# Missing Data for Nicole

Michelle Byrne

Check missing data patterns and MCAR.

My example data I'm working with here is longitudinal (panel) data with four time points, and we'll need to work with both long and wide format. To start by testing if the data is MCAR, I'll need it to be long format. This is because it's hard to test MCAR if you have separate variables that are too colinear, and repeated measures from the same person usually are. Making this long format allows each row (i.e., time point) to have truly separate observed variables and test MCAR.

But then consider carefully what variables you want imputed (for example, if a repeated measure was not collected at all during one wave/time point, you may not want it imputed). For that reason, once we're ready to actually impute the data, we'll switch back to wide format.

```
workdir='C:/Users/michelle/Dropbox/academic/collaboration/giuliani_missing/'
data <- read.csv(file.path(workdir,"AllData_wide_v3_excl.csv", fsep="")) # Start here with wide format to see how it's moved to long format
library(panelr)
```

```
## Warning: package 'panelr' was built under R version 3.6.3
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'panelr'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
data_long <- long_panel(data, prefix = "T", label_location = "beginning", begin = 1, end = 4)
```

Remember as a first step to think of the possibility that the missing values in your dataset could theoretically be dependent on their missingness - in other words, the missing data could have significantly different values than the non-missing data, if we were able to know what the missing values were. We can't know that, so whatever you decide is only an assumption. If you don't think there is a good reason why this might be the case, you could assume the data is missing at random. Then you can run a test on top of that to see if it is also missing *completely* at random, which means the missingness is also not related to the other observed variables (that we do know the values of).

The MissMech package can test for MCAR and runs both parametric and non-parametric tests.

