.234  -0.158  0.119 -0.613  0.000
## ab.halfb -0.086  0.532 -0.767 -0.095 -1.788   0.401   0.109  0.798 -0.665  0.000   0.00

# How many are greater than 1.96?
#  - divided by 2 bc this is a full matrix not just a tri
sum(abs(lavResiduals(sr.fit, type = "cor")$cov.z) > 1.96) / 2

## [1] 1
```

One after thought

```
two.mod <- "
del =~ del.mom + del.self + del.cop
ab =~ ab.halfa + ab.halfb
"
two.fit <- cfa(two.mod, sample.cov = cov.mat, sample.nobs = 709)
summary(two.fit)

## lavaan 0.6-3 ended normally after 129 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         11
##
##   Number of observations                           709
##
##   Estimator                                         ML
##   Model Fit Test Statistic                       1.771
##   Degrees of freedom                                 4
##   P-value (Chi-square)                           0.778
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard Errors                             Standard
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   del =~
##     del.mom           1.000
##     del.self          1.187    0.097   12.244    0.000
```

Dealing with mean structures

- So far this semester we have been dealing entirely with covariance or correlation matrices.
- This results in the implicit assumption that the means of all of our latent variables are zero (i.e., they are mean-centered).
- But what if these are of substantitive interest?
  - What if we want to know how the mean of a latent variable changes over time?
  - What if we want to know if one group is higher or lower on a latent variable?

Revisiting regression

Let's say we want to predict a participant's depression given their hypochondriasis.

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

and

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

where Y_i is participant i depression and X_i is their hypochondriasis.

How do we calculate $\hat{\beta}_0$ and $\hat{\beta}_1$?

Estimating the parameters

We know that $\hat{\beta}_1$ is a function of the **covariance** between depression and hypochondriasis.

$\hat{\beta}_1 = r_{xy}\frac{s_Y}{s_x} = \frac{cov_{xy}}{s_X^2}$

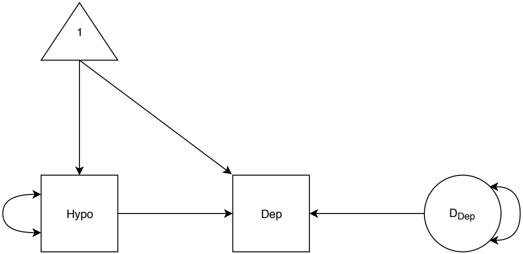
We know that $\hat{\beta}_0$ is a function (partially) of the **means** of depression and hypochondriasis.

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

```r
# wuschiz data on course website
wuschiz <- read.csv("https://tinyurl.com/y36sdp42")
# slope
b1 <- with(wuschiz, cov(D, Hs) / var(Hs)); b1

## [1] 0.7613852

# intercept
b0 <- mean(wuschiz$D) - b1*mean(wuschiz$Hs); b0

## [1] 17.51529

lm(D ~ 1 + Hs, wuschiz)

##
## Call:
## lm(formula = D ~ 1 + Hs, data = wuschiz)
##
## Coefficients:
## (Intercept)           Hs
##     17.5153       0.7614
```

Constants

What if we add a constant (1)?



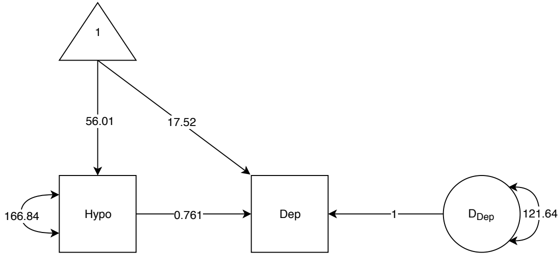
In this model, we have both a covariance and a mean structure.

Fitting constants in lavaan

```r
library(lavaan)
mod <- "
D ~ 1 + Hs
Hs ~ 1

# disturbance
D ~~ D

# variance of hypo
Hs ~~ Hs
"
fit <- lavaan(model = mod, data = wuschiz)
parameterestimates(fit)
```

```
##   lhs op rhs     est     se      z pvalue ci.lower ci.upper
## 1   D ~1      17.515  2.531  6.921      0   12.555   22.476
## 2   D  ~  Hs   0.761  0.044 17.291      0    0.675    0.848
## 3  Hs ~1      56.005  0.666 84.077      0   54.700   57.311
## 4   D ~~   D 121.635  8.871 13.711      0  104.248  139.022
## 5  Hs ~~  Hs 166.835 12.168 13.711      0  142.987  190.683
```

Constants



1. What is the direct effect of the constant on hypochondriasis (the exogenous variable)?
2. What is the direct effect of the constant on depression (the endogenous variable)?
3. What is the total effect of the constant on depression?

Understanding the constant

1. 56.005
2. 17.515
3. 17.515 + 56.005 * .761 = 60.135

```
#1
mean(wuschiz$Hs)

## [1] 56.00532

#2
coef(lm(D ~ 1 + Hs, wuschiz))["(Intercept)"]

## (Intercept)
##    17.51529

#3
mean(wuschiz$D)

## [1] 60.15691
```

Recap

- The mean of an endogenous variable is a function of: 1) intercept, 2) the unstandardized path coefficient, and 3) the mean the exogenous variable.
- The predicted mean for exo/endo variables is the total effect of the constant.
- For exogenous continuous variables in an SEM, the unstandardized path coefficient for the direct effect of the constant is a mean.
- For endogenous continuous variables in an SEM, the direct effect of the constant is the intercept but the total effect is a mean.

New rule for counting observations

When we include covariances and means, we now have a new rule for the number of observations:

$$v(v + 3)/2$$

The gives the 1) number of variances, 2) number of means, and 3) number of covariances.

For the previous path diagram, how many observations do we have?

More on identification

- Identification of the means is separate from the covariances.
- For example, overidentified covariances won't help underidentified means.
- If the means are just-identified,
  1. The predicted means (the total effects of the constants) will exactly equal the observed means
  2. The fit of the model with just covariance structure will be identical to the one with both a mean and a covariance structure.

Overidentified means

When means are overidentified then ...

- Predicted means could differ from observed means, resulting in **mean residuals** (observed mean - predicted mean).
- A mean residual over the estimated SE of that residual, **standardized mean residual**, is approx. a z-test in a large sample.

The independence model

- Recall the independence model is that the variables uncorrelated with one another. (This is the second model you see printed in when you ask for fit.measures = TRUE).
- Some fit measures use this independence model (as you'll see on Thursday).
- An independence model that fixed covariances and means to zero is very unrealistic.
- We'll discuss this when we fit specific models.