# Estimation and Model Fit

SDS 390 Structural Equation Modeling

Monday Mar 25, 2019

# Statistical and Data Sciences

## Presentation of the Major

- Undeclared? Come hear about the SDS major!
- All current majors and minors are welcome!
- Free lunch served from Teapot!

**Tuesday, March 26**

**12:15 pm – 1:05 pm**

**Ford Hall Atrium**

# Women in Statistics and Data Science!

- The Women in Statistics and Data Science conference is an annual conference, and the deadline for submitting an abstract is coming up: ***April 19*** is the last day to submit. Several Smith SDS students attended the conference in 2017 and again in 2018. It was a huge success!

- **Where and When:** The conference is October 3-October 5 in Bellevue Washington.
- **Funding:** For those of you not graduating in 2019, there is funding for SDS students to travel to this conference. ***Priority will be given to fund travel for students who will be presenting at the conference (which requires you submit an abstract by April 19)***, and we may also consider seniority.

# Agenda

- Mid-semester assessment feedback
- Estimation
- Model fit
- Code example
- Lab 5 time
  - Time on Wed too
  - Due date on Thursday night

# How’s it going?

- “I think it's going pretty well!”
- “It has been a challenge, but not a impossible one. I appreciate the push”
- “I think it's going pretty well, sometimes I feel a little confused during an initial lecture and then I'll start to understand later, so I don't get too worried when I don't understand right away.”
- “I think I'm grabbing small chunks of information gradually, however, not enough connection or information have been received so that I feel comfortable about the big picture of SEM in general.”
- “I think SDS 390 is going better than I initially anticipated. It wasn't until very recently that I felt confident in the lab work."

# What's going well?

- "I enjoy doing the labs and I feel like the monday lecture wednesday lab model works well. I'm excited to work more on the project."
- "I like the structure of the class (lecture on one day, lab on the other). In addition, I think the course material is interesting, and presented in a relatively accessible manner. I also enjoy the show-and-tell format and the exposure to different fields that it brings."
- "Working with other people during lab exercise is super helpful."
- "I like the lecture and lab pairings; so lecturing one day and then using it in lab is very helpful."

# What's not going so well?

- "The book is challenging to understand, so I try to do the reading but sometimes don't feel like I've learned anything from it."
- "I think sometimes we go off on tangents about topics that aren't particularly important to the class - maybe minimizing those a little bit would be helpful."
- "The book is kind of confusing, but given it's a graduate level book it's understandable. The bad thing is that I don't know when is it NOT ok to not understand something vs. when is it ok to let go of something I do not understand."
- "Sometimes I feel that the wording in labs is not covered in lecture but it's in the textbook except the wording is somewhat different so that can be a bit of a challenge to figure out."
- "I find myself needing more and more clarifications on the labs or that after turning in the lab that the question being asked of me was not what I answered, so probably having more time for lab during class would be good."
- "We have been jumping around from chapter to chapter and we have used a different name for things. The inconsistency is a little confusing."

# What can YOU do?

- “I think I should probably do a better job of reading the textbook, and maybe asking more questions when I'm not totally sure of things?”
- “I could attend **office hours** and read the readings more slowly.”
- “I think that I could engage a little bit more with the material in the labs. I tend to just go through them semi-mindlessly but think that applying myself more would be very useful.”
- “Since it's the labs that I certainly need more work with is making it to **office hours** more often and working more with my classmates outside of class.”

# What can I change?

- More time for labs! Although lots of class time will now be devoted to projects…
- *Fewer* tangents (no promises!)
- More examples…
  - Today: Prepared SEM example code
- Faster with feedback!

# Parameter Estimation in SEM

# Estimation in SEM

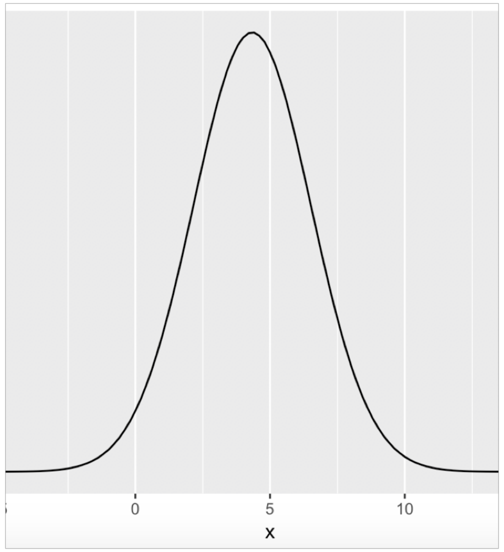
- The default estimation technique in SEM is Maximum likelihood (ML) estimation.
- Assumptions
  - Multivariate normality of the endogenous variables
  - Other stuff…
- What is ML estimation?

