Structural Equation Modeling in Stata, R en Mplus

Harrie Jonkman

18 juni 2017

Opzet

- Wat is Structural Equation Modeling
- SEM in STATA
- SEM in R
- SEM in Mplus

Wat is Structural Equation Modeling

- Korte geschiedenis
- Pad diagrammen
- Concepten, jargon en aannames
- Model fit vaststellen
- Het SEM-proces

Korte geschiedenis 1

Factor analyse heeft de wortels in de psychologie:

- Charels Spearman (1904) ontwikkelde een algemeen factormodel. Correlaties tussen verschillende mentale mogelijkheden konden worden verklaard door een algemene factor die mogelijkheden representeert.
- Thurnstone presenteerde multiple factor modellen (hij was tegen een factormodel dat ten grondslag ligt aan alle intelligenties).
- Anderson en Rubin (1956) schreven over factor analyse en J?reskog(1969) introduceerde CFA en het schatten via maximum likelihood estimation, het testen van hypothesen van aantal factoren en hoe ze gerelateerd zijn aan geobserveerde variabelen.

Korte geschiedenis 2

Pad analyses ontwikkelden zich verder ook in genetica, econometrie en later ook sociologie:

- Sewall Wright (genetica).
- Haavelmo en Koopmans (economie).
- Blalock en Duncan (sociologie).

Korte geschiedenis 3

- Nieuwe methodes ontwikkelden zich verder voor nieuwe modellen, om de identificatie te evalueren, om model fit vast te stellen ed.
- Ook programma's (LISREL, AMOS, EQS maar ook in STATA, R en Mplus)

SEM is . . .

- . . . een groep statistische technieken die ons in staat stelt om hypotheses op te stellen en te toetsen over de relatie tussen variabelen.
- . . . een analyse van covariantie structuren. Het maakt modellen op basis van geobserveerde covarianties en, zo mogelijk, gemiddelden.
- . . . een methode die ook andere methodes omvat zoals correlatie, liniaire regressie en factor analyse.
- . . . een multivariate techniek die ons instaat stelt een systeem van vergelijkingen te schatten. Variabelen hierbinnen mogen met fouten zijn gemeten. En er kunnen variabelen in zitten die niet direct gemeten kunnen worden.

SEM is ook raamwerk

SEM is een wetenschappelijk raamwerk voor het opbouwen en evalueren van hypothesen over oorzaak-effect verbanden in systemen.

- Daar heb je statistische en mathematische gereedschappen voor.
- Binnen SEM wordt de methode van causale analyse gebruikt.
- Om oorzaken in het netwerk en de systemen zelf beter te begrijpen.

Het is netwerkanalyse

Het gaat er om abstracte systemen als causale en probabilistische netwerken en om op een effici?nte en effectieve manier daarbinne relaties te begrijpen.

Maar het is vooral ook een vorm van grafisch modelleren

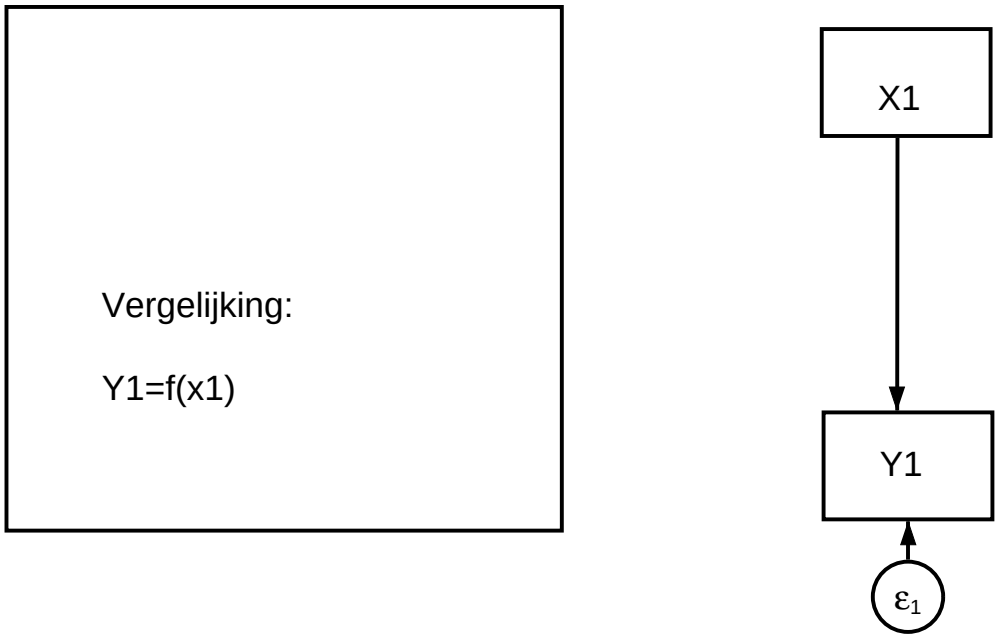
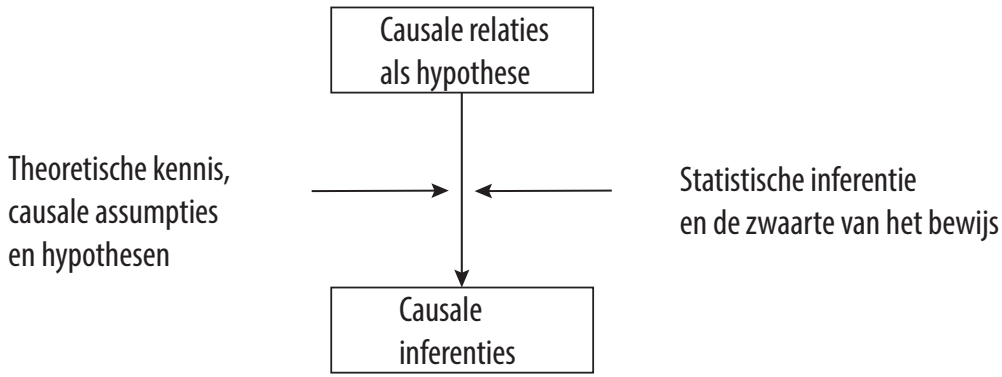


Figure 1

Het is een poging om causale effecten te kunnen schatten



"Geen oorzaken erin, geen causale schattingen eruit."

Figure 2

Structureel model en Meet Model

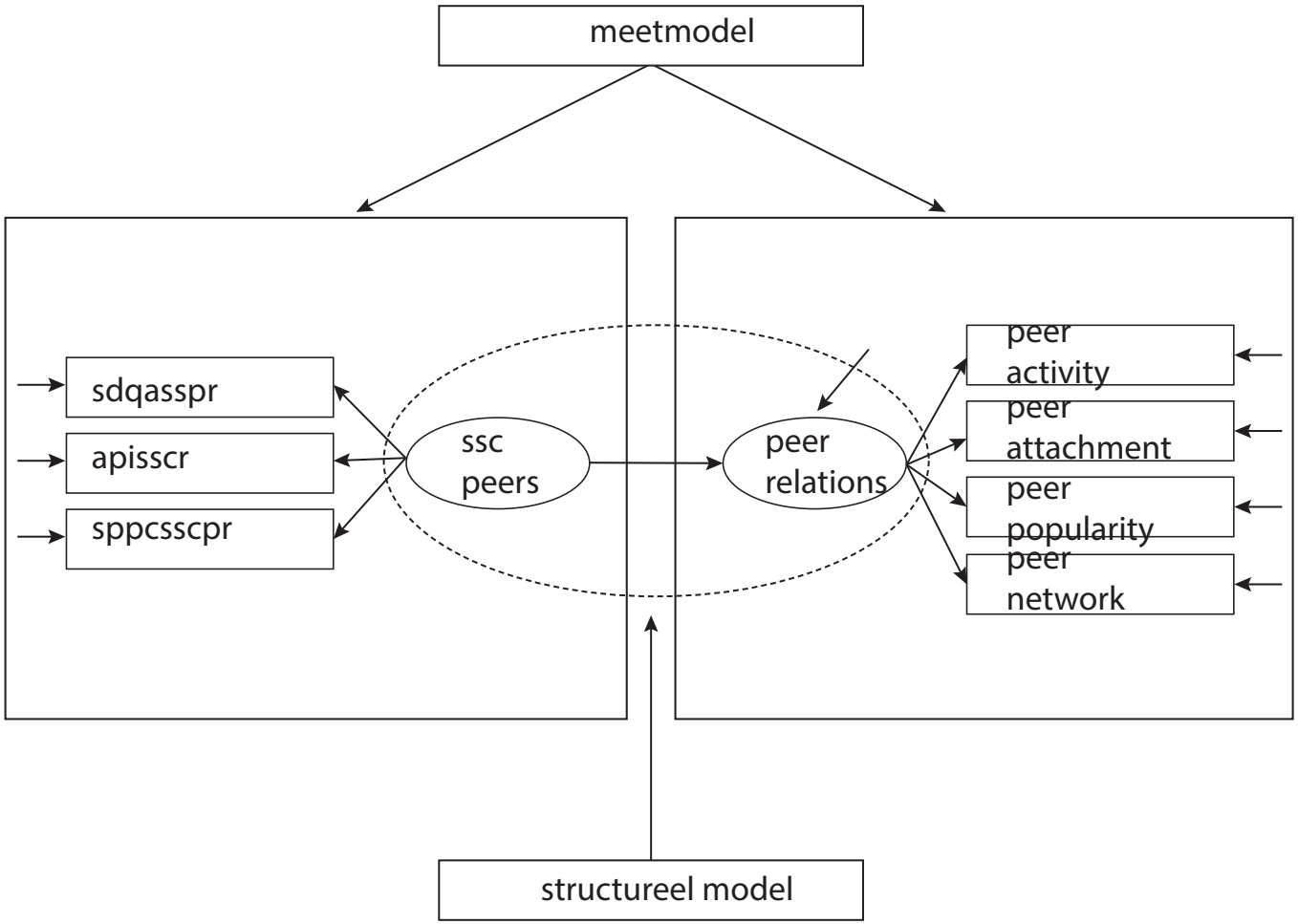


Figure 3

“Als ik hier druk, voel ik het daar”

Het concept van causale relatie is dat *A* een oorzaak is van *B*
als verandering van *A* leidt tot een reactie in *B*

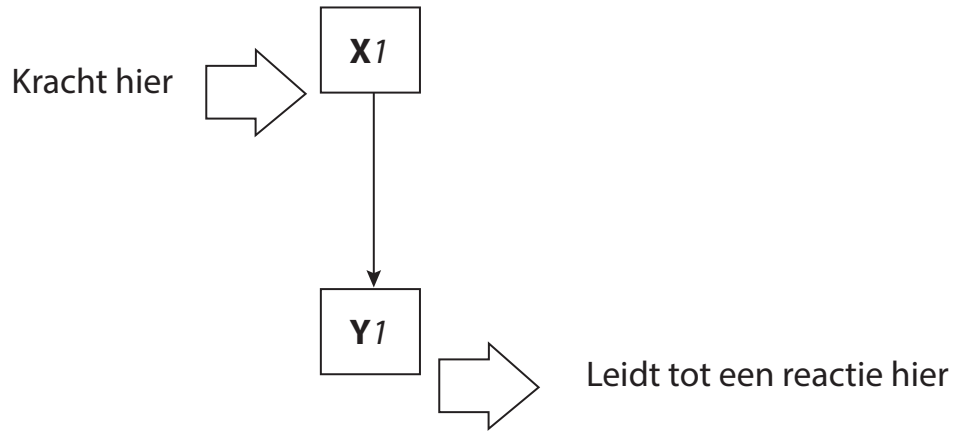


Figure 4

Vooral dus om causale hypothesen te onderzoeken

"SEM results should not be taken as a proof of causal claims, but instead as evaluation or tests of models representing causal hypotheses" (J. Grace)

