Vertiefte Forschungsmethodik der Psychologie und Psychotherapie

II. Multivariate Statistische Verfahren

Dr. Leonard Pleschberger

Heinrich-Heine-Universität Düsseldorf

WS 25/26





Übersicht

II.1. Multivariate Statistik

II.2. Abhängigkeitsmodelle

- II.2.1. Multiple Regression
- II.2.2. Mehrebenenanalyse (HLM)
- II.2.3. Multivariate Varianzanalyse (MANOVA)

II.3. Interdependenzmodelle

- II.3.1. Exploratorische Faktorenanalyse (EFA)
- II.3.2. Konfirmatorische Faktorenanalyse (CFA)
- II.3.3. Clusteranalyse

II.4. Strukturgleichungsmodelle (SEM)

II.5. Meta-Analyse





II.1. Multivariate Statistik

- Bisher haben wir uns **univariat** mit einzelnen Merkmalen oder höchstens **bivariat** mit der Beziehung zweier Variablen beschäftigt.
- In der multivariaten Statistik betrachten wir nun die Beziehungen mehrerer Variablen gleichzeitig.
- Damit werden unsere statistischen Modelle größer und komplexer aber auch realitätsnäher, denn psychologische Konstrukte hängen selten nur von einer einzigen Einflussgröße ab.



II.2. Abhängigkeitsmodelle





II.2.1. Multiple Regression

Fragestellung: Wie kann der lineare Einfluss gleich mehrerer Prädiktoren X_1, \ldots, X_p auf eine Zielvariable Y erklärt werden?

Das Modell der multiplen (linearen) Regression lautet also

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + \varepsilon.$$

Hier bezeichnen

- b₀ den Achsenabschnitt (Intercept).
- b_1, \ldots, b_p die Koeffizienten sie bestimmen den individuellen Einfluss von X_1, \ldots, X_p .
- ullet arepsilon einen zufälligen Fehlerterm.

Ziel: Wir möchten die Koeffizienten b_0, \ldots, b_p mit Hilfe von R so schätzen, dass sie unsere Daten am besten fitten.





Multiple Regression: Theorie

Bei n Messungen mit 1 Zielvariable und p Prädikatoren ergibt sich

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

was wir in fetter Vektorschreibweise noch kompakter schreiben können als

$$Y = Xb + \varepsilon$$
.

Hier wird \mathbf{X} als Designmatrix bezeichnet. Nun kommen in jeder Zeile immer genau die gleichen Koeffizienten \mathbf{b} vor und der Rest besteht ausschließlich aus Messungen und Residuen.



Multiple Regression: Theorie

Wir wollen einen Schätzer $\hat{\mathbf{b}}$ bestimmen, sodass die vorhergesagten Werte $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$ möglichst nahe bei den gemessenen Werten \mathbf{Y} liegen. Dazu rechnen wir das Minimum der Summe der quadrierten Vorhersagefehler Q mittels Kurvendiskussion aus:

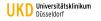
$$Q\left(\hat{\mathbf{b}}\right) = \left|\mathbf{Y} - \hat{\mathbf{Y}}\right|^2 = \left|\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}\right|^2 = \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}\right)^{\top} \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}\right)$$

Wie leiten nach dem Schätzer ab und setzen die Ableitung gleich null:

$$\frac{dQ}{d\hat{\mathbf{b}}} = -2\mathbf{X}^{\top}\mathbf{Y} + 2\mathbf{X}^{\top}\mathbf{X}\hat{\mathbf{b}} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}^{\top}\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}^{\top}\mathbf{Y}.$$

Die rechte Gleichung heißt *Normalengleichung*. Wir invertieren noch Matrix auf der linken Seite und erhalten den Schätzer

$$\hat{\mathbf{b}} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{Y}.$$





Beispiel Multiple Regression: Depressionsschwere

Fragestellung: Wie beeinflussen die Prädiktoren *Stresslevel* (X_1) , *Schlafqualität* (X_2) und *Soziale Unterstützung* (X_3) die Zielvariable *Depressionsschwere* (Y) linear?

Wir stellen das folgende multiple lineare Regressionsmodel auf:

Depression =
$$b_0 + b_1 \cdot \mathsf{Stress} + b_2 \cdot \mathsf{Schlaf} + b_3 \cdot \mathsf{Soziales} + \varepsilon$$

Dies schreiben wir als

$$Y = Xb + \varepsilon$$
.

- $Y (n \times 1)$: Vektor der Zielvariablen *Depression*.
- X ($n \times 4$): Designmatrix für Intercept, Stress, Schlaf und Soziales.
- **b** (4 × 1): Koeffizienten.
- ε ($n \times 1$): Residuen.

