Confirmatory Factor Analysis

Author: Joanna Moody Date: September 4, 2019 Last modified: November 2, 2019 (converted R code to R markdown file)

```
#Initialization
rm(list=ls(all=TRUE))
library(lavaan)
```

```
## This is lavaan 0.6-4
```

```
## lavaan is BETA software! Please report any bugs.
```

```
library(knitr)

#Load input data
setwd("/Users/jcmoody/Dropbox (MIT)/07_Latent_attitude_and_AV_perference/")
data <-read.csv("data/individual_new.csv",header=TRUE, sep = ",", stringsAsFactors = FALSE)
data[data == -1] <- NA #-1 indicates missing data

#Note: indicators labeled with "PERC" are "Enjoyment"

#Function to wrap text in "knitted" R markdown document
hook_output = knit_hooks$get('output')
knit_hooks$set(output = function(x, options) {
  # this hook is used only when the linewidth option is not NULL
  if (!is.null(n <- options$linewidth)) {
    x = knitr:::split_lines(x)
    # any lines wider than n should be wrapped
    if (any(nchar(x) > n)) x = strwrap(x, width = n)
    x = paste(x, collapse = '\n')
  }
  hook_output(x, options)
})
```

CONFIRMATORY FACTOR ANALYSIS

Treating 7-Point scale as continuous (using Maximum Likelihood estimation)

WALK

Model 0: Baseline model with all 5 indicators, no correlated errors

```
walk.model0 <- 'Walk =~ WSAFE + WCOMF + WRELY + WEASY + WPERC'
fit.walk.model0 <- cfa(walk.model0, data=data, missing='fiml')

#Model fit has TLI and RMSEA outside of recommended bounds:
#  chi-square(N = 2003, df = 5) = 208.247,
#  CFI = 0.939, TLI = 0.879,
#  RMSEA = 0.142, SRMR = 0.042
#Factor loadings: lowest standardized loadings is 0.5; all others > 0.6
summary(fit.walk.model0, fit.measures=TRUE, standardized=TRUE)
```

```
## lavaan 0.6-4 ended normally after 27 iterations
##
## Optimization method NLMINB
## Number of free parameters 15
```

```
##
## Number of observations 2003
## Number of missing patterns 1
##
## Estimator ML
## Model Fit Test Statistic 208.247
## Degrees of freedom 5
## P-value (Chi-square) 0.000
##
## Model test baseline model:
##
## Minimum Function Test Statistic 3368.343
## Degrees of freedom 10
## P-value 0.000
##
## User model versus baseline model:
##
## Comparative Fit Index (CFI) 0.939
## Tucker-Lewis Index (TLI) 0.879
##
## Loglikelihood and Information Criteria:
##
## Loglikelihood user model (H0) -16567.666
## Loglikelihood unrestricted model (H1) -16463.543
##
## Number of free parameters 15
## Akaike (AIC) 33165.332
## Bayesian (BIC) 33249.368
## Sample-size adjusted Bayesian (BIC) 33201.712
##
## Root Mean Square Error of Approximation:
##
## RMSEA 0.142
## 90 Percent Confidence Interval 0.126 0.159
## P-value RMSEA <= 0.05 0.000
##
## Standardized Root Mean Square Residual:
##
## SRMR 0.042
##
## Parameter Estimates:
##
## Information Observed
## Observed information based on Hessian
## Standard Errors Standard
##
## Latent Variables:
## Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## Walk =~
## WSAFE 1.000 0.630 0.510
## WCOMF 1.784 0.095 18.870 0.000 1.124 0.631
## WRELY 1.389 0.067 20.736 0.000 0.875 0.677
## WEASY 2.001 0.096 20.787 0.000 1.260 0.736
## WPERC 1.971 0.094 20.876 0.000 1.241 0.845
```

```
##
## Intercepts:
## Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## .WSAFE 5.770 0.028 209.133 0.000 5.770 4.673
## .WCOMF 4.104 0.040 103.092 0.000 4.104 2.303
## .WRELY 5.936 0.029 205.679 0.000 5.936 4.596
## .WEASY 5.615 0.038 146.853 0.000 5.615 3.281
## .WPERC 5.455 0.033 166.298 0.000 5.455 3.716
## Walk 0.000 0.000 0.000
##
## Variances:
## Estimate Std.Err z-value P(>|z|) Std.lv Std.all
## .WSAFE 1.128 0.039 29.139 0.000 1.128 0.740
## .WCOMF 1.912 0.069 27.644 0.000 1.912 0.602
## .WRELY 0.904 0.035 25.694 0.000 0.904 0.542
## .WEASY 1.341 0.055 24.329 0.000 1.341 0.458
## .WPERC 0.615 0.038 16.100 0.000 0.615 0.285
## Walk 0.397 0.036 11.089 0.000 1.000 1.000
```

```
#Modification indices: suggest that model fit could be greatly improved if
#  the error terms of the following indicators are correlated:
#  WSAFE ~~ WRELY are correlated (157.7); WCOMF ~~ WPERC (96.7)
mi.walk.model0 <- modindices(fit.walk.model0)
mi.walk.model0
```

```
## lhs op rhs mi epc sepc.lv sepc.all sepc.nox
## 18 WSAFE ~~ WCOMF 4.318 -0.077 -0.077 -0.052 -0.052
## 19 WSAFE ~~ WRELY 157.677 0.328 0.328 0.325 0.325
## 20 WSAFE ~~ WEASY 0.936 -0.033 -0.033 -0.027 -0.027
## 21 WSAFE ~~ WPERC 56.491 -0.219 -0.219 -0.263 -0.263
## 22 WCOMF ~~ WRELY 31.687 -0.205 -0.205 -0.156 -0.156
## 23 WCOMF ~~ WEASY 16.493 -0.196 -0.196 -0.122 -0.122
## 24 WCOMF ~~ WPERC 96.700 0.434 0.434 0.400 0.400
## 25 WRELY ~~ WEASY 4.953 0.079 0.079 0.072 0.072
## 26 WRELY ~~ WPERC 31.608 -0.189 -0.189 -0.254 -0.254
## 27 WEASY ~~ WPERC 5.136 0.110 0.110 0.121 0.121
```