Info: http://www.jstatsoft.org/v56/i06/ (http://www.jstatsoft.org/v56/i06/)

```
library(MissMech)

#First, get an overview of how much missing data you have for each variable (and you'll want to report N and/or percentage of missing for each variable)
Missinginfo_long <- OrderMissing(data_long, del.lesscases = 0) #You can change del.lesscases if you want a variable with more than x missing values to be deleted from the dataset.

summary(Missinginfo_long) #Remember if this is for the long format data. If you want to just see how many are missing for each time point variable, change the dataset in this function back to the wide one (probably more useful for reporting).
```

```
##        id      wave         sub      condition  gender         age
##  1      :  4   1:103   1001   :  4   0:200      0  : 68   36     : 52
##  10     :  4   2:103   1002   :  4   1:212      1  :308   39     : 52
##  100    :  4   3:103   1003   :  4            NA's: 36   35     : 44
##  101    :  4   4:103   1004   :  4                       40     : 44
##  102    :  4           1005   :  4                       41     : 32
##  103    :  4           1006   :  4                       (Other):184
##  (Other):388           (Other):388                       NA's   :  4
##    ethnicity       LCLNC_striatum       LCLNC_vmpfc         RCLC_dlpfc
##   0     :336    0.74723590:  8     0.542058945:  8   -0.009449408:  4
##   1     : 28   -0.01381681:  4    -0.022824374:  4   -0.018887077:  4
##   6     : 16   -0.03049383:  4    -0.029416798:  4   -0.027481690:  4
##   2     : 12   -0.06878383:  4    -0.041777771:  4   -0.055349381:  4
##   3     :  8   -0.07329476:  4    -0.046066853:  4   -0.056318798:  4
##  (Other):  8   (Other)    :328    (Other)     :328   (Other)     :332
##  NA's   :  4   NA's       : 60    NA's        : 60   NA's        : 60
##          RCLC_IFG          RCLC_dacc          RCLC_vmpfc
##  -0.009562889:  4   -0.002671357:  4   -0.007318982:  4
##  -0.036576266:  4   -0.013088338:  4   -0.083939536:  4
##  -0.038043520:  4   -0.013793688:  4   -0.090048814:  4
##  -0.039604656:  4   -0.060225819:  4   -0.100193828:  4
##  -0.051841465:  4   -0.064147823:  4   -0.106396197:  4
##  (Other)     :332   (Other)     :332   (Other)     :332
##  NA's        : 60   NA's        : 60   NA's        : 60
##       RCLC_Lparahip        RCLC_Lifg        RCLC_Lsupra
##  -0.011398216:  4   -0.007187735:  4   -0.06252652:  4
##  -0.013299922:  4   -0.138466311:  4   -0.15016707:  4
##  -0.015496031:  4   -0.185287938:  4   -0.22274314:  4
##  -0.036103453:  4   -0.202465450:  4   -0.22856910:  4
##  -0.070725424:  4   -0.293391171:  4   -0.26654001:  4
##  (Other)     :332   (Other)     :332   (Other)    :332
##  NA's        : 60   NA's        : 60   NA's       : 60
##        RCLC_Lcereb          RCLC_na          RCLC_Rsupra
##  -0.034521754:  4   -0.012654899:  4   -0.007541319:  4
##  -0.044274783:  4   -0.034664621:  4   -0.010057313:  4
##  -0.058275625:  4   -0.093410205:  4   -0.012385620:  4
##  -0.058506077:  4   -0.121920979:  4   -0.030012352:  4
##  -0.058561510:  4   -0.138900107:  4   -0.031528568:  4
##  (Other)     :332   (Other)     :332   (Other)     :332
##  NA's        : 60   NA's        : 60   NA's        : 60
##     RCLC_RpostmedFront       bmi          unhelCrv        helCrv
##  -0.008622812:  4      30.62139:  2   1.391905:  2   1.000000:  8
##  -0.049828218:  4      22.74189:  1   1.494286:  2   1.066667:  6
##  -0.058109016:  4      23.65890:  1   1.630476:  2   1.133333:  5
##  -0.137461571:  4      23.86597:  1   1.790476:  2   1.200000:  4
##  -0.183359167:  4      24.13467:  1   2.001429:  2   1.344444:  4
##  (Other)     :332      (Other) :229   (Other) :300   (Other) :283
##  NA's        : 60      NA's    :177   NA's    :102   NA's    :102
##     unhelLike       helLike        HEItotal          KCal
##  2.607619:  3   2.273333:  3   67.51103:  3    203.2500:  1
##  2.752857:  3   2.284444:  3   49.97599:  2    342.7693:  1
##  2.186667:  2   2.424444:  3   15.87570:  1    460.6819:  1
##  2.305238:  2   3.348889:  3   30.19267:  1    532.1559:  1
##  2.325238:  2   2.086667:  2   31.23568:  1    535.6765:  1
##  (Other) :288   (Other) :286   (Other) :253   (Other)  :257
##  NA's    :112   NA's    :112   NA's    :151   NA's     :150
##         FVavg            empty
##  2.50000000: 11   20.0000000: 21
##  5.00000000:  6    0.0000000:  7
##  1.25000000:  4   13.6550455:  3
##  3.36740216:  3    5.5055010:  2
##  0.00000000:  2    0.2500976:  1
##  (Other)   :236   (Other)   :228
##  NA's      :150   NA's      :150
```