# Maximum likelihood estimation

| x |
| ---: |
| 3 |

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

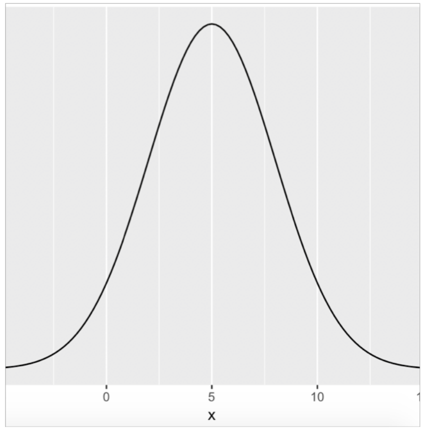
$$f(3|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(3-\mu)^2}{2\sigma^2}}$$



# Maximum likelihood estimation

| x |
|---|
| 3 |
| 5 |
| 5 |
| 6 |
| 7 |
| 1 |
| 5 |
| 6 |
| 1 |

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$




$$\prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

# Other Assumptions of ML estimation

- Multivariate normality of endogenous variables
- Variables are unstandardized
- Assumes no missing data when using the raw data file
- Independence of observations
- Independence of exogenous variables and errors (equal error variance)
- When exogenous variables are measured (not latent), we're assuming they are measured without error
- The model is correctly specified – model misspecification can propagate

