LATENT VARIABLE MODELS

Session 1 – Introduction

Course prerequisites

- Knowledge of statistics
  - Statistical testing (e.g., chi-square & normal distributions)
  - Regression (GLMs)
  - Var, cov, cor, mean
- Knowledge of psychometrics
  - Validity
  - PCA, EFA, CFA
  - Reliability
  - IRT
- Programming in R

Course materials

Book(s):

- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*.
  - Good as a starting guide, not an authorative standard
- Kaplan, D. (2009). *Structural Equation Modeling: Foundations and Extensions.*
  - Authorative standard. But more technical and not focused on specific software, not practical for current course.

GitHub repo:

- Lecture slides
- Markdown files with examples, exercises and answers

Book examples

- You are strongly advised to copy and run R code from examples in Beaujean book and from GitHub repo:
  - They give you a step-by-step guide on how to perform analyses
  - They give you a starter for making the exercises
  - If you make a mistake, you will get an error or warning message, from which you can learn alot! (But only if you try to decipher it, <u>red = good!)</u>

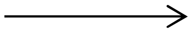
Structural Equation Modeling

- **Structural**: We aim to *explain associations between* (that is, the structure of) *observed variables*
- **Equations**: Associations between variables are described using mathematical formulae
- **Modeling**: we construct models (hypotheses, theories) of reality. The models can be statistically tested. That is: rejected by the data (or not), but never proven 'true' or 'right'.
  - All models are wrong, but some are useful.

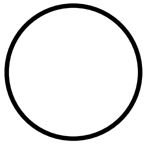
Structural Equation Modeling

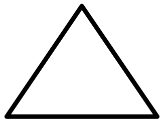
SEMS are graphically represented using these building blocks:


Observed (manifest) variable


Directional relationship (regression relationship)


Non-directional relationship (correlation/(co)variance)


Latent variable


Constant term (i.e., not a variable, e.g., intercept)

Structural Equation Modeling

- Arrows in SEM denote regression relationships
- All generalized linear models (GLMs) can be formulated as a SEM:
  - t-test
  - ANOVA
  - Multiple linear regression
  - Multiple logistic regression
  - …..
- Also, SEM can be used to models for multilevel or longitudinal data (i.e., GLMMs)

Example dataset

Variables in the model:

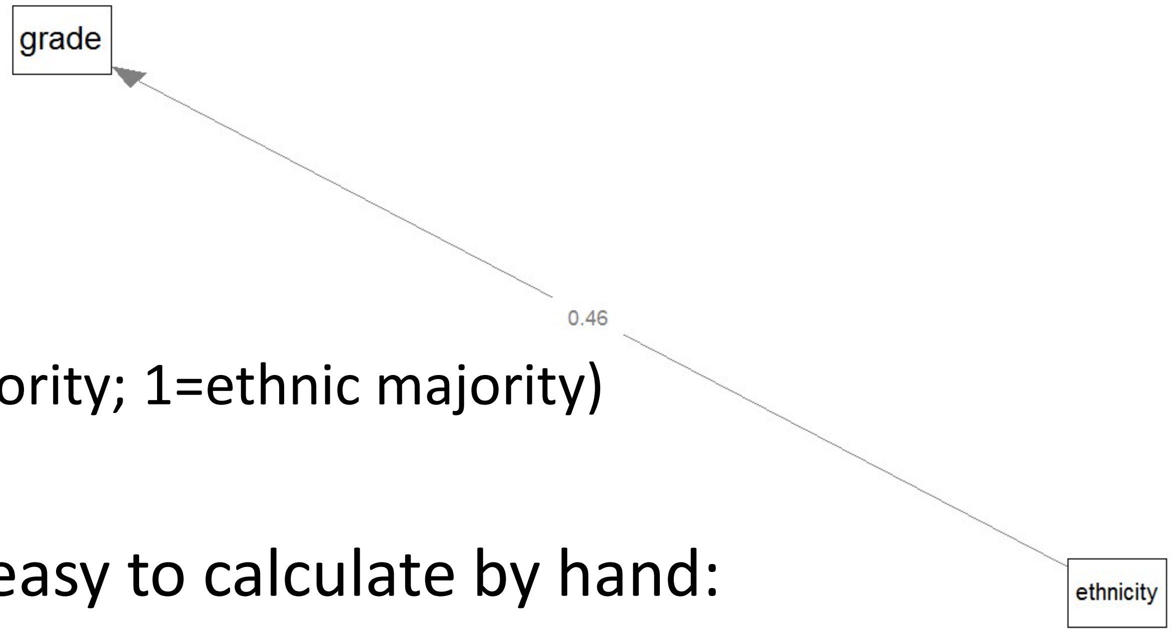
grade | ethnicity | homework | SES | Prev_ach

- GPA in 10th grade
- Ethnicity
- Homework (8th grade)
- Socio-economic status
- Previous achievement (8th grade)

- Sample covariance matrix **S**:

	grade	homwrk	prv_ch	ethnct	SES
grade	2.185				
homework	0.335	0.649			
prev_ach	6.429	2.067	79.092		
ethnicity	0.081	0.028	1.201	0.175	
SES	0.338	0.176	3.541	0.106	0.690

Model: Univariate regression

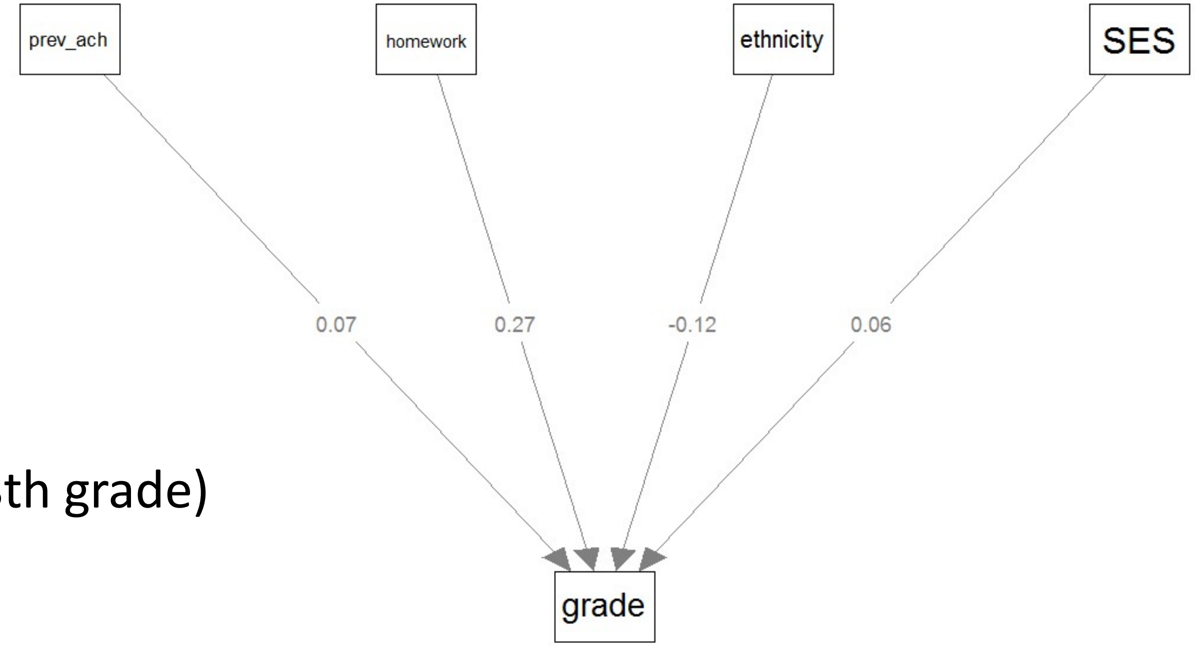


- Dependent:
  - GPA in 10th grade
- Independent:
  - ethnicity (0=ethnic minority; 1=ethnic majority)

- Regression coefficient easy to calculate by hand:
  - $\hat{b}_{xy} = \frac{cov_{x,y}}{var_x} = \frac{0.0814}{0.1752} = 0.4646$
  - standardized $\hat{b}_{xy} = \hat{\rho}_{xy} = \frac{cov_{x,y}}{s_x s_y} = s_x \frac{\hat{b}_{xy}}{s_y} = 0.132$
- Measure of fit or (strength of) association: $\hat{\rho}_{xy}{}^2$

