Solution for the assignment of the eleventh class

Kovacs Marton

9/29/2021

Importing data

```
processed <- read_tsv("data/boldog_processed.tsv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   neme = col_character(),
##   isk = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Data exploration

```
skimr::skim(processed) %>%
  kable()
```

skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.empty	character.n_unique	character.whitespace	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
character	neme	0	1.000	2	5	0	2	0	NA	NA	NA	NA	NA	NA	NA	NA
character	isk	0	1.000	7	11	0	4	0	NA	NA	NA	NA	NA	NA	NA	NA
numeric	index	0	1.000	NA	NA	NA	NA	NA	972.34	486.43	27.16	598.00	1070.00	1361.75	1700.00	<U+2585><U+2585><U
numeric	eletkora	0	1.000	NA	NA	NA	NA	NA	52.52	20.07	25.78	45.00	52.50	61.00	86.00	<U+2581><U+2585><U
numeric	gyermekek	0	1.000	NA	NA	NA	NA	NA	1.75	1.02	0.37	1.00	2.00	2.00	9.00	<U+2585><U+2587><U
numeric	anyagi	0	1.000	NA	NA	NA	NA	NA	3.10	0.56	1.44	2.96	3.00	3.00	5.00	<U+2581><U+2581><U
numeric	pic_elmeny_percent	0	1.000	NA	NA	NA	NA	NA	58.46	20.04	18.35	40.00	60.00	80.00	90.00	<U+2582><U+2585><U
numeric	testi_fi	3	0.994	NA	NA	NA	NA	NA	4.16	0.95	1.72	4.00	4.00	5.00	6.00	<U+2581><U+2583><U
numeric	alt_lelki	4	0.992	NA	NA	NA	NA	NA	4.16	1.00	1.72	3.62	4.00	4.00	5.00	<U+2582><U+2583><U
numeric	alt_eg_all	7	0.986	NA	NA	NA	NA	NA	4.12	1.04	1.95	3.61	4.00	5.00	6.00	<U+2582><U+2583><U
numeric	fizero	6	0.988	NA	NA	NA	NA	NA	4.15	1.04	1.86	3.07	4.00	5.00	6.00	<U+2582><U+2585><U
numeric	aicocska	0	1.000	NA	NA	NA	NA	NA	2.57	1.04	1.43	1.89	2.00	3.00	7.00	<U+2587><U+2585><U
numeric	aiggodalo	0	1.000	NA	NA	NA	NA	NA	2.77	1.04	1.32	2.23	2.00	3.50	4.00	<U+2587><U+2583><U
numeric	idoeges	0	1.000	NA	NA	NA	NA	NA	2.49	1.03	1.37	1.61	1.00	2.00	3.00	<U+2587><U+2582><U
numeric	feszult	0	1.000	NA	NA	NA	NA	NA	2.61	1.03	1.39	1.97	1.00	2.00	3.00	<U+2587><U+2583><U
numeric	nyugtala	0	1.000	NA	NA	NA	NA	NA	2.40	1.04	1.57	1.80	1.00	2.00	3.00	<U+2587><U+2582><U

skim_type	variable	missing	complete	n	min	max	empty	n_unique	whitespace	mean	sd	p0	p25	p50	p75	p100	hist
numeric	diener1	0	1.000	NA	NA	NA	NA	NA		5.514	0.328	5.00	5.0000	6.0000	6.0000	7	<U+2581><U+2581><U
numeric	diener2	0	1.000	NA	NA	NA	NA	NA		5.228	0.321	3.46	4.0000	6.0000	6.0000	7	<U+2581><U+2581><U
numeric	diener3	0	1.000	NA	NA	NA	NA	NA		5.574	0.231	7.88	5.0000	6.0000	6.0000	7	<U+2581><U+2581><U
numeric	diener4	0	1.000	NA	NA	NA	NA	NA		5.418	0.261	1.32	5.0000	6.0000	6.0000	7	<U+2581><U+2581><U
numeric	diener5	0	1.000	NA	NA	NA	NA	NA		5.756	0.072	7.74	5.0000	6.0000	7.0000	7	<U+2581><U+2581><U
numeric	diener6	0	1.000	NA	NA	NA	NA	NA		5.768	0.245	0.02	5.0000	6.0000	7.0000	7	<U+2581><U+2581><U
numeric	diener7	0	1.000	NA	NA	NA	NA	NA		5.620	0.373	1.76	5.0000	6.0000	7.0000	7	<U+2581><U+2581><U