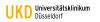




Beispiel Multiple Regression: Daten

Bei n = 8 Personen wurden die folgenden Daten erhoben:

ID	Stress	Schlaf	Soziales	Depression
1	80	4	2	30
2	65	6	3	22
3	40	8	5	15
4	90	3	1	35
5	55	7	4	18
6	70	5	2	25
7	85	4	2	29
8	60	6	3	21





Beispiel Multiple Regression: Modell in R

Wir laden die Daten in R und verwenden das Im()-Modell wie folgt:

```
# Import the data
data <- data.frame(
stress = c(80, 65, 40, 90, 55, 70, 85, 60),
sleep = c(4, 6, 8, 3, 7, 5, 4, 6),
social = c(2, 3, 5, 1, 4, 2, 2, 3),
depression = c(30, 22, 15, 35, 18, 25, 29, 21)
# Multiple lineare Regression
modell <- lm(depression ~ stress + sleep + social,
   data = data)
# Results
summary(modell)
```



Beispiel Multiple Regression: Ergebnisse in R

R liefert die folgende Ausgabe:

```
Residuals:
-0.30 0.65 0.30 1.15 0.20 -0.35 -0.85 -0.80
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
                     16.8919 3.440
(Intercept)
         58.1000
                                     0.0263 *
         -0.0900 0.1378 -0.653 0.5492
stress
sleep -5.8500 1.7250 -3.391 0.0275 *
social 1.4000 1.2546 1.116 0.3270
Signif. codes:
0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
```





Beispiel Multiple Regression: Interpretation

Des Weiteren:

```
Residual standard error: 0.9287 on 4 degrees of freedom
Multiple R-squared: 0.9889, Adjusted R-squared: 0.9806
F-statistic: 119.2 on 3 and 4 DF, p-value: 0.0002286
```

• Koeffizienten: Die Schätzungen ergeben die Regressionsgleichung

$$Depression = 58.10 - 0.1 \cdot Stress - 5.9 \cdot Schlaf + 1.4 \cdot Soziales.$$

Lediglich zwei Merkmale stehen signifikant zum Niveau $\alpha=5\%$ im Zusammenhang mit Depression (gekennzeichnet mit *): Das Intercept mit p=0.03<0.05 und der Schlaf mit p=0.03<0.05. Stress und soziale Unterstützung haben keinen signifikanten Einfluss in diesem Modell.



Multiple Regression: Interpretation

 Residuen: Die Werte geben die Abweichung zwischen den prognostizierten und den tatsächlichen Werten an. Die Vorzeichen sind so zu deuten: Bei -0.30 überschätzt das Modell den Wert, i.e. man muss vom Prognosewert 0.30 abziehen, um auf den tatsächlichen Messwert zu kommen. Analog deutet man ein positives Vorzeichen. Deren Durchschnitt ergibt den Residuen-Standardfehler 0.93 mit

$$\begin{aligned} & \mathsf{df} = \#\{\mathsf{Beobachtungen}\} - \#\{\mathsf{Pr\ddot{a}dikatoren}\} - \#\{\mathsf{Intercept}\} \\ &= 8 - 3 - 1 = 4 \end{aligned}$$

Freiheitsgraden. Die Residuen sind relativ klein \Rightarrow das Model fittet die Daten gut.



Beispiel Multiple Regression: Interpretation

 Bestimmtheitsmaße: Mit Pearson's Korrelationskoeffizienten r ergibt sich

$$R^2 = r^2 = 0.99.$$

Das sagt aus, dass das Modell extrem gut fittet. Unter Beachtung der Anzahl der Prädikatoren ergibt sich

$$R_{\rm adjusted}^2 = 0.98,$$

woraus ein sehr geringes Overfitting resultiert.

• Statistiken: Mit t-Tests wird die Signifikanz für jeden einzelnen Prädiktor getestet und ein F-Test klärt die Signifikanz des gesamten Regressionsmodells. Deshalb sind die empirischen t- und F-Statistikenaufgeführt. Der Wert F=119.2 zusammen mit p=0.0002 sagt aus, dass das Modell insgesamt hochsignifikant ist.