Model 1: Baseline model + correlated error between WSAFE and WRELY

```
walk.model1 <- 'Walk =~ WSAFE + WCOMF + WRELY + WEASY + WPERC
                WSAFE ~~ WRELY'
fit.walk.model1 <- cfa(walk.model1, data=data, missing='fiml')

#All model fit indices now within recommended bounds:
#  chi-square(N = 2003, df = 4) = 50.299,
#  CFI = 0.986, TLI = 0.966,
#  RMSEA = 0.076, SRMR = 0.020
#Factor loadings: standardized loading for WSAFE dropped to 0.454 (okay, not great)
summary(fit.walk.model1, fit.measures=TRUE, standardized=TRUE)
```

```
## lavaan 0.6-4 ended normally after 31 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         16
##
##   Number of observations                          2003
##   Number of missing patterns                         1
##
##   Estimator                                         ML
##   Model Fit Test Statistic                      50.299
##   Degrees of freedom                                 4
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             3368.343
##   Degrees of freedom                                10
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.986
##   Tucker-Lewis Index (TLI)                       0.966
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -16488.692
##   Loglikelihood unrestricted model (H1)     -16463.543
##
##   Number of free parameters                         16
##   Akaike (AIC)                               33009.384
##   Bayesian (BIC)                             33099.022
##   Sample-size adjusted Bayesian (BIC)        33048.189
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.076
##   90 Percent Confidence Interval          0.058  0.095
##   P-value RMSEA <= 0.05                          0.009
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.020
##
## Parameter Estimates:
##
##   Information                                 Observed
##   Observed information based on                Hessian
##   Standard Errors                             Standard
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Walk =~
##     WSAFE              1.000                               0.560    0.454
##     WCOMF              2.031    0.116   17.502    0.000    1.138    0.639
##     WRELY              1.470    0.071   20.789    0.000    0.823    0.637
##     WEASY              2.219    0.119   18.720    0.000    1.243    0.726
##     WPERC              2.297    0.123   18.737    0.000    1.287    0.876
##
## Covariances:
```

```
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  .WSAFE ~~
##    .WRELY             0.334    0.029   11.457    0.000    0.334    0.305
##
## Intercepts:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .WSAFE             5.770    0.028  209.133    0.000    5.770    4.673
##    .WCOMF             4.104    0.040  103.092    0.000    4.104    2.303
##    .WRELY             5.936    0.029  205.679    0.000    5.936    4.596
##    .WEASY             5.615    0.038  146.853    0.000    5.615    3.281
##    .WPERC             5.455    0.033  166.298    0.000    5.455    3.716
##     Walk              0.000                               0.000    0.000
##
## Variances:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .WSAFE             1.211    0.041   29.890    0.000    1.211    0.794
##    .WCOMF             1.880    0.068   27.818    0.000    1.880    0.592
##    .WRELY             0.991    0.037   27.007    0.000    0.991    0.594
##    .WEASY             1.384    0.058   23.788    0.000    1.384    0.473
##    .WPERC             0.499    0.040   12.346    0.000    0.499    0.232
##     Walk              0.314    0.032    9.756    0.000    1.000    1.000
```

```
#Modification indices: MI for WCOMF ~~ WPERC fell to 47.039, so probably not
#  worth the loss in degrees of freedom to include another correlated error term
mi.walk.model1 <- modindices(fit.walk.model1)
mi.walk.model1
```

```
##      lhs op   rhs     mi    epc sepc.lv sepc.all sepc.nox
## 19 WSAFE ~~ WCOMF  1.761  0.046   0.046    0.031    0.031
## 20 WSAFE ~~ WEASY  2.313  0.048   0.048    0.037    0.037
## 21 WSAFE ~~ WPERC  5.230 -0.066  -0.066   -0.084   -0.084
## 22 WCOMF ~~ WRELY 13.411 -0.126  -0.126   -0.092   -0.092
## 23 WCOMF ~~ WEASY 17.529 -0.215  -0.215   -0.133   -0.133
## 24 WCOMF ~~ WPERC 47.039  0.360   0.360    0.372    0.372
## 25 WRELY ~~ WEASY 25.938  0.168   0.168    0.144    0.144
## 26 WRELY ~~ WPERC  3.052 -0.056  -0.056   -0.079   -0.079
## 27 WEASY ~~ WPERC  8.623 -0.182  -0.182   -0.219   -0.219
```

