Úvod do strukturního modelování

Michal Jabůrek, Stanislav Ježek, Hynek Cígler, Adam Ťápal

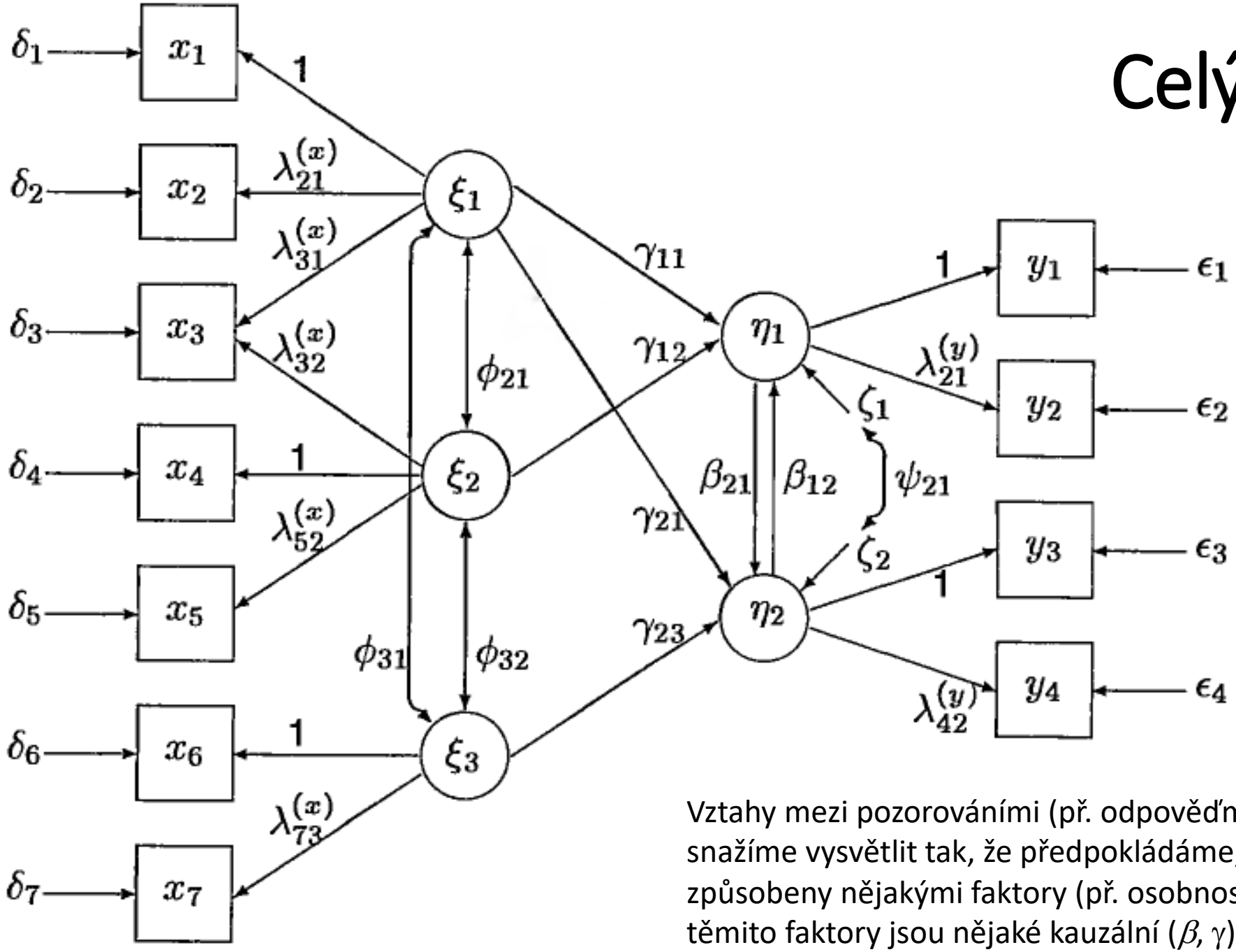
PSY028_E – Statistická analýza dat v psychologii
FRMU podpořený kurz

Cíle kurzu

Získat základní schopnost využít analytické možnosti, které SEM nabízí

- Konkretizace uvažování o kauzálních vztazích mezi více proměnnými
  - PATH ANALYSIS
- Práce s latentními proměnnými namísto součtových skórů (SEM)
  - očištění vztahů mezi proměnnými o některé nedokonalosti měření proměnných
- Konkretizace uvažování o vztahu mezi pozorovanými indikátory a konstrukty (CFA)
  - reflexe kvalit měření

Celý SEM model

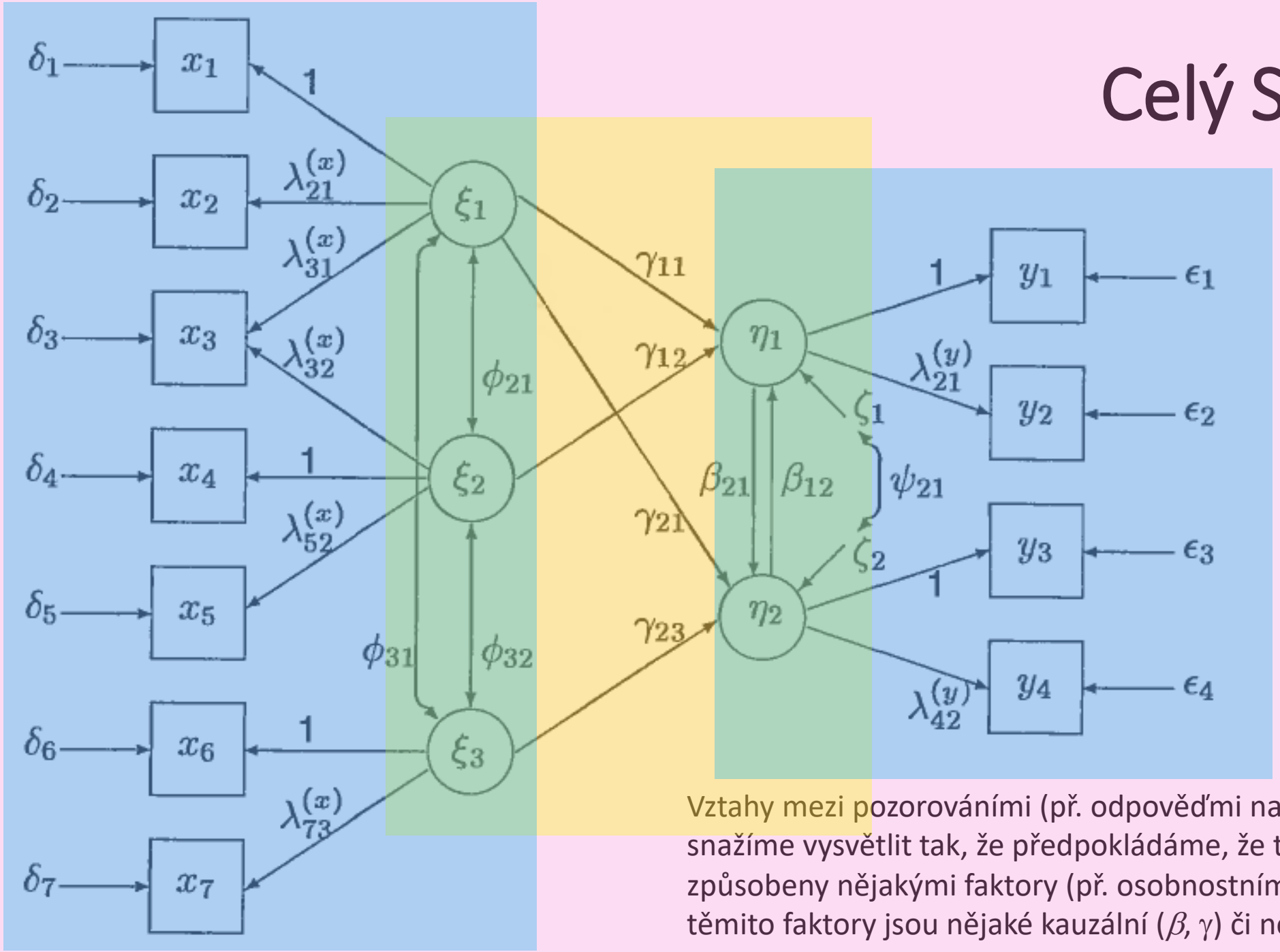


Vztahy mezi pozorováními (př. odpověďmi na položkami – *x*, *y*) se snažíme vysvětlit tak, že předpokládáme, že tyto odpovědi jsou způsobeny nějakými faktory (př. osobnostními – ξ, η) a, že mezi těmito faktory jsou nějaké kauzální (β, γ) či nekauzání (ϕ, ψ) vztahy.