```
Missinginfo <- OrderMissing(data, del.lesscases = 0)
summary(Missinginfo)
```

```
##       sub         condition          gender            age
##  Min.   :1001   Min.   :0.0000   Min.   :0.0000   Min.   :33.00
##  1st Qu.:1028   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:36.00
##  Median :1054   Median :1.0000   Median :1.0000   Median :39.00
##  Mean   :1054   Mean   :0.5146   Mean   :0.8191   Mean   :39.17
##  3rd Qu.:1080   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:41.75
##  Max.   :1105   Max.   :1.0000   Max.   :1.0000   Max.   :46.00
##                                  NA's   :9        NA's   :1
##    ethnicity      LCLNC_striatum      LCLNC_vmpfc
##  Min.   :0.0000   Min.   :-1.90128   Min.   :-2.323590
##  1st Qu.:0.0000   1st Qu.:-0.32763   1st Qu.:-0.276074
##  Median :0.0000   Median : 0.04495   Median : 0.007598
##  Mean   :0.5098   Mean   :-0.01796   Mean   :-0.036671
##  3rd Qu.:0.0000   3rd Qu.: 0.34813   3rd Qu.: 0.234072
##  Max.   :6.0000   Max.   : 1.26207   Max.   : 1.653087
##  NA's   :1        NA's   :15         NA's   :15
##    RCLC_dlpfc          RCLC_IFG          RCLC_dacc
##  Min.   :-0.77451   Min.   :-0.72356   Min.   :-0.56345
##  1st Qu.:-0.03445   1st Qu.: 0.04185   1st Qu.: 0.01049
##  Median : 0.17532   Median : 0.23153   Median : 0.18647
##  Mean   : 0.20178   Mean   : 0.27461   Mean   : 0.25973
##  3rd Qu.: 0.46018   3rd Qu.: 0.52565   3rd Qu.: 0.47133
##  Max.   : 1.69886   Max.   : 1.67390   Max.   : 1.53164
##  NA's   :15         NA's   :15         NA's   :15
##    RCLC_vmpfc      RCLC_Lparahip        RCLC_Lifg       RCLC_Lsupra
##  Min.   :-1.6684   Min.   :-1.31034   Min.   :-1.379   Min.   :-0.5217
##  1st Qu.:-0.2297   1st Qu.:-0.02065   1st Qu.: 0.610   1st Qu.: 0.4238
##  Median : 0.1892   Median : 0.30061   Median : 1.452   Median : 0.8598
##  Mean   : 0.1033   Mean   : 0.29709   Mean   : 1.528   Mean   : 0.8467
##  3rd Qu.: 0.4214   3rd Qu.: 0.63415   3rd Qu.: 2.236   3rd Qu.: 1.1695
##  Max.   : 2.1945   Max.   : 1.83934   Max.   : 4.604   Max.   : 3.0482
##  NA's   :15        NA's   :15         NA's   :15       NA's   :15
##   RCLC_Lcereb         RCLC_na         RCLC_Rsupra      RCLC_RpostmedFront
##  Min.   :-1.3928   Min.   :-1.1772   Min.   :-0.8638   Min.   :-0.9346
##  1st Qu.:-0.0747   1st Qu.:-0.1005   1st Qu.: 0.1278   1st Qu.: 0.2772
##  Median : 0.2600   Median : 0.3009   Median : 0.6172   Median : 0.5666
##  Mean   : 0.2741   Mean   : 0.3618   Mean   : 0.5820   Mean   : 0.6077
##  3rd Qu.: 0.6318   3rd Qu.: 0.7419   3rd Qu.: 0.9231   3rd Qu.: 0.8462
##  Max.   : 1.9305   Max.   : 1.9742   Max.   : 2.2889   Max.   : 2.2174
##  NA's   :15        NA's   :15        NA's   :15        NA's   :15
##      T1bmi           T2bmi           T4bmi         T1unhelCrv
##  Min.   :23.87   Min.   :24.13   Min.   :22.74   Min.   :1.102
##  1st Qu.:28.71   1st Qu.:28.74   1st Qu.:27.72   1st Qu.:2.055
##  Median :30.73   Median :31.03   Median :30.62   Median :2.410
##  Mean   :31.33   Mean   :31.38   Mean   :31.06   Mean   :2.442
##  3rd Qu.:33.52   3rd Qu.:33.69   3rd Qu.:33.71   3rd Qu.:2.803
##  Max.   :41.96   Max.   :41.75   Max.   :43.90   Max.   :4.229
##  NA's   :9       NA's   :19      NA's   :46
##     T1helCrv      T1unhelLike      T1helLike       T1HEItotal
##  Min.   :1.000   Min.   :1.586   Min.   :1.116   Min.   :15.88
##  1st Qu.:1.481   1st Qu.:2.430   1st Qu.:2.289   1st Qu.:48.29
##  Median :1.969   Median :2.693   Median :2.598   Median :57.09
##  Mean   :2.098   Mean   :2.688   Mean   :2.646   Mean   :57.59
##  3rd Qu.:2.536   3rd Qu.:2.960   3rd Qu.:3.018   3rd Qu.:66.37
##  Max.   :3.920   Max.   :3.674   Max.   :3.711   Max.   :88.70
##                  NA's   :2       NA's   :2       NA's   :3
##      T1KCal          T1FVavg         T1empty         T2unhelCrv
##  Min.   : 203.2   Min.   :0.000   Min.   : 0.000   Min.   :1.069
##  1st Qu.:1707.0   1st Qu.:1.424   1st Qu.: 7.342   1st Qu.:1.711
##  Median :2009.6   Median :2.500   Median :12.672   Median :2.157
##  Mean   :2112.3   Mean   :2.541   Mean   :11.664   Mean   :2.186
##  3rd Qu.:2604.5   3rd Qu.:3.501   3rd Qu.:16.014   3rd Qu.:2.576
##  Max.   :4054.0   Max.   :5.000   Max.   :20.000   Max.   :3.855
##  NA's   :3        NA's   :3       NA's   :3        NA's   :20
##     T2helCrv      T2unhelLike      T2helLike       T2HEItotal
##  Min.   :1.000   Min.   :1.497   Min.   :1.800   Min.   :31.24
##  1st Qu.:1.400   1st Qu.:2.302   1st Qu.:2.284   1st Qu.:47.43
##  Median :1.989   Median :2.638   Median :2.588   Median :56.84
##  Mean   :1.986   Mean   :2.602   Mean   :2.630   Mean   :57.62
##  3rd Qu.:2.453   3rd Qu.:2.920   3rd Qu.:2.862   3rd Qu.:65.77
##  Max.   :3.613   Max.   :3.434   Max.   :3.822   Max.   :85.63
##  NA's   :20      NA's   :23      NA's   :23      NA's   :22
##      T2KCal          T2FVavg           T2empty         T3unhelCrv
##  Min.   : 535.7   Min.   :0.02594   Min.   : 0.000   Min.   :1.174
##  1st Qu.:1233.3   1st Qu.:1.72346   1st Qu.: 7.868   1st Qu.:1.819
##  Median :1785.2   Median :2.59872   Median :11.479   Median :2.252
##  Mean   :1796.9   Mean   :2.59359   Mean   :11.703   Mean   :2.224
##  3rd Qu.:2216.5   3rd Qu.:3.53147   3rd Qu.:16.461   3rd Qu.:2.598
##  Max.   :3345.4   Max.   :5.00000   Max.   :20.000   Max.   :3.662
##  NA's   :21       NA's   :21        NA's   :21       NA's   :39
##     T3helCrv      T3unhelLike      T3helLike       T3HEItotal
```

```
##  Min.   :1.000   Min.   :1.349   Min.   :1.664   Min.   :38.57
##  1st Qu.:1.451   1st Qu.:2.360   1st Qu.:2.224   1st Qu.:47.66
##  Median :1.941   Median :2.634   Median :2.538   Median :58.96
##  Mean   :2.016   Mean   :2.617   Mean   :2.600   Mean   :58.31
##  3rd Qu.:2.487   3rd Qu.:2.920   3rd Qu.:2.890   3rd Qu.:67.38
##  Max.   :3.536   Max.   :3.668   Max.   :3.840   Max.   :77.86
##  NA's   :39      NA's   :41      NA's   :41      NA's   :60
##     T3KCal          T3FVavg          T3empty         T4unhelCrv
##  Min.   : 957.2   Min.   :0.1518   Min.   : 0.000   Min.   :1.129
##  1st Qu.:1362.2   1st Qu.:1.3636   1st Qu.: 9.345   1st Qu.:1.660
##  Median :1740.8   Median :2.7367   Median :13.230   Median :2.190
##  Mean   :1888.6   Mean   :2.5676   Mean   :12.268   Mean   :2.162
##  3rd Qu.:2221.4   3rd Qu.:3.3877   3rd Qu.:15.684   3rd Qu.:2.551
##  Max.   :3913.0   Max.   :5.0000   Max.   :20.000   Max.   :3.713
##  NA's   :60       NA's   :60       NA's   :60       NA's   :43
##    T4helCrv      T4unhelLike      T4helLike       T4HEItotal
##  Min.   :1.000   Min.   :1.392   Min.   :1.758   Min.   :30.19
##  1st Qu.:1.386   1st Qu.:2.317   1st Qu.:2.164   1st Qu.:53.12
##  Median :1.929   Median :2.620   Median :2.440   Median :58.12
##  Mean   :1.925   Mean   :2.608   Mean   :2.561   Mean   :59.01
##  3rd Qu.:2.494   3rd Qu.:2.855   3rd Qu.:2.791   3rd Qu.:64.35
##  Max.   :3.067   Max.   :3.846   Max.   :3.880   Max.   :87.67
##  NA's   :43      NA's   :46      NA's   :46      NA's   :66
##     T4KCal          T4FVavg          T4empty
##  Min.   : 342.8   Min.   :0.4039   Min.   : 0.000
##  1st Qu.:1318.3   1st Qu.:1.9751   1st Qu.: 6.924
##  Median :1748.7   Median :2.6734   Median :12.727
##  Mean   :1786.0   Mean   :2.7918   Mean   :11.662
##  3rd Qu.:2225.1   3rd Qu.:3.5798   3rd Qu.:15.291
##  Max.   :3903.2   Max.   :4.9779   Max.   :20.000
##  NA's   :66       NA's   :66       NA's   :66
```

```
write.csv (summary(Missinginfo), file.path(workdir,"MissingInfo.csv", fsep=""))
```