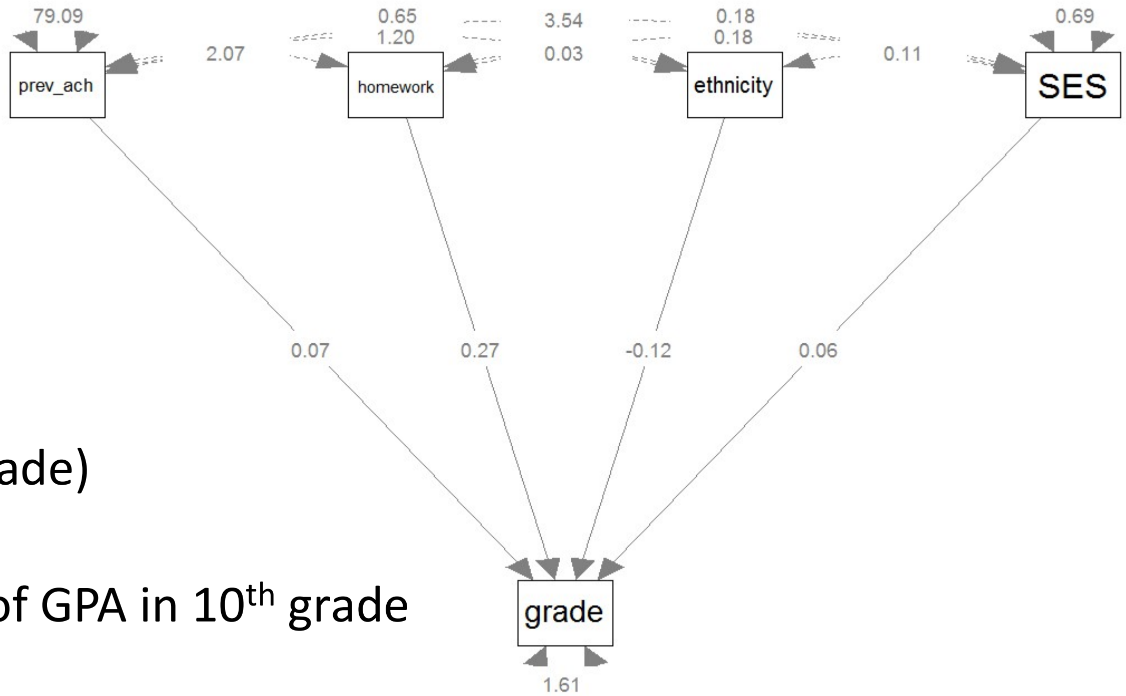
Model: Multiple regression

- Dependent:
  - GPA in 10th grade
- Independent:
  - Ethnicity
  - Homework (8th grade)
  - Previous achievement (8th grade)
  - Socio-economic status



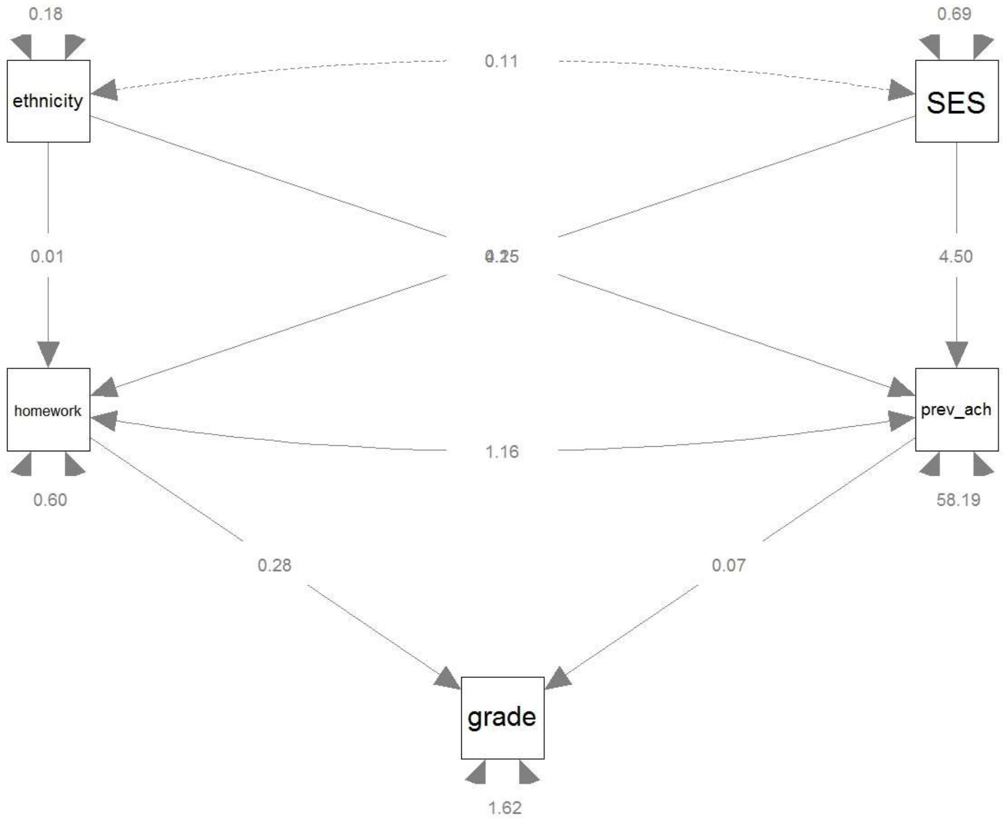
- Regression estimates are now a vector of partial regression coefficients, need matrix algebra to compute: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- Measure of fit: multiple correlation (R=.512), or variance explained (R^2=.262)
- Measure of (strength of) association: $\hat{b}_{xy}$ or standardized $\hat{b}^*_{xy} = s_x \frac{\hat{b}_{xy}}{s_y}$ (where $\hat{b}_{xy}$ is a partial regression coefficient)

Model: SEM



- *Endogenous* variables:
  - GPA in 10th grade
- *Exogenous* variables:
  - Ethnicity
  - Homework (8th grade)
  - Previous achievement (8th grade)
  - Socio-economic status
  - (disturbance/error/residual) of GPA in 10th grade
- Regression estimates are still a vector of partial regression coefficients, need matrix algebra and optimization to compute
- Measure of (strength of) associations: Partial regression coefficients
- Overall model fit: How well are the observed variables' (co)variances reproduced by the model?
  - Quantified by a χ^2value and model fit indices

Model: SEM



- *Endogenous* variables:
  - Homework (8th grade)
  - Previous achievement (8th grade)
  - GPA in 10th grade
- *Exogenous* variables:
  - Ethnicity
  - Socio-economic status
  - (disturbances/errors/residuals of
    - Homework (8th grade)
    - Previous achievement (8th grade)
    - GPA in 10th grade
- Measure of (strength of) associations: partial regression coefficients
- Overall model fit: How well are the observed variables' (co)variances reproduced by the model?
  - Quantified by a χ^2value and model fit indices

Model: SEM

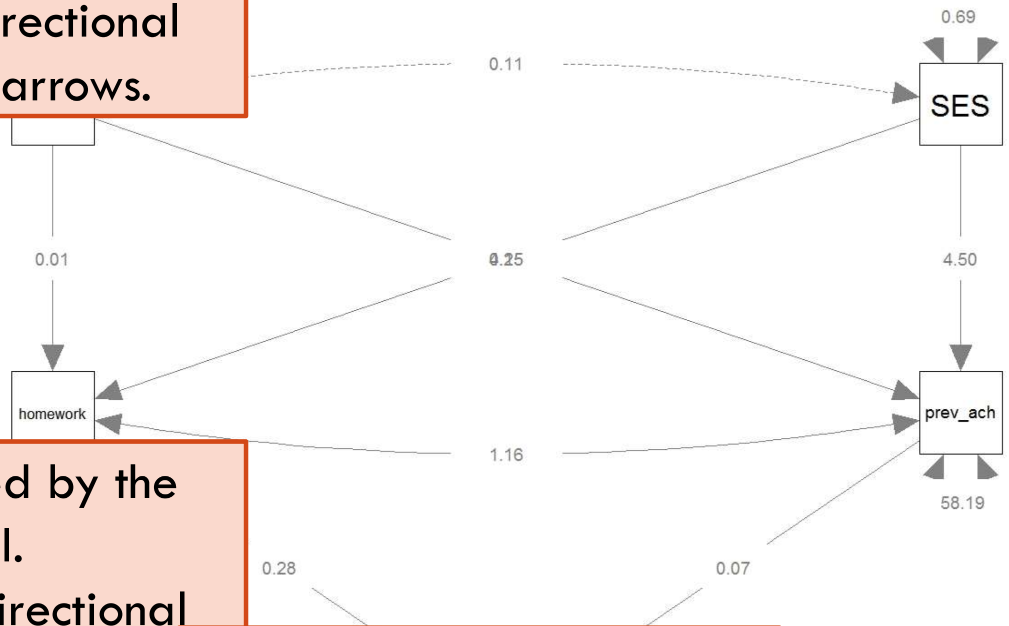
From Greek *endo*, meaning 'inside', and *gignomai*

Explained by the model.
Have unidirectional incoming arrows.