Modifikovaný program

- Setkání 1 – Úseková (path) analýza jako extenze lineární regrese
  - dnes
- Setkání 2 – Latentní proměnné vysvětlující pozorované indikátory
  - zítra
- Setkání 3 – Praktická rozšíření CFA modelů
  - příští pátek
- Setkání 4 – SEM jako úseková analýza latentních proměnných
  - příští sobotu

Celý SEM model



Vztahy mezi pozorováními (př. odpověďmi na položkami – *x*, *y*) se snažíme vysvětlit tak, že předpokládáme, že tyto odpovědi jsou způsobeny nějakými faktory (př. osobnostními – ξ, η) a, že mezi těmito faktory jsou nějaké kauzální (β, γ) či nekauzání (ϕ, ψ) vztahy.

Program pro 1. setkání

Cíl: realizovat a interpretovat úsekovou analýzu v **R** pomocí balíku **lavaan**

- Lineární regrese jako SEM model
  - prvky modelu a jejich zobrazení
  - specifikace modelu a odhad jeho parametrů pomocí funkcí lavaan
  - modelem implikovaná korelační (kovarianční matice)
- Úsekové modely
  - jednoduchá mediace
  - parciální korelace
  - složitější modely

Lineární regrese

Na základě korelací mezi proměnnými lze jednu z nich predikovat ostatními

$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_k X_k + e$

Přiřazujeme tím proměnným funkční role:

závislá (endogenní) a nezávislá (exogenní)

Je na nás a teorii, kterou proměnnou predikovat a které proměnné budou prediktory

Dáváme tím korelacím nějaký konkrétní smysl

Lineární regrese – příklad
GPA – SES – nAch – IQ

Korelace → **Regrese**

	GPA	SES	nAch	IQ
GPA	1,00	0,24	0,45	0,47
SES	0,24	1,00	0,30	0,24
nAch	0,45	0,30	1,00	0,16
IQ	0,47	0,24	0,16	1,00

$GPA' = 0{,}03SES + 0{,}37nAch + 0{,}40IQ$

$R^2 = 0{,}37$

Kdyby SES, nAch a IQ **způsobovaly** GPA tímto způsobem, pozorovali bychom mezi nimi tyto korelace

(ano, ty korelace se zadí z regresních parametrů zase zpětně spočítat, ale o tom později)

Lineární regrese v R

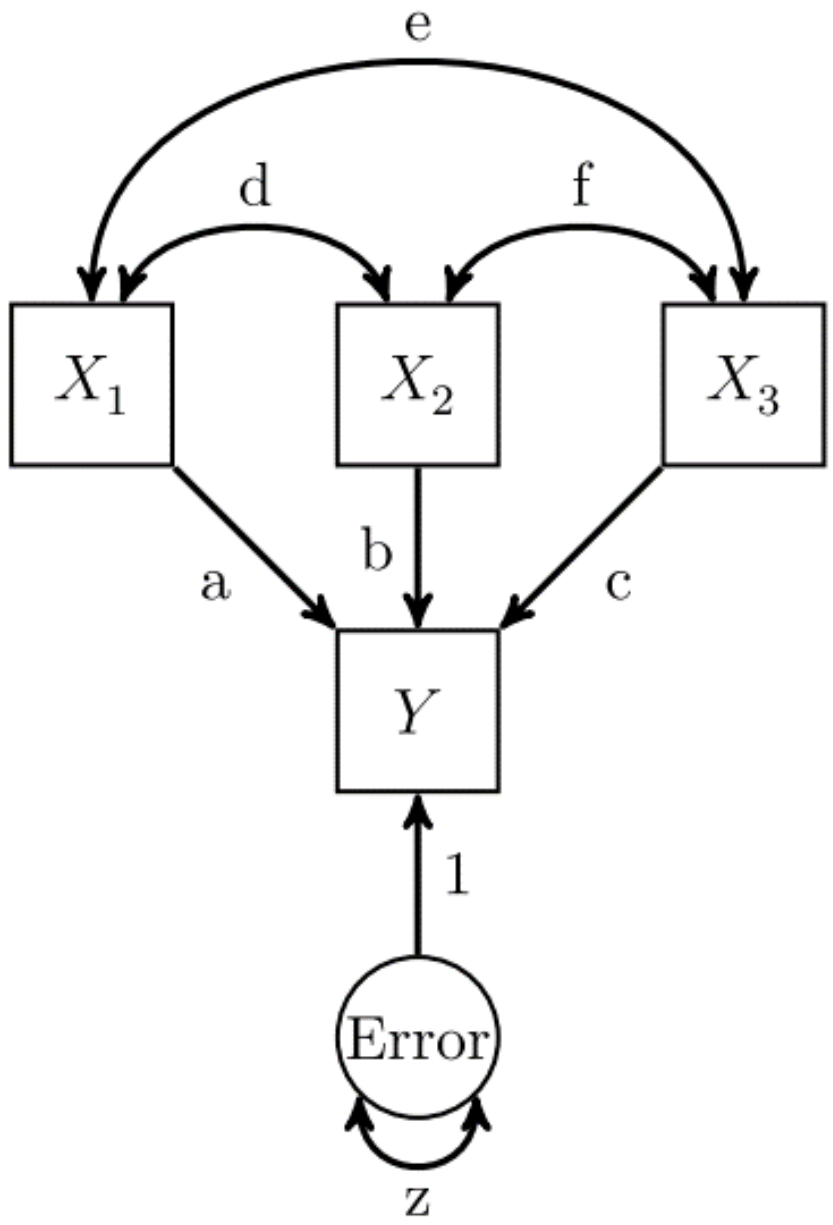
- Data Pedhazur

```
Coefficients:
             Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept) -17.47656      0.00000    2.02594  -8.626   <2e-16 ***
SES           0.13846      0.03476    0.10777   1.285    0.199
nAch          0.29912      0.37277    0.02134  14.019   <2e-16 ***
IQ            0.24911      0.40273    0.01617  15.406   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.471 on 996 degrees of freedom
Multiple R-squared:  0.3654,  Adjusted R-squared:  0.3635
F-statistic: 191.2 on 3 and 996 DF,  p-value: < 2.2e-16
```

LR jako úsekový model

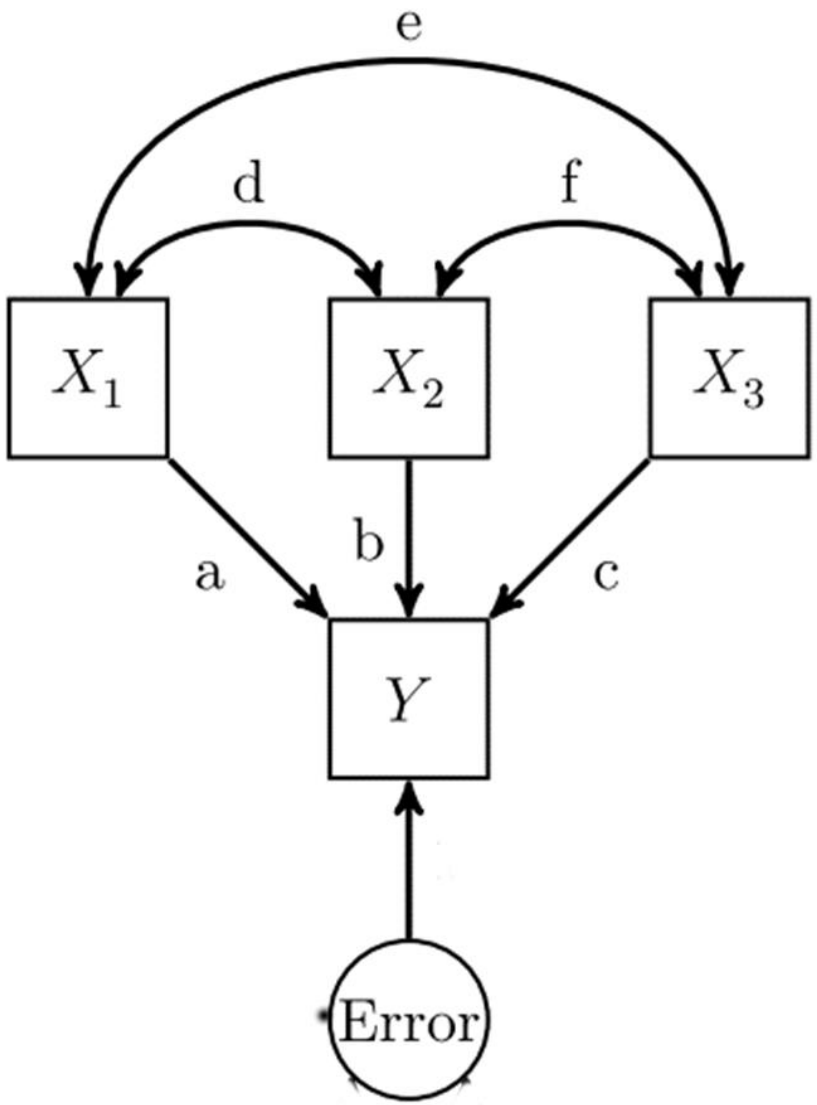
- Model tentýž, jen explicitně specifikujeme **každý** jeho prvek
- Co nezmíníme, v modelu není
- Úsekový model je **kauzální** – kauzalita efektu X na Y musí být teoreticky plauzibilní, hypotetizovaná
- Prvky modelu mají svou konvenční grafickou podobu v **path (structural) diagramech**