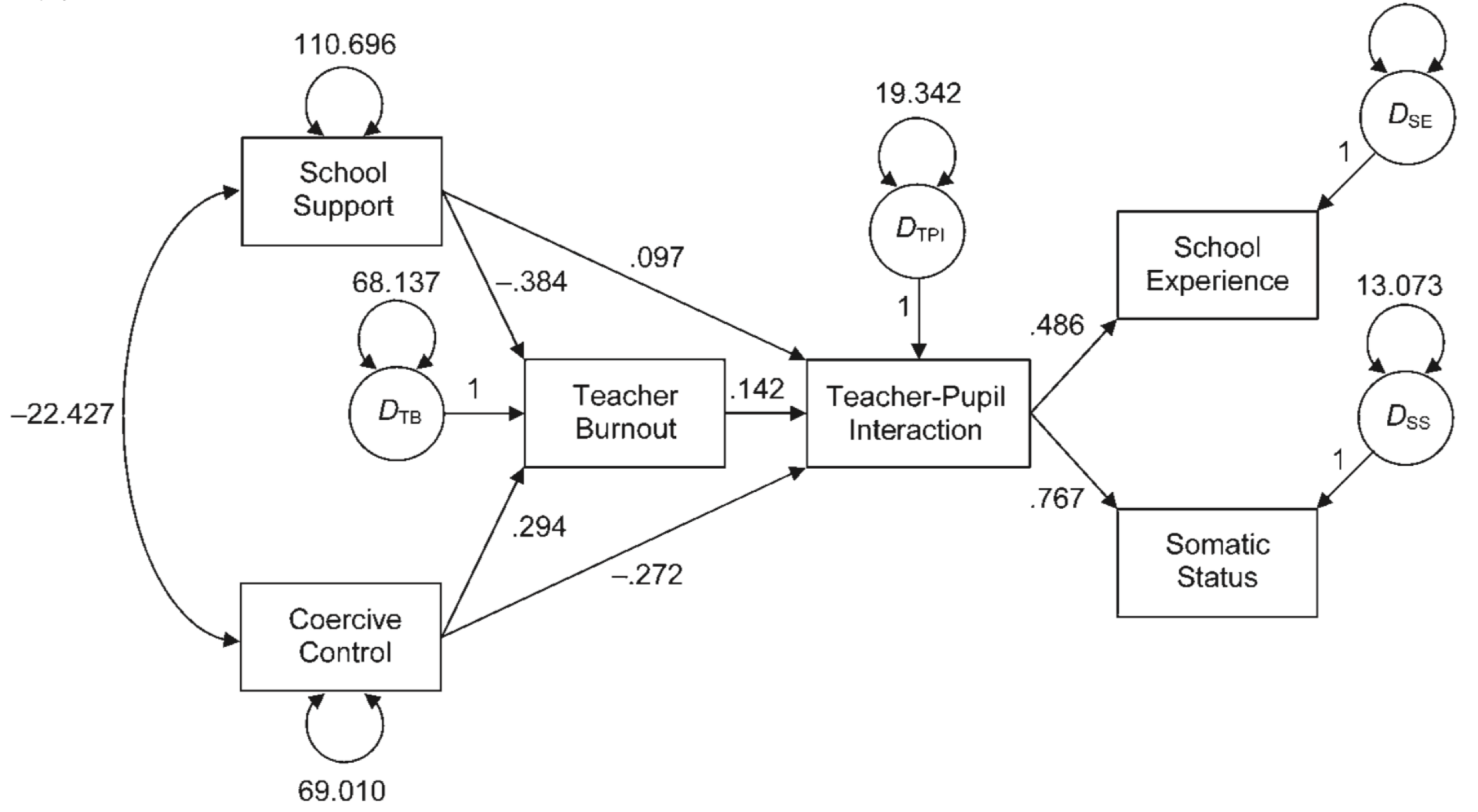
# Interpretations

- Path estimates interpreted like regression coefficients
- Endogenous error variance and
  - "Squared multiple correlations" are like $R^2$ values
  - For each endogenous variable, Squared multiple correlation = 1 – standardized residual variance
- Standardized factor loadings