Het is om een netwerk en processen te onderzoeken

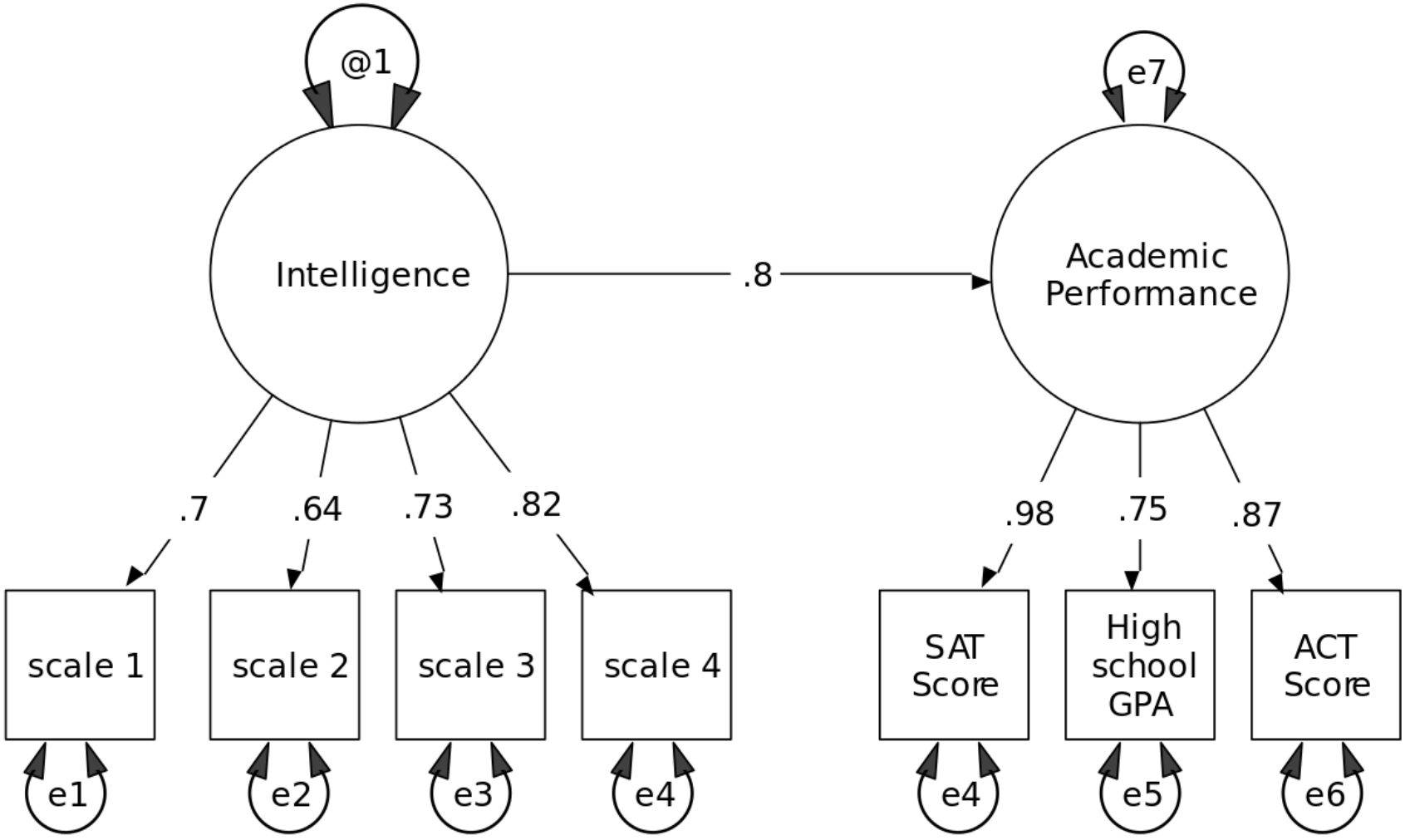


Figure 5

Dat kan best ingewikkeld zijn

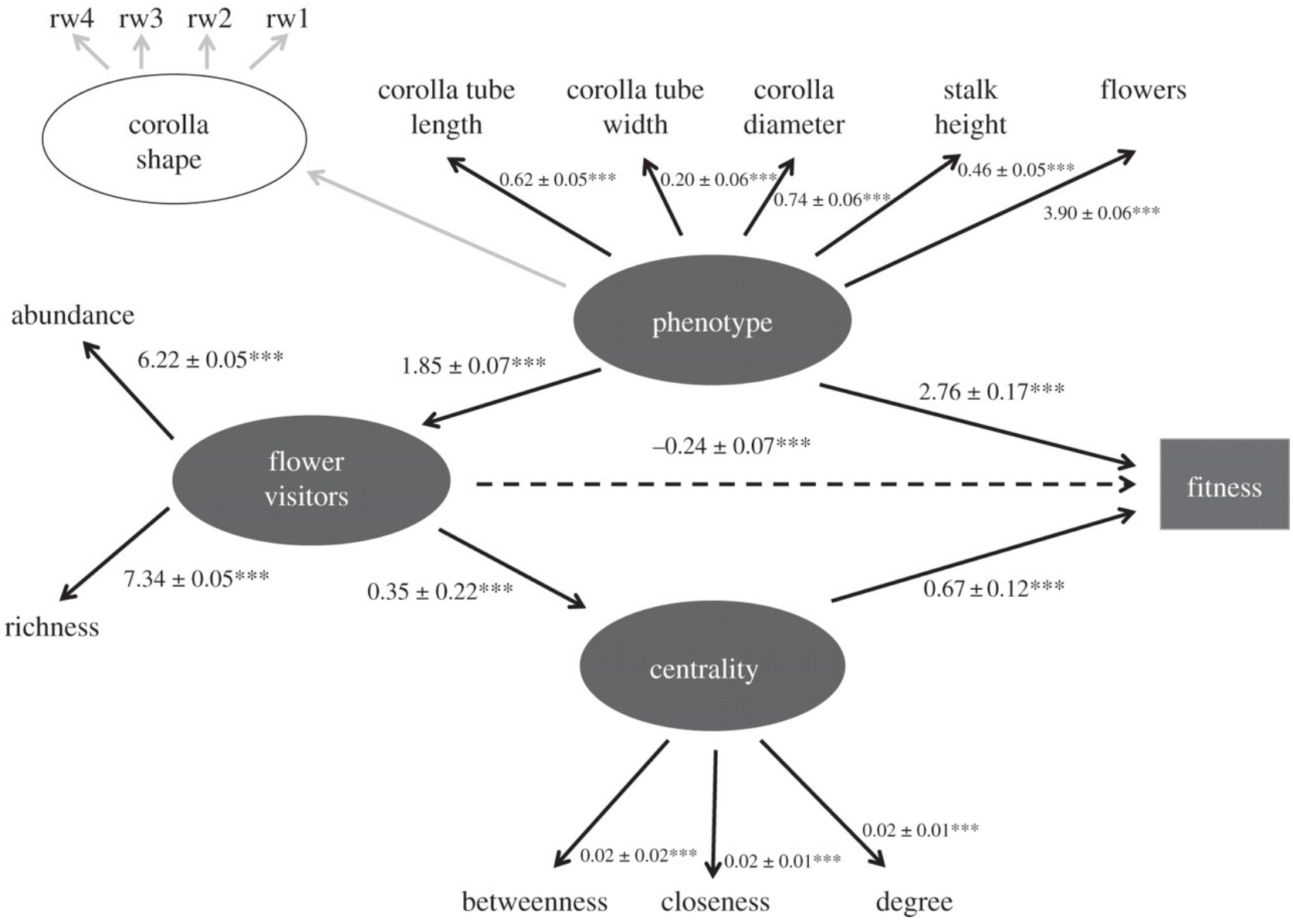
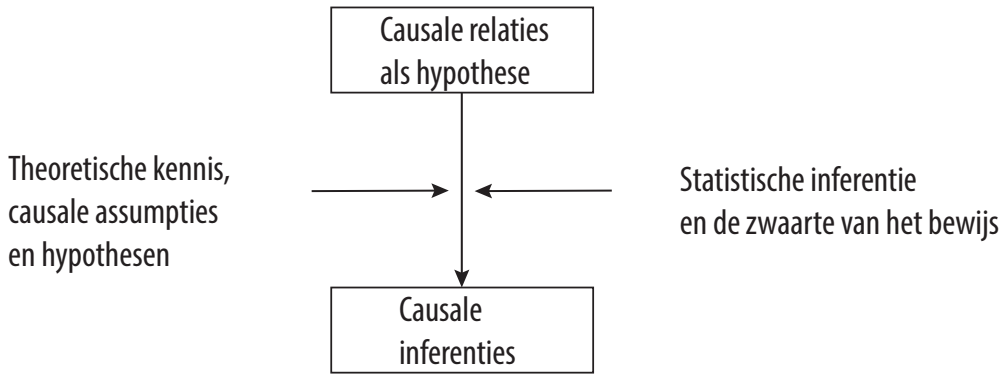


Figure 6

Kennis en methode in samenhang



"Geen oorzaken erin, geen causale schattingen eruit."