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}$

LR jako úsekový model
Prvky modelu

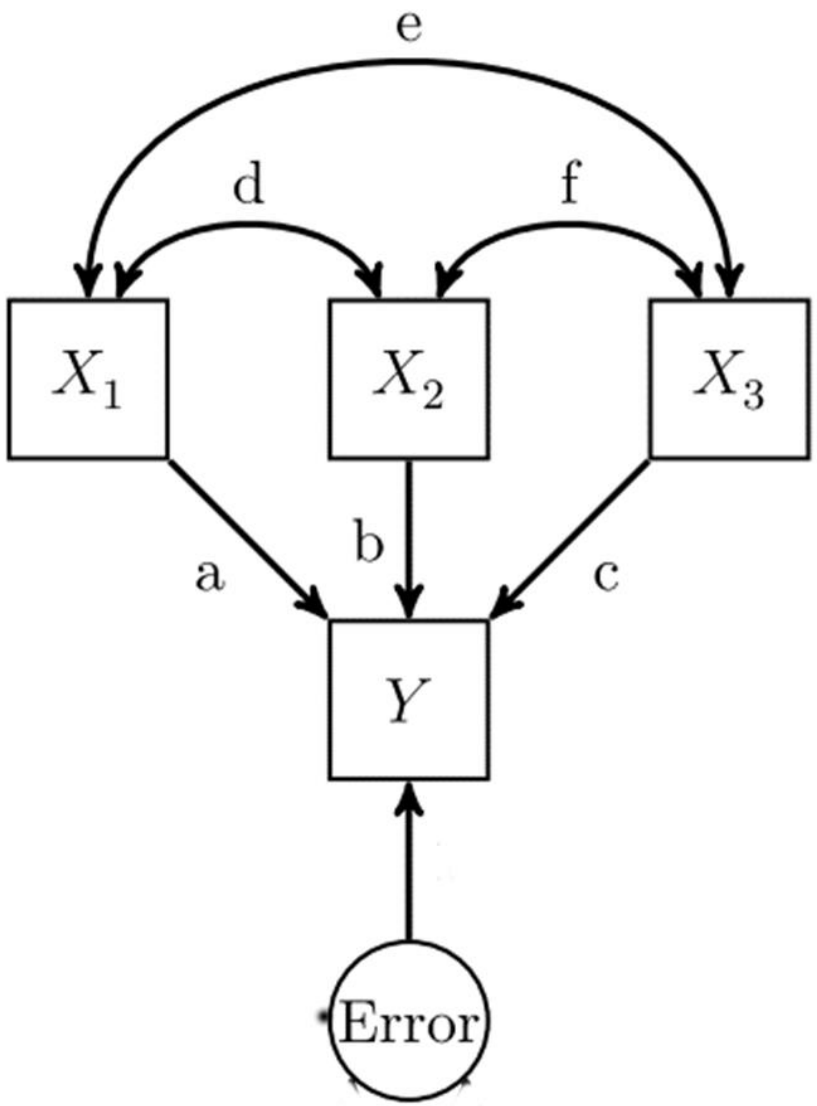
- Měřené (pozorované) proměnné – **MANIFESTNÍ** proměnné
- Zde *Y, X_1*…
- **Čtverec nebo obdélník**



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

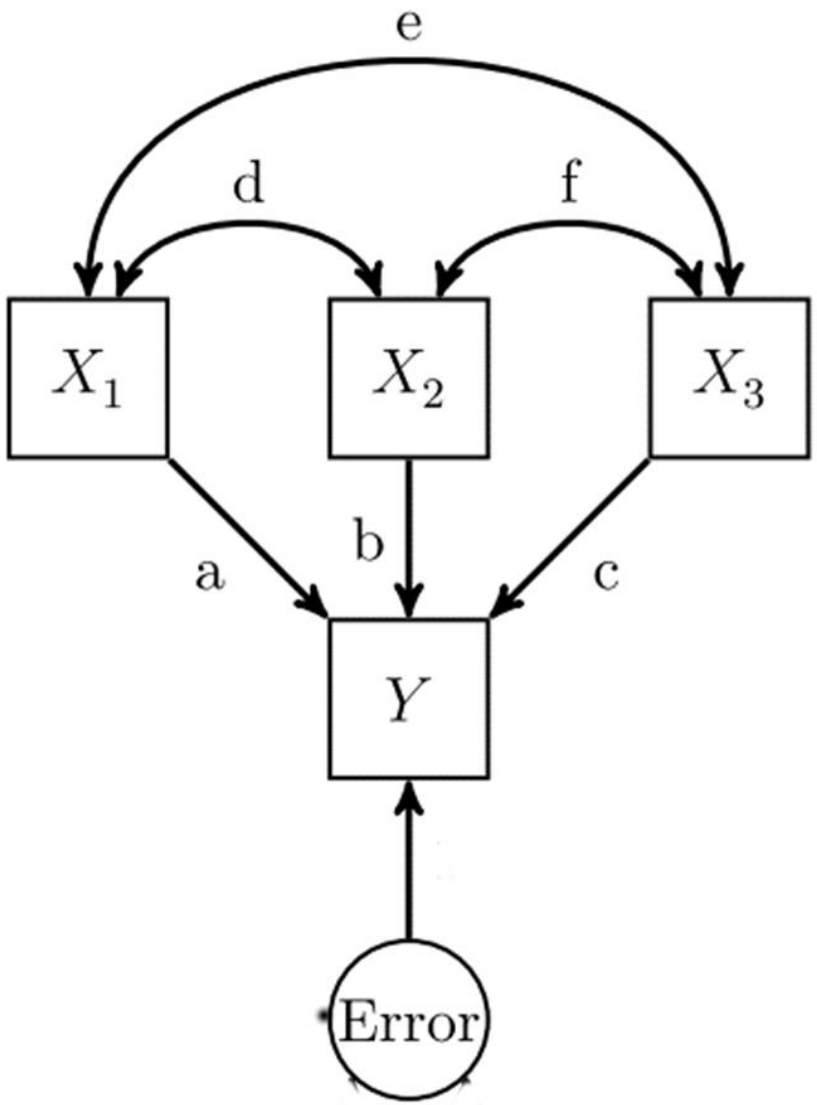
- Přímý efekt X na Y – **direct effect DI**
- Jednosměrná rovná šipka z X do Y (např. a
- Úsekový (path) koeficient nese stejný význam jako regresní koeficient – změna vyvolaná jednotkovou změnou Y
- $p_{YX} = b_{YX}$ (v subskriptu se začíná závislou!)
- V úsekové analýze obvykle ve standardizované podobě: $p_{YX} = \beta_{YX}$



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

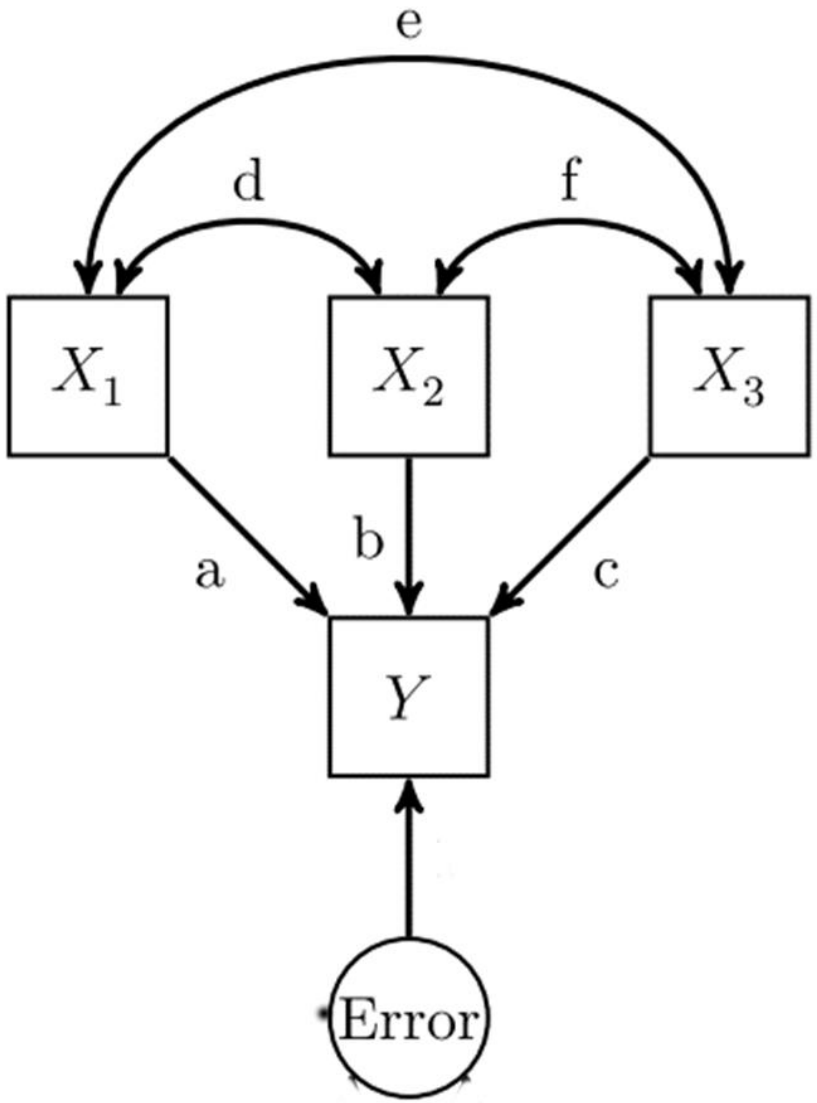
- Nekauzální vztah mezi proměnnými – NON-CAUSAL, **UNANALYZED**
- Obousměrná zakřivená šipka (např. e)
- Obvykle mezi exogenními proměnnými nebo mezi rezidui
- Korelace, kovariance
- Víme, že proměnné korelují, ale příčina této korelace leží mimo náš model, není analyzována