  - Correlation between that individual item and the factor (shared variance)

# Extended Example, Ch 7 pg. 179

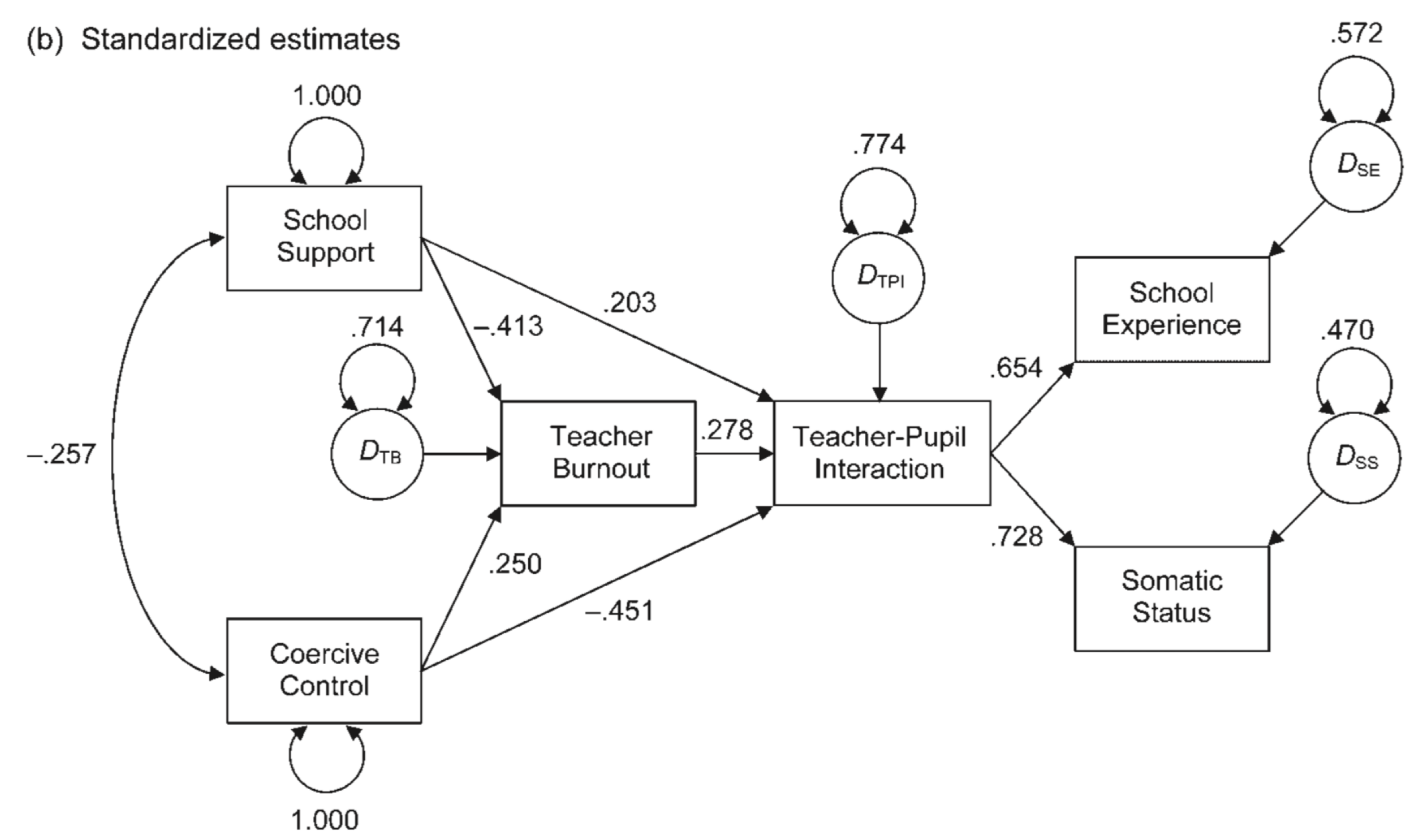
- Sava (2002): perceived school support, burnout, and extent of a coercive view of student discipline
- N = 109 high school teachers
- Student responses were averaged