Figure 7

Sem als denkproces

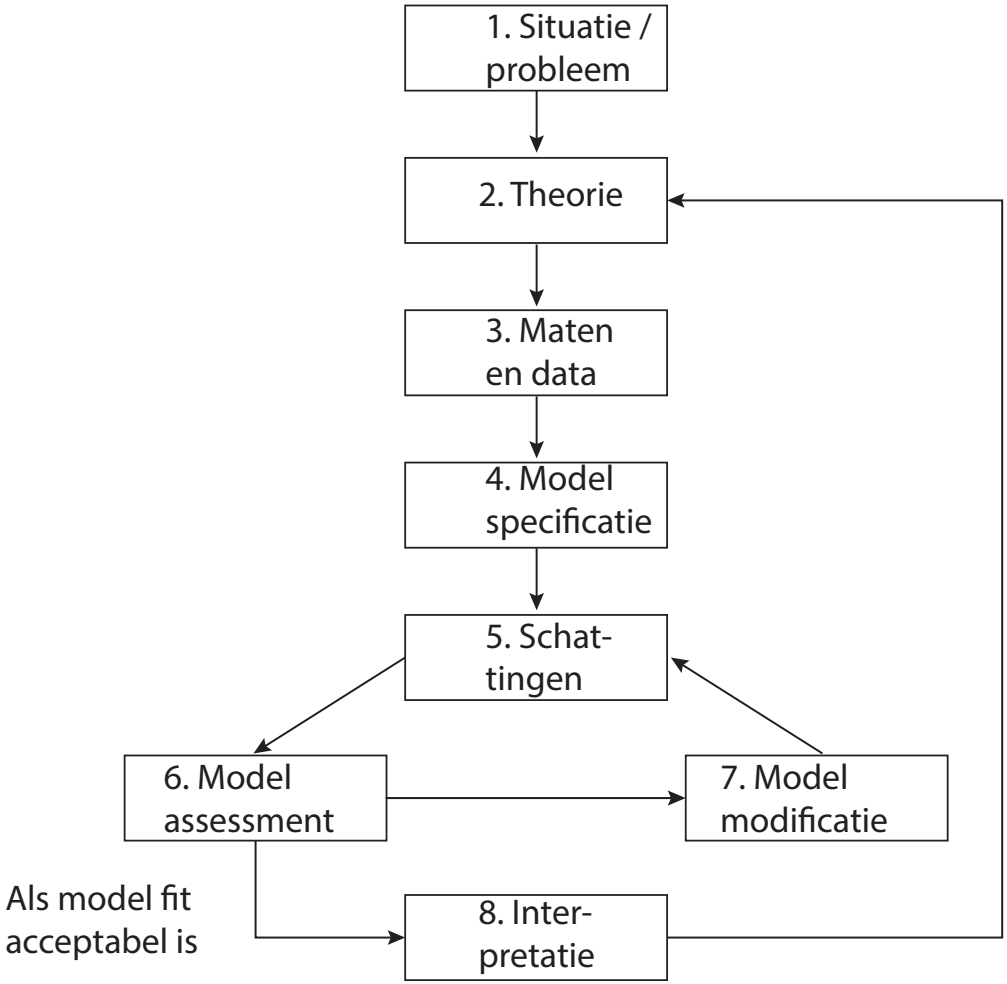


Figure 8

Sem modellen vaak getekend als Paddiagrammen

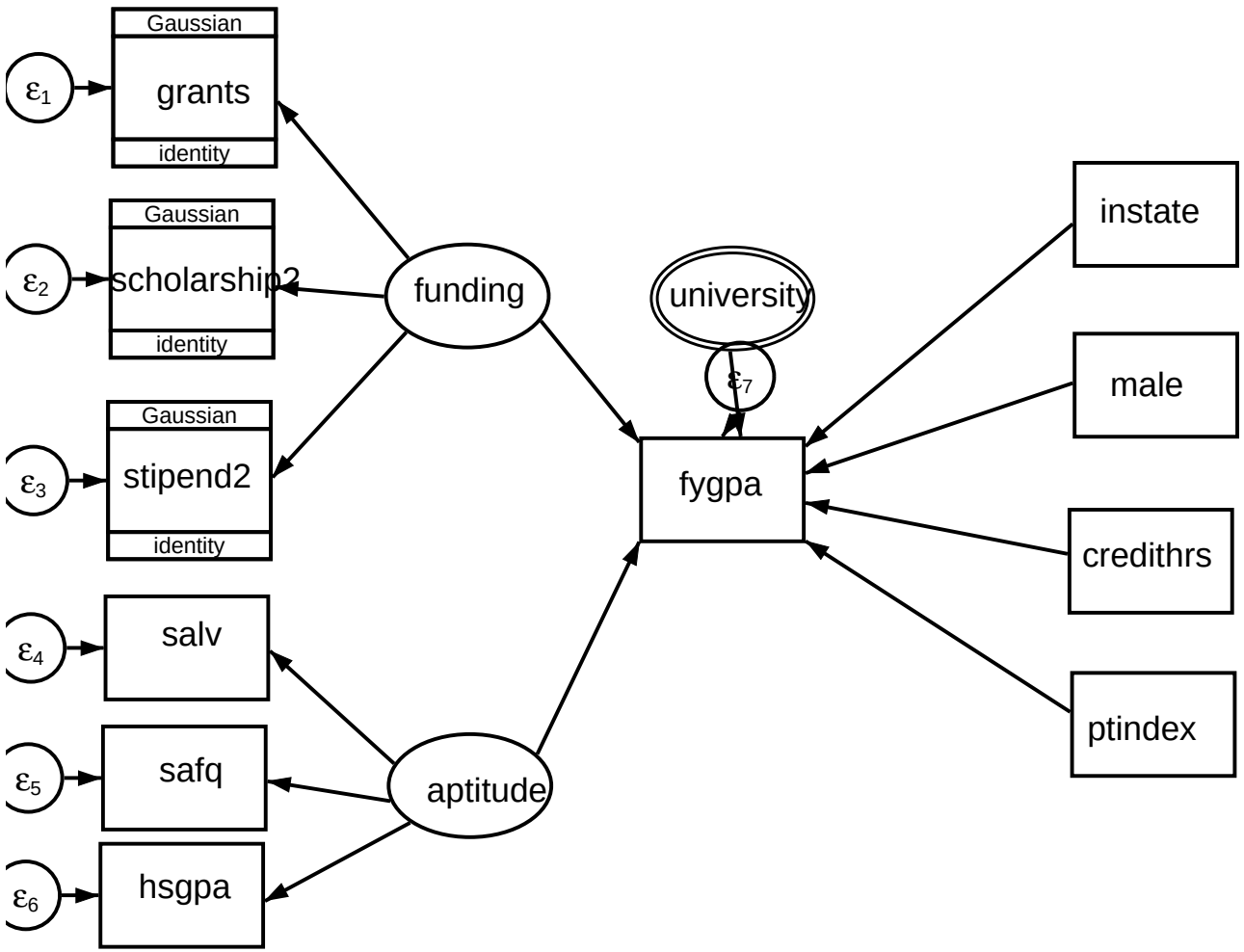


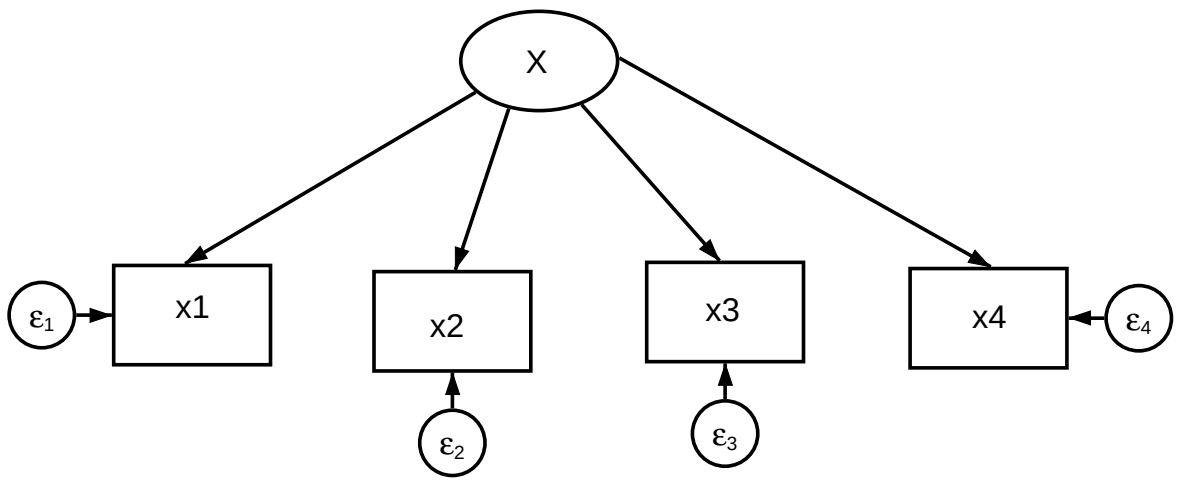
Figure 9

Jargon

- Geobserveerde en latente variabelen
- Paden en covarianties
- Endogene en exogene variabelen
- Recursieve en niet-recursieve modellen

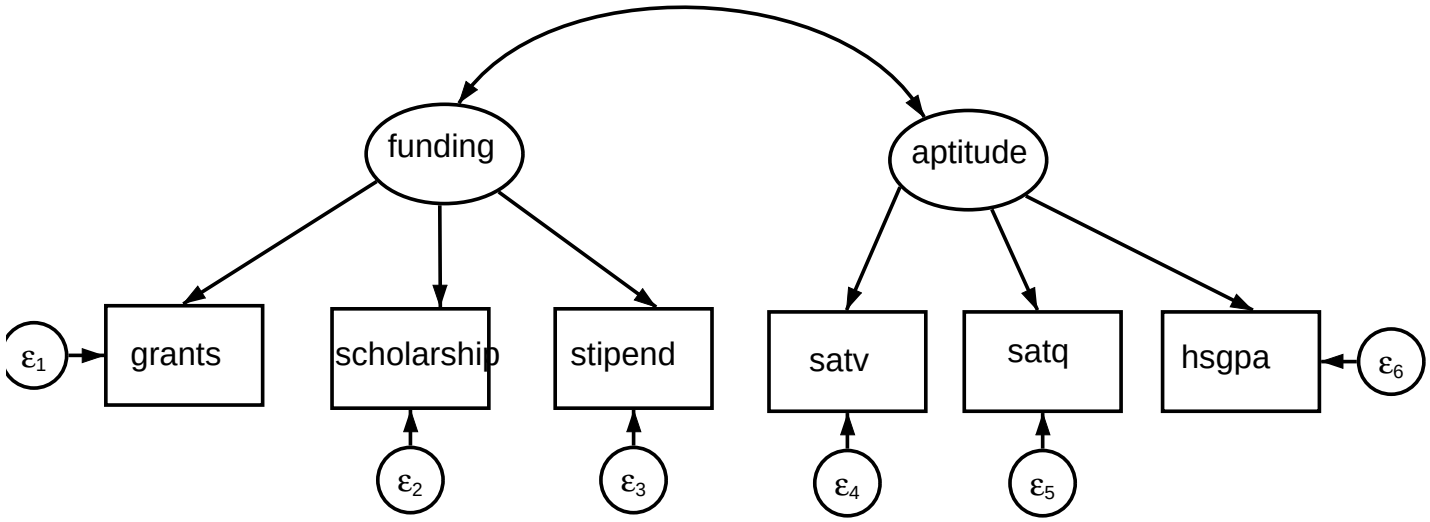
Geobserveerde en latente variabelen

- **Geobserveerde variabelen** zijn variabelen die in onze dataset zitten. Ze worden met rechthoeken uitgebeeld. Hier gaat het om x1, x2, x3 en x4.
- **Latente variabelen** zijn ongeobserveerde variabelen die we hadden willen observeren. Ze worden samengesteld door andere variabelen en als ovaal afgebeeld, zo hier de X.



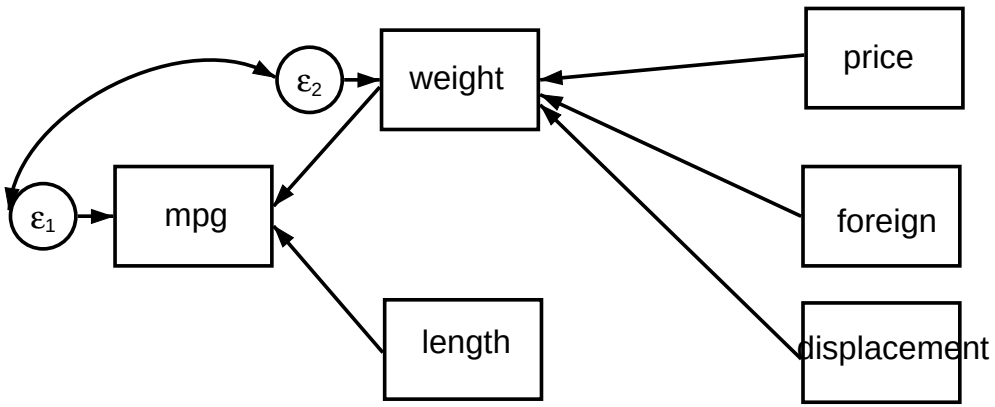
Paden en covarianties

- **Paden** zijn directe relaties tussen variabelen. Geschatte padcoeffici?nten zijn vergelijkbaar met regressiecoeffici?nten. Ze worden door een rechte pijl gerepresenteerd.
- **Covariantie** specificeert dat twee latente variabelen of fouten termen covarieren. Ze worden door een kromme wederzijdse pijl gerepresenteerd.



Exogene en endogene variabelen

- **Exogene** variabelen zijn gedetermineerd buiten het systeem van vergelijkingen. Er gaan geen paden (pijlen) naar toe. Hieronder *price*, *foreign* en *displacement*.
- **Endogene** variabelen zijn gedetermineerd binnen het vergelijkingssysteem. Er gaat in ieder geval een pad (pijl) naar toe. Hieronder zijn *weight* en *mpg* endogeen.

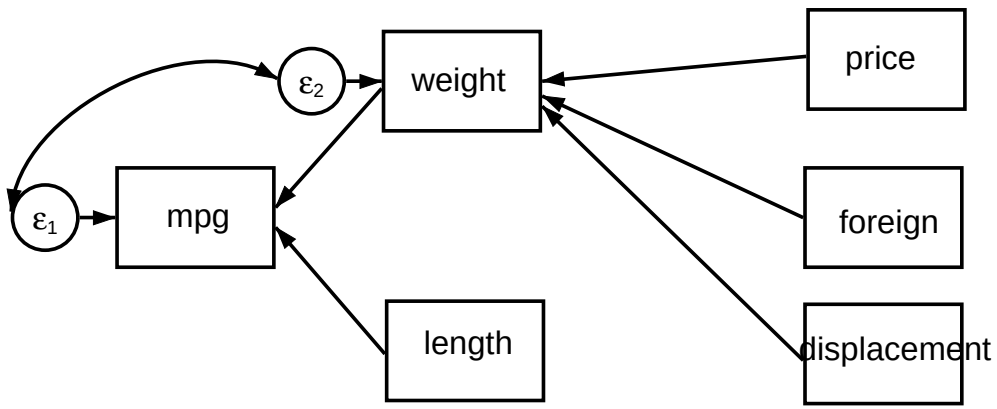


Nogmaals

- **Geobserveerd exogeen** is een variabele in een dataset die als exogeen is behandeld in het model.
- **Latent exogeen** is een ongeobserveerde variabele die als exogeen is behandeld in het model.
- **Geobserveerde endogeen** is een variabele in een dataset die als endogeen is behandeld in het model.
- **Latent endogeen** is een ongeobserveerde variabele die als endogeen is behandeld in het model.

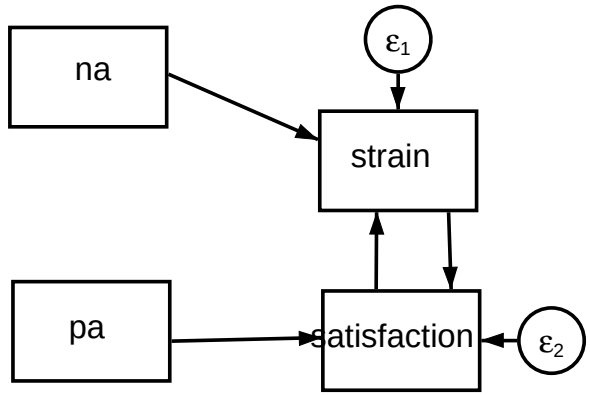
Recursieve modellen

- **Recursieve modellen** hebben geen feedback loops (wederkeringe richtingen).



Nonrecursieve modellen

- **Non-recursieve modellen** hebben wel feedback loops. In deze modellen zijn er paden in beide richtingen tussen een of meer paren van endogene variabelen.



Model Goodness of Fit vaststellen

- Likelihood Ratio Chi-Squared Test
- Akaike's Information Criterion (AIC)
- Swartz's Bayesian Information Criterium (BIC)
- Verklaarde variantie (R^2)
- Root Mean Squared Error of Approximation (RMSEA)
- Comparative Fit Index (CFI)
- Tucker-Lewis Index (TLI)
- Standarized Root Mean Square Residual (SRMR)

Model Goodness of Fit vaststellen 1

Likelihood Ratio Chi-Squared Test
Goede fit aangegeven door p-value$>$0.05

AIC en BIC
Goede fit aangegeven door - twee modellen te vergelijken - kleiner (in absolute betekenis) is beter

R^2 (Verklaarde variantie)
Goede fit aangegeven door waarde die dichter bij 1 ligt.