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

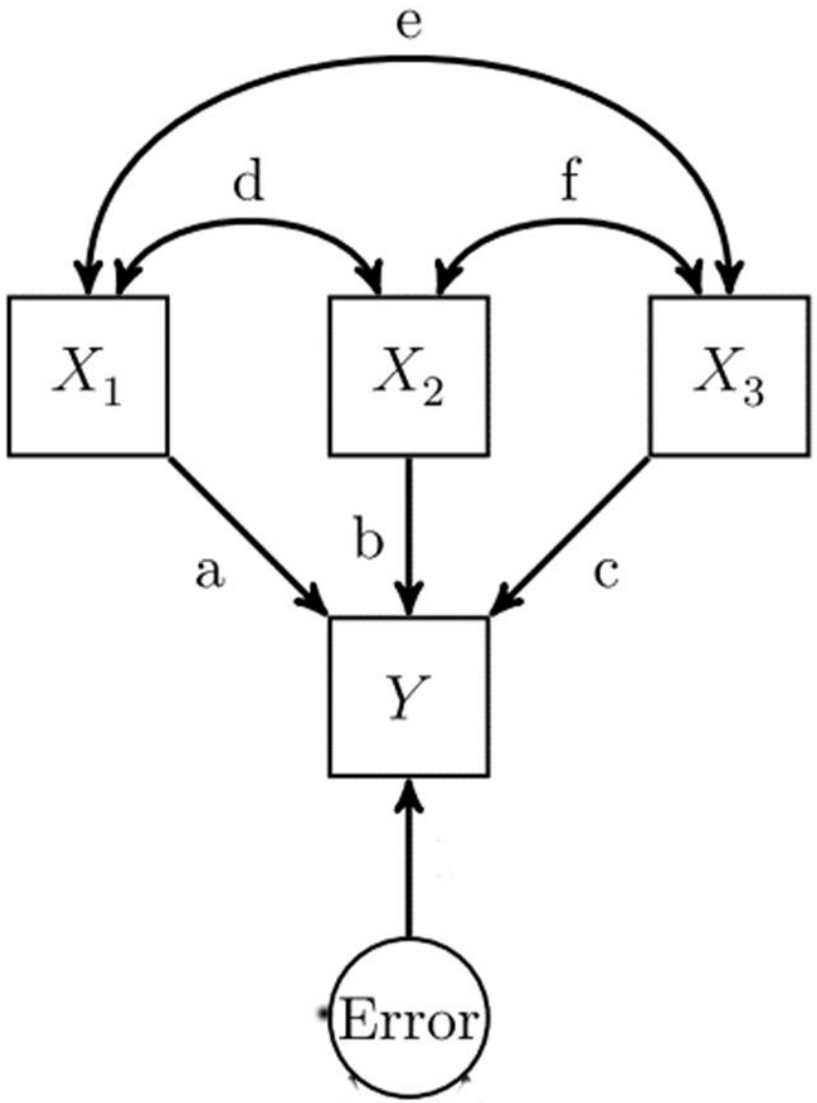
- Nepozorované proměnné reprezentující kauzální činitele, které na základě teorie předpokládáme či potřebujeme – **LATENTNÍ PROMĚNNÉ**
- Ústřední prvek SEM modelů
- Zde jen v podobě proměnné reprezentují reziduální rozptyl Y – proměnná reprezentující všechny vlivy ovlivňující Y, které nejsou v modelu



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

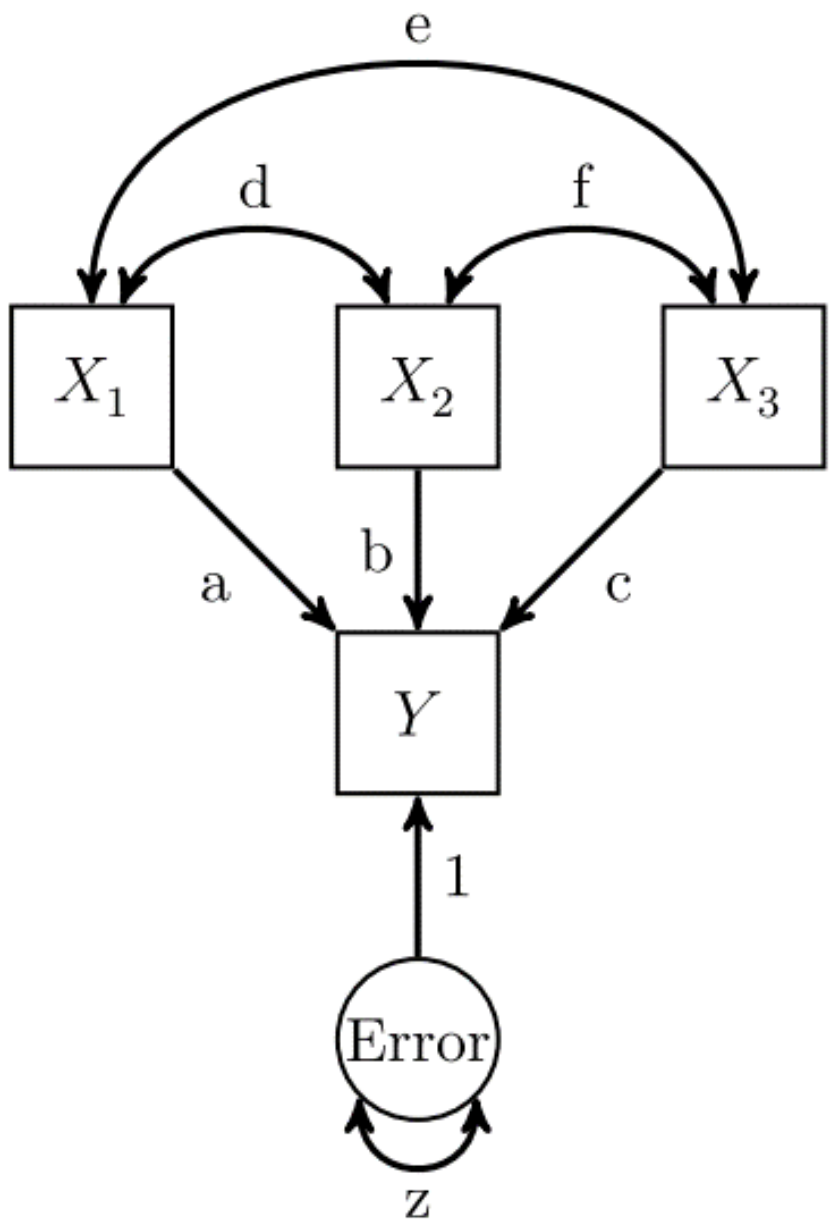
- Každá endogenní proměnná – každá, do které míří alespoň jedna kauzální šipka – má svou proměnnou reprezentující její modelem nevysvětlený rozptyl
- Rezidua, **disturbance**