numeric	diener8	0	1.000	NA	NA	NA	NA	NA		5.532	0.214	7.07	5.0000	6.0000	6.0000	7	<U+2581><U+2581><U
numeric	jollet	0	1.000	NA	NA	NA	NA	NA		4.488	0.661	0.84	4.0000	4.6667	5.3333	6	<U+2581><U+2582><U
numeric	savor	0	1.000	NA	NA	NA	NA	NA		4.514	0.713	3.60	4.0000	4.6667	5.0000	6	<U+2581><U+2582><U
numeric	vhat	0	1.000	NA	NA	NA	NA	NA		4.595	0.939	0.64	4.0000	4.8000	5.2000	6	<U+2581><U+2581><U
numeric	reg	0	1.000	NA	NA	NA	NA	NA		4.217	0.221	3.68	3.3333	4.3333	5.0833	6	<U+2582><U+2583><U
numeric	zil	0	1.000	NA	NA	NA	NA	NA		4.044	0.721	0.53	3.3333	4.0000	4.6667	6	<U+2581><U+2583><U
numeric	flow	0	1.000	NA	NA	NA	NA	NA		4.773	0.890	7.98	4.3333	5.0000	5.3333	6	<U+2581><U+2581><U
numeric	jerz	0	1.000	NA	NA	NA	NA	NA		4.152	0.031	7.75	3.4000	4.2000	5.0000	6	<U+2581><U+2583><U
numeric	jpszi	0	1.000	NA	NA	NA	NA	NA		4.282	0.989	2.83	3.7500	4.5000	5.0000	6	<U+2581><U+2582><U
numeric	jszoc	0	1.000	NA	NA	NA	NA	NA		3.949	0.178	9.09	3.0000	4.0000	5.0000	6	<U+2582><U+2585><U
numeric	jspir	0	1.000	NA	NA	NA	NA	NA		4.245	0.200	3.07	3.5000	4.2500	5.2500	6	<U+2582><U+2583><U
numeric	jerzpsz	0	1.000	NA	NA	NA	NA	NA		4.217	0.972	0.34	3.6187	5.3375	4.9250	6	<U+2581><U+2582><U
numeric	jspszoc	0	1.000	NA	NA	NA	NA	NA		4.097	0.038	2.01	3.3750	4.1250	5.0000	6	<U+2581><U+2583><U
numeric	mm	0	1.000	NA	NA	NA	NA	NA		3.027	3.426	7.87	2.6667	3.1667	3.5000	4	<U+2581><U+2582><U
numeric	av	0	1.000	NA	NA	NA	NA	NA		3.175	0.623	1.89	2.7500	3.2500	3.7500	4	<U+2581><U+2581><U
numeric	onr	0	1.000	NA	NA	NA	NA	NA		3.064	0.791	2.81	2.6667	3.3333	3.6667	4	<U+2581><U+2582><U
numeric	rez	0	1.000	NA	NA	NA	NA	NA		3.150	0.874	9.24	2.6667	3.3333	3.6667	4	<U+2581><U+2582><U
numeric	poz_erz	0	1.000	NA	NA	NA	NA	NA		21.742	5.027	5.89	20.0000	23.0000	26.0000	30	<U+2581><U+2581><U
numeric	elmely	0	1.000	NA	NA	NA	NA	NA		22.360	4.958	8.41	20.0000	23.0000	26.0000	30	<U+2581><U+2582><U
numeric	poz_kapc	1	1.000	NA	NA	NA	NA	NA		22.442	5.067	1.07	19.0000	24.0000	26.0000	30	<U+2581><U+2581><U
numeric	ert_cel	0	1.000	NA	NA	NA	NA	NA		23.374	4.679	9.81	21.0000	24.0000	27.0000	30	<U+2581><U+2581><U
numeric	telj	0	1.000	NA	NA	NA	NA	NA		22.716	4.342	0.99	21.0000	24.0000	26.0000	30	<U+2581><U+2581><U
numeric	boldog	0	1.000	NA	NA	NA	NA	NA		7.404	2.021	7.46	7.0000	8.0000	9.0000	10	<U+2581><U+2581><U
numeric	egeszs	0	1.000	NA	NA	NA	NA	NA		22.266	5.022	8.43	19.0000	23.0000	26.0000	30	<U+2581><U+2581><U
numeric	neg_erz	0	1.000	NA	NA	NA	NA	NA		9.766	5.708	7.79	5.0000	9.0000	13.0000	30	<U+2587><U+2587><U
numeric	magany	0	1.000	NA	NA	NA	NA	NA		3.162	3.056	5.07	0.0000	2.0000	5.0000	10	<U+2587><U+2582><U
numeric	erma	0	1.000	NA	NA	NA	NA	NA		169.373	30.267	22.81	155.0000	175.5000	191.0000	225	<U+2581><U+2581><U

1. Run CFA on the best model from the sixths task in the ninth assignments best model.

Get the data for the model.