II.2.2. Mehrebenenanalyse (HLM)

- Die Mehrebenenanalyse (engl. Multilevel Modeling oder Hierarchical Linear Modeling, HLM) ist eine statistische Methode, die verwendet wird, wenn Daten hierarchisch oder geschachtelt organisiert sind – etwa Personen innerhalb von Gruppen, Patient:innen innerhalb von Kliniken, Messzeitpunkte innerhalb von Individuen, etc.
- Die Mehrebenenanalyse ist eine Erweiterung der multiplen Regression, die die Abhängigkeit von Beobachtungen innerhalb übergeordneter Gruppen berücksichtigt. Dabei werden gruppenspezifische Intercepts modelliert, um Unterschiede zwischen Gruppen zu erfassen. Das Residuum eines Patienten setzen sich aus dem individuellen Residuum und einem Gruppenresiduum zusammen.



Beispiel Mehrebenenanalyse: Depression in Kliniken

Fragestellung: Wie erstelle ich eine multiple Regression mit der Zusatzinformation, dass die Patienten in Untergruppen (wie Kliniken) aufgeteilt sind?

Level 1: Personenebene

- Zielvariable: Depressionswert via BDI-II (0-63)
- Prädiktoren: Stress im Alltag (0–100), Schlafqualität (1–10)

Level 2: Klinikebene (Untergruppen)

- Patienten stammen aus verschiedenen Kliniken (e.g. 5 Kliniken in Deutschland).
- Jede Klinik hat eigene Strukturen (Therapiephilosophie, Personal, Ressourcen).
- **Prädiktor**: Durchschnittliche Anzahl an Psychotherapeut:innen pro 10 Patienten (als Indikator für Therapieintensität).





Beispiel Mehrebenenanalyse: Depression in Kliniken

Gemeinsamkeit zur multiplen Regression: Der Depressionswert des *i*-ten Patienten aus der *j*-ten Klinik wird u.a. durch eine Linearkombination von durchschnittlichem Intercept, sowie den Prädikatoren Stress, Schlaf und Therapieintensität (gleich für Patienten der Klinik *j*) ermittelt:

 $Depression_{ij} = b_0 + b_1 \cdot Stress_{ij} + b_2 \cdot Schlaf_{ij} + b_3 \cdot Intensität_j + Fehlerterm$

Unterschied zur multiplen Regression: Durch die verfeinerte Struktur wird es möglich den Fehlerterm durch Residuen auf Personenebene ε_{ij} (Level 1) und Residuen auf Klinikebene κ_j (Level 2) zu modellieren.

 \Rightarrow In Vektorschreibweise ergibt sich für alle Patienten der Klinik j gerade

$$\mathbf{Y}_{j} = \mathbf{X}_{j}\mathbf{b} + \boldsymbol{\varepsilon} + \boldsymbol{\kappa}_{j}.$$





Beispiel Mehrebenenanalyse: Daten erzeugen

Erzeuge n normalverteilte Zufallszahlen mit Mittelwert mu und Standardabweichung s:

```
rnorm(n, mean = mu, sd = s) # Default: rnorm(n, 0, 1)
```

```
set.seed(12) # Reproducibility
N <- n_clinics * n_patients
# Clinic IDs
clinic <- factor(rep(1:n_clinics, each = n_patients))</pre>
# Level-2 predictor: Therapy intensity per clinic
therapy_intensity \leftarrow rep(c(1.5, 2.0, 2.5, 3.0, 3.5),
                   each = n_patients)
```



Beispiel Mehrebenenanalyse: Daten erzeugen

```
# Level-1 predictors (patient characteristics)
stress <- round(rnorm(N, mean = 60, sd = 15), 1)
           # Daily stress (0-100)
sleep \leftarrow round(rnorm(N, mean = 6.5, sd = 1.2), 1)
           # Sleep quality (1-10)
# True effects for simulation
b0 <- 30 # Baseline depression score
b1 <- 0.25  # Effect of stress (per stress point)
b2 <- -1.5  # Effect of sleep (per hour)
g01 <- -2.0 # Effect of th. intensity (level 2)
# Random intercepts for clinics (level 2)
u0 <- rnorm(n_clinics, 0, 2) # each clinic has a</pre>
                                  # different baseline
rand_intercepts <- rep(u0, each = n_patients)
```





Beispiel Mehrebenenanalyse: Daten erzeugen

```
# Simulate depression
depression <- b0 + b1 * stress + b2 * sleep + g01 *
  therapy_intensity + rand_intercepts + rnorm(N, 0, 3)
# Clinical and individual error

# Build final dataset
data <- data.frame(clinic, therapy_intensity, stress, sleep, depression)
head(data, 3)</pre>
```

Ausgabe:

	clinic	therapy_intensity	stress	sleep	depression
1	l 1	1.5	80.6	6.9	36.800
2	2 1	1.5	51.5	5.6	31.475
3	3 1	1.5	65.4	8.4	30.750