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

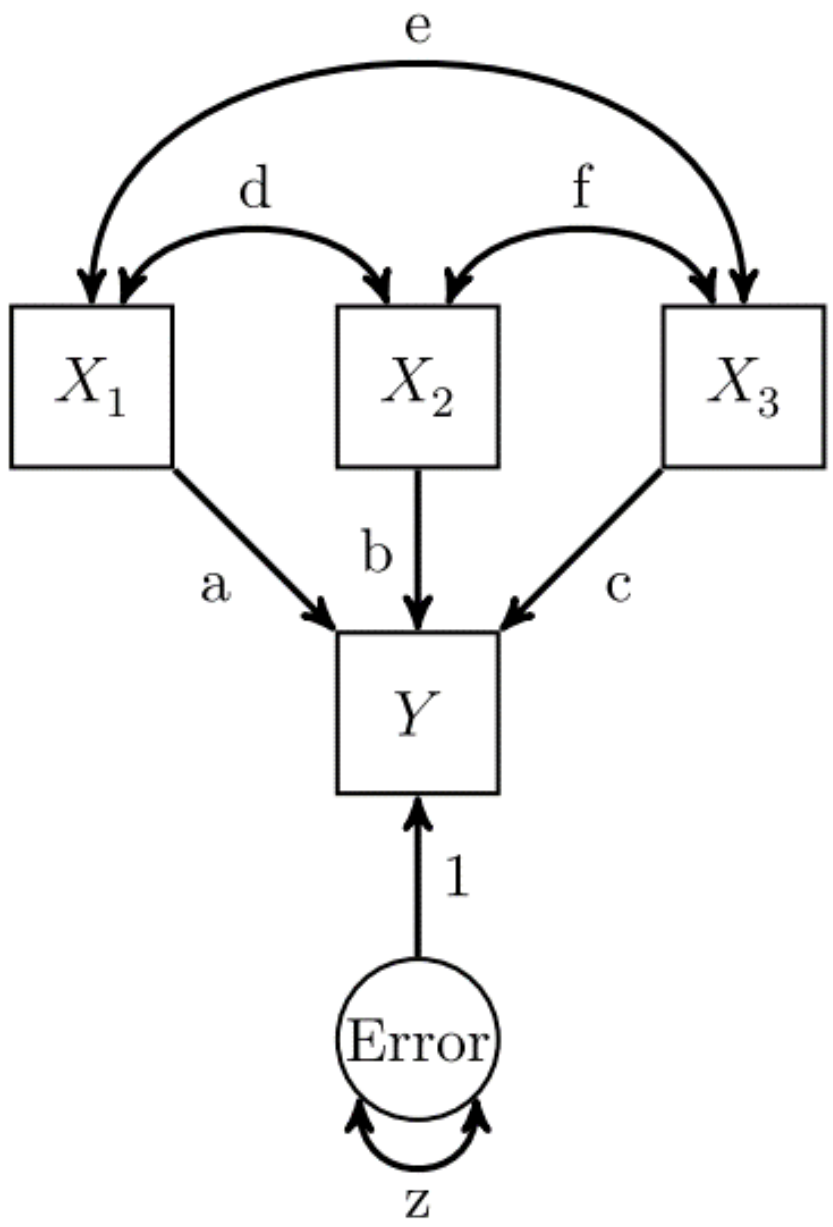
- **Rozptyly** proměnných (pokud neanalyzujeme korelační matici – standardizovaný model)
  - Ve standardizovaném modelu jsou rozptyly=1
- Oboustranná šipka z proměnné do ní samotné
  - V diagramu často chybí
- Každá exogenní (manifestní i latentní) má rozptyl (vč. reziduí/disturbancí)
- Endogenní ho nemají – je plně vysvětlen



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

LR jako úsekový model
Prvky modelu

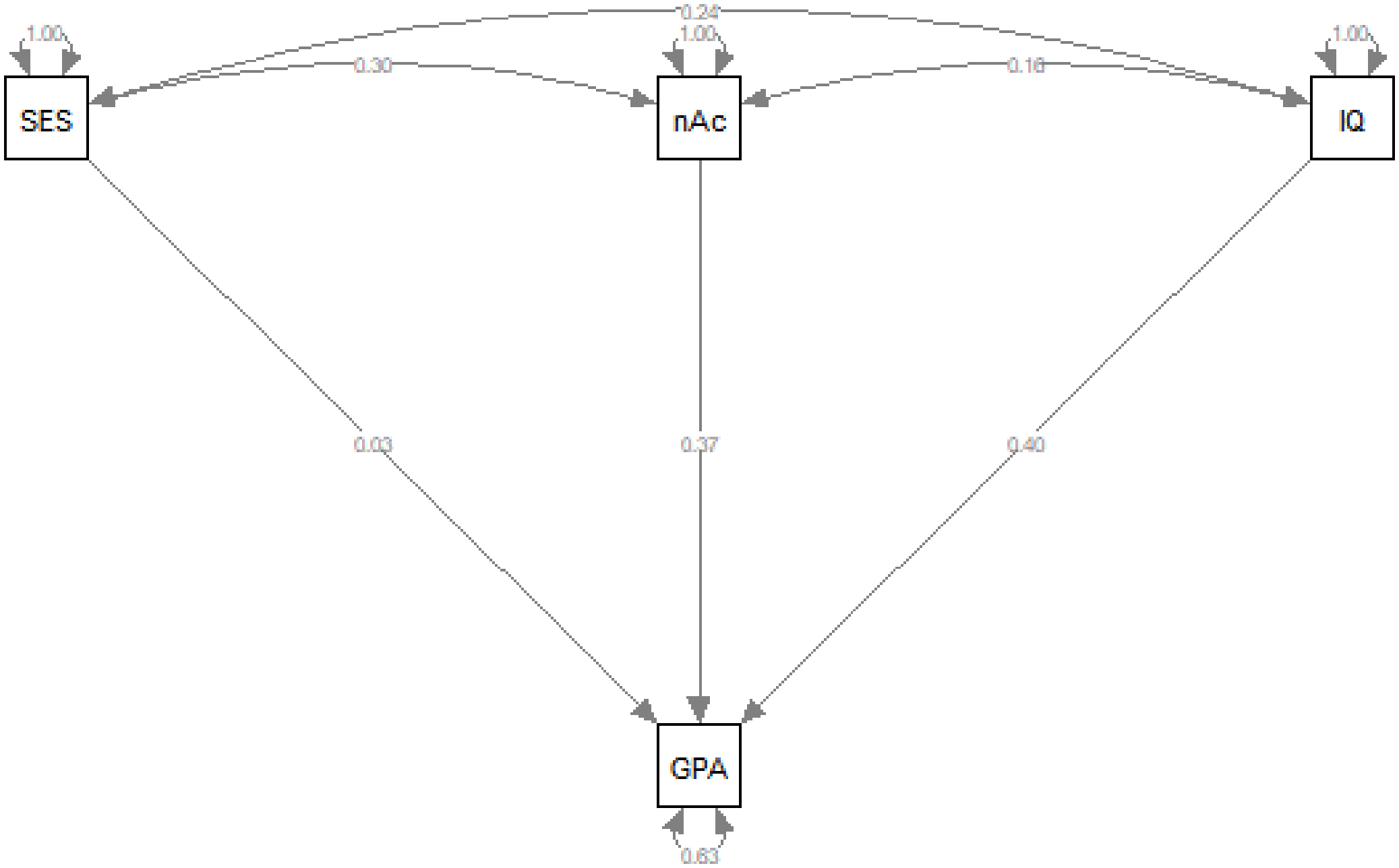
- **Absence vztahu** – když mezi proměnnými není specifikován přímý efekt nebo korelace – tak to znamená, že hypotetizujeme, že je vztah má skutečně hodnotu 0.



(a) Standardized model:
$Y = aX_1 + bX_2 + cX_3 + \text{Error}.$

Pojďme si nyní sestrojit úsekovou variantu předchozího modelu

```
# Regressions:
#                Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
# GPA ~
#   SES      (b1)   0.138    0.007   19.471    0.000    0.125    0.152    0.138    0.035
#   nAch     (b2)   0.299    0.000  639.316    0.000    0.298    0.300    0.299    0.373
#   IQ       (b3)   0.249    0.000  925.045    0.000    0.249    0.250    0.249    0.403
#
# Covariances:
#                Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
# SES ~~
#   nAch     (r12)   8.257    0.032  260.982    0.000    8.195    8.319    8.257    0.301
#   IQ       (r13)   8.643    0.032  273.194    0.000    8.581    8.705    8.643    0.243
# nAch ~~
#   IQ       (r23)  28.494    0.032  900.604    0.000   28.432   28.556   28.494    0.161
#
# Variances:
#                Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
#   SES      (v1)   5.525    0.032  174.632    0.000    5.463    5.587    5.525    1.000
#   nAch     (v2) 136.169    0.032 4303.898    0.000  136.107  136.231  136.169    1.000
#   IQ       (v3) 229.161    0.032 7243.086    0.000  229.099  229.223  229.161    1.000
#  .GPA      (e1)  55.641    0.042 1337.512    0.000   55.559   55.722   55.641    0.635
```