(a) Unstandardized estimates



# Extended Example, Ch 7 pg. 179

- Sava (2002): perceived school support, burnout, and extent of a coercive view of student discipline
- N = 109 high school teachers
- Student responses were averaged



# Model implied covariances

- Tracing rule

A model-implied correlation is the sum of all the causal effects and noncausal associations from all valid tracings between two variables in a recursive model. A "valid" tracing means that a variable is not (Rule 7.1)

1. Entered through an arrowhead and exited by the same arrowhead, nor
2. Entered twice in the same tracing.

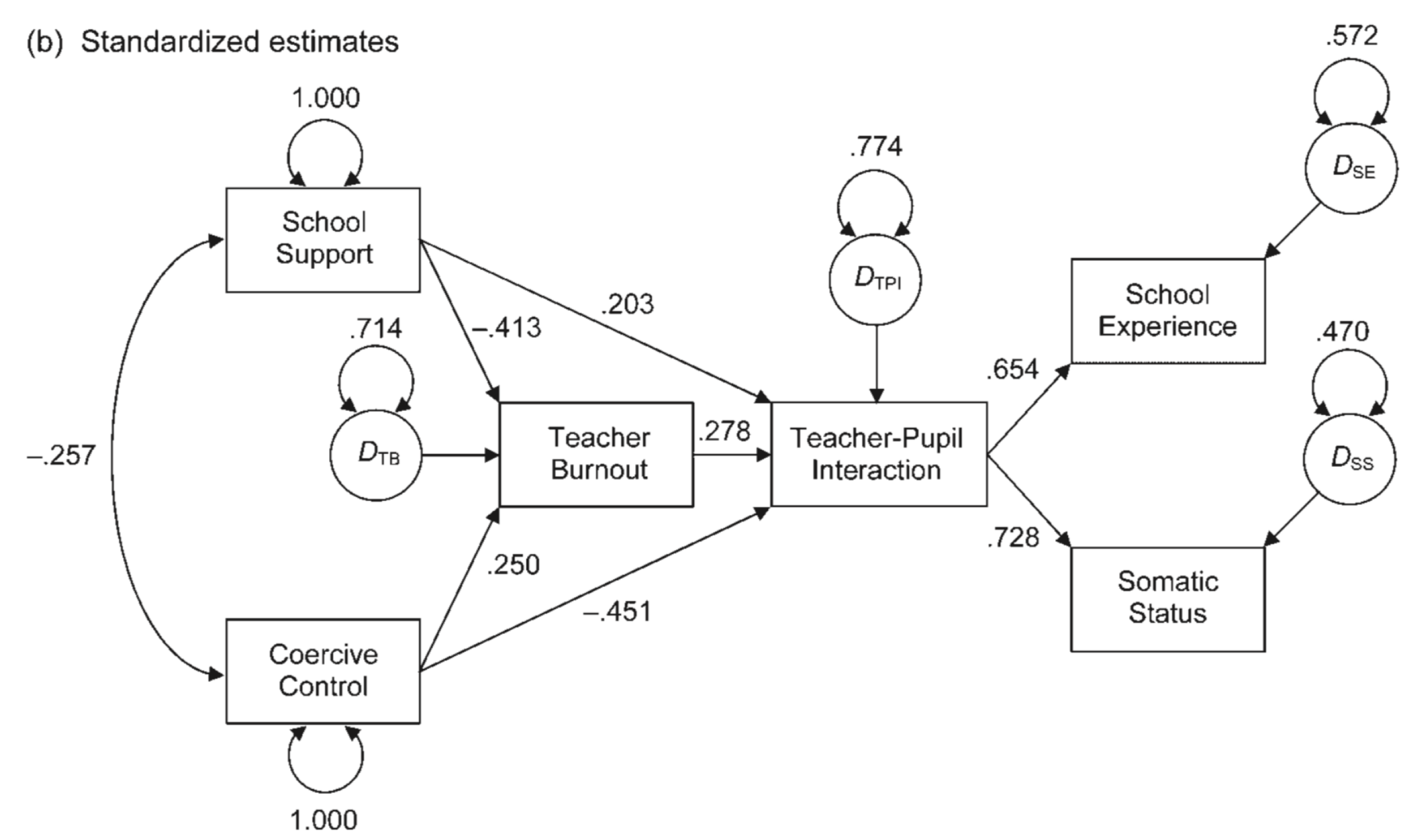
# Model implied covariances

- Model implied relationship between coercive control and school experience – four valid tracings:

1. single mediator (teacher–pupil interactions)
2. two mediators (teacher burnout, teacher–pupil interactions)
3. Coercive Control $\leftrightarrow$ School Support $\rightarrow$ Teacher–Pupil Interactions $\rightarrow$ School Experience
4. Coercive Control $\leftrightarrow$ School Support $\rightarrow$ Teacher Burnout $\rightarrow$ Teacher–Pupil Interactions $\rightarrow$ School Experience

# Extended Example, Ch 7 pg. 179

- Sava (2002): perceived school support, burnout, and extent of a coercive view of student discipline
- N = 109 high school teachers
- Student responses were averaged



# Alternative estimators

- Hot area of research! Get in there!
- **Unweighted least squares** (ULS)
  - An OLS estimation technique that minimizes sum of squared errors between sample and model implied covariance matrices
  - Not as efficient as ML estimation (worse standard errors)
- **Generalized least squares** (GLS)
  - Can be used for non-normal data
  - Think logistic regression, Poisson regression
- Bootstrapping
  - There's always bootstrapping…

# Model Fit

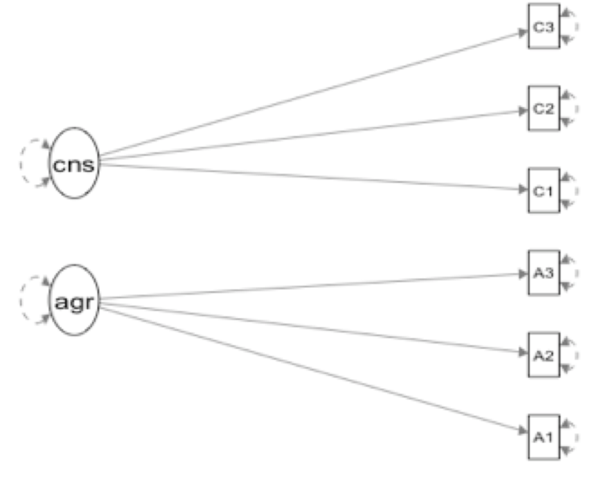
# Confirmatory Factor Analysis (CFA)

- Does the model we have in our heads actually fit the data?

