Lösung Zettel 6

2023-06-02

Aufgabe 1

Wenn die Aussage falsch ist, begründe wieso.

Wahr oder falsch?

Im Model $y_1 = \alpha + \gamma x_1 + \zeta_1$ y_1 steht dafür, dass es nur eine Beobachtung gibt.

Antwort: Falsch. Die 1 in dem Modell steht jeweils für die erste Variable y. Wir können in SEM Modellen mehrere y-Variablen haben.

Wahr oder falsch?

Die Regressionskoeffizienten einer gewöhnlichen Regression mit kleinsten Quadraten entsprechen der einer Maximum-Likelihood Schätzung, aber die Residualvarianz unterscheidet sich zwischen den beiden.

Antwort: Wahr. (Begründung hier nicht gebraucht da wahr, aber der Grund dafür ist der dass der KQ Schätzer der Varianz die Form $\hat{\sigma} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{N-K}$, mit N = Anzahl der Beobachtungen, K = Anzahl der Parameter ($\boldsymbol{\beta}$) hat. Wir wissen aus der Theorie, dass die Schätzer das korrekte Ergebnis liefert. der ML Varianz-Schätzer sieht hingegen wie folgt aus $\hat{\sigma} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{N}$ und entspricht somit nicht dem "korrekten" Ergebnis und ist damit Verzerrt (engl. Biased).)

Wahr oder falsch?

In einem Modell gibt es zwei Regressions Koeffizienten γ weil es zwei exogene Variablen gibt. Bei demselben Modell würde sich dies auch nicht ändern, wenn wir den Stichprobenumfang erhöhen.

Antwort: Wahr. (Begründung hier nicht gebraucht da wahr, aber der Grund dafür ist, dass die Anzahl der Regressions Koeffizienten, also unsere Einfluss-parameter, nichts damit zu tun haben wie viele Beobachtungen wir haben. Bsp.: Eisverkauf. Wir schreiben uns die Temperatur und den Preis pro Kugel auf. Wir haben als zwei Koeffzienten. Jetzt notieren wir die Werte für 30 Tage. Wir haben also nun 30 Beobachtungen. Schreiben wir jetzt noch eine weitere Woche auf, dann erhöht sich die Anzahl der Beobachtungen auf 37 Tage, aber die Anzahl der Parameter Temperatur und Preis bleibt gleich.)

Aufgabe 2

Vorbreitung:

Wir wollen SEMs (**S**tructural **E**quation **M**odels = **S**trukturgleichungs**m**odelle) bauen. Dafür brauchen wir das `lavaan` R-Package. UND WICHTIG! Wir setzen einen seed.

Hinweis zu SEM Syntax:

1. `=~` Measurement Modell, wir bauen uns eine neue Variable.
2. `~` Regressionsmodell, wir machen eine klassische Regression.
3. `~~` Covarianz. Wir Modellieren eine spezifische Covarianz, inkl. Correlation.
   3.1. Sonderfall: `~~ 0` wir setzen eine Covarianz, inkl Correlation, gleich Null und entfernen sie aus dem Modell.

```
library(lavaan)
```

```
## This is lavaan 0.6-15
## lavaan is FREE software! Please report any bugs.
```

```
set.seed(42)
```

a)

Lade zunächst den Datensatz ''PoliticalDemocracy' aus dem Paket lavaan.

```
data("PoliticalDemocracy")
```

Hinweis: Ihr werdet in der Klausur ein Datensatz gestellt bekommen, ihr müsst diesen über die Funktion zum Einlesen der Daten nutzen.

b)

Wir bauen ein Measurement Modell, welche erst einmal aus den erklärenden Variablen x_1, x_2 und x_3 eine neue latente Größe "misst" die wir `Var1` nennen. Erst einmal nehmen wir keine speziellen Annahmen an über die Correlationen.

```
m_gleichung <- 'Var1 =~ x1 + x2 + x3'
m_b <- sem(m_gleichung, data=PoliticalDemocracy)
```

Erklärung: Wir schreiben also die Modellgleichung auf. Die Syntax des Ganzen ist genau wie bei einer normalen Regression wie mit `lm()`, nur nutzen wir `=~` was R sagt, dass wir eine NEUE Variable bauen.

c)

Wir erweitern nun das erste Modell, um eine weitere Measurement Modellgleichung, welche die Größen $y1 + y2 + y3 + y4$ nutzt um eine weitere neue Größe `Var2` zu bemessen. Füge dem Modell noch eine Regressions Komponente `y ~ Variable 1 + ... + Variable k` hinzu, bei der du `Var1` nutzt um `Var2` zu erklären.

```
m_gleichung <- 'Var1 =~ x1 + x2 + x3
                Var2 =~ y1 + y2 + y3 + y4
                Var2 ~ Var1'
m_c <- sem(m_gleichung, data=PoliticalDemocracy)
```