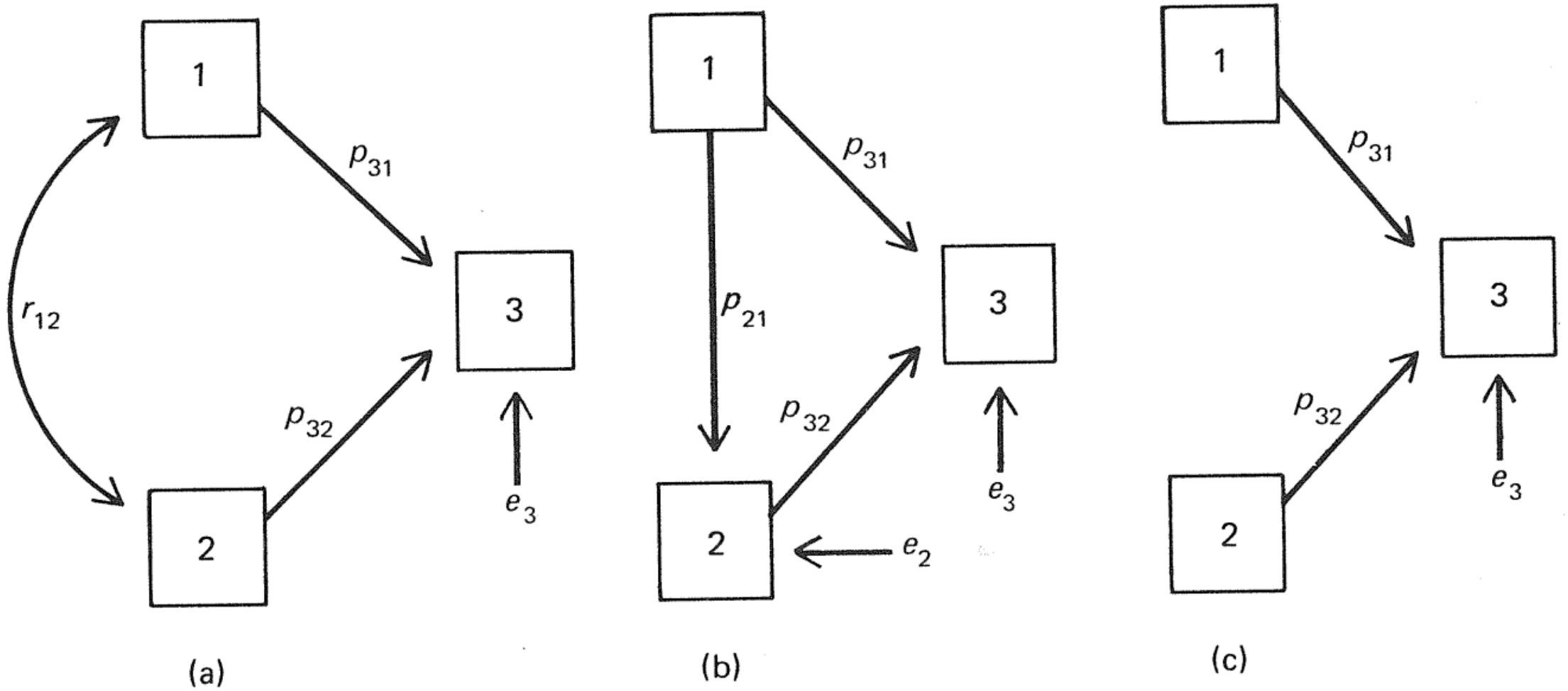

1.00
1.00
1.00
0.24
0.30
0.16
SES
nAc
IQ
0.03
0.37
0.40
GPA
0.63

Dekompozice korelační matice jako účel strukturního modelování

Čím to je, že spolu proměnné korelují? (to je to, co můžeme pozorovat)

- Protože jedna způsobuje druhou (aspoň z části) – **přímý efekt (DI)**
  - b_{YX}
- Protože jedna způsobuje druhou nepřímo – **nepřímý efekt (IE)**
  - mediace, X→M→Y, efekt se skládá z přímých efektů
  - $b_{YM} . b_{MX}$
- Protože mají obě stejnou *příčinu* – **zdánlivá korelace (S)**purious
  - Příčina musí být v modelu
- Protože sdílí nemodelované příčiny – **neanalyzovaná (U)**nanalyzed
  - Mohou zahrnovat i nepřímé efekty zahrnující jednu ne-kauzální cestu

Jedna korelace může mít více zdrojů. Protože jsou nezávislé, sčítají se.


1
2
3
p_{31}
p_{32}
r_{12}
e_3
(a)
Correlated Causes
1
2
3
p_{31}
p_{21}
p_{32}
e_2
e_3
(b)
Mediated Cause
1
2
3
p_{31}
p_{32}
e_3
(c)
Independent Causes

Interpretujme náš regresní – úsekový model

Interpretujme náš regresní – úsekový model *GPA – SES – nAch – IQ*

Korelace → **Úsekový model**

	GPA 1	SES 2	nAch 3	IQ 4
GPA	1,00	0,24	0,45	0,47
SES	0,24	1,00	0,30	0,24
nAch	0,45	0,30	1,00	0,16
IQ	0,47	0,24	0,16	1,00

r_{12}=?

DI: β_{12}=0,035

UN1: $r_{23} * \beta_{13}$= 0,30 * 0,37 = 0,11

UN2: $r_{24} * \beta_{14}$= 0,24 * 0,40 = 0,10

r_{12}= DI + UN1 + UN2 = 0,245

SES tedy s GPA koreluje převážně z neznámých příčin, přímý efekt má minimální.

Wrightovy „tracing rules“

- Trace all paths between two variables (or a variable back to itself), multiplying all the coefficients along a given path.
- You can start by going backwards along a single-headed arrow, but once you start going forward along these arrows you can no longer go backwards.
- No loops! That is, you cannot go through the same variable more than once for a given path.
- At *maximum*, there can be one double-headed arrow included in a path.
- After tracing all the paths for a given relationship, sum all the paths.

Figure 2.5 Tracing rules for a path model without a constant term.

Interpretujme náš regresní – úsekový model *GPA – SES – nAch – IQ*

Korelace → **Úsekový model**

	GPA 1	SES 2	nAch 3	IQ 4
GPA	1,00	0,24	0,45	0,47
SES	0,24	1,00	0,30	0,24
nAch	0,45	0,30	1,00	0,16
IQ	0,47	0,24	0,16	1,00

r_{12}=?

DI: β_{12}=0,035

UN1: $r_{23} * \beta_{13}$= 0,30 * 0,37 = 0,11

UN2: $r_{24} * \beta_{14}$= 0,24 * 0,40 = 0,10

r_{12}= DI + UN1 + UN2 = 0,245

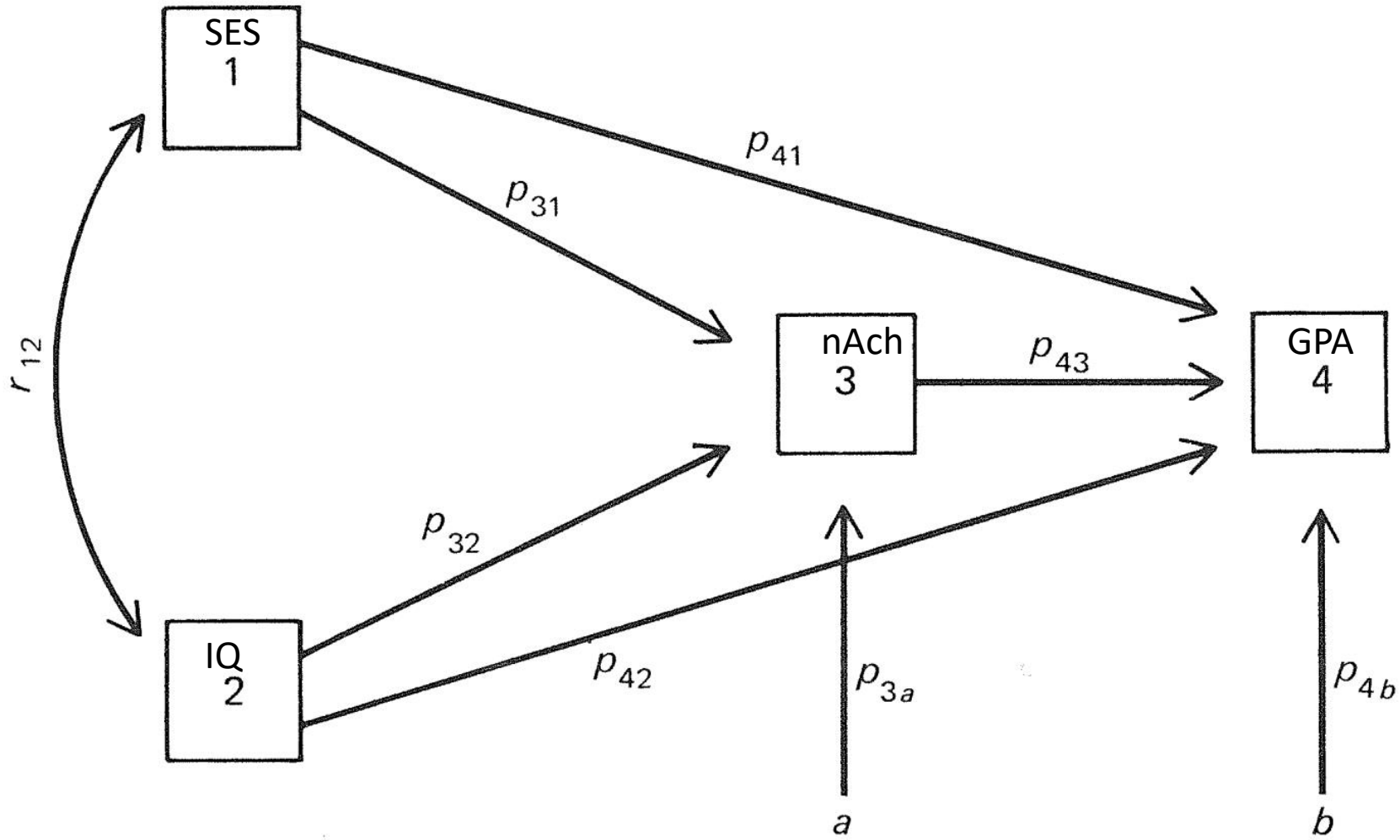
SES tedy s GPA koreluje převážně z neznámých příčin, přímý efekt má minimální.