Now you're ready to test if your data is MCAR. If the tests are significant, it is not.

NOTE: If your data matrix is singular (you'll get the dread pirate "system is computationally singular" message), TestMCARNormality won't work. Try removing extra variables that may be colinear. Definitely never have variables in your dataset that are derived from raw variables (e.g., transformed or totaled. Impute the raw data first and then re-transform or re-total). If you have longitudinal data and try this one wide format, it probably won't work.

```
data.nummat <- data.matrix(data_long, rownames.force = NA)
data.out <- TestMCARNormality(data.nummat)
print(data.out)
```

```
## Call:
## TestMCARNormality(data = data.nummat)
##
## Number of Patterns:  7
##
## Total number of cases used in the analysis:  360
##
##  Pattern(s) used:
##           id   wave   sub   condition   gender   age   ethnicity
## group.1    1      1     1           1        1     1           1
## group.2    1      1     1           1        1     1           1
## group.3    1      1     1           1        1     1           1
## group.4    1      1     1           1        1     1           1
## group.5    1      1     1           1        1     1           1
## group.6    1      1     1           1        1     1           1
## group.7    1      1     1           1       NA     1           1
##           LCLNC_striatum   LCLNC_vmpfc   RCLC_dlpfc   RCLC_IFG   RCLC_dacc
## group.1                1             1            1          1           1
## group.2                1             1            1          1           1
## group.3                1             1            1          1           1
## group.4                1             1            1          1           1
## group.5                1             1            1          1           1
## group.6               NA            NA           NA         NA          NA
## group.7               NA            NA           NA         NA          NA
##           RCLC_vmpfc   RCLC_Lparahip   RCLC_Lifg   RCLC_Lsupra
## group.1            1               1           1             1
## group.2            1               1           1             1
## group.3            1               1           1             1
## group.4            1               1           1             1
## group.5            1               1           1             1
## group.6           NA              NA          NA            NA
## group.7           NA              NA          NA            NA
##           RCLC_Lcereb   RCLC_na   RCLC_Rsupra   RCLC_RpostmedFront   bmi
## group.1             1         1             1                    1     1
## group.2             1         1             1                    1    NA
## group.3             1         1             1                    1     1
## group.4             1         1             1                    1    NA
## group.5             1         1             1                    1    NA
## group.6            NA        NA            NA                   NA     1
## group.7            NA        NA            NA                   NA    NA
##           unhelCrv   helCrv   unhelLike   helLike   HEItotal   KCal
## group.1          1        1           1         1          1      1
## group.2          1        1           1         1         NA     NA
## group.3          1        1           1         1         NA     NA
## group.4         NA       NA          NA        NA         NA     NA
## group.5          1        1           1         1          1      1
## group.6          1        1           1         1          1      1
## group.7         NA       NA          NA        NA         NA     NA
##           FVavg   empty   Number of cases
## group.1       1       1               180
## group.2      NA      NA                28
## group.3      NA      NA                21
## group.4      NA      NA                58
## group.5       1       1                39
## group.6       1       1                13
## group.7      NA      NA                21
##
##
##     Test of normality and Homoscedasticity:
##   ------------------------------------------
##
## Hawkins Test:
##
##     P-value for the Hawkins test of normality and homoscedasticity:  1.912707e-11
##
##     Either the test of multivariate normality or homoscedasticity (or both) is rejected.
##     Provided that normality can be assumed, the hypothesis of MCAR is
##     rejected at 0.05 significance level.
##
## Non-Parametric Test:
##
##     P-value for the non-parametric test of homoscedasticity:  0.003404727
##
##     Hypothesis of MCAR is rejected at  0.05 significance level.
##     The multivariate normality test is inconclusive.
```

```
# Here's some code to quickly check what's mega correlated in your dataset if you get the singular thing:
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.6.1
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

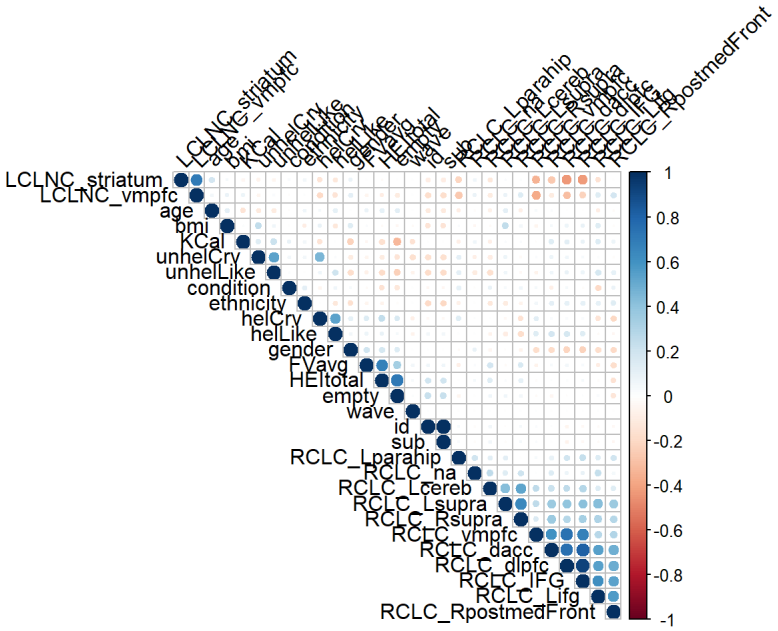
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.1
```

```
## corrplot 0.84 loaded
```

```
corrs <- rcorr(as.matrix(data_long))
corrplot(corrs$r, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```




If it's not MCAR, it is ok, but you should figure out what's going on and report differences in observed variables between your missing and non-missing main outcomes of interest. (main predictor and outcome variables). Here's an example. Maybe helCrv missingness is dependent on age:

```
data_long_missing <- data_long

# Group the missingness of whatever variable you want:
for (i in 1:nrow(data_long_missing)) {
  if (is.na(data_long_missing$helCrv[i])) {
  data_long_missing$helcrv_missing[i] = 1
} else {
  data_long_missing$helcrv_missing[i] = 0
}
}
```

```
## Warning: Unknown or uninitialised column: 'helcrv_missing'.
```

```
# ANOVA to check if age significant differs between missing helcrv and not missing helcrv:
helcrv_aov <- aov(age ~ helcrv_missing, data = data_long_missing)
summary (helcrv_aov) # Spoiler, it doesn't.
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## helcrv_missing   1      2   2.084   0.174  0.677
## Residuals      406   4863  11.977
## 4 observations deleted due to missingness
```

```
# You could also check this in wide format, too, to see if a variable's missingness at a certain time point is dependent on
 some other observed variable.
```

{r impute} # # Multiple Imputation using Amelia and Zelig. P

{r em models} # # Linear models with Single Imputed data EM

```{r fiml models}

# FIML using Lavaan

# fit <- sem(model, data, missing='fiml')

library("lavaan")

# Create descriptive model object

model1 <- '

# Note that fixed.x=FALSE in the sem may eliminate need to estimate variances and covariances of predictors (??)

fit1 <- sem(model=model1, data=data, missing='fiml', fixed.x=FALSE)

summary(fit1, fit.measures=TRUE, rsquare=TRUE, standardize=TRUE)

# To select the best fitting model, The model with the smallest AIC and BIC is chosen.

#Reminder: CFI>0.9, TLI>0.9, RMSEA<0.08, SRMR<0.08 (Marsh, et al. (2010). Psychol Assess 22:471)
```