Erklärung: Wir schreiben also die Modellgleichung wie in (b) auf, schreiben dann nur unter der ersten Gleichung die anderen zwei, wobei ein Measurement Modell `=~` und eine Regression `~` betrachtet wird .

d)

Wir erweitern nun das letzte Modell, um noch eine weitere Measurement Modellgleichung, welche die Größen $y5+y6+y7+y8$ nutzt um eine weitere neue Größe `Var3` zu bemessen. Füge dem Modell noch eine Regressions Komponente `y ~ Variable 1 + ... + Variable k` hinzu, bei der du `Var1` und `Var2` nutzt um `Var3` zu erklären.

```
m_gleichung <- 'Var1 =~ x1 + x2 + x3
                Var2 =~ y1 + y2 + y3 + y4
                Var3 =~ y5+ y6 + y7 + y8
                Var3 ~ Var1 + Var2'
m_d <- sem(m_gleichung, data=PoliticalDemocracy)
```

e)

Wir erweitern nun das letzte Modell, um noch eine explizite residuale Korrelationsstruktur. Unterstellte hierbei die folgende Korrelationsstruktur zwischen:

- `y1` und `y5`
- `y2` und `y4` und `y6`
- `y3` und `y7`
- `y4` und `y8`
- `y5` und `y8`

```
m_gleichung <- 'Var1 =~ x1 + x2 + x3
                Var2 =~ y1 + y2 + y3 + y4
                Var3 =~ y5+ y6 + y7 + y8
                Var3 ~ Var1 + Var2
                y1 ~~ y5
                y2 ~~ y4 + y6
                y3 ~~ y7
                y4 ~~ y8
                y5 ~~ y8'
m_e <- sem(m_gleichung, data=PoliticalDemocracy)
```

f)

Vergleiche die Modelle aus d) und e). Wenn du dich für ein Modell entscheiden müsstest, welches würdest du nehmen? Begründe wieso?

Entscheidung:

- Das zweite Modell scheint sich besser dafür zu eignen die unterliegenden Daten zu analysieren.

Begründung:

1. Informationskriterien: Alle drei Kriterien (AIC, BIC und SABIC) zeigen geringere Werte für das zweite Modell. (*Hinweis:* Kleine Informationskriterien = Besser.)
2. Root Mean Square Error of Approximation: Der Fehler der Approximation spricht für das zweite Modell, da diese Werte geringer ausfallen.(*Hinweis:* Kleine RMSE = Besser.)