Model Goodness of Fit Vervolg

RMSEA

Goede fit aangegeven door RMSEA < 0.06 ook wel <0.05 (goede fit) tussen 0.05 en 0.08 (adequaat) >0.1 (slechte fit)

CFI

Goede fit aangegeven door CFI > 0.95 (soms > 0.90)

TLI

Goede fit aangegeven door TLI >0.95

SRMS

Goed fit aangegeven door SRMS < 0.08

SEM in STATA

- SEM in STATA
- Modellen voor continue uitkomsten met *sem*
- Modellen met andere uitkomsten met *gsem*
- Enkele voorbeelden

Sem in STATA

- STATA
- De sembuilder
- De **sem**syntax
- De **gsem**syntax
- Verschillen tussen **sem** and **gesem**

De interface van STATA

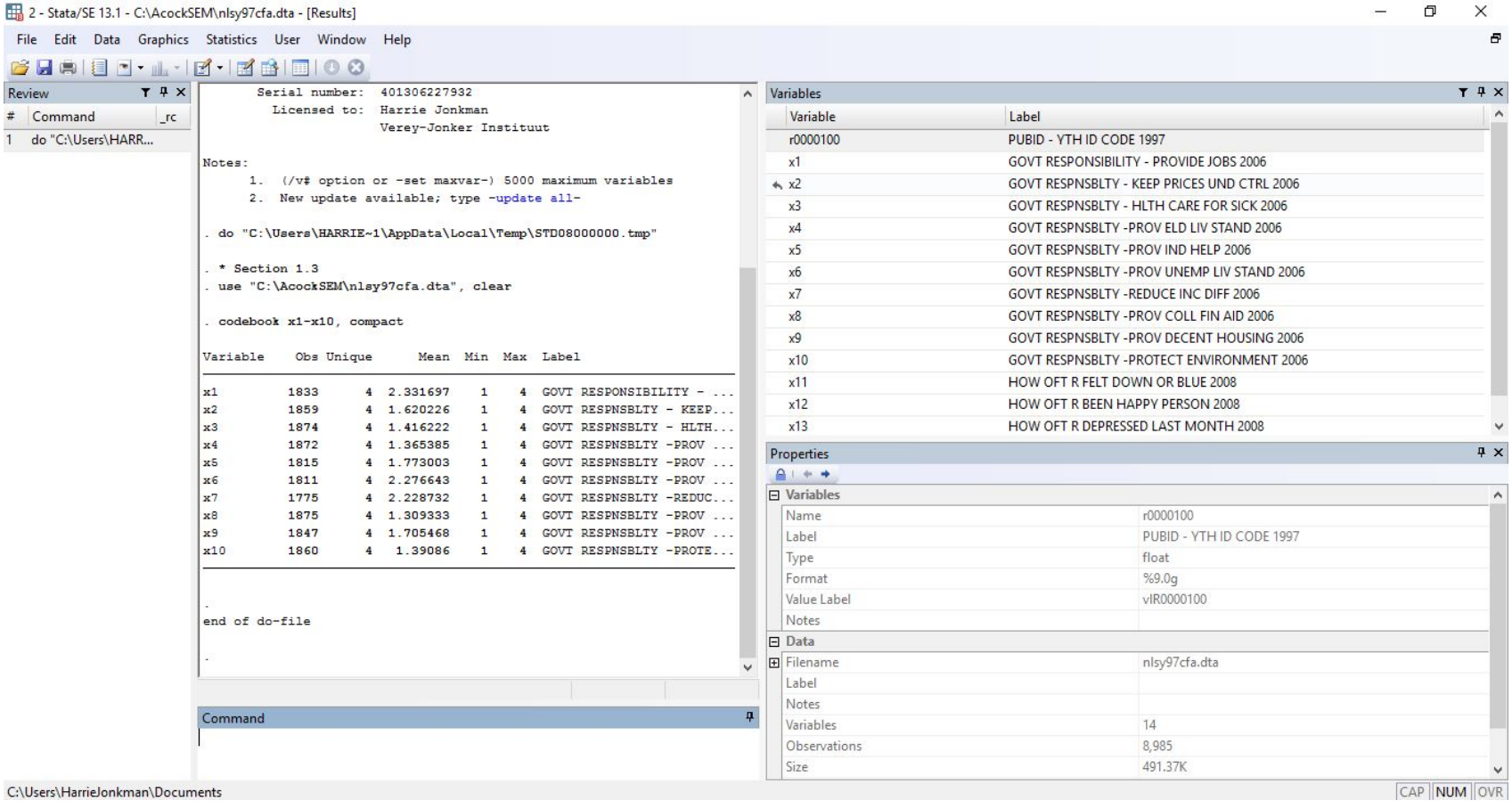


Figure 10

De menu's en dialoog boxen

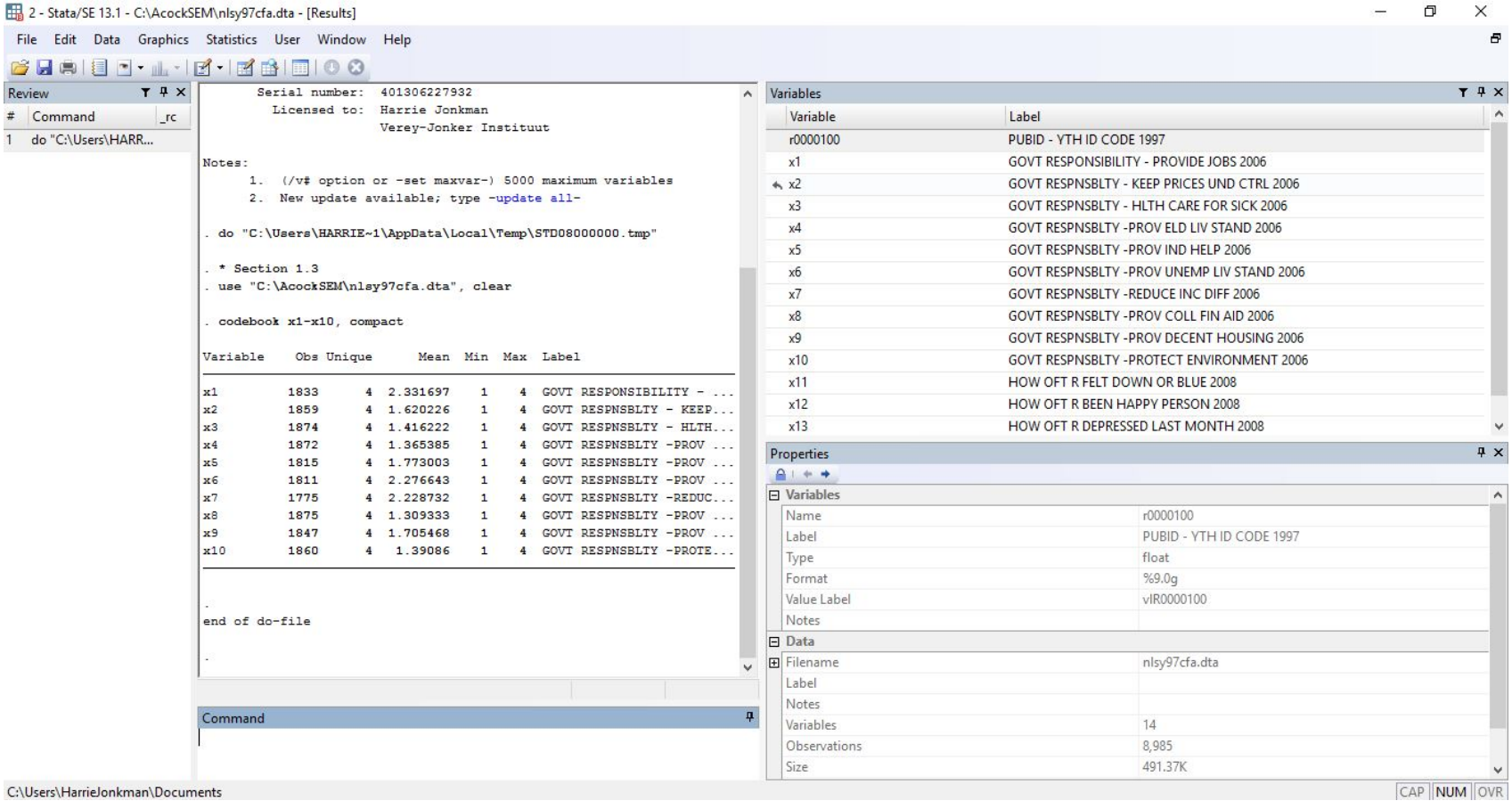


Figure 11

De data editor

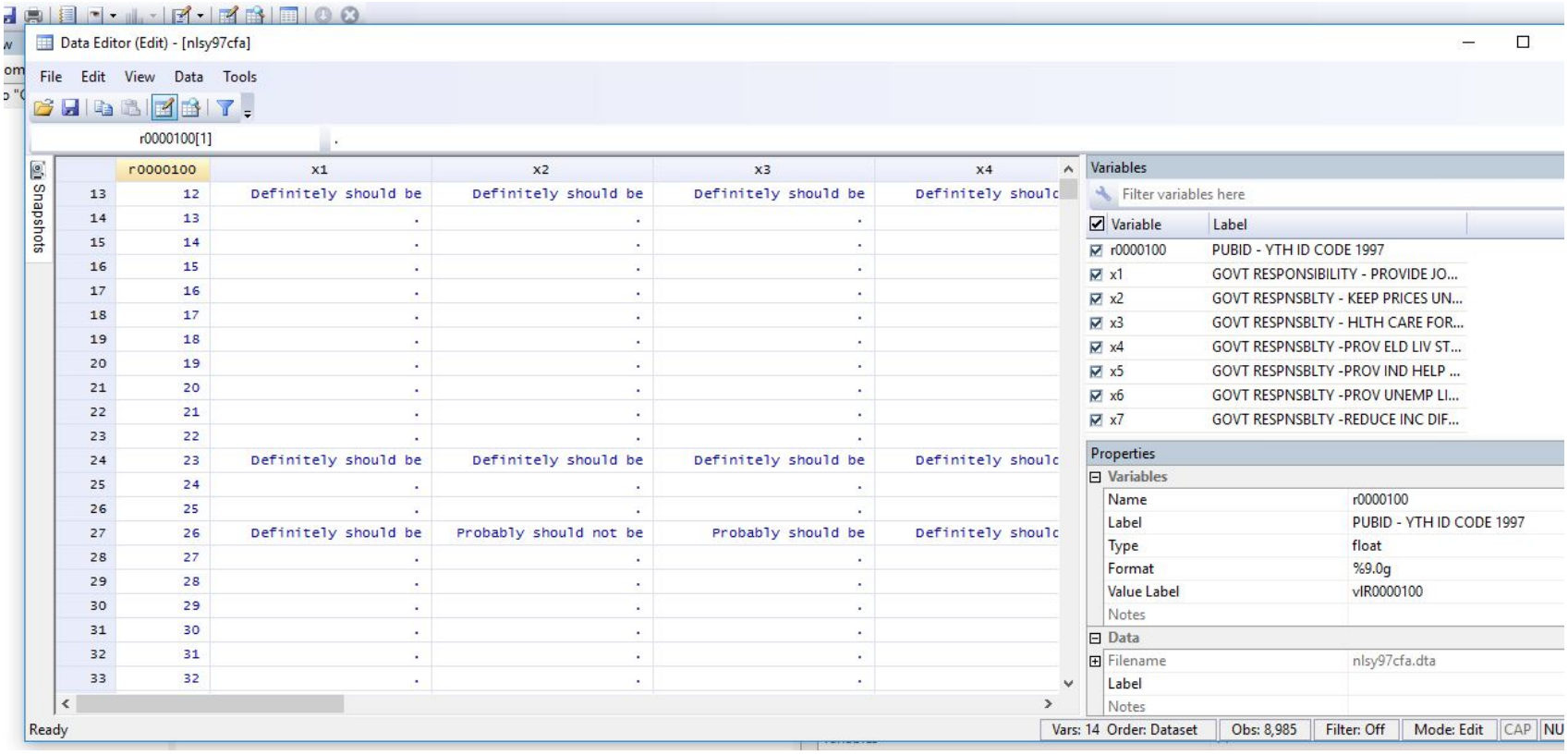


Figure 12

De do-file editor

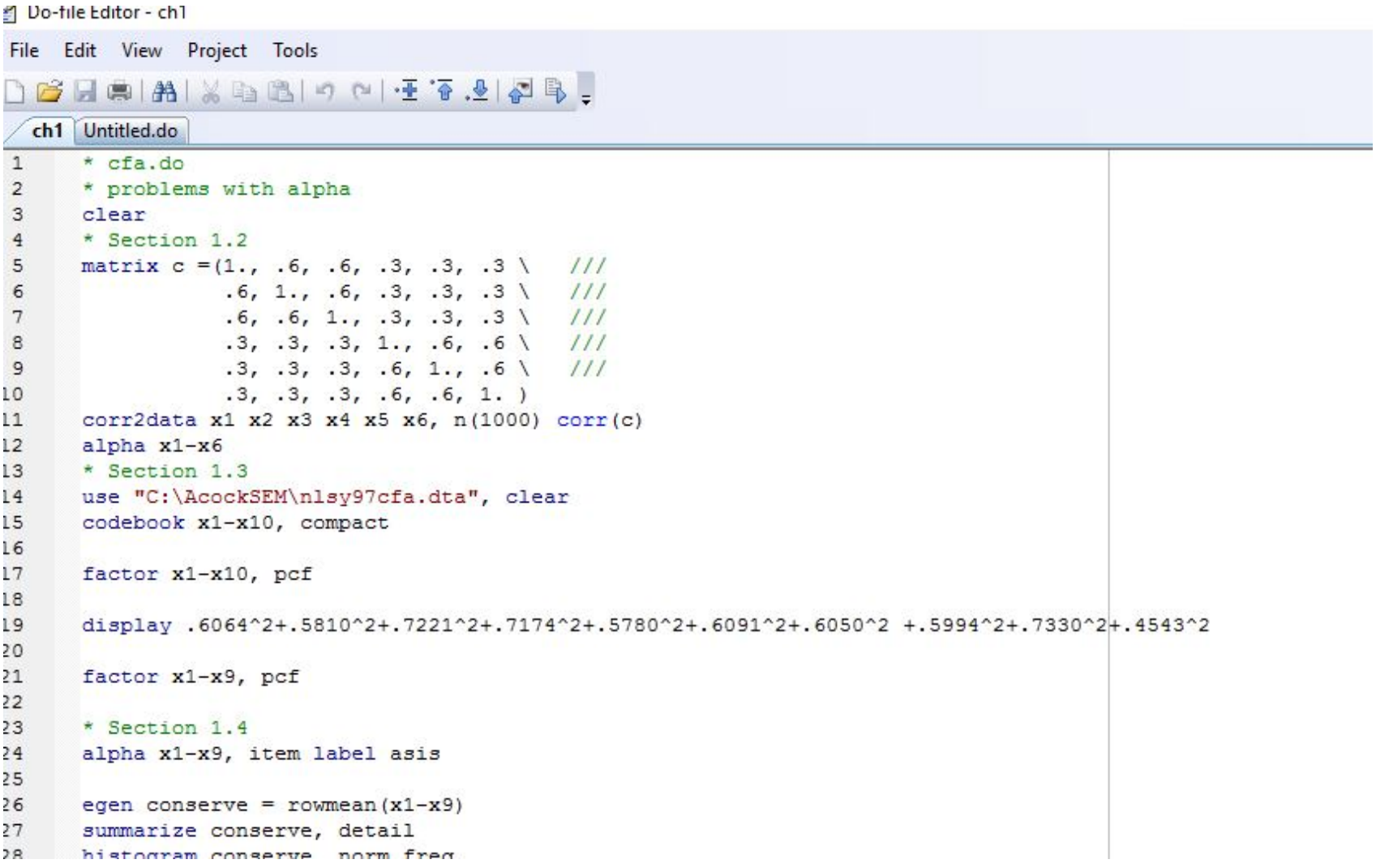


Figure 13

CFA in Stata: de variabelen

x1	GOVT RESPONSIBILITY - PROVIDE JOBS 2006
x2	GOVT RESPNSBLTY - KEEP PRICES UND CTRL 2006