- **_Endogenous_ variables:**
  - Homework (8th grade)
  - Previous achievement (8th grade)
  - GPA in 10th grade
- **_Exogenous_ variables:**
  - Ethnicity
  - Socio-economic sta
  - (disturbances/error
    - Homework (8th gr
    - Previous achievement (8th grade)
    - GPA in 10th grade

Not explained by the model.
Have no unidirectional incoming arrow(s).

From Greek exo, meaning 'outside', and *gignomai*, meaning 'to produce'



- Measure of (strength of) associations: partial regression coefficients
- Overall model fit: How well are the observed variables' (co)variances reproduced by the model?
  - Quantified by a χ^2value and model fit indices

SEM using lavaan

To fit a SEM in R with lavaan, we need two things:

1. Data, which can be:
   - Raw data, which is often an external file (e.g., .sav, .xls) which needs to be loaded into R (most common case in practice)
   - Covariance or correlation matrix, which can be an external file, or can be entered manually (most often the case in book's examples and exercises)
2. Model specification:
   - A long character string that specifies whether population parameters are fixed (e.g., to a constant like 1 or 0) restricted (e.g., to be equal to another parameter) or should be freely estimated, using lavaan model syntax

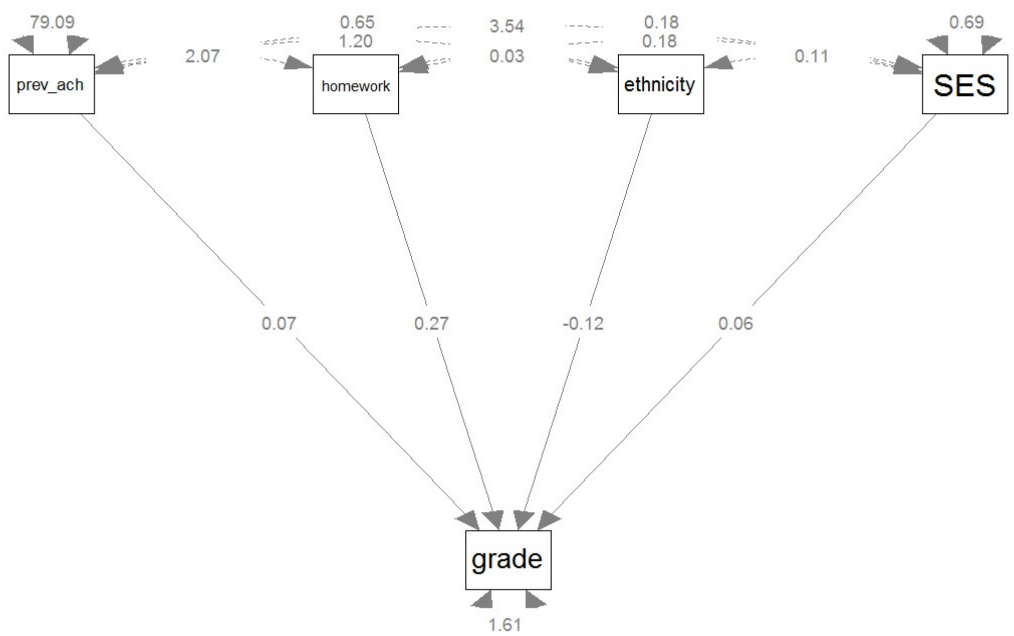
Lavaan model syntax

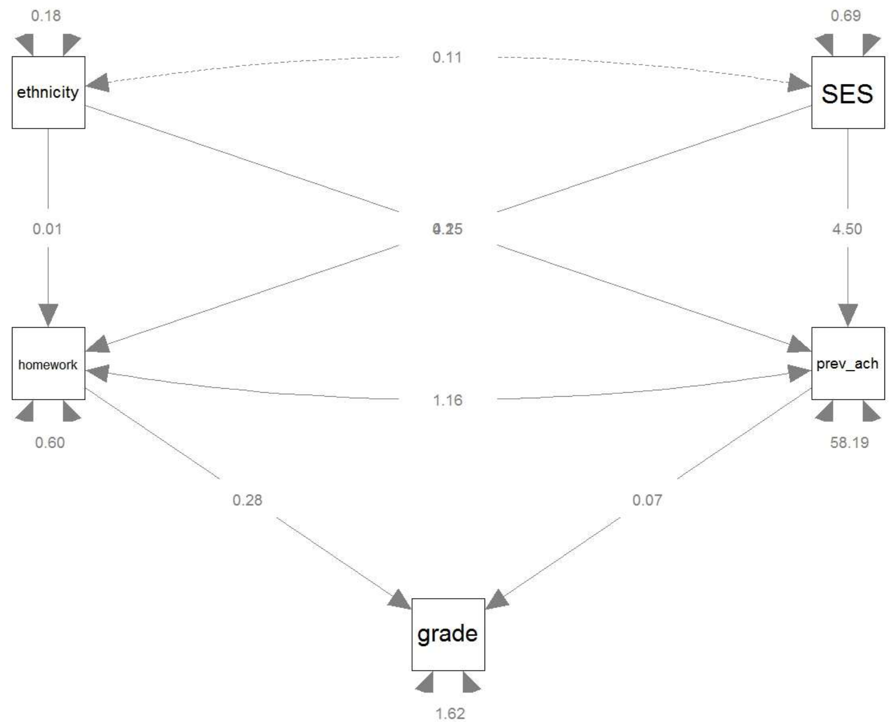
<table>
<tr><th>Syntax</th><th>Command</th><th>Example</th></tr>
<tr><td><code>~</code></td><td>Regress onto</td><td>Regress B onto A: <code>B ~ A</code></td></tr>
<tr><td><code>~~</code></td><td>(Co)variance</td><td>Variance of A: <code>A ~~ A</code>
Covariance of A and B: <code>A ~~ B</code></td></tr>
<tr><td><code>~1</code></td><td>Constant/mean/intercept</td><td>Regress B onto A, and and include the intercept in the model: <code>B ~ 1 + A</code> or
<code>B ~ A</code>
<code>B ~ 1</code></td></tr>
<tr><td><code>=~</code></td><td>Define reflective latent variable</td><td>Define Factor 1 by A-D:
<code>F1 =~A+B+C+D</code></td></tr>
<tr><td><code><~</code></td><td>Define formative latent variable</td><td>Define Factor 1 by A-D:
<code>F1 <~ 1*A+B+C+D</code></td></tr>
<tr><td><code>:=</code></td><td>Define non-model parameter</td><td>Define parameter u2 to be twice the square of u:
<code>u2 := 2*(u^2)</code></td></tr>
<tr><td><code>*</code></td><td>Label parameters (the label has to be pre-multiplied)</td><td>Label the regression of Z onto X as b:
<code>Z ~b*X</code></td></tr>
<tr><td><code>|</code></td><td>Define the number of thresholds (for categorical endogenous variables)</td><td>Variable <code>u</code> has three thresholds:
<code>u | t1 + t2 + t3</code></td></tr>
</table>

grade | ethnicity | homework | SES | Prev_ach

Syntax	Command	Example
`~`	Regress onto	Regress B onto A: `B ~ A`
`~~`	(Co)variance	Variance of A: `A ~~ A`
		Covariance of A and B: `A ~~ B`

Q: How do we specify these models in lavaan syntax?





Computation time!