Předpoklady

- Aby výše uvedené interpretace vztahů mohly platit, předpokládáme poměrně hodně věcí:

- Vztahy v modelu jsou **lineární, aditivní a kauzální**. Tedy nelineární vztahy a interakce do modelu nepatří.
- Rezidua nekorelují s proměnnými, které jim v modelu předcházejí. Obecně rezidua mohou korelovat pouze s jinými rezidui, které jsou v kauzálním modelu na stejné úrovni.
- V modelu jsou všechny relevantní proměnné. Tedy neměřené proměnné, které nejsou reprezentovány rezidui, nekorelují s proměnnými v modelu (LOVE, omission error, specification error)
- Proměnné jsou měřeny na intervalové škále (ideálně spojité, normálně rozložené)
- Proměnné jsou měřené bez chyby

„Opravdový“ kauzální model



Specifikace úsekového modelu

- Každá endogenní proměnná má svou regresní rovnici – z **DI**
  - nAch = p_{31}SES + p_{32}IQ (+ p_{3a}a)
  - GPA = p_{41}SES + p_{42}IQ + p_{43}nAch (+ p_{4b}b)
- Exogenní proměnné (na téže úrovni kauzality) mohou korelovat (kovariovat) **- UN**
  - $r_{SES\text{-}IQ}$
  - $r_{SES\text{-}a}$ a $r_{IQ\text{-}a}$ jsou v souladu s předpoklady 0, totéž platí pro $r_{SES\text{-}b}$ $r_{IQ\text{-}b}$ a $r_{a\text{-}b}$
- Exogenní proměnné mají rozptyly
  - exogenní manifestní: σ^2_{SES}, σ^2_{IQ}
  - disturbance: σ^2_{a}, σ^2_{b}
- Modré koeficienty(parametry) chceme stanovit, zelené z nich pak vyplynou. Modrých je 10.

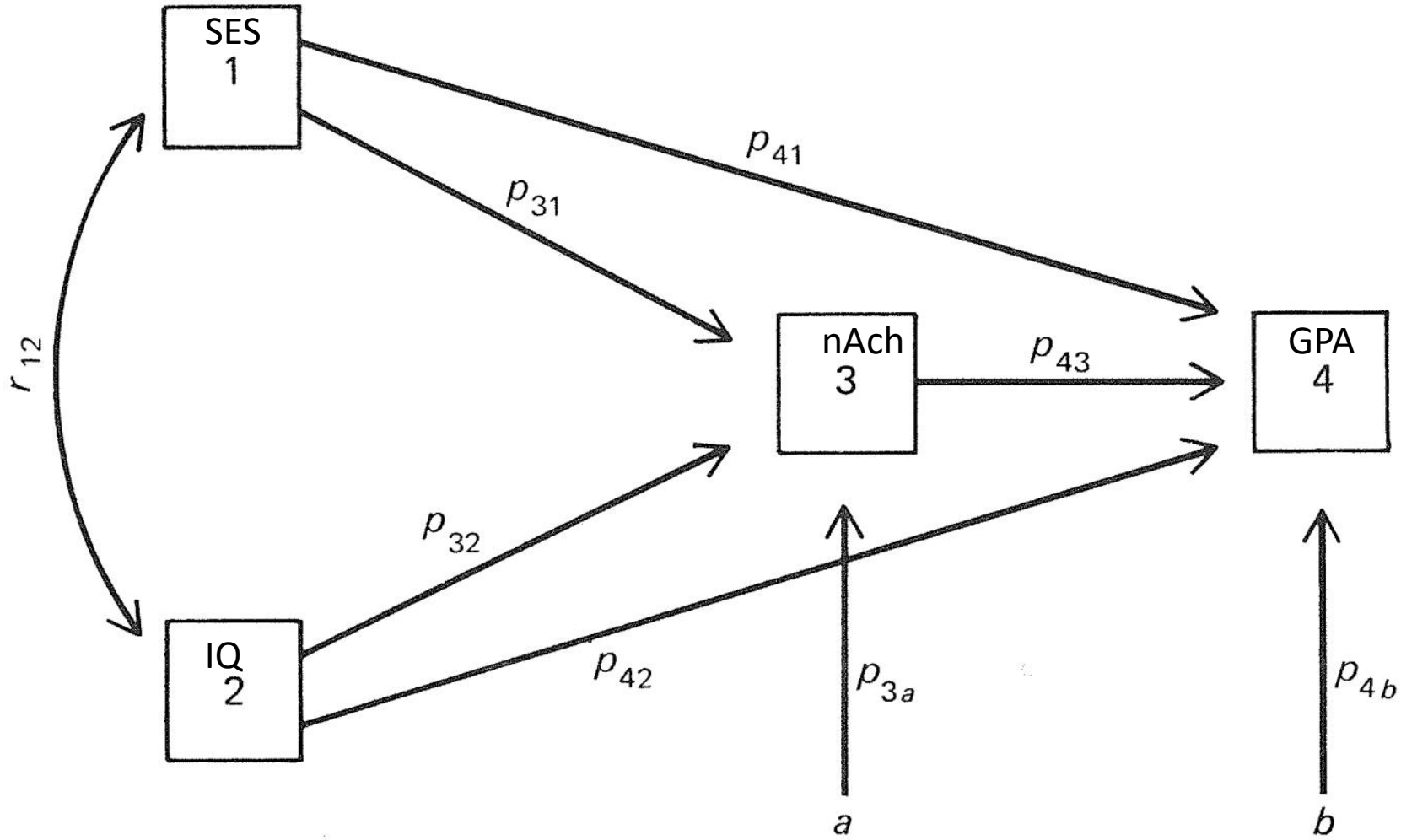
Specifikace úsekového modelu - lavaan

- Každá endogenní proměnná má svou regresní rovnici – z **DI**
  - `nAch ~ SES + IQ`
  - `GPA ~ SES + IQ + nAch`
- Exogenní proměnné (na téže úrovni kauzality) mohou korelovat (kovariovat) **- UN**
  - `SES ~~ IQ`
  - $r_{SES\text{-}a}$ a $r_{IQ\text{-}a}$ jsou v souladu s předpoklady 0, totéž platí pro $r_{SES\text{-}b}$ $r_{IQ\text{-}b}$ a $r_{a\text{-}b}$
- Exogenní proměnné mají rozptyly
  - exogenní manifestní: `SES ~~ SES, IQ~~IQ`
  - disturbance: `nAch ~~ nAch,  GPA ~~ GPA`
- Jednotlivé parametry lze můžeme pojmenovat (jmeno*) před název proměnné na pravé straně výrazu

Poj

Pojďme si parametry odhadnout

„Opravdový“ kauzální model



Interpretujme náš úsekový model
GPA – nAch – IQ – SES

Korelace → **Úsekový model**

	GPA 4	SES 1	nAch 3	IQ 2
GPA	1,00	0,24	0,45	0,47
SES	0,24	1,00	0,30	0,24
nAch	0,45	0,30	1,00	0,16
IQ	0,47	0,24	0,16	1,00

r_{GPA_nAch}= 0,45

DI: β = 0,37

S1: $\beta_{41} * \beta_{31}$= 0,28 * 0,03 = 0,01

S2: $\beta_{42} * \beta_{32}$ = 0,09 * 0,40 = 0,04

UN1: β_{41} * r_{12}* β_{32} = 0,03*0,24*0,09 = 0,001

UN1: β_{42} * r_{12}* β_{31} = 0,40*0,24*0,28 = 0,03

r_{GPA_nAch} = DI + S1 + S2 + UN1 + UN2 = 0, 45

nAch tedy s GPA koreluje především díky přímému kauzálnímu efektu, ale *r* je nadhodnocována společnými příčinami