x3	GOVT RESPNSBLTY - HLTH CARE FOR SICK 2006
x4	GOVT RESPNSBLTY -PROV ELD LIV STAND 2006
x5	GOVT RESPNSBLTY -PROV IND HELP 2006
x6	GOVT RESPNSBLTY -PROV UNEMP LIV STAND 2006
x7	GOVT RESPNSBLTY -REDUCE INC DIFF 2006
x8	GOVT RESPNSBLTY -PROV COLL FIN AID 2006
x9	GOVT RESPNSBLTY -PROV DECENT HOUSING 2006

Figure 14

Stata commando

sem(Conservative->x1-x9) standarized estat gof, stats(all)

sem x1-x9 estat framework, fitted

estimates restore hold estat mindices codebook x3 x4, compact

sem (Conservative -> x1 x3-x7 x9), covariance(e.x3*e.x4) sem, standardized estat gof, stats(all) estat mindices sem (Conservative->x1 x3-x7 x9), covariance(e.x3*e.x4) variance(Conservative@1)

De sembuilder

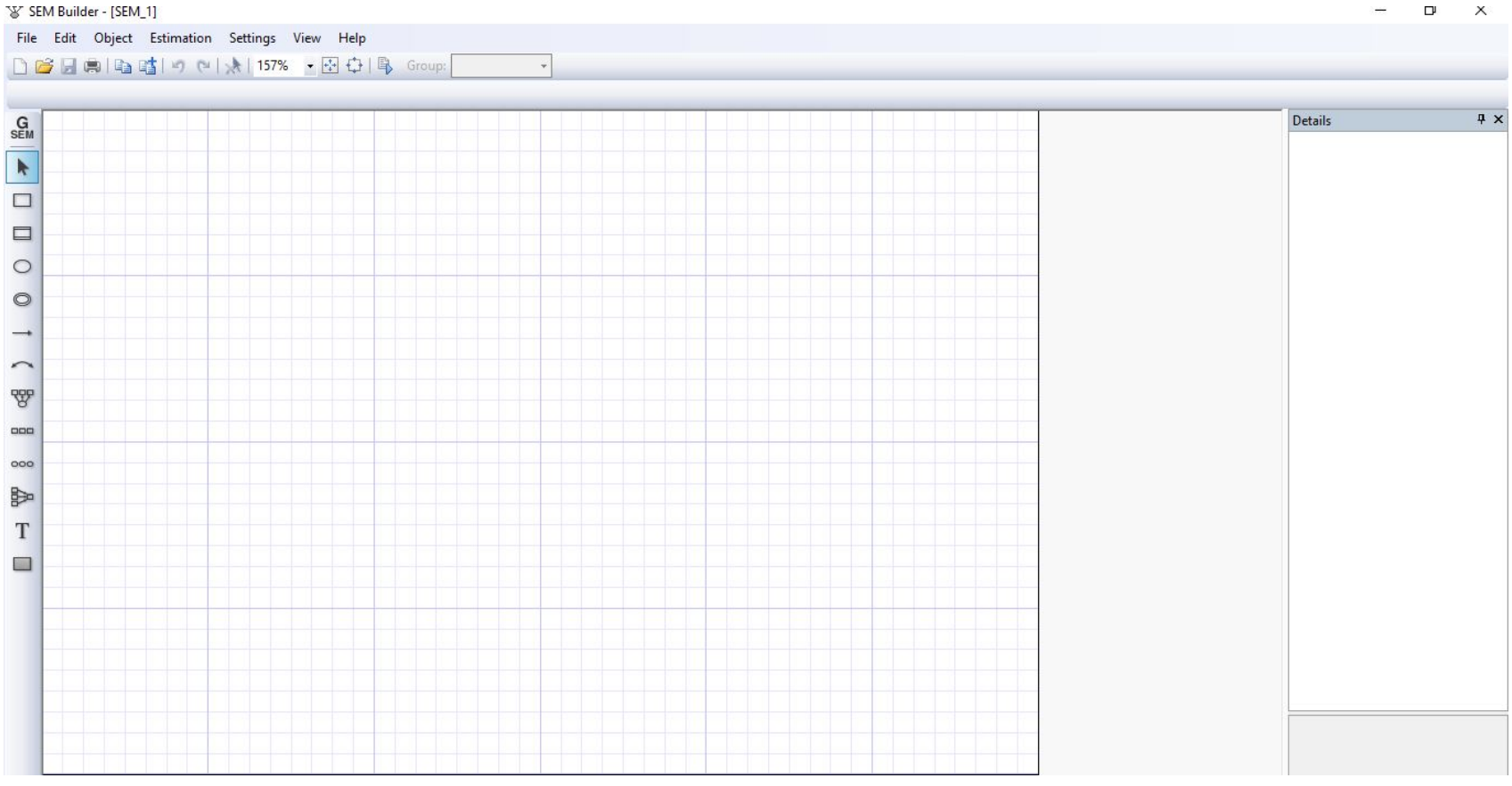


Figure 15

Two factor model

sem (Depress-> x11-x13) sem (Conservative -> x1-x9) sem (Conservative -> x1 x3-x7 x9), /// covariance(e.x3*e.x4)

sem (Depress-> x11-x13) /// (Conservative -> x1 x3-x7 x9), /// covariance(e.x3*e.x4)