Example 2.4.1

- get PDF from Github

Make Exercise 2.1:

- Get Exercises_week_1.pdf from Github (adapted version of the exercises in the Beaujean book)

Structural Equation Modeling

- Fitted model is used to explain associations between observed variables
- I.e., to explain sample covariances between observed variables:

$$\mathrm{cov}_{xy} = \left(\frac{1}{N-1}\right)\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\mathrm{cov}_{xy} = r_{xy} SD_x SD_y$$

  - Note: means, and skewness & kurtosis can be also be involved in SEM (discussed later in course)

Structural Equation Modeling

- With SEM, we obtain a fitted model that minimizes the difference between
  - sample matrix of observed covariances $\mathbf{S}$ and
  - population matrix of model-implied covariances $\widehat{\boldsymbol{\Sigma}}$
    - In addition, we try to keep the model parsimoneous through applying restrictions (i.e., specifying the model) so that not all possible paths are estimated
- These covariance matrices contain all (co)variances of the observed variables in the model. Note that:
  - Covariance matrices are always symmetric, because cov(x,y)=cov(y,x)
  - Covariance matrices have the variance of the observed variables on the diagonal. I.e., cov(x,x) = var(x)

Model-implied (co)variances

- Variables in the model:

grade	ethnicity	homework	SES	Prev_ach

- Observed covariance matrix **S**:

	grade	homwrk	prv_ch	ethnct	SES
grade	2.185				
homework	0.335	0.649			
prev_ach	6.429	2.067	79.092		
ethnicity	0.081	0.028	1.201	0.175	
SES	0.338	0.176	3.541	0.106	0.690

- Once the model is estimated, the model-implied covariance matrix $\hat{\Sigma}$ can be calculated using path analysis, or equivalently, matrix algebra

Model-implied (co)variances

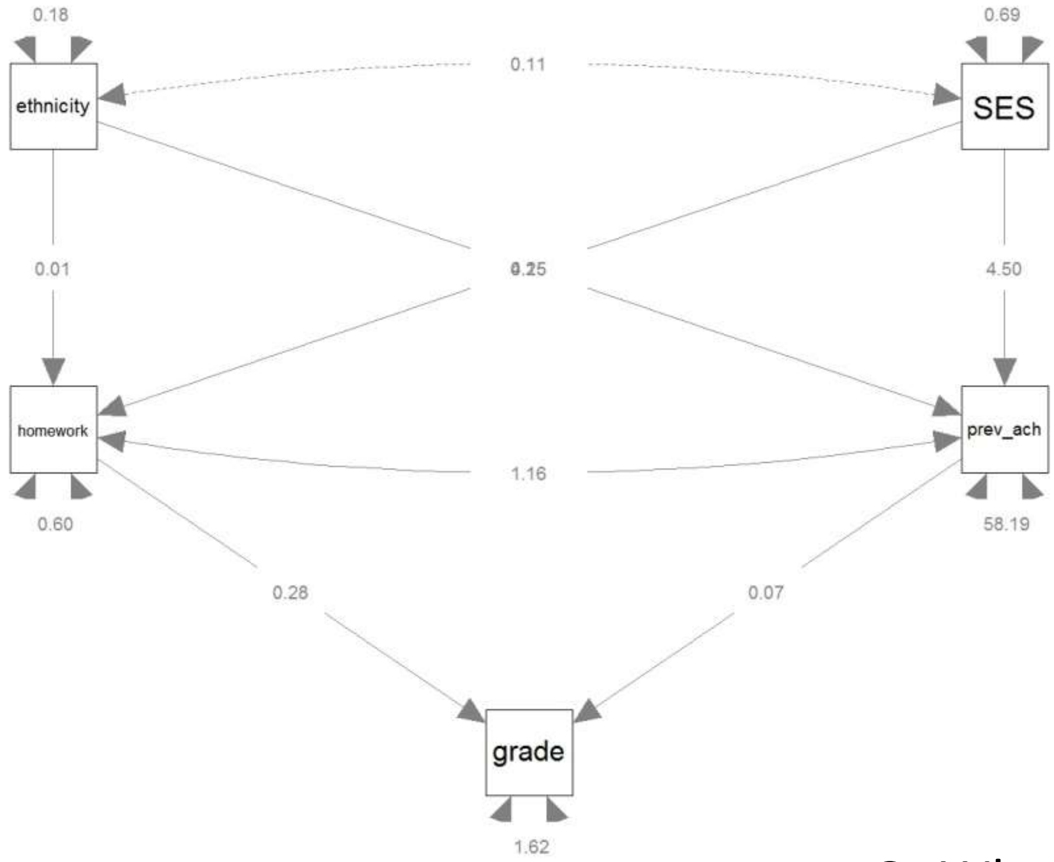
Path analysis:

- Model-implied covariance between variables X and Y can be computed as follows:
  - Find all paths leading from X to Y
  - Multiply all parameter values along a given path from X to Y, but:
    - No loops: may not go through same variable more than once
    - May switch forward/backward direction only once within a path
    - May go through double-headed arrow only once within a path
  - Summing all values thus obtained

- Variances of variables are calculated as follows:
  - For exogenous variables, model-implied variances are equal to sample variances, so are given (not computed)
  - For endogenous variables, variances are computed like covariances (rules above)

Model-implied (co)variances

Model:

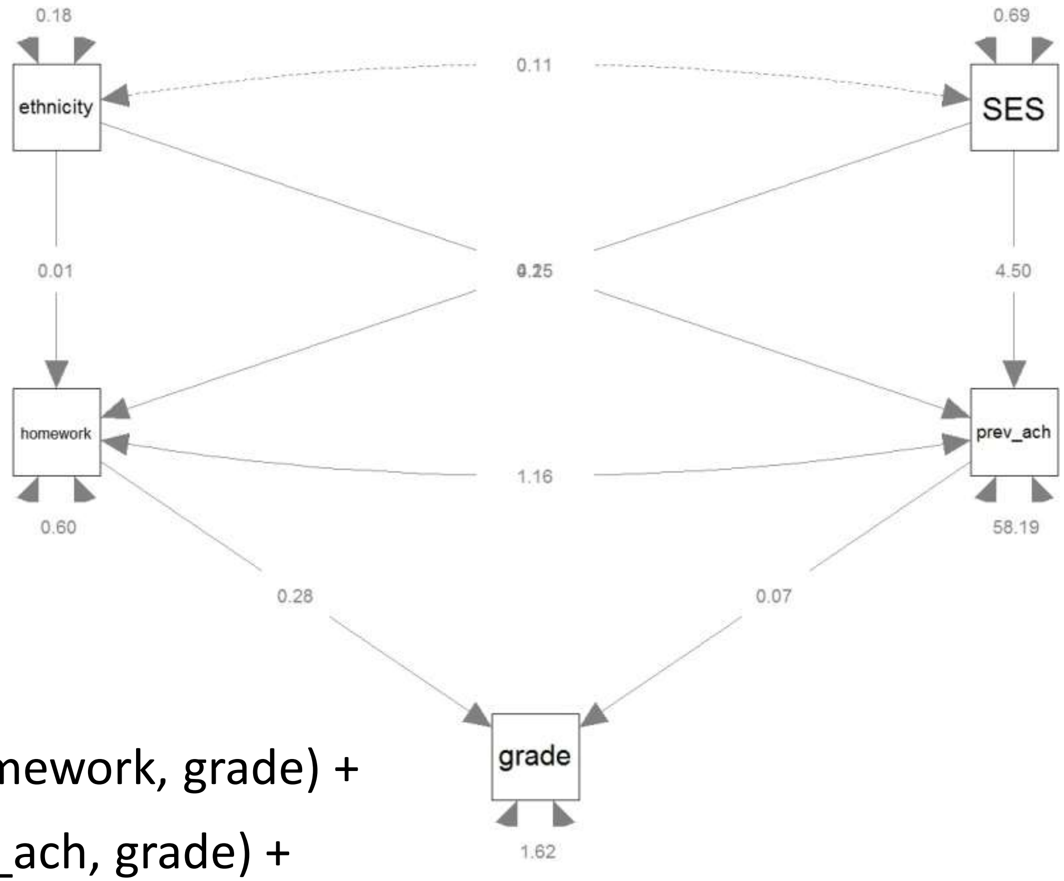


Parameter estimates:

```
      lhs op       rhs    est
    grade  ~  prev_ach  0.074
    grade  ~  homework  0.281
 homework  ~ ethnicity  0.007
 homework  ~       SES  0.254
 prev_ach  ~ ethnicity  4.147
 prev_ach  ~       SES  4.496
 homework ~~  prev_ach  1.158
    grade ~~     grade  1.616
 homework ~~  homework  0.604
 prev_ach ~~  prev_ach 58.190
ethnicity ~~ ethnicity  0.175
ethnicity ~~       SES  0.106
      SES ~~       SES  0.690
```

Q: What is the model-implied cov(SES, grade)?