```
vars <- c("p_elmeny_percent", "testi_fi", "alt_lelki", "alt_eg_all", "fizero", "arcocska", "aggodalo",

task_data <-
  processed %>%
  select(all_of(vars))
```

First I will run the EFA as I did not run it in the *ninth_assignment*.

With 2 factors.

```r
efa2 <- fa(task_data, nfactors = 2, rotate = "varimax", fm = "ml")

efa2
```

```
## Factor Analysis using method =  ml
## Call: fa(r = task_data, nfactors = 2, rotate = "varimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                    ML1   ML2   h2   u2 com
## p_elmeny_percent -0.50  0.24 0.31 0.69 1.4
## testi_fi         -0.11  0.88 0.79 0.21 1.0
## alt_lelki        -0.42  0.51 0.44 0.56 1.9
## alt_eg_all       -0.16  0.86 0.77 0.23 1.1
## fizero           -0.12  0.84 0.71 0.29 1.0
## arcocska          0.53 -0.40 0.44 0.56 1.9
## aggodalo          0.86 -0.07 0.75 0.25 1.0
## ideges            0.92 -0.12 0.86 0.14 1.0
## feszult           0.94 -0.12 0.90 0.10 1.0
## nyugtala          0.66 -0.18 0.47 0.53 1.1
##
##                        ML1  ML2
## SS loadings           3.68 2.77
## Proportion Var        0.37 0.28
## Cumulative Var        0.37 0.64
## Proportion Explained  0.57 0.43
## Cumulative Proportion 0.57 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  45  and the objective function was  6.96 with Chi Squa
## The degrees of freedom for the model are 26  and the objective function was  0.25
##
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic number of observations is  496 with the empirical chi square  73.41  with prob <  2.1e-0
## The total number of observations was  500  with Likelihood Chi Square =  124.08  with prob <  9.5e-15
##
## Tucker Lewis Index of factoring reliability =  0.95
## RMSEA index =  0.087  and the 90 % confidence intervals are  0.072 0.103
## BIC =  -37.5
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                 ML1  ML2
## Correlation of (regression) scores with factors   0.97 0.95
## Multiple R square of scores with factors          0.95 0.91
## Minimum correlation of possible factor scores     0.90 0.81
```

With 3 factors.