sem, standardized estat gof, stats(all) estat mindices

Insert in sembuilder

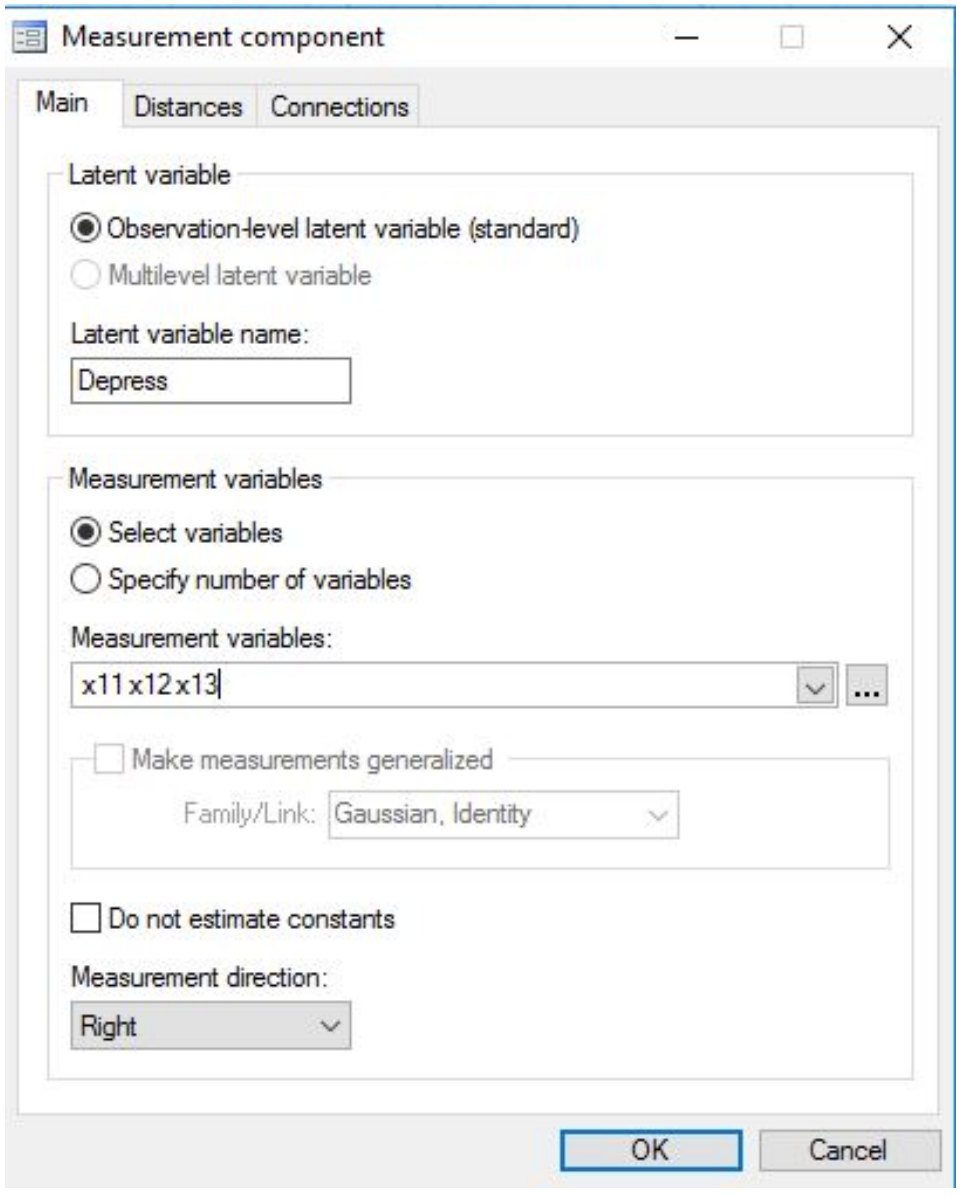


Figure 16

Two factor model

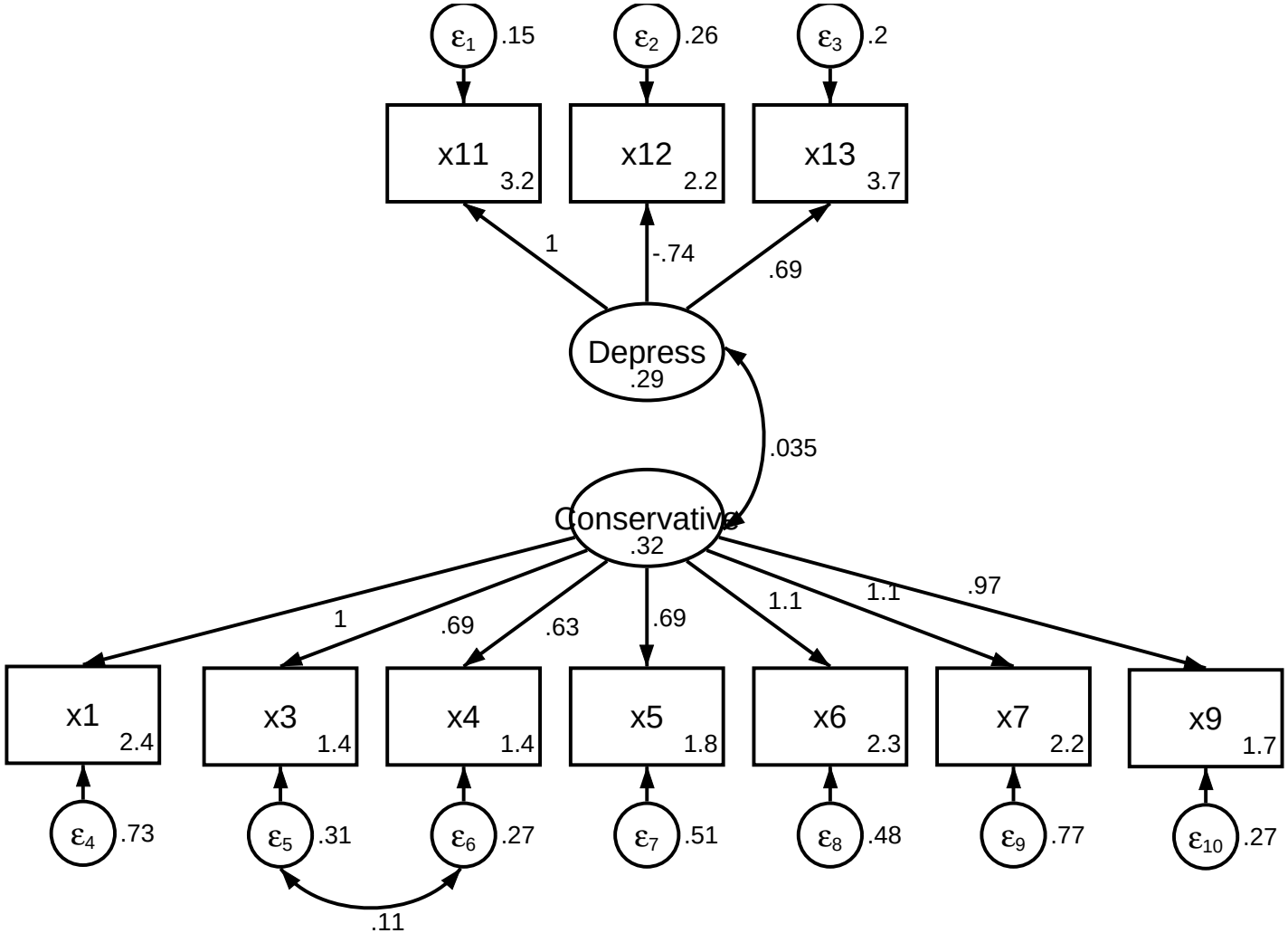


Figure 17

Change variablesettings

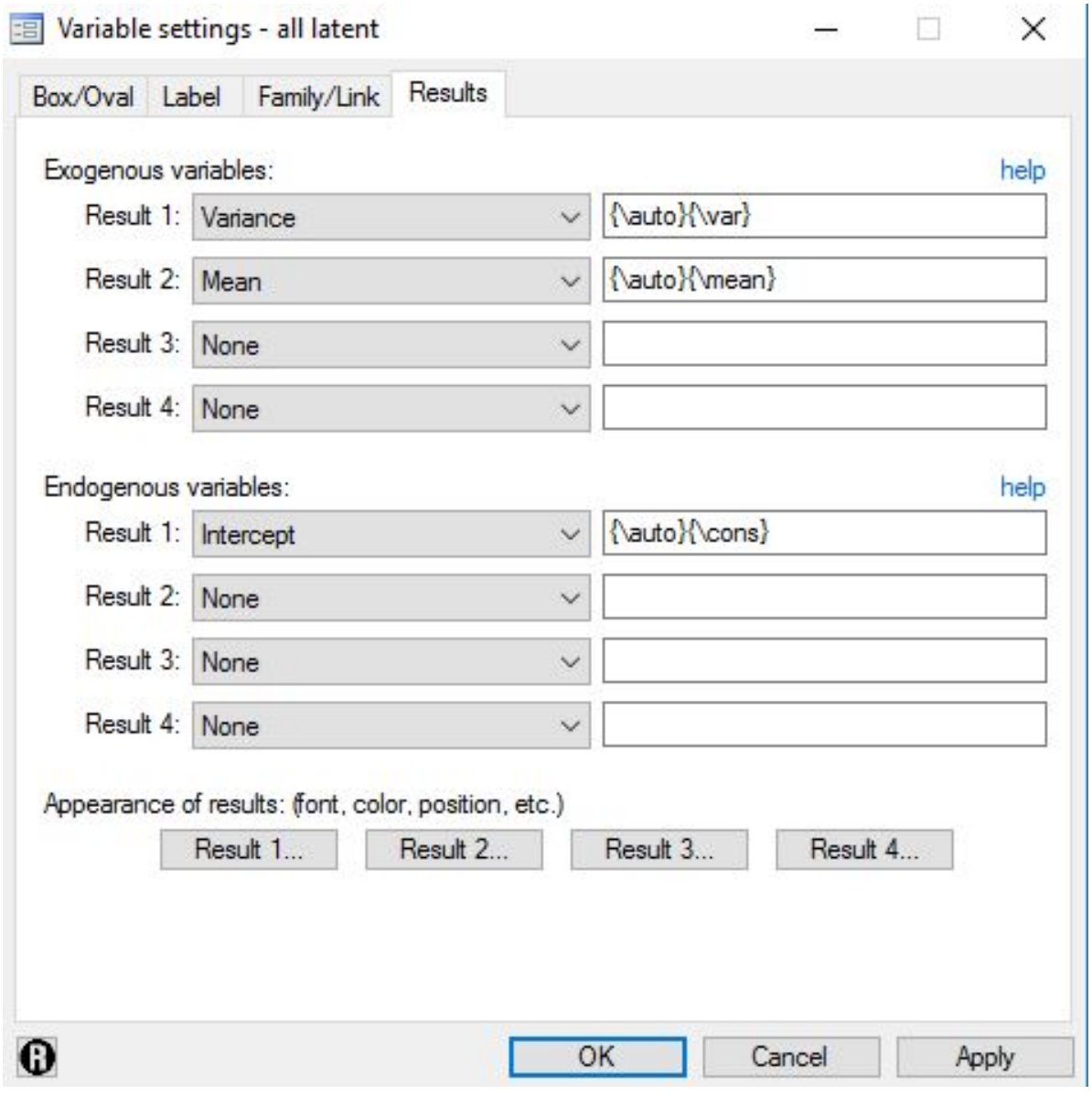


Figure 18

Two factor model again

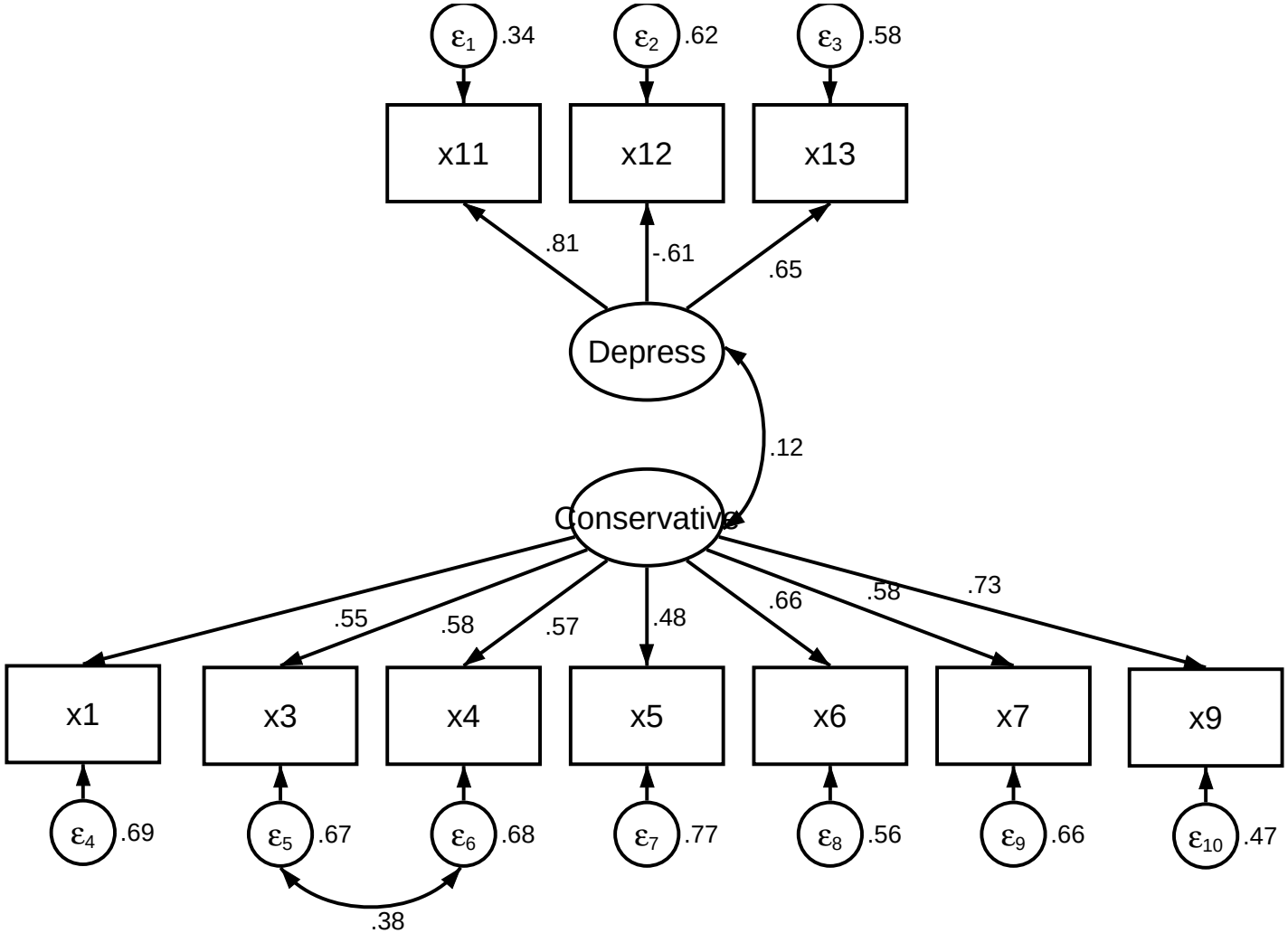


Figure 19

Tekst erbij

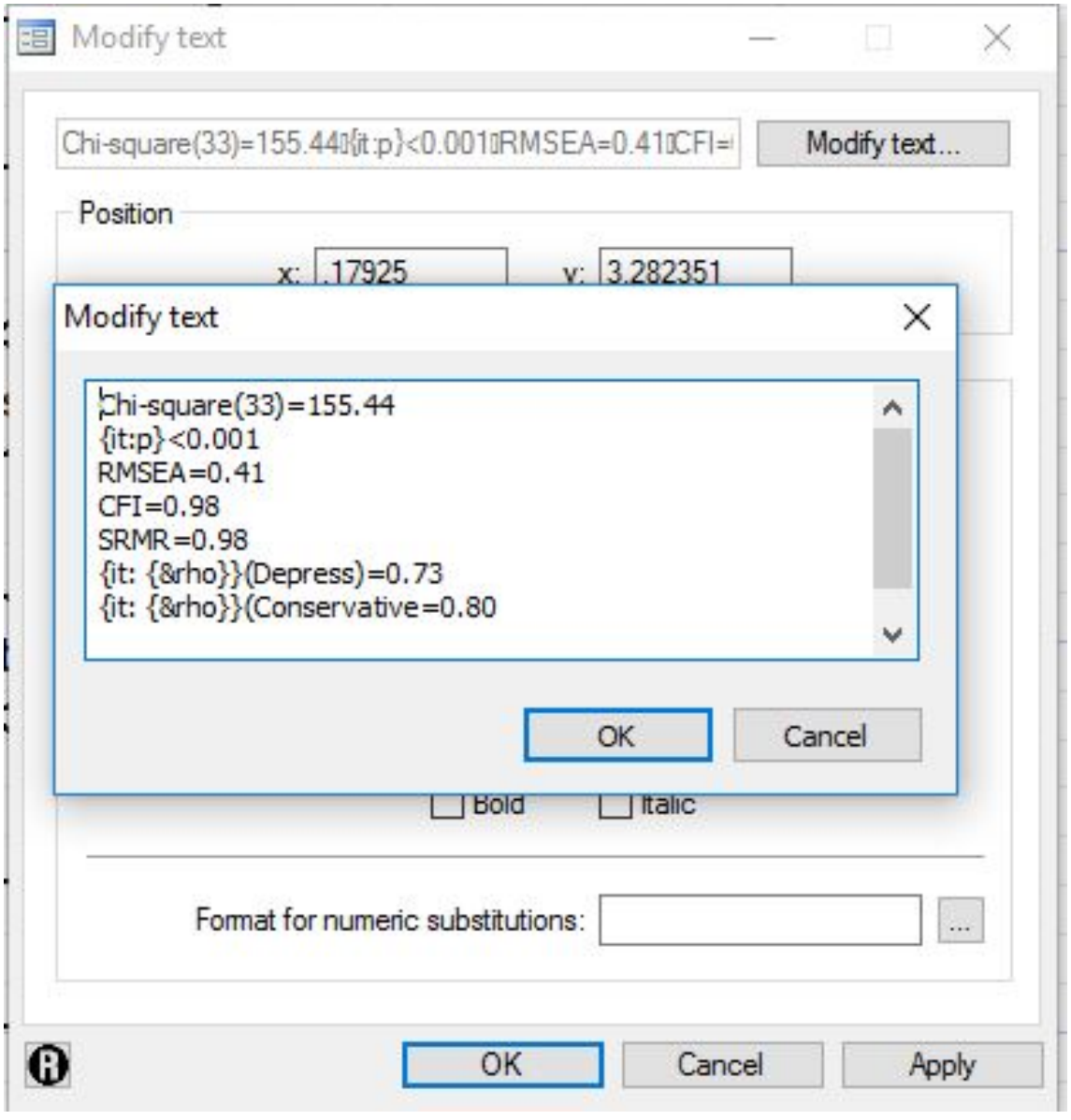


Figure 20

Two factor final

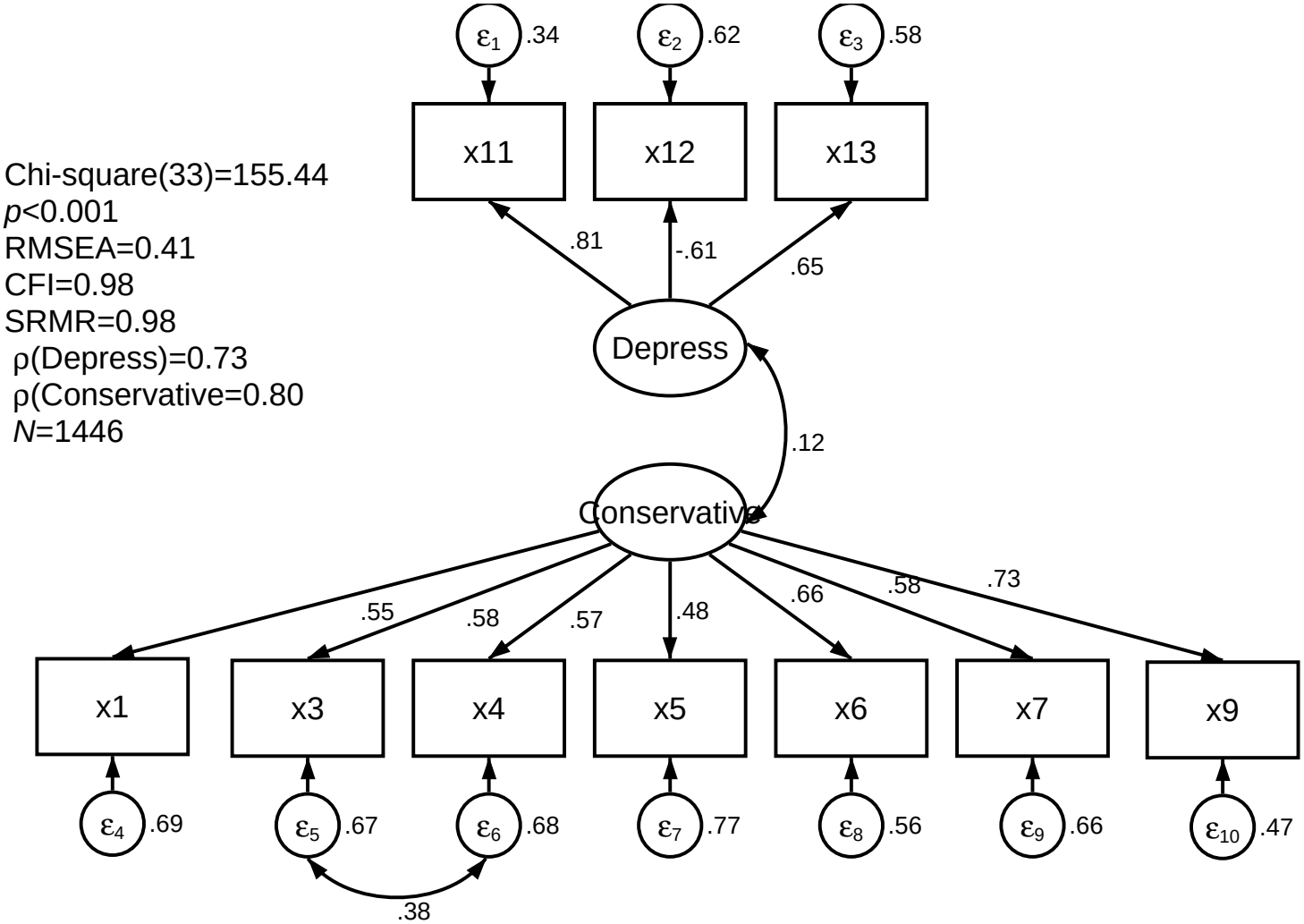
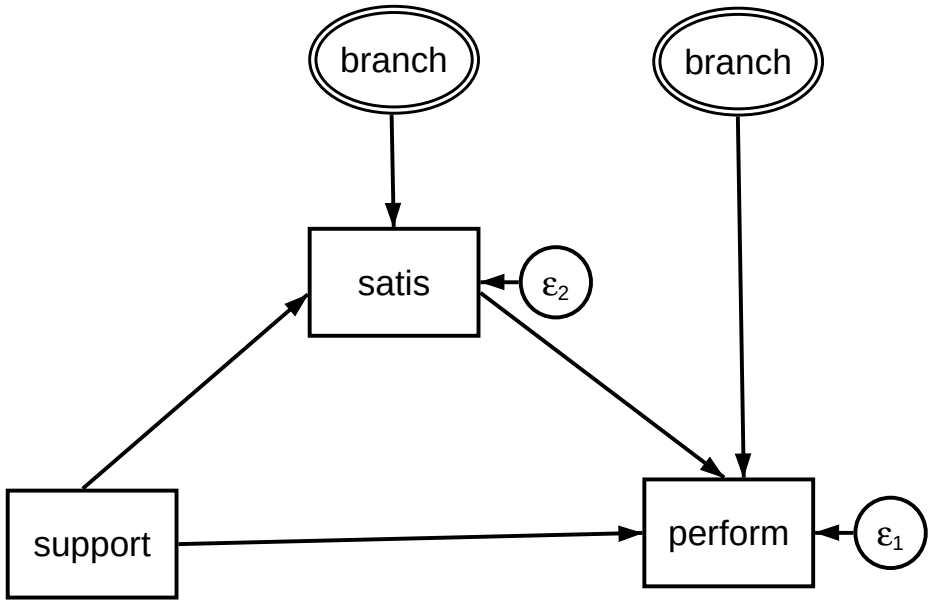


Figure 21

gsem

For multilevel models we have to use gsem



gesem commando

gsem (perform <- satis support M1[branch]) (satis <-support M2[branch]), cov(M1[branch]*M2[branch]@0)

Indirecte effecten

```
nlcom _b[perform:satis]*_b[satis:support]

      _nl_1:  _b[perform:satis]*_b[satis:support]
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.1627062	.0141382	11.51	0.000	.1349958	.1904165

totale effecten

```
nlcom _b[perform: support] +_b[perform:satis]*_b[satis:support]

  _nl_1:  _b[perform: support] +_b[perform:satis]*_b[satis
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	.8608587	.0257501	33.43	0.000	.8103894	.911328

SEM in R

Voordeel hiervan is dat het programma gratis is: Lavaan is dan het pakket

`http://lavaan.ugent.be/`

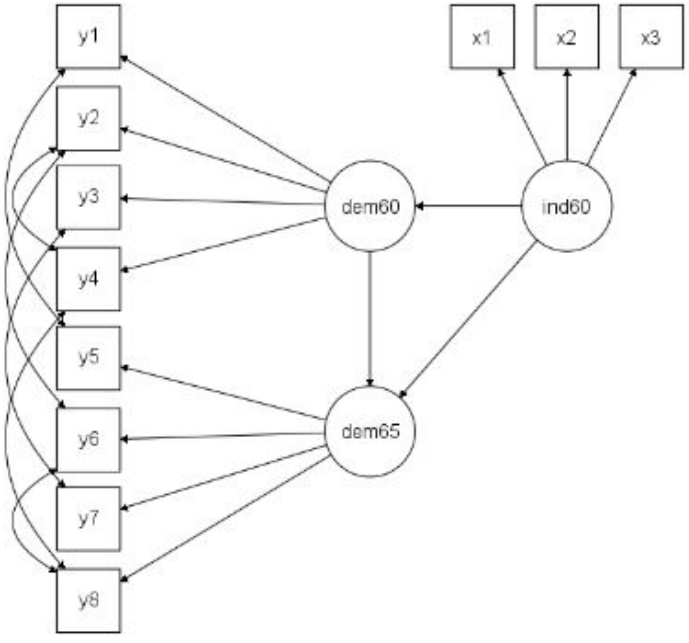
Figure 22

Wat zeggen de data

Data hieronder komen van Bollen (1989). Het zijn data van 75 ontwikkelingslanden rond latente variabelen democratie (1960 en 1965) en mate van industrialisatie (1960).

- Latente variabele *democratie (1960)* gebaseerd op: y1=vrijheid van pers; y2=vrijheid van politieke oppositie; y3=eerlijkheid van verkiezingen; y4=effectiviteit van gekozen wetgeving.
- Latente variabele *democratie (1965)*: y5=vrijheid van pers; y6=vrijheid van politieke oppositie; y7=eerlijkheid van verkiezingen; y8=effectiviteit van de gekozen wetgeving.
- Latente variabele *mate van industrialisatie (1960)*: x1=Bruto nationaal product; x2=Consumptie energie; x3=Percentage werkkracht in de industrie.

SEM in R gaat via pakket lavaan




```
model <- '
   # latent variables
     ind60 =~ x1 + x2 + x3
     dem60 =~ y1 + y2 + y3 + y4
     dem65 =~ y5 + y6 + y7 + y8
   # regressions
     dem60 ~ ind60
     dem65 ~ ind60 + dem60
   # residual covariances
     y1 ~~ y5
     y2 ~~ y4 + y6
     y3 ~~ y7
     y4 ~~ y8
     y6 ~~ y8
'
fit <- sem(model,