PUBLIC TRANSIT

```
## Model 0: Baseline model with all 5 indicators, no correlated errors
pt.model0 <- 'PT =~ PTSAFE + PTCOMF + PTRELY + PTEASY + PTPERC'
fit.pt.model0 <- cfa(pt.model0, data=data, missing='fiml')
summary(fit.pt.model0, fit.measures=TRUE, standardized=TRUE)
#Model fit: chi-square(N = 2003, df = 5) = 94.823, CFI = 0.973, TLI = 0.946, RMSEA = 0.095, SRMR = 0.028
#   Model fit is not terrible, but should inspect modification indices just in case (for slighly high R
#Factor loadings: PTSAFE has standardized loading of 0.451, the rest are over 0.65
mi.pt.model0 <- modindices(fit.pt.model0)
mi.pt.model0
#Modification indices: Highest MIs are around 45 for PTSAFE ~~ PTEASY and PTSAFE ~~ PTPERC (probably not
```

RIDEHAILING

```
## Model 0: Baseline model with all 5 indicators, no correlated errors
rh.model0 <- 'RH =~ CARSAFE + CARCOMF + CARRELY + CAREASY + CARPERC'
fit.rh.model0 <- cfa(rh.model0, data=data, missing='fiml')
```

```
summary(fit.rh.model0, fit.measures=TRUE, standardized=TRUE)
#Model fit: chi-square(N = 2003, df = 5) = 54.086, CFI = 0.988, TLI = 0.976, RMSEA = 0.070, SRMR = 0.018
#   Good model fit
#Factor loadings: All indicators have standardized loadings of > 0.65. Great!
mi.rh.model0 <- modindices(fit.rh.model0)
mi.rh.model0 #super low MIs, so no need for correlated errors (supported by already good model fit)



### DRIVE
#Here is where we run into issues with missingness: individuals who do not have access to a car were not
#   asked these questions; there is no way to estimate a factor score for them because they are missing
#   all indicators systematically
#Only estimating for 953 of the 2003 individuals

## Model 0: Baseline model with all 5 indicators, no correlated errors
drive.model0 <- 'Drive =~ DRSAFE + DRCOMF + DRRELY + DREASY + DRPERC'
fit.drive.model0 <- cfa(drive.model0, data=data, missing='fiml')
summary(fit.drive.model0, fit.measures=TRUE, standardized=TRUE)
#Model fit: chi-square(N = 953, df = 5) = 10.634, CFI = 0.997, TLI = 0.995, RMSEA = 0.034, SRMR = 0.011
#   Ridiculously good model fit
#Factor loadings: All indicators have standardized loadings of > 0.60!
mi.drive.model0 <- modindices(fit.drive.model0)
mi.drive.model0 #super low MIs, so no need for correlated errors; can't improve model fit much over curr


#Have lavaan calculate R^2 for the indicators in the final model specifications
inspect(fit.walk.model1, 'r2')
inspect(fit.pt.model0, 'r2')
inspect(fit.rh.model0, 'r2')
inspect(fit.drive.model0, 'r2')

### ESTIMATE AND APPEND FACTOR SCORES
# With continuous indicators, the possible options for method = are "regression" or "Bartlett"
data$WALK_LV <- lavPredict(fit.walk.model1, type = "lv", method = "regression")
data$PT_LV <- lavPredict(fit.pt.model0, type = "lv", method = "regression")
data$RH_LV <- lavPredict(fit.rh.model0, type = "lv", method = "regression")
data$DRIVE_LV <- lavPredict(fit.drive.model0, type = "lv", method = "regression")
write.csv(data, "data/individual_new_LV.csv")


### SIMULTANEOUS ESTIMATION?
full.model <- ' Walk  =~ WSAFE + WCOMF + WRELY + WEASY + WPERC
                WSAFE ~~ WRELY
                PT =~ PTSAFE + PTCOMF + PTRELY + PTEASY + PTPERC
                RH =~ CARSAFE + CARCOMF + CARRELY + CAREASY + CARPERC
                Drive =~ DRSAFE + DRCOMF + DRRELY + DREASY + DRPERC'

fit <- cfa(full.model, data=data, missing="fiml")
summary(fit, fit.measures=TRUE, standardized=TRUE)
#Model fit is poor: chi-square(2003, 163) = 3967.0, CFI = 0.785, TLI = 0.750, RMSEA = 0.108, SRMR = 0.07
#Currently the model is using FIML to fill in for more than half of the sample missing data on the drive
#   personal car indicators -- this is extremely suspect!
```

```
fit.listwisedelete <- cfa(full.model, data=data) #use default of listwise deletion
summary(fit.listwisedelete, fit.measures=TRUE, standardized=TRUE)
#Model fit is still poor: chi-square(953, 163) = 2381.3, CFI = 0.783, TLI = 0.747, RMSEA = 0.120, SRMR =
#I suspect there is correlation among the error terms for the same indicators across modes; we can chec
mi.fit.listwisedelete <- modindices(fit.listwisedelete)
mi.fit.listwisedelete
#Modification indices:
#  Thankfully, we see very few low cross-loadings: RH =~  PTSAFE (64.7), which means our factor structu
#  However, we see a lot of correlated error terms: as expected WSAFE ~~  PTSAFE (192.7), WCOMF ~~  PTC
#      WEASY ~~ PTEASY (131.9)

#But in general, the results above suggest that we should estimate/extract the factor scores for each
#individual from the separate CFAs rather than from the combined measurement model
```