Model-implied (co)variances



model-implied cov(SES, grade)=

var(SES) * b(SES, homework) * b(homework, grade) +

var(SES) * b(SES, prev_ach) * b(prev_ach, grade) +

cov(SES, ethnicity) * b(ethnicity, homework) * b(homework, grade) +

cov(SES, ethnicity) * b(ethnicity, prev_ach) * b(prev_ach, grade)

Model-implied (

Note that Beaujean's examples in section 2.1.3 seem more simple, because he uses the standardized solution. Then all variances of exogenous variables equal 1 and can be omitted, which simplifies calculations a lot.

model-implied cov(SES,grade)=

var(SES) * b(SES, homework) * b(homework, grade) +

var(SES) * b(SES, prev_ach) * b(prev_ach, grade) +

cov(SES, ethnicity) * b(ethnicity, homework) * b(homework, grade) +

cov(SES, ethnicity) * b(ethnicity, prev_ach) * b(prev_ach, grade) =

.690 * .254 * .281 +

.690 * 4.496 * .074 +

.106 * .007 * .281 +

.106 * 4.147 * .074 =

0.3115514

lhs	op	rhs	est
grade	~	prev_ach	0.074
grade	~	homework	0.281
homework	~	ethnicity	0.007
homework	~	SES	0.254
prev_ach	~	ethnicity	4.147
prev_ach	~	SES	4.496
homework	~~	prev_ach	1.158
grade	~~	grade	1.616
homework	~~	homework	0.604
prev_ach	~~	prev_ach	58.190
ethnicity	~~	ethnicity	0.175
ethnicity	~~	SES	0.106
SES	~~	SES	0.690

Model-implied (co)variances

- A SEM is a system of linear equations, which we can represent by matrices
  - Although non-linear SEM also exists, but outside the scope of this course
- The path tracing rules represent matrix algebra but more tedious/confusing/error prone
- Beaujean’s book hardly involves formulas, and no matrix notation. To get a good understanding of SEM, you need to know about underlying matrices and vectors

Model-implied (co)variances

- In lavaan, the (co)variance stucture of a fitted model is given by four parameter matrices
- Matrix algebra gives us the model-implied covariance matrix:

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}(\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\psi}\left[(\mathbf{I} - \boldsymbol{\beta})^{-1}\right]^{\mathbf{T}}\boldsymbol{\Lambda}^{\mathbf{T}} + \boldsymbol{\Theta}$$

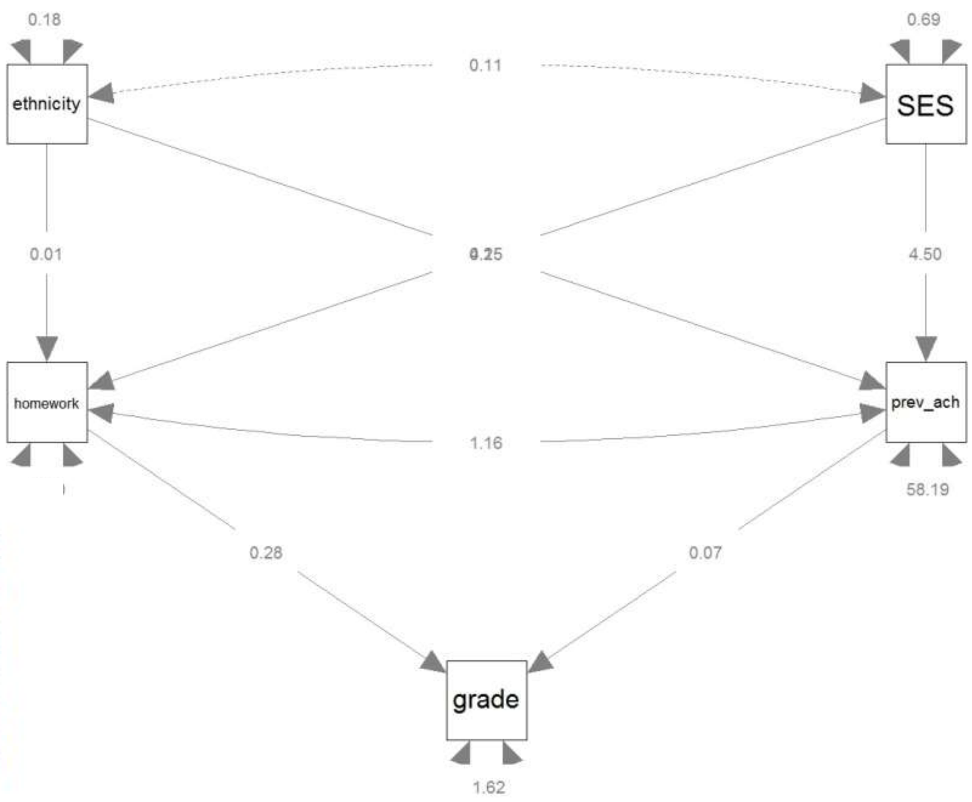
- Today, our models assume no measurement error, so $\boldsymbol{\Lambda}$ is an identity matrix and $\boldsymbol{\Theta}$ all zeros. Thus, the above formula simplifies to:

$$\widehat{\boldsymbol{\Sigma}} = (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\psi}\left[(\mathbf{I} - \boldsymbol{\beta})^{-1}\right]^{\mathbf{T}}$$

Model-implied (co)variances

- Let p be the number of observed variables in the model
- If we have observed variables only:
  - $\boldsymbol{\beta}$ is a p x p matrix of regression coefficients, relating predictor to criterion variables
    - 'Contains' single-headed (directed) arrows, therefore non-symmetric
    - The columns reflect the variables as predictors, the rows reflect the variables as responses
  - $\boldsymbol{\psi}$ is a p x p matrix of (co)variances not explained by the regression equations
    - 'Contains' double headed (undirected) arrows, therefore symmetric
- $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ describe the **structural** model
- Often, SEM models also involve a **measurement** model (described by $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$, which will be introduced next session)

Model-implied (co)variances




```
> beta
          grade homwrk prv_ch ethnct   SES
grade         0  0.281  0.074  0.000 0.000
homework      0  0.000  0.000  0.007 0.254
prev_ach      0  0.000  0.000  4.147 4.496
ethnicity     0  0.000  0.000  0.000 0.000
SES           0  0.000  0.000  0.000 0.000
> psi
          grade  homwrk prv_ch ethnct SES
grade      1.616
homework   0.000  0.604
prev_ach   0.000  1.158 58.190
ethnicity  0.000  0.000  0.000  0.175
SES        0.000  0.000  0.000  0.106  0.690
```

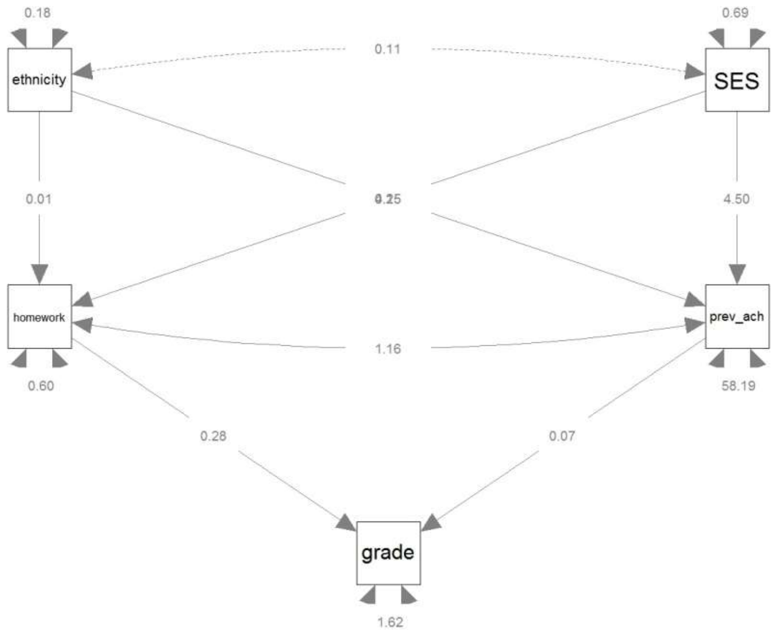
Structural and measurement model

Two main components of SEMs:

- the ***<u>structural model</u>*** contains *causal* regression relationships between endogenous and exogenous variables
  - path models (without measurement errors) can be viewed as SEMs that contain only the structural model
- the ***<u>measurement model</u>*** contains the associations between latent variables and their indicators
  - confirmatory factor analysis models contain only the measurement part

Model-implied (co)variances

Model:



Parameter estimates:

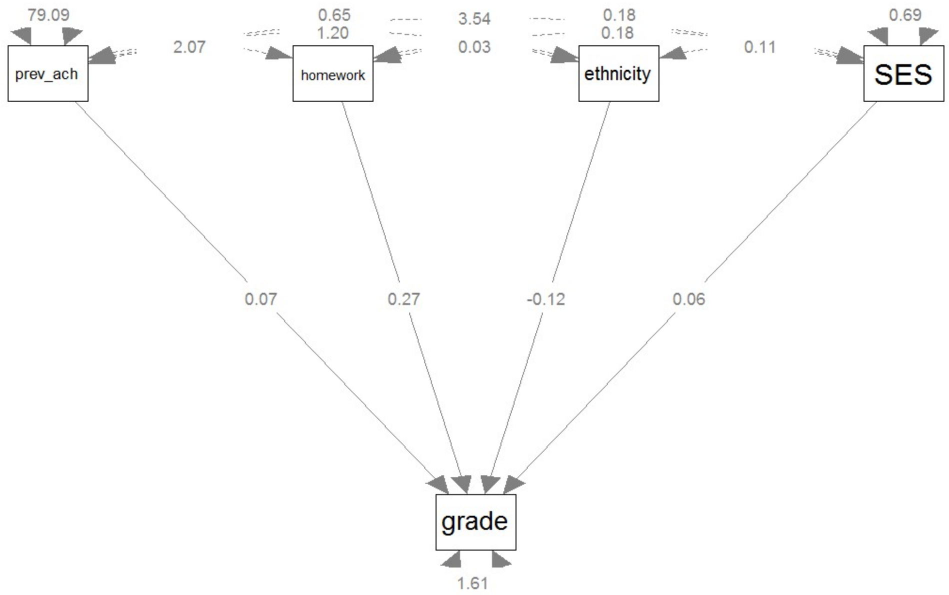
lhs	op	rhs	est
grade	~	prev_ach	0.074
grade	~	homework	0.281
homework	~	ethnicity	0.007
homework	~	SES	0.254
prev_ach	~	ethnicity	4.147
prev_ach	~	SES	4.496
homework	~~	prev_ach	1.158
grade	~~	grade	1.616
homework	~~	homework	0.604
prev_ach	~~	prev_ach	58.190
ethnicity	~~	ethnicity	0.175
ethnicity	~~	SES	0.106
SES	~~	SES	0.690

Model-implied covariance matrix $\widehat{\boldsymbol{\Sigma}}$:

	grade	homwrk	prv_ch	ethnct	SES
grade	2.185				
homework	0.335	0.649			
prev_ach	6.429	2.067	79.092		
ethnicity	0.097	0.028	1.201	0.175	
SES	0.311	0.176	3.541	0.106	0.690

Model-implied (co)variances

Model:



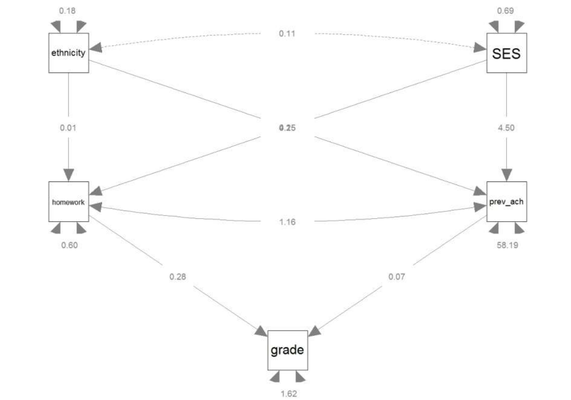
Parameter estimates:

lhs	op	rhs	est
grade	~	prev_ach	0.073
grade	~	homework	0.271
grade	~	ethnicity	-0.119
grade	~	SES	0.063
grade	~~	grade	1.612
prev_ach	~~	prev_ach	79.092
prev_ach	~~	homework	2.067
prev_ach	~~	ethnicity	1.201
prev_ach	~~	SES	3.541
homework	~~	homework	0.649
homework	~~	ethnicity	0.028
homework	~~	SES	0.176
ethnicity	~~	ethnicity	0.175
ethnicity	~~	SES	0.106
SES	~~	SES	0.690

Model-implied covariance matrix $\widehat{\boldsymbol{\Sigma}}$:

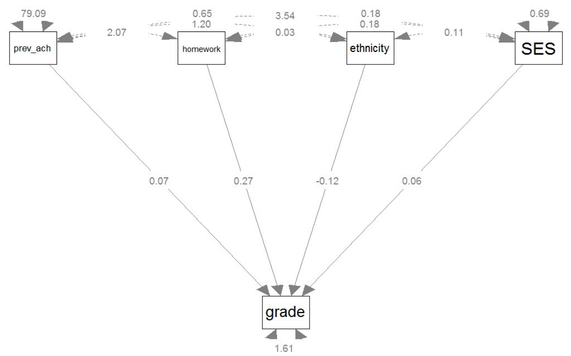
	grade	prv_ch	homwrk	ethnct	SES
grade	2.185				
prev_ach	6.429	79.092			
homework	0.335	2.067	0.649		
ethnicity	0.081	1.201	0.028	0.175	
SES	0.338	3.541	0.176	0.106	0.690

Model-implied (co)variances



$\hat{\Sigma} =$

	grade	homwrk	prv_ch	ethnct	SES
grade	2.185				
homework	0.335	0.649			
prev_ach	6.429	2.067	79.092		
ethnicity	0.097	0.028	1.201	0.175	
SES	0.311	0.176	3.541	0.106	0.690



$\hat{\Sigma} =$

	grade	prv_ch	homwrk	ethnct	SES
grade	2.185				
prev_ach	6.429	79.092			
homework	0.335	2.067	0.649		
ethnicity	0.081	1.201	0.028	0.175	
SES	0.338	3.541	0.176	0.106	0.690

Which model fits data best (i.e., approximates sample covariances best)?

Which is most parsimonious (i.e., estimates lowest number of population parameters)?

S =

	grade	homwrk	prv_ch	ethnct	SES
grade	2.185				
homework	0.335	0.649			
prev_ach	6.429	2.067	79.092		
ethnicity	0.081	0.028	1.201	0.175	
SES	0.338	0.176	3.541	0.106	0.690