3. Standardized Root Mean Square Residual: Auch hier sind die Werte geringer für das zweite Modell und somit besser.
4. User Model versus Baseline Model: `CFI` und `TLI` weisen bessere Werte für das zweite Modell auf. (*Hinweis:* Diese Größen vergleichen nur mit dem BASELINE Modell. Da die beiden betrachteten Modelle die selben Varialen nutzen, sind die Baseline Modelle für beide gleich, sodass diese Werte auch zum Vergleich beider unserer Modelle erlaubt.)

```
summary(m_d, fit.measures=T)
```

```
## lavaan 0.6.15 ended normally after 39 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        25
##
##   Number of observations                            75
##
## Model Test User Model:
##
##   Test statistic                                72.462
##   Degrees of freedom                                41
##   P-value (Chi-square)                           0.002
##
## Model Test Baseline Model:
##
##   Test statistic                               730.654
##   Degrees of freedom                                55
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.953
##   Tucker-Lewis Index (TLI)                       0.938
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1564.959
##   Loglikelihood unrestricted model (H1)      -1528.728
##
##   Akaike (AIC)                                3179.918
##   Bayesian (BIC)                              3237.855
##   Sample-size adjusted Bayesian (SABIC)       3159.062
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.101
##   90 Percent confidence interval - lower         0.061
##   90 Percent confidence interval - upper         0.139
##   P-value H_0: RMSEA <= 0.050                    0.021
##   P-value H_0: RMSEA >= 0.080                    0.827
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.055
```

```
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   Var1 =~
##     x1               1.000
##     x2               2.182    0.139   15.714    0.000
##     x3               1.819    0.152   11.956    0.000
##   Var2 =~
##     y1               1.000
##     y2               1.354    0.175    7.755    0.000
##     y3               1.044    0.150    6.961    0.000
##     y4               1.300    0.138    9.412    0.000
##   Var3 =~
##     y5               1.000
##     y6               1.258    0.164    7.651    0.000
##     y7               1.282    0.158    8.137    0.000
##     y8               1.310    0.154    8.529    0.000
##
## Regressions:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   Var3 ~
##     Var1             0.453    0.220    2.064    0.039
##     Var2             0.864    0.113    7.671    0.000
##
## Covariances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   Var1 ~~
##     Var2             0.660    0.206    3.202    0.001
##
## Variances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##    .x1               0.082    0.020    4.180    0.000
##    .x2               0.118    0.070    1.689    0.091
##    .x3               0.467    0.090    5.174    0.000
##    .y1               1.942    0.395    4.910    0.000
##    .y2               6.490    1.185    5.479    0.000
##    .y3               5.340    0.943    5.662    0.000
##    .y4               2.887    0.610    4.731    0.000
##    .y5               2.390    0.447    5.351    0.000
##    .y6               4.343    0.796    5.456    0.000
##    .y7               3.510    0.668    5.252    0.000
##    .y8               2.940    0.586    5.019    0.000
##     Var1             0.448    0.087    5.169    0.000
##     Var2             4.845    1.088    4.453    0.000
##    .Var3             0.115    0.200    0.575    0.565
```

```
summary(m_e, fit.measures=T)
```

```
## lavaan 0.6.15 ended normally after 70 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        31
##
##   Number of observations                            75
##
## Model Test User Model:
##
##   Test statistic                                42.381
##   Degrees of freedom                                35
##   P-value (Chi-square)                           0.183
##
## Model Test Baseline Model:
##
##   Test statistic                               730.654
##   Degrees of freedom                                55
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.989
##   Tucker-Lewis Index (TLI)                       0.983
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1549.919
##   Loglikelihood unrestricted model (H1)      -1528.728
##
##   Akaike (AIC)                                3161.838
##   Bayesian (BIC)                              3233.680
##   Sample-size adjusted Bayesian (SABIC)       3135.976
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.053
##   90 Percent confidence interval - lower         0.000
##   90 Percent confidence interval - upper         0.103
##   P-value H_0: RMSEA <= 0.050                    0.436
##   P-value H_0: RMSEA >= 0.080                    0.217
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.046
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
```

```
##   Var1 =~
##     x1                1.000
##     x2                2.179    0.138   15.754    0.000
##     x3                1.818    0.152   11.968    0.000
##   Var2 =~
##     y1                1.000
##     y2                1.256    0.184    6.820    0.000
##     y3                1.045    0.152    6.880    0.000
##     y4                1.275    0.147    8.647    0.000
##   Var3 =~
##     y5                1.000
##     y6                1.234    0.165    7.477    0.000
##     y7                1.264    0.160    7.924    0.000
##     y8                1.341    0.170    7.912    0.000
##
## Regressions:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   Var3 ~
##     Var1             0.567    0.220    2.576    0.010
##     Var2             0.817    0.100    8.208    0.000
##
## Covariances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##  .y1 ~~
##    .y5               0.629    0.361    1.741    0.082
##  .y2 ~~
##    .y4               1.394    0.698    1.997    0.046
##    .y6               2.197    0.754    2.913    0.004
##  .y3 ~~
##    .y7               1.186    0.609    1.947    0.052
##  .y4 ~~
##    .y8               0.328    0.458    0.716    0.474
##  .y5 ~~
##    .y8              -0.653    0.351   -1.857    0.063
##   Var1 ~~
##     Var2             0.668    0.209    3.187    0.001
##
## Variances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##    .x1               0.081    0.019    4.177    0.000
##    .x2               0.120    0.070    1.727    0.084
##    .x3               0.467    0.090    5.179    0.000
##    .y1               1.896    0.450    4.216    0.000
##    .y2               7.470    1.384    5.395    0.000
##    .y3               5.202    0.963    5.401    0.000
##    .y4               3.114    0.754    4.131    0.000
##    .y5               2.273    0.496    4.581    0.000
##    .y6               4.398    0.821    5.356    0.000
##    .y7               3.589    0.704    5.095    0.000
##    .y8               2.321    0.609    3.809    0.000
##     Var1             0.449    0.087    5.175    0.000
##     Var2             4.909    1.114    4.407    0.000
##    .Var3             0.427    0.221    1.932    0.053
```

```
AIC(m_d, m_e)
```

```
##     df      AIC
## m_d 25 3179.918
## m_e 31 3161.838
```