```r
efa3 <- fa(task_data, nfactors = 3, rotate = "varimax", fm = "ml")

efa3
```

```
## Factor Analysis using method =  ml
## Call: fa(r = task_data, nfactors = 3, rotate = "varimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                    ML1   ML2   ML3   h2    u2 com
## p_elmeny_percent -0.41  0.17 -0.43 0.38 0.619 2.3
## testi_fi         -0.08  0.88 -0.16 0.80 0.198 1.1
## alt_lelki        -0.31  0.44 -0.48 0.52 0.478 2.7
## alt_eg_all       -0.13  0.85 -0.16 0.77 0.228 1.1
## fizero           -0.09  0.82 -0.18 0.71 0.290 1.1
## arcocska          0.36 -0.27  0.78 0.81 0.189 1.7
## aggodalo          0.84 -0.06  0.22 0.75 0.248 1.1
## ideges            0.90 -0.11  0.22 0.87 0.133 1.1
## feszult           0.92 -0.11  0.22 0.90 0.099 1.1
## nyugtala          0.63 -0.16  0.21 0.47 0.534 1.4
##
##                        ML1  ML2  ML3
## SS loadings           3.16 2.52 1.30
## Proportion Var        0.32 0.25 0.13
## Cumulative Var        0.32 0.57 0.70
## Proportion Explained  0.45 0.36 0.19
## Cumulative Proportion 0.45 0.81 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are  45  and the objective function was  6.96 with Chi Squa
## The degrees of freedom for the model are 18  and the objective function was  0.04
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.01
##
## The harmonic number of observations is  496 with the empirical chi square  3.76  with prob <  1
## The total number of observations was  500  with Likelihood Chi Square =  20.35  with prob <  0.31
##
## Tucker Lewis Index of factoring reliability =  0.998
## RMSEA index =  0.016  and the 90 % confidence intervals are  0 0.044
## BIC =  -91.51
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 ML1  ML2  ML3
## Correlation of (regression) scores with factors   0.96 0.94 0.85
## Multiple R square of scores with factors          0.92 0.89 0.72
## Minimum correlation of possible factor scores     0.85 0.78 0.44
```

The BIC is smaller for the 2 factor model so I choose that one. However, three items have a high complexity and they load for both factors.

Specifying the model for CFA.

```
model <- '
factor1 =~ p_elmeny_percent + arcocska + aggodalo + ideges + feszult + nyugtala
factor2 =~ testi_fi + alt_lelki + fizero  + alt_eg_all'
```

Fitting the model.

```r
cfa <- cfa(model, data = task_data)