Beispiel Mehrebenenanalyse: Daten auswerten

Zunächst installieren wir einmal das Package für die Mehrebenenanalyse:

```
> install.packages("lme4")
```

- Zielvariable: depression.
- Prädikatoren: stress, sleep, therapy_intensity.
- (1 | clinic) \rightarrow Füfälliger Interceopt für jedes clinic.





```
Linear mixed model fit by REML ['lmerMod']
Formula:
depression ~ stress + sleep + therapy_intensity +
            (1 | clinic)
Data: data
REML criterion at convergence: 255.3
Scaled residuals:
    Min 1Q Median
                               3 Q
                                       Max
-1.86553 -0.66244 0.03221 0.68334 1.93486
```

- Die Residuen sollten idealerweise gleichmäßig um 0 verteilt sein.
- Median $\approx 0 \Rightarrow$ Die Modellfehler mitteln sich nahezu aus.
- Fazit: Das Modell passt die Daten gut an.





Abweichungen zwischen den Kliniken (Level 2) und auf Patientenebene (Level 1):

```
Random effects:
Groups Name Variance Std.Dev.
clinic (Intercept) 3.999 2.000
Residual 8.722 2.953
Number of obs: 50, groups: clinic, 5
```

- clinic (Intercept): Die Kliniken (Level 2) unterscheiden sich leicht im Grundniveau der Depression (zwischen-Gruppen-Varianz).
- Residual: Die gruppenunspezifische Reststreuung auf Patientenebene (Level 1); das, was durch die Prädikatoren nicht erklärt wird.





Nun die Schätzungen der Koeffizienten und deren Signifikanz (Daumenregel: $|t\text{-Wert}| > 2 \Rightarrow \text{Merkmal ist signifikant}$):

```
Fixed effects:
                  Estimate Std. Error t value
(Intercept)
                   38, 1546
                                5.1406
                                         7.422
                                0.0335 8.263
stress
                    0.2768
sleep
                    -1.9853
                                0.4335
                                        -4.580
                   -4.0916
                                1.4002
                                        -2.922
therapy_intensity
```

- Pro 1 Stresspunkt: Zunahme von 0.3 Depressionspunkten.
- Pro 1 Stunde Schlaf: Abnahme von 2 Depressionspunkten.
- Pro 1 Therapeut: Abnahme von 4 Depressionspunkten.



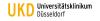


Zuletzt betrachten wir noch die Korrelationen zwischen den geschätzten Prädiktoren:

```
Correlation of Fixed Effects:

(Intr) stress sleep
stress -0.376
sleep -0.618 0.049
therapy_intensity -0.701 -0.041 0.065
```

Eine geringe Korrelation bedeutet eine gute Trennbarkeit der Prädiktoren.





II.2.3. Multivariate Varianzanalyse (MANOVA)

Fragestellung: Wie beeinflussen Stress und Schlaf die gleichzeitigen Ausprägungen mehrerer psychologischer Zielgrößen wie z.B. Depressivität, Ängstlichkeit und Erschöpfung?



II.3. Interdependenzmodelle

Interdependenzmodelle decken Zusammenhänge zwischen Items oder Fällen auf, ohne zwischen unabhängigen und abhängigen Variablen zu unterscheiden. Wir betrachten Faktoren- und Clusteranalysen.

Es gibt zwei Arten von Faktorenanalysen:

- Exploratorische Faktorenanalyse (EFA): Man möchte die Dimensionsstruktur von gegebenen Items aufdecken und darüber Hypothesen generieren (Modellgewinnung).
- Konfirmatorische Faktorenanalyse (CFA): Es werden vorab konkrete Hypothesen über die Dimensionsstruktur der Items formuliert und diese werden dann getestet (Modellprüfung).

Die **Clusteranalyse** hat das Ziel, Fälle (e.g. Personen) anhand von Ähnlichkeiten zu gruppieren – *ohne* vorherige Annahmen.



II.3.1. Exploratorische Faktorenanalyse (EFA)

Fragestellung: Welche übergeordneten, latenten Konstrukte (Faktoren) liegen den beobachteten Merkmalen (Items) zu Grunde? *Oder*: Welche Items laden auf denselben Faktor und messen somit dasselbe latente Merkmal?

Ziele:

- Skalenbildung: Welche Items messen tatsächlich denselben Faktor und können somit sinnvoll zu einer Skala zusammengefasst werden?
- Theoriegewinn: Die EFA generiert Hypothesen; durch die mathematische Reduktion können sich neue, strukturelle Zusammenhänge ergeben.
- Konstruktvalidierung: Messen die Items tatsächlich das, was sie messen sollen? Welche Dimensionsstruktur ergibt sich für die Faktoren?