           data=PoliticalDemocracy)
summary(fit)
```

Figure 23

Commando in R

fit <- sem(model, data=PoliticalDemocracy) summary(fit, standardized=TRUE)

Resultaten 1

```
 Number of observations                         75

 Estimator                                      ML
 Minimum Function Test Statistic            38.125
 Degrees of freedom                             35
 P-value (Chi-square)                        0.329

arameter estimates:

 Information                             Expected
 Standard Errors                         Standard

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
atent variables:
 ind60 =~
   x1                 1.000                               0.670    0.920
   x2                 2.180    0.139   15.742    0.000    1.460    0.973
   x3                 1.819    0.152   11.967    0.000    1.218    0.872
 dem60 =~
   y1                 1.000                               2.223    0.850
   y2                 1.257    0.182    6.889    0.000    2.794    0.717
   y3                 1.058    0.151    6.987    0.000    2.351    0.722
   y4                 1.265    0.145    8.722    0.000    2.812    0.846
 dem65 =~
   y5                 1.000                               2.103    0.808
   y6                 1.186    0.169    7.024    0.000    2.493    0.746
   y7                 1.280    0.160    8.002    0.000    2.691    0.824
   y8                 1.266    0.158    8.007    0.000    2.662    0.828
```

Figure 24

Resultaten 2

dem60 ~						
ind60	1.483	0.399	3.715	0.000	0.447	0.447
dem65 ~						
ind60	0.572	0.221	2.586	0.010	0.182	0.182
dem60	0.837	0.098	8.514	0.000	0.885	0.885
Covariances:						
y1 ~~						
y5	0.624	0.358	1.741	0.082	0.624	0.296
y2 ~~						
y4	1.313	0.702	1.871	0.061	1.313	0.273
y6	2.153	0.734	2.934	0.003	2.153	0.356
y3 ~~						
y7	0.795	0.608	1.308	0.191	0.795	0.191
y4 ~~						
y8	0.348	0.442	0.787	0.431	0.348	0.109
y6 ~~						
y8	1.356	0.568	2.386	0.017	1.356	0.338
Variances:						
x1	0.082	0.019			0.082	0.154
x2	0.120	0.070			0.120	0.053
x3	0.467	0.090			0.467	0.239
y1	1.891	0.444			1.891	0.277
y2	7.373	1.374			7.373	0.486
y3	5.067	0.952			5.067	0.478
y4	3.148	0.739			3.148	0.285
y5	2.351	0.480			2.351	0.347
y6	4.954	0.914			4.954	0.443
y7	3.431	0.713			3.431	0.322
y8	3.254	0.695			3.254	0.315
ind60	0.448	0.087			1.000	1.000
dem60	3.956	0.921			0.800	0.800
dem65	0.172	0.215			0.039	0.039