### Data Cor matrix

| | A1 | A2 | A3 | C1 | C2 | C3 |
|---|---|---|---|---|---|---|
| A1 | 1.000 | -0.340 | -0.265 | 0.028 | 0.016 | -0.019 |
| A2 | -0.340 | 1.000 | 0.485 | 0.092 | 0.136 | 0.192 |
| A3 | -0.265 | 0.485 | 1.000 | 0.097 | 0.141 | 0.132 |
| C1 | 0.028 | 0.092 | 0.097 | 1.000 | 0.428 | 0.308 |
| C2 | 0.016 | 0.136 | 0.141 | 0.428 | 1.000 | 0.356 |
| C3 | -0.019 | 0.192 | 0.132 | 0.308 | 0.356 | 1.000 |

### Model




### Model implied Cor matrix

| | A1 | A2 | A3 | C1 | C2 | C3 |
|---|---|---|---|---|---|---|
| A1 | 1.000 | | | | | |
| A2 | -0.337 | 1.000 | | | | |
| A3 | -0.256 | 0.492 | 1.000 | | | |
| C1 | -0.063 | 0.122 | 0.093 | 1.000 | | |
| C2 | -0.074 | 0.143 | 0.109 | 0.418 | 1.000 | |
| C3 | -0.056 | 0.108 | 0.082 | 0.316 | 0.370 | 1.000 |

**Fit?**

# Model Identification and Fit

- Assigning factors scales (marker variables)
  - Underidentification
  - Just-identified or saturated
  - **Over-identified:**

$$\begin{cases} a + b = 6 \\ 2a + b = 10 \\ 3a + b = 12 \end{cases}$$

Find values of *a* and *b* that are positive and yield total scores such that the sum of the squared differences between the observations (6, 10, 12) and these totals is as small as possible.

No single solution:
- (a = 4, b = 2)
- (a = 2, b = 6)

# Model Identification and Fit

- Assigning factors scales (marker variables)
  - Underidentification - **no estimates! Booo**
  - Just-identified or saturated - **estimates but *perfect* fit**
  - Over-identified - **estimates and fit stats! Yay!**

# Model Fit Statistics

- Model Fit
  - Chi-square statistic and test (p-value)
    - $(N - 1)\, F_{ML}$ , where $F_{ML}$ is the fit function that was minimized during the ML estimation.
    - Is distributed as a chi-square, and as the sample size gets larger this statistics gets larger.
  - CFI - Bentler Comparative Fit Index
    - $> 0.95$
  - RMSEA - Steiger–Lind root mean square error of approximation
    - $< 0.08$
  - GFI - Jöreskog–Sörbom Goodness of Fit Index
    - $> 0.95$
    - proportion of covariances in the sample data matrix explained by the model

$$\mathrm{CFI} = 1 - \frac{\chi^2_{\mathrm{M}} - df_{\mathrm{M}}}{\chi^2_{\mathrm{B}} - df_{\mathrm{B}}}$$

$$\mathrm{RMSEA} = \sqrt{\frac{\chi^2_{\mathrm{M}} - df_{\mathrm{M}}}{df_{\mathrm{M}}(N-1)}}$$

$$\mathrm{GFI} = 1 - \frac{\mathbf{C}_{\mathrm{res}}}{\mathbf{C}_{\mathrm{tot}}}$$

# Incremental Fit: Comparing Models

- Determining whether one model fits the data better than another.
- Models are **nested** when one can be obtained by imposing constraints on the other.
- Chi-squares can be directly contrasted to test whether one model fits the data better than the other.
  - This is a likelihood ratio test! Follow these steps:
  1. Compute the difference in the chi-square statistics associated with each model = $\Delta\chi^2$.
  2. Compute the difference in the df for each = $\Delta df$.
  3. Evaluate the $\Delta\chi^2$ as if it were an ordinary chi-square, using $\Delta df$ as the df for the significance test.

# Incremental Fit: Comparing Models

- If $\Delta\chi 2$ is significant, then the model with the smaller individual $\chi 2$ (lower df) is considered to provide a relative improvement in fit over the other.
- The most persuasive case that a given model has been correctly specified is created when a researcher finds a differentially better fit of that model in comparison to numerous other models.
- If models are NOT nested compare the models' AIC's and BIC's
  - Smallest wins!

# To R!

Copy the Structural Equation Modeling code into an .Rmd file