Die EFA geht davon aus, dass n beobachtete Items X_1, \ldots, X_n durch eine Linearkombination aus p gemeinsamen Faktoren F_1, \ldots, F_p und einem spezifischen Fehlerterm $\varepsilon_1, \ldots, \varepsilon_n$ erklärt werden. In Zeichen:

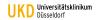
$$X_i = \lambda_{i1}F_1 + \cdots + \lambda_{ip}F_p + \varepsilon_i,$$

was gerade

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{np} \end{pmatrix} \begin{pmatrix} F_1 \\ \vdots \\ F_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ergibt. In fetter Vektorschreibweise mit der Ladungsmatrix Λ können wir das noch kompakter schreiben als

$$X = \Lambda F + \varepsilon$$
.





Wir interessieren dafür, wie stark die Items miteinander zusammenhängen. Wir berechnen also Ihre **empirischen Kovarianzen**. Diese werden in der Kovarianzmatrix

$$\mathsf{Cov}(\mathbf{X}) = \Sigma = egin{pmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots \\ \mathsf{Cov}(X_1, X_2) & \mathsf{Var}(X_1) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

zusammengefasst. Wir setzen nun $\mathsf{X} = \mathsf{\Lambda}\mathsf{F} + arepsilon$ oben ein. Man erhält

$$\Sigma = \mathsf{Cov}(X) = \mathsf{Cov}(\mathbf{\Lambda}\mathbf{F} + \varepsilon) = \mathbf{\Lambda}\mathsf{Cov}(F)\mathbf{\Lambda}^\top + \mathsf{Cov}(\varepsilon) = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbf{\Psi}$$

mit

- Φ: Kovarianzmatrix der Faktoren.
- Ψ: Diagonalmatrix der Fehlervarianzen der Items.





Nimmt man an, dass die Faktoren unabhängig voneinander sind, so ergibt sich die Einheitsmatrix $\Phi = I_n$. In diesem Fall gilt:

$$\mathbf{\Sigma} = \mathbf{\Lambda} \mathbf{\Lambda}^{\top} + \mathbf{\Psi}$$

Das bedeutet: Die beobachtete Gesamtvarianz lässt sich zerlegen in

- eine gemeinsame Varianz (erklärt durch Faktoren),
- eine spezifische Varianz (nicht erklärbar, z.B. Messfehler).

Das wäre der Idealfall: Wir könnten die Gesamtvarianz durch möglichst wenige Ursachen (Faktoren und Residuen) erklären. Also versuchen wir in der EFA mathematisch $\hat{\Lambda}$ und $\hat{\Psi}$ so zu schätzen, dass sich gerade

$$\mathbf{\Sigma} pprox \mathbf{\hat{\Lambda}} \mathbf{\hat{\Lambda}}^{ op} + \hat{\mathbf{\Psi}}$$

ergibt.





"Ein Vergleich mit der Multiplen Regression macht die unterschiedlichen Denkweisen bei Abhängigkeits- und Interdependenzmodell deutlich: Bei der Multiplen Regression ist die Zielvariable (latentes Merkmal) zusammengesetzt aus gewichteten Prädikatoren (manifeste Merkmale); bei der EFA hingegen lädt das Item (manifestes Merkmal) auf die Faktoren (latenten Merkmale) – eine inverse Logik!"



II.3.2. Konfirmatorische Faktorenanalyse (CFA)

Es werden vorab konkrete Hypothesen über die Dimensionsstruktur der Items formuliert und diese werden dann getestet (Modellprüfung).



II.3.3. Clusteranalyse

Die **Clusteranalyse** hat das Ziel, Fälle (e.g. Personen) anhand von Ähnlichkeiten zu gruppieren – *ohne* vorherige Annahmen.

Wir wollen anhand von standardisierten Beobachtungen

- Symptom-Profile clustern,
- Persönlichkeitsprofile gruppieren,
- Antwortmuster auf Skalen entdecken,
- Subtypen von Therapieansprachen finden.



II.4. Strukturgleichungsmodelle (SEM)

Ein Strukturgleichungsmodell besteht aus zwei Teilen:

- Messmodel: Wie werden latente Konstrukte / Faktoren durch beobachtbare Items gemessen? Prinzip: Konfirmatorische Faktorenanalyse (CFA).
- Strukturmodell: Wie werden Pfade (gerichtete Zusammenhänge)
 zwischen latenten Konstrukten beschrieben? Prinzip: (Multivariate)
 Regression mit latenten Konstrukten als Prädikatoren und
 Zielvariablen.





II.5. Meta-Analyse