Variances of exogenous variables often not explicitly depicted

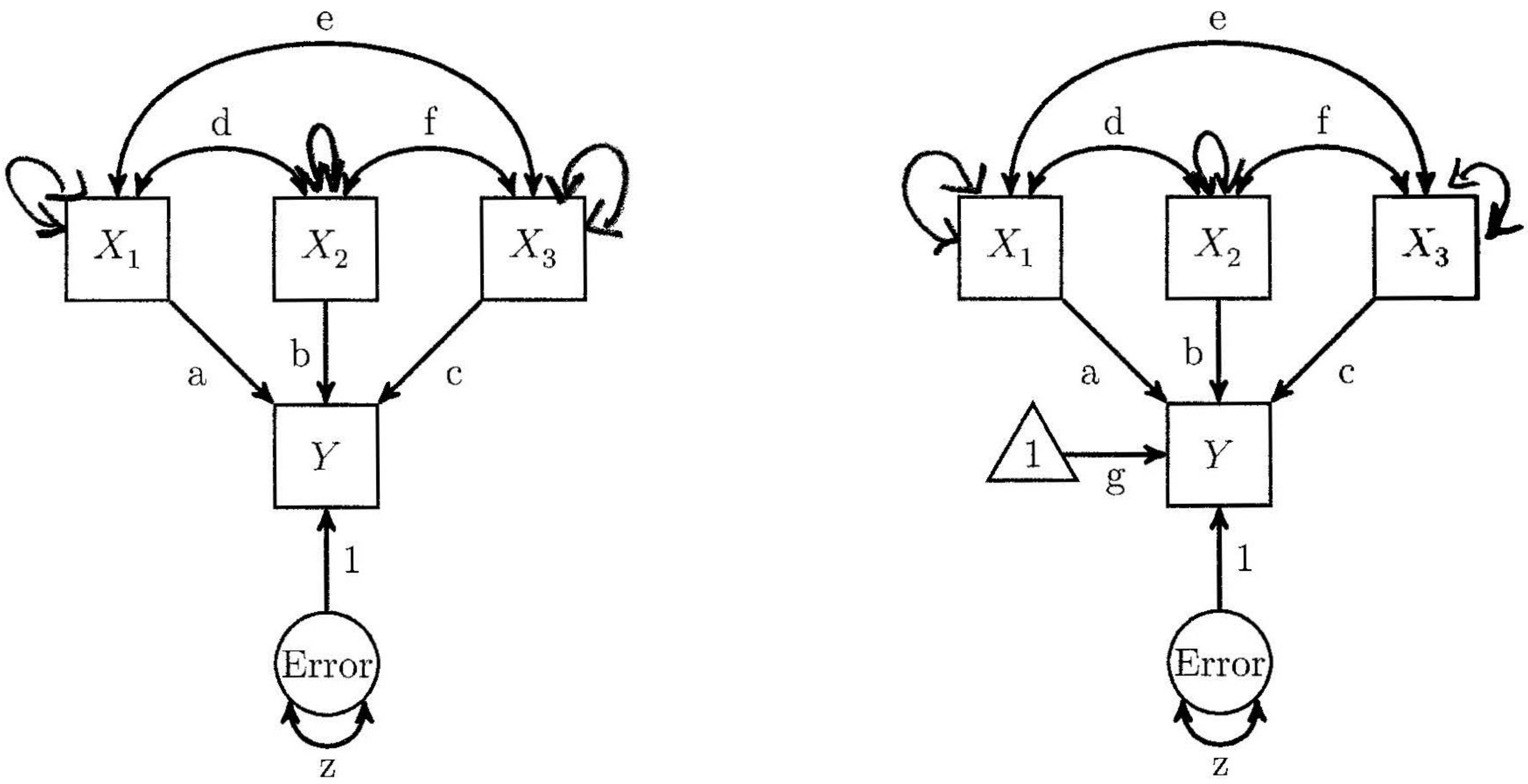
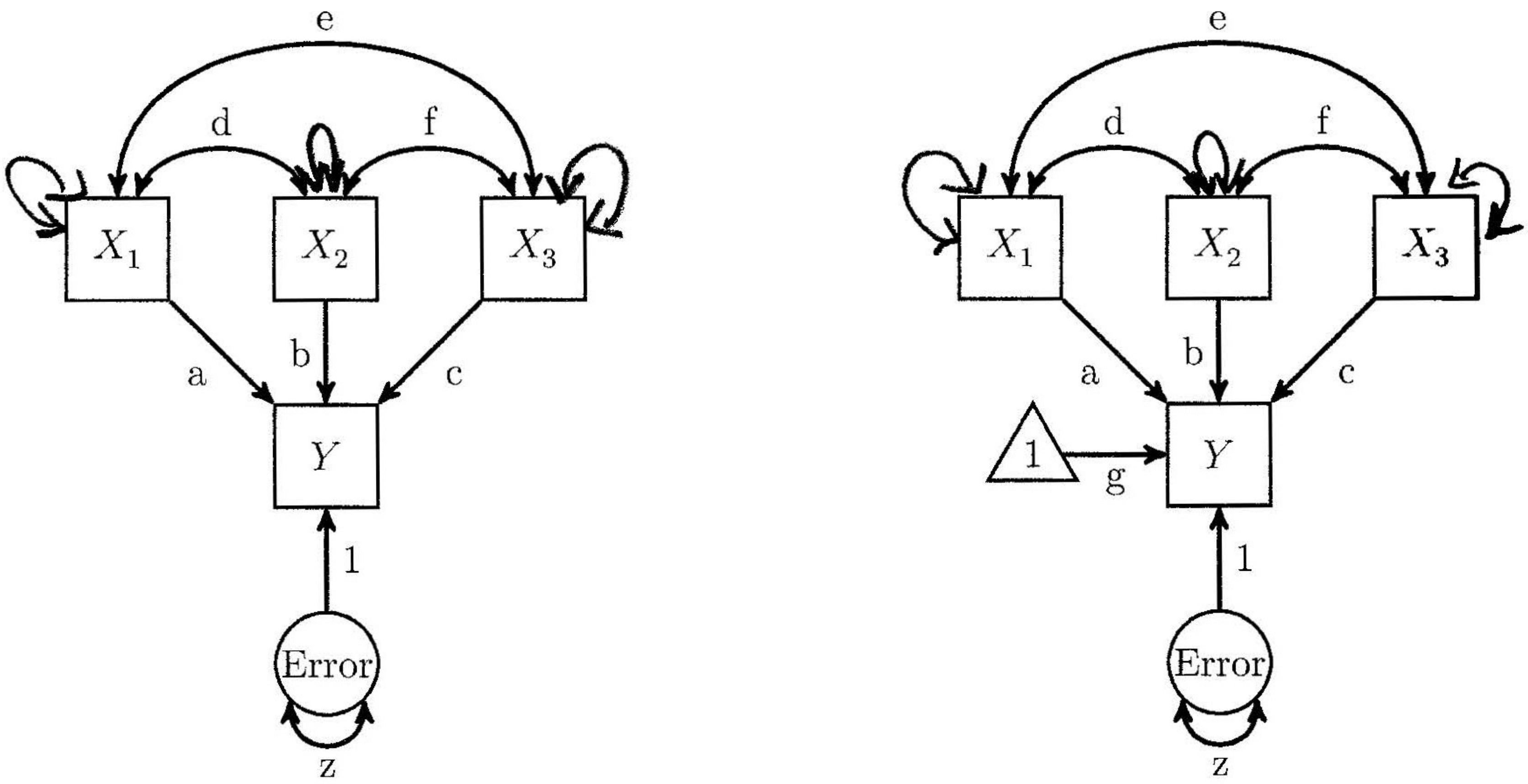


Figure 2.2 Path model of a multiple regression with three predictor (exogenous) variables.

Mean structure often omitted



(co)variance structure only
all means omitted (i.e., assumed zero)
$Y = aX_1 + bX_2 + cX_3 + error$

(co)variance and mean structure
means freely estimated
$Y = g + aX_1 + bX_2 + cX_3 + error$

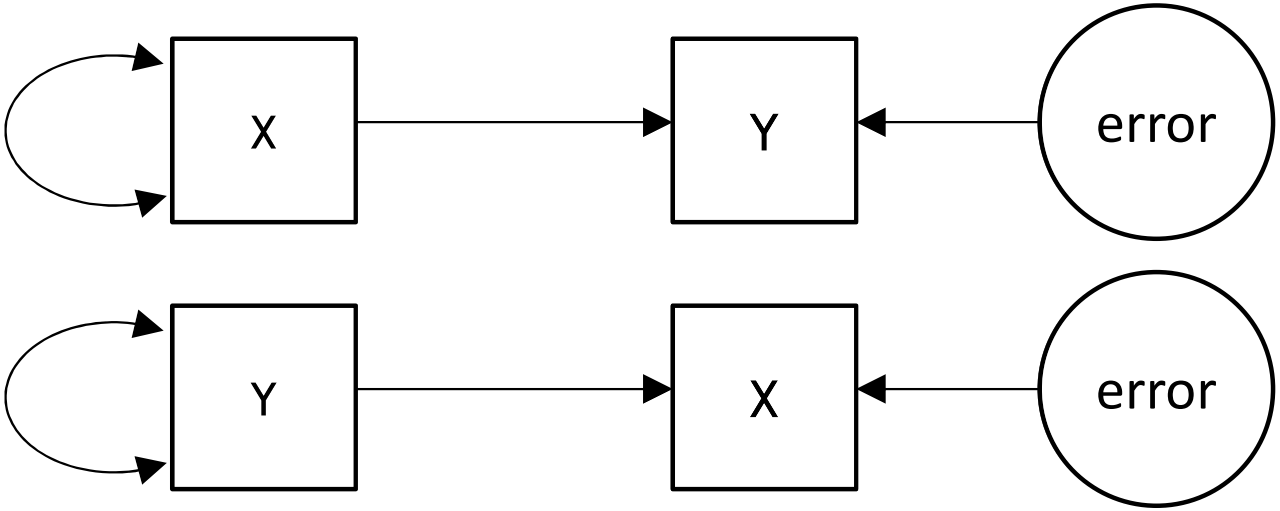
Figure 2.2 Path model of a multiple regression with three predictor (exogenous) variables.