Interpretujme náš úsekový model
GPA – nAch – IQ – SES

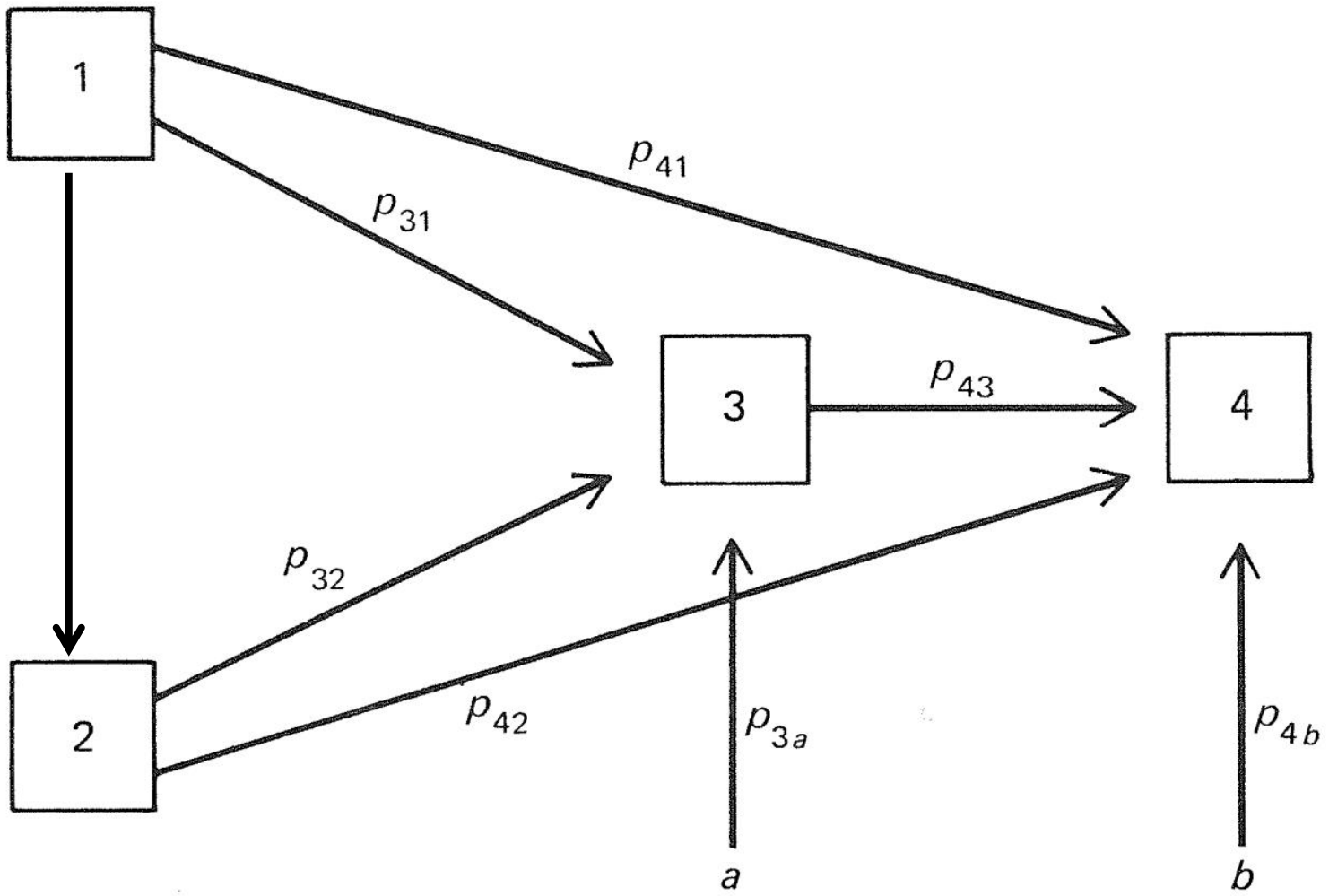
Korelace → **Úsekový model**

r_{GPA_SES}= 0,24

	GPA 4	SES 1	nAch 3	IQ 2
GPA	1,00	0,24	0,45	0,47
SES	0,24	1,00	0,30	0,24
nAch	0,45	0,30	1,00	0,16
IQ	0,47	0,24	0,16	1,00

Test nepřímého efektu

Alternativní model:



Identifikace

- U všech předchozích modelů platilo, že všechny pozorované korelace byly modelem přesně zreplikovány. To není samozřejmost.
- Bylo to tím, že počet odhadovaných parametrů byl přesně roven počtu „vstupních informací“ . počtu korelací, které jsme chtěli dekomponovat
- Takovým modelům se říká „právě identifikované“ **just identified**
- Vstupní informace = počet jedinečných prvků korelační nebo kovarianční matice (podle toho, co analyzujeme)
  - Korelační: k(k-1)/2 např. při 4 proměnných je to 4-3/2 = 6
  - **Kovarianční: k(k+1)/2** (tj. počítáme navíc i rozptyly) … 4*5/2=10
  - Když jsou součástí modelu i průměry, tak ještě přidáme jedničku za každou proměnnou.
- *Počet kousků informace* – počet odhadovaných parametrů = stupně volnosti – **degrees of freedom**
- **Právě identifikovaný model je model, který má df = 0. Takový model, bez ohledu na svou strukturu, přesně replikuje r_{obs}.**

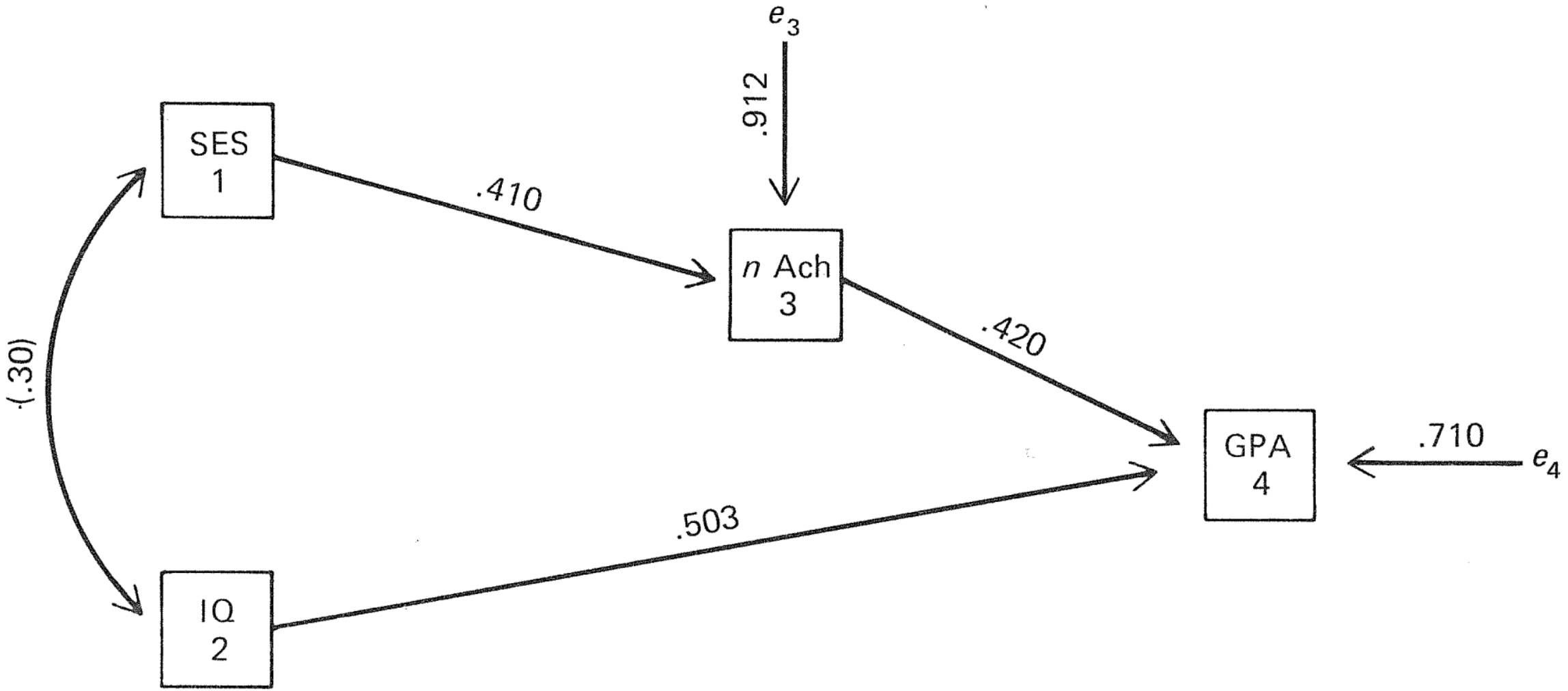
Pod(ne)identifikovaný model - underidentified

- df<0
- Když specifikujeme více parametrů, než pro kolik máme informace
- Neexistuje jedinečné řešení

Overidentified model – nový cíl!

- df > 0
- Existuje jedinečné řešení
- Pozorované korelace nebudou modelem přesně replikovány
- Příležitost pro porovnávání overidentified modelů pod vlajkou PARSIMONIE
- Cílem již není pouze dekomponovat korelace na jejich zdroje
- Nový cíl: **Vysvětlit korelace mezi proměnnými co nejjednodušším modelem**

Dokáže model bez přímého kauzálního efektu SES na GPA a IQ na nAch zachytit podobně dobře jako právě identifikovaný model?



Metrika srovnání modelů

- rozdíly mezi pozorovanou kovarianční maticí a modelem implikovanou kovarianční maticí
- Chí-kvadrát modelu $\chi^2_M = (N-1)F_{ML}$ vyjadřuje shodu modelu s daty
  - Má chí-kvadrát rozložení s tolika stupni volnosti, kolik jich má testovaný model
  - Rozdíl chí-kvadrátů dvou modelů má df = rozdílu df těchto modelů
- Odpovídá náš model datům?
  - Srovnání našeho modelu s právě identifikovaným (jehož df=0)
- Odpovídá model A datům lépe než model B?
- Výše uvedené jsou statistické testy hypotéz se všemi jejich nedostatky

Další ukazatele fitu

- RMSEA – Čím menší, tím lepší. Chceme <0,08. Horní mez 90%intervalu spolehlivosti by neměla přesahovat 0,10. Trestá za komplexitu.
- CFI – Liberální, čím vyšší tím lepší. Chceme >0,95
- SRMR – vychází ze standardizovaných reziduí, čím menší, tím lepší, chceme <0,08

Maticový zápis lineární regrese

- https://onlinecourses.science.psu.edu/stat501/node/382/