Figure 25

SEM in Mplus

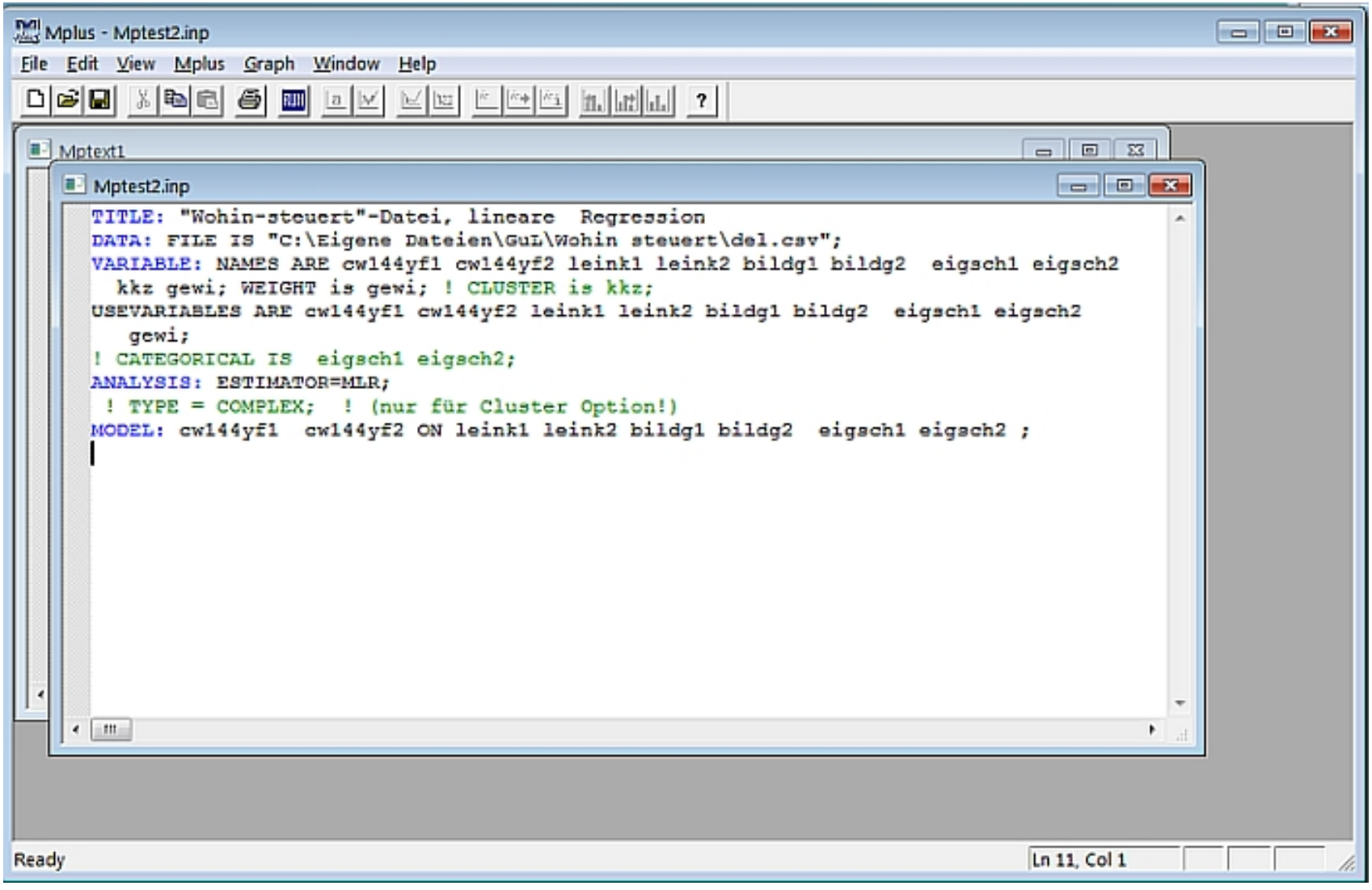


Figure 26

LCA-voorbeeld in Mplus

1. I like to drink
2. I drink hard liquor
3. I have drank in the morning
4. I have drank at work
5. I drink to get drunk
6. I like the taste of alcohol
7. I drink help me sleep
8. Drinking interferes with my relationships
9. I frequently visit bars

Figure 27

Mplus data

```
list id item1-item9 in 1/10
```

	id	item1	item2	item3	item4	item5	item6	item7	item8	item9
1.	1	1	0	0	0	0	0	0	0	0
2.	2	1	1	0	1	1	1	1	0	0
3.	3	1	0	0	0	0	0	0	0	0
4.	4	1	0	0	0	0	1	1	0	0
5.	5	1	0	0	0	1	0	0	0	1
6.	6	0	1	0	0	0	1	0	0	0
7.	7	1	1	0	0	0	0	0	0	1
8.	8	1	0	1	0	0	0	0	0	0
9.	9	1	0	0	0	0	0	0	1	0
10.	10	0	0	0	0	0	1	0	0	0

Figure 28

Mplus code

```
Title:
  Fictitous Latent Class Analysis.
Data:
  File is http://stats.idre.ucla.edu/wp-content/uploads/2016/02/lca1.dat ;
Variable:
  names        = id item1 item2 item3 item4 item5 item6 item7 item8 item9;
  usevariables = item1 item2 item3 item4 item5 item6 item7 item8 item9;
  categorical  = item1 item2 item3 item4 item5 item6 item7 item8 item9;
  classes = c(3);
Analysis:
  Type=mixture;
Plot:
  type is plot3;
  series is item1 (1) item2 (2) item3 (3) item4 (4) item5 (5)
            item6 (6) item7 (7) item8 (8) item9 (9);
Savedata:
  file is lca1_save.txt ;
  save is cprob;
  format is free;
output:
  tech11 tech14;
```

Figure 29

Verdelen van items over de drie klassen

	Class 1	Class 2	Class 3	Item Label
ITEM1	0.908	0.312	0.923	I like to drink
ITEM2	0.337	0.164	0.546	I drink hard liquor
ITEM3	0.067	0.036	0.426	I have drank in the morning
ITEM4	0.065	0.056	0.418	I have drank at work
ITEM5	0.219	0.044	0.765	I drink to get drunk
ITEM6	0.320	0.183	0.471	I like the taste of alcohol
ITEM7	0.113	0.098	0.512	I drink help me sleep
ITEM8	0.140	0.110	0.619	Drinking interferes with my relationships
ITEM9	0.325	0.188	0.349	I frequently visit bars

Figure 30

Hoe zien de drie klassen eruit

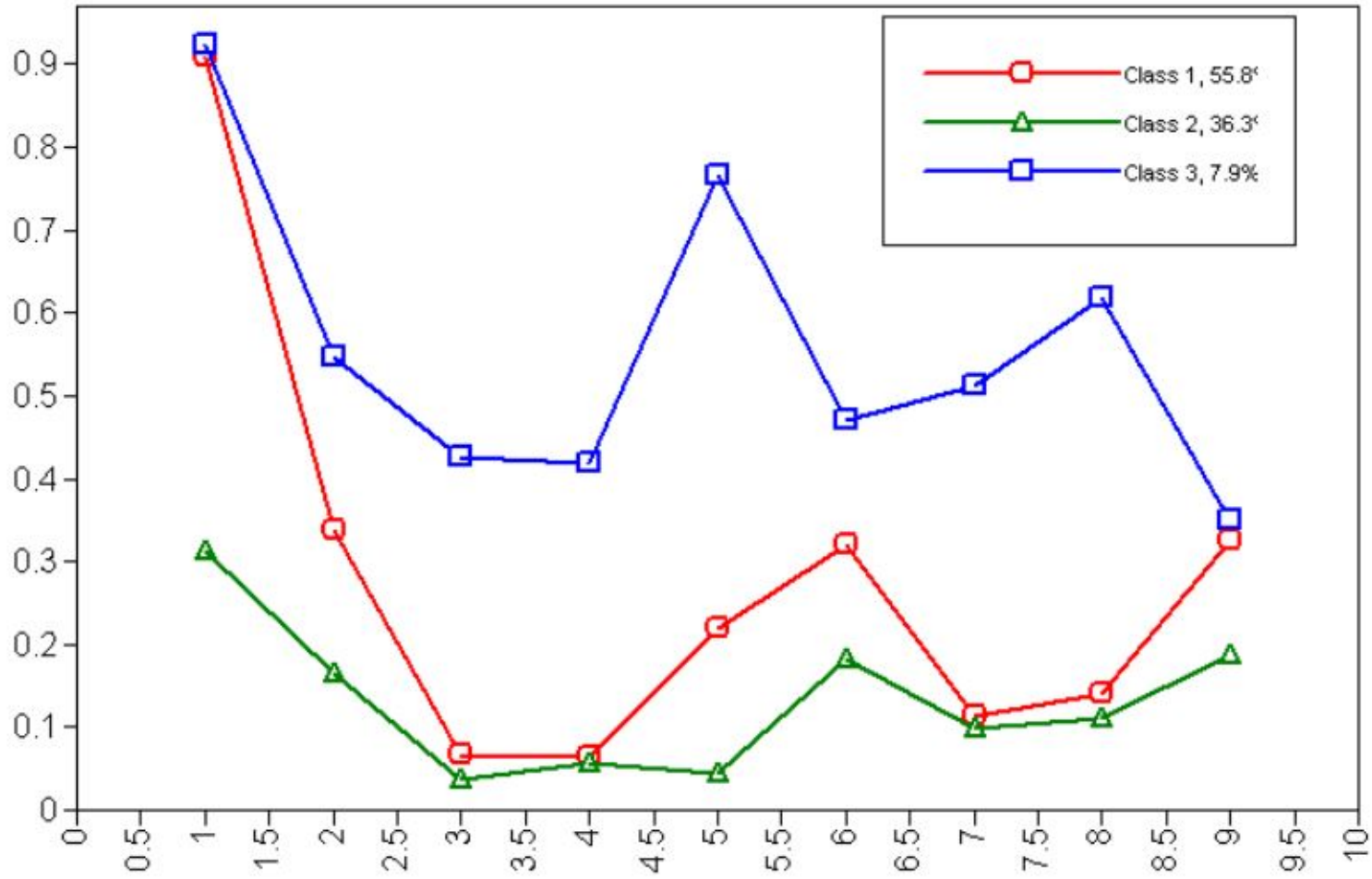


Figure 31

Zo wordt het opgeslagen

Items 1 - 9											
									P(c1)	P(c2)	P(c3)	Class
1	0	0	0	0	0	0	0	0	0.645	0.354	0.001	1
1	1	0	1	1	1	1	0	0	0.098	0.001	0.901	3
1	0	0	0	0	0	0	0	0	0.645	0.354	0.001	1
1	0	0	0	0	1	1	0	0	0.797	0.177	0.026	1
1	0	0	0	1	0	0	0	1	0.934	0.041	0.025	1
0	1	0	0	0	1	0	0	0	0.312	0.686	0.002	2
1	1	0	0	0	0	0	0	1	0.903	0.092	0.005	1
1	0	1	0	0	0	0	0	0	0.766	0.218	0.017	1
1	0	0	0	0	0	0	1	0	0.696	0.290	0.014	1
0	0	0	0	0	1	0	0	0	0.149	0.850	0.000	2

Figure 32

Verdeling over klassen

CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

Latent Classes		
1	646	0.64600
2	288	0.28800
3	66	0.06600

##

Verdeling per percentages

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASS PATTERNS
BASED ON ESTIMATED POSTERIOR PROBABILITIES

Latent Classes		
1	557.56836	0.55757
2	363.13989	0.36314
3	79.29175	0.07929

3 Klassen tov 2 klassen (0-hypothese)

```
TECHNICAL 11 OUTPUT

     VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES
          H0 Loglikelihood Value                          -4251.208
          2 Times the Loglikelihood Difference               39.025
          Difference in the Number of Parameters                 10
          Mean                                               20.255
          Standard Deviation                                 22.224
          P-Value                                            0.1457

     LO-MENDELL-RUBIN ADJUSTED LRT TEST
          Value                                              38.468
          P-Value                                            0.1500

TECHNICAL 14 OUTPUT

     BOOTSTRAPPED PARAMETRIC LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES
          H0 Loglikelihood Value                          -4251.208
          2 Times the Loglikelihood Difference               39.025
          Difference in the Number of Parameters                 10
          Approximate P-Value                                0.0000
```

Figure 33

Waar SEM zoal geschikt voor is

- Regressies en padanalyses
- Exploratieve en Conformatieve Factor Analyse
- Groei modellen en survival analyse
- Mixture modelling (zowel cross sectioneel als longitudinaal)
- Multilevel modelling (nog niet in Lavaan)
- Bayesiaans (Mplus en Blavaan, maar ook in WinBUGS)

Literatuur

Acock, A.C. (2013). Discovering Structural Equation. Modeling Using Stata. Revised edition. College Station: StataCorp

Chuck Huber: http://www.stata.com/meeting/italy13/abstracts/materials/it13_huber.pdf

Jim Grace website (heel inzichtelijk): https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation?qt-science_center_objects=1#qt-science_center_objects

Stata Corp (2013): http://www.stata.com/manuals13/sem.pdf

Rosseel, Y. (2017). The lavaan tutorial: http://lavaan.ugent.be/tutorial/tutorial.pdf

https://stats.idre.ucla.edu/mplus/dae/latent-class-analysis/