Error terms

- Errors are also latent variables: they are hypothetical, not directly observed
- Error is defined as the difference between observed (sample) variance and variance explained by other variables in the model
  - Therefore, a variable that has an error/disturbance term is an endogenous variable (vice versa)
  - Errors/disturbance terms are always exogenous (i.e., no incoming directional arrows)

Causation

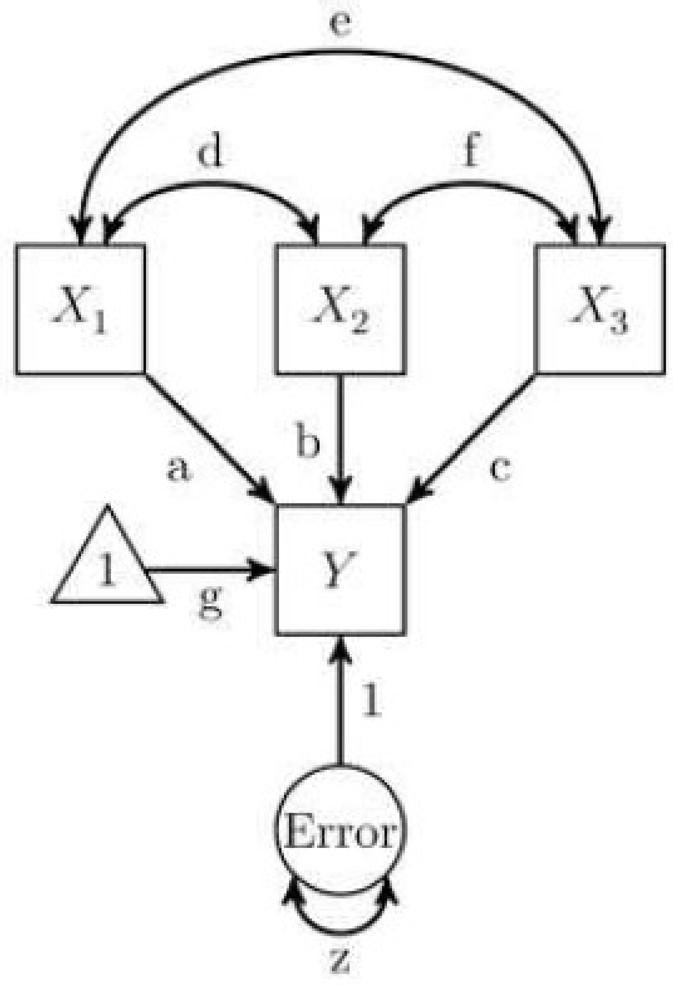
- Causality needs to be ensured through the research design, cannot be statistically proven



- Both models above will fit the observed data equally well, it is up to the researcher to decide on the direction of the arrows!
  - In the SEM model, it is merely a matter of scaling:
    $b_x = \frac{cov_{xy}}{var_x}$ and $b_y = \frac{cov_{xy}}{var_y}$

Path & partial regression coefficients

- Path coefficients (a, b, c, g and 1) are partial regression coefficients
- That is, the expected increase in the response variable, when the predictor variable increases by 1, controlling for (= keeping constant) all the other predictor variables
  - Note that the intercept is always 1, so cannot in- or decrease



Standardized coefficients

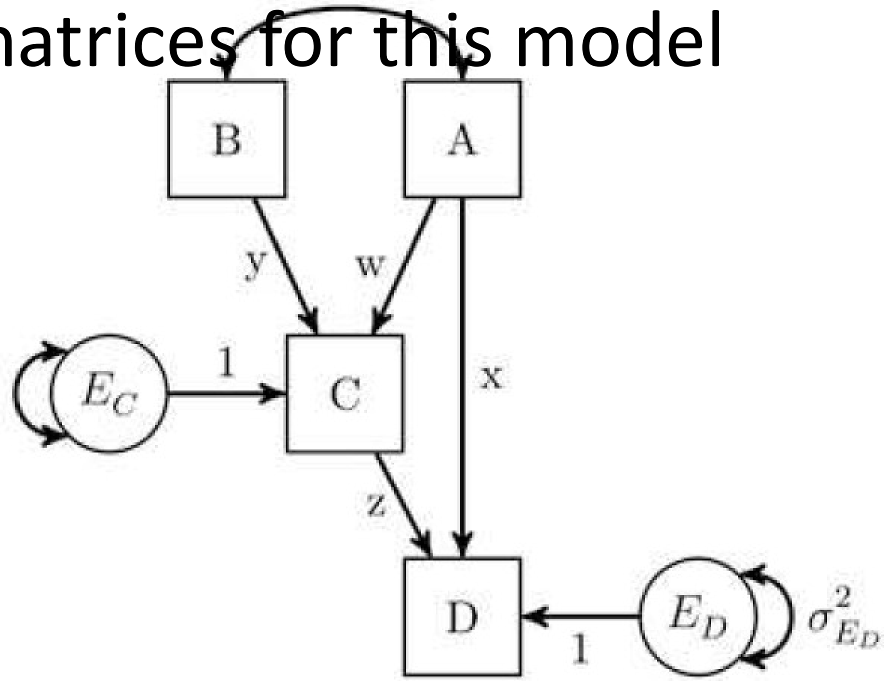
- Parameter estimates (path coefficients) can be standardized and unstandardized
  - Unstandardized: Interpret like regression coefficients
    - Expected increase in Y if X increases by 1
  - Standardized: Interpret like correlation coefficients
    - Expected increase in SDs of Y if X increases by 1 SD
    - 0: no linear association; -1: perfect negative association; 1: perfect positive association
    - squared standardized coefficient = prop. of variance in Y explained by X (vice versa)

Lavaan model syntax

<table>
<tr><th>Syntax</th><th>Command</th><th>Example</th></tr>
<tr><td>~</td><td>Regress onto</td><td>Regress B onto A: B ~ A</td></tr>
<tr><td>~~</td><td>(Co)variance</td><td>Variance of A: A ~~ A
Covariance of A and B: A ~~ B</td></tr>
<tr><td>~1</td><td>Constant/mean/intercept</td><td>Regress B onto A, and and include the intercept in the model: B ~ 1 + A or
B ~ A
B ~ 1</td></tr>
<tr><td>=~</td><td>Define reflective latent variable</td><td>Define Factor 1 by A-D:
F1 =~A+B+C+D</td></tr>
<tr><td><~</td><td>Define formative latent variable</td><td>Define Factor 1 by A-D:
F1 <~ 1*A+B+C+D</td></tr>
<tr><td>:=</td><td>Define non-model parameter</td><td>Define parameter u2 to be twice the square of u:
u2 := 2*(u^2)</td></tr>
<tr><td>*</td><td>Label parameters (the label has to be pre-multiplied)</td><td>Label the regression of Z onto X as b:
Z ~b*X</td></tr>
<tr><td>|</td><td>Define the number of thresholds (for categorical endogenous variables)</td><td>Variable u has three thresholds:
u | t1 + t2 + t3</td></tr>
</table>

Lavaan syntax exercise

1) How do we write the model below in lavaan syntax?
2) How can we label and refer to the indirect effect from A on D via C in lavaan syntax?
3) What do the beta and psi matrices for this model look like?



Note that Beaujean often labels paths in lavaan syntax, but that is not required - I never do it, unless there are indirect effects that I want to explicitly define in the model. Labeling or not does not make a difference for the estimated parameters and model fit.

Homework

- Exercises 2.2 and 2.3 (see PDF on GitHub repo)
- See Example-2.4.1.pdf on GitHub repo for instructions on extracting beta and psi matrices