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

```
## lavaan 0.6-9 ended normally after 86 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        21
##
##                                                   Used       Total
##   Number of observations                           483         500
##
## Model Test User Model:
##
##   Test statistic                               305.427
##   Degrees of freedom                                34
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                              3394.908
##   Degrees of freedom                                45
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.919
##   Tucker-Lewis Index (TLI)                       0.893
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -7686.417
##   Loglikelihood unrestricted model (H1)      -7533.703
##
##   Akaike (AIC)                               15414.833
##   Bayesian (BIC)                             15502.614
##   Sample-size adjusted Bayesian (BIC)        15435.961
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.129
##   90 Percent confidence interval - lower         0.116
##   90 Percent confidence interval - upper         0.142
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.129
##
## Parameter Estimates:
##
##   Standard errors                             Standard
```

```
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   factor1 =~
##     p_elmeny_prcnt    1.000                               11.364    0.528
##     arcocska         -0.059    0.006  -10.107    0.000   -0.669   -0.586
##     aggodalo         -0.109    0.009  -12.535    0.000   -1.240   -0.866
##     ideges           -0.112    0.009  -12.928    0.000   -1.277   -0.931
##     feszult          -0.118    0.009  -13.019    0.000   -1.335   -0.950
##     nyugtala         -0.087    0.008  -11.038    0.000   -0.985   -0.677
##   factor2 =~
##     testi_fi          1.000                                0.857    0.888
##     alt_lelki         0.713    0.053   13.515    0.000    0.611    0.570
##     fizero            1.172    0.049   24.045    0.000    1.005    0.843
##     alt_eg_all        1.164    0.045   25.771    0.000    0.997    0.884
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   factor1 ~~
##     factor2           3.075    0.550    5.594    0.000    0.316    0.316
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .p_elmeny_prcnt  333.961   21.905   15.246    0.000  333.961    0.721
##    .arcocska          0.857    0.057   15.142    0.000    0.857    0.657
##    .aggodalo          0.514    0.039   13.144    0.000    0.514    0.250
##    .ideges            0.251    0.025    9.838    0.000    0.251    0.133
##    .feszult           0.194    0.025    7.897    0.000    0.194    0.098
##    .nyugtala          1.150    0.077   14.893    0.000    1.150    0.542
##    .testi_fi          0.197    0.021    9.424    0.000    0.197    0.212
##    .alt_lelki         0.774    0.052   14.809    0.000    0.774    0.675
##    .fizero            0.410    0.035   11.576    0.000    0.410    0.289
##    .alt_eg_all        0.277    0.029    9.626    0.000    0.277    0.218
##     factor1         129.134   21.076    6.127    0.000    1.000    1.000
##     factor2           0.734    0.061   12.054    0.000    1.000    1.000
```

2. How good is the model fit?

The model has a significantly good fit with $X\hat{}2(34, 500) = 305.43$, $p < 0.001$.

3. What about the other test statistics?

The RMSEA = 0.129 CI90[0.116, 0.142] with $p < 0.001$ for alpha <= 0.05.

4. What about CFI, SRMR, and TLI?

The Standardized Root Mean Square Residual (SRMR) is 0.129. The value ranges between 0 and 1, and as it is higher than the usually accepted 0.08 cutoff value, it does not indicate a good fit.

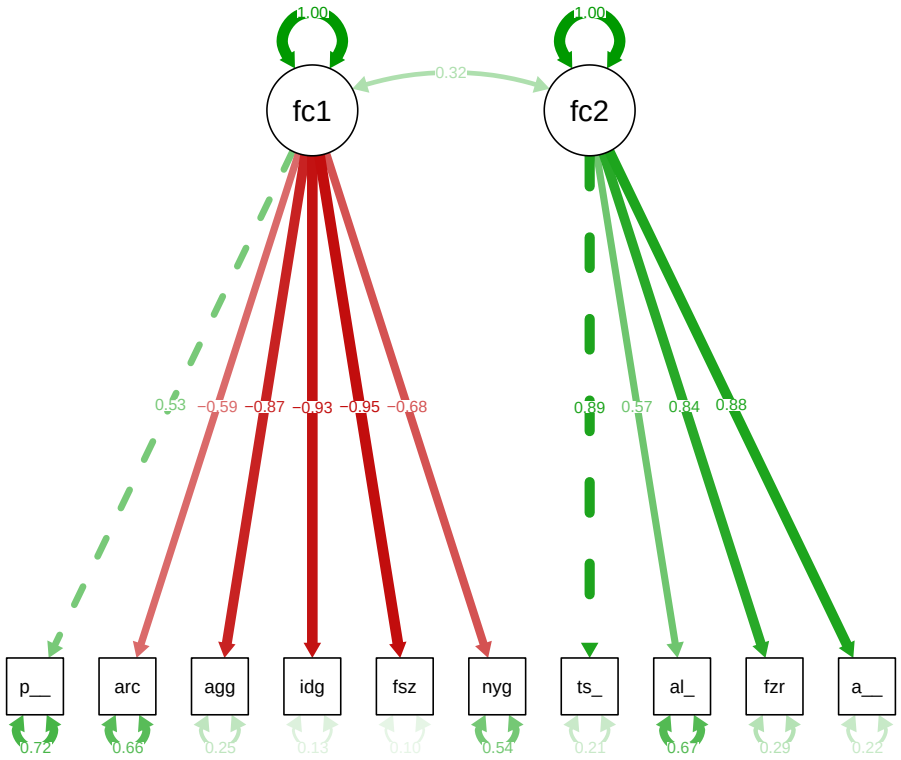
The TLI has a penalty for adding more parameters. It is 0.893 which is below the .9 cutoff point for a good model.

Tha CFI is usually highly correlated with the TLI. The value here is 0.919 which means it is not good either.

5. Create a graph for the latent structure.

Plotting the model with standardized model parameter estimates.

```
semPlot::semPaths(cfa, "std")
```



6. Which variables should be dropped?

I suspect that if I drop the two variables with high complexity the model will be better.

Lets test it for fun!

```
task_data_filtered <-
  task_data %>%
  select(-alt_lelki, arcocska)

model_filtered <- '
factor1 =~ p_elmeny_percent + aggodalo + ideges + feszult + nyugtala
```

```
factor2 =~ testi_fi + fizero  + alt_eg_all'

cfa_filtered <- cfa(model_filtered, data = task_data_filtered)

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

```
## lavaan 0.6-9 ended normally after 83 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        17
##
##                                                   Used       Total
##   Number of observations                            486         500
##
## Model Test User Model:
##
##   Test statistic                                37.665
##   Degrees of freedom                                19
##   P-value (Chi-square)                           0.007
##
## Model Test Baseline Model:
##
##   Test statistic                              2769.544
##   Degrees of freedom                                28
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.993
##   Tucker-Lewis Index (TLI)                       0.990
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6442.754
##   Loglikelihood unrestricted model (H1)      -6423.922
##
##   Akaike (AIC)                               12919.509
##   Bayesian (BIC)                             12990.674
##   Sample-size adjusted Bayesian (BIC)        12936.717
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.045
##   90 Percent confidence interval - lower         0.023
##   90 Percent confidence interval - upper         0.066
##   P-value RMSEA <= 0.05                          0.625
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.049
##
## Parameter Estimates:
```

```
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   factor1 =~
##     p_elmeny_prcnt    1.000                               11.048    0.515
##     aggodalo         -0.112    0.009  -12.166    0.000   -1.236   -0.863
##     ideges           -0.116    0.009  -12.548    0.000   -1.278   -0.932
##     feszult          -0.121    0.010  -12.627    0.000   -1.337   -0.951
##     nyugtala         -0.089    0.008  -10.789    0.000   -0.984   -0.675
##   factor2 =~
##     testi_fi          1.000                                0.866    0.898
##     fizero            1.161    0.048   24.084    0.000    1.005    0.844
##     alt_eg_all        1.137    0.045   25.239    0.000    0.984    0.874
##
## Covariances:
##                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   factor1 ~~
##     factor2           2.600    0.523    4.967    0.000    0.272    0.272
##
## Variances:
##                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .p_elmeny_prcnt  338.618   22.117   15.310    0.000  338.618    0.735
##    .aggodalo          0.525    0.040   13.211    0.000    0.525    0.256
##    .ideges            0.245    0.026    9.474    0.000    0.245    0.130
##    .feszult           0.188    0.025    7.457    0.000    0.188    0.095
##    .nyugtala          1.157    0.077   14.936    0.000    1.157    0.544
##    .testi_fi          0.179    0.022    8.321    0.000    0.179    0.193
##    .fizero            0.409    0.036   11.347    0.000    0.409    0.288
##    .alt_eg_all        0.299    0.030    9.826    0.000    0.299    0.236
##     factor1         122.047   20.463    5.964    0.000    1.000    1.000
##     factor2           0.750    0.061   12.238    0.000    1.000    1.000
```

Indeed the CFI and TLI values are higher!