Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Allgemeines lineares Modell

Grundlagen

Das allgemeine lineare Modell (general linear model) ist ein abstraktes statistisches Modell der Beziehung von quantitativen (stetigen) und/oder qualitativen (kategorialen) unabhängigen Variablen zu einer (oder mehrerer) abhängigen Variablen

Es umfasst verschiedene bekannte statistische Verfahren als Spezialfälle, wie beispielsweise

- (Multiple) Regressionsanalyse
- (Multivariate) Varianzanalyse
- (Multivariate) Kovarianzanalyse

Grundlagen

Vorteile der Formulierung des allgemeinen linearen Modells

- Es bietet einen formalen Rahmen, innerhalb dessen sich die für verschiedene statistische Verfahren grundlegenden Techniken und Formeln zur Parameterschätzung und Hypothesenprüfung unabhängig von speziellen Problemstellungen entwickeln lassen
- Diese allgemeinen Ergebnisse müssen für eine konkrete Anwendung dann lediglich spezifiziert werden

Diese Vorteile werden auch bei der Implementierung statistischer Verfahren durch die einschlägigen Softwarepakete genutzt (z. B. R, SPSS, SAS)

Die grundsätzliche Vorgehensweise bei der formalen Integration von Regressionsanalyse und Varianzanalyse wird nachfolgend an einem Beispiel erläutert

Integration der multiplen Regressionsanalyse

Zur Vorhersage der Kriteriumsvariablen Y werden bei der multiplen Regression $p \geq 1$ Prädiktorvariablen $X_1, \ldots, X_p$ betrachtet

Bestimmt man für sämtliche Variablen die entsprechenden Realisierungen in einer Stichprobe vom Umfang n, so ergibt sich für das Modell der multiplen Regressionsanalyse

$$
\begin{aligned}
y_1 &= \beta_0 x_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + e_1 \\
y_2 &= \beta_0 x_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + e_2 \\
&\vdots \\
y_n &= \beta_0 x_{n0} + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + e_n
\end{aligned}
$$

mit $x_{i0} = 1$ für alle i

Integration der multiplen Regressionsanalyse

Damit hat man folgende Listen bzw. Schemata von Werten

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Nach Einführung einer kompakten Notation und entsprechender Berechnungsvorschriften (Vektor- und Matrizenrechnung) kann man dieses Schema notieren durch

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{e}$$

Als Kleinste-Quadrate-Schätzer für den Parametervektor ergibt sich

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{mit} \quad \mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Integration der Varianzanalyse

Wir betrachten eine einfaktorielle Varianzanalyse mit drei Faktorstufen

Dann gelten für die Realisierungen von drei Untersuchungseinheiten, die jeweils aus einer der Gruppen 1, 2 bzw. 3 stammen, folgende Modellgleichungen

$$
\begin{array}{lcccccccc}
y_1 & = & \mu & + & \alpha_1 & & & + & e_1 \\
 & \vdots & & & & & & & \\
y_i & = & \mu & & + & \alpha_2 & & + & e_i \\
 & \vdots & & & & & & & \\
y_n & = & \mu & & & & + \ \alpha_3 & + & e_n
\end{array}
$$

Integration der Varianzanalyse: Dummykodierung

Um dieses System von Gleichungen wiederum in eine einzige Gleichung zusammenfassen zu können, definieren wir

$$\alpha_1 := 0 \quad \text{und} \quad x_{i0} = 1 \text{ für alle } i$$

sowie

$$x_{ij} = \begin{cases} 1 & \text{wenn } i \text{ in Gruppe } j \text{ ist} \\ 0 & \text{sonst} \end{cases}$$

Dann gilt

$$\begin{aligned} y_1 &= x_{10}\,\mu + x_{12}\,\alpha_2 + x_{13}\,\alpha_3 + e_1 \\ &\vdots \\ y_i &= x_{i0}\,\mu + x_{i2}\,\alpha_2 + x_{i3}\,\alpha_3 + e_i \\ &\vdots \\ y_n &= x_{n0}\,\mu + x_{n2}\,\alpha_2 + x_{n3}\,\alpha_3 + e_n \end{aligned}$$

Integration der Varianzanalyse: Dummykodierung

Man hat also auch hier folgende Schemata von Werten

$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \quad \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}$$

Auch die Varianzanalyse lässt sich daher durch eine Vektorgleichung der folgenden Form darstellen

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Dabei bezeichnen μ den Erwartungswert der Referenzkategorie und α_2, α_3 jeweils die Differenz zur Referenzkategorie

Integration der Varianzanalyse: Effektkodierung

Alternativ läßt sich das Gleichungssystem durch folgende Definitionen repräsentieren

$$\alpha_3 = -(\alpha_1 + \alpha_2) \quad \text{und} \quad x_{i0} = 1 \text{ für alle } i$$

sowie

$$x_{ij} = \begin{cases} 1 & \text{wenn } i \text{ in Gruppe } j \text{ ist} \\ -1 & \text{wenn } i \text{ in Gruppe } J \text{ ist} \\ 0 & \text{sonst} \end{cases}$$

Dann gilt

$$\begin{aligned} y_1 &= x_{10}\,\mu + x_{11}\,\alpha_1 + x_{12}\,\alpha_2 + e_1 \\ &\vdots \\ y_i &= x_{i0}\,\mu + x_{i1}\,\alpha_1 + x_{i2}\,\alpha_2 + e_i \\ &\vdots \\ y_n &= x_{n0}\,\mu + x_{n1}\,\alpha_1 + x_{n2}\,\alpha_2 + e_n \end{aligned}$$

Integration der Varianzanalyse: Effektkodierung

Man hat also auch hier folgende Schemata von Werten

$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \end{pmatrix} \quad \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \quad \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}$$

bzw. die Vektorgleichung

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Dabei bezeichnen μ den Gesamterwartungswert und α_1, α_2 jeweils die Differenz zum Gesamterwartungswert

Integration der Varianzanalyse: Beispiel

Dummykodierung

```
options(contrasts = c("contr.treatment",
                      "contr.poly"))
m <- lm(weight ~ group, PlantGrowth)

model.matrix(m) |> unique()  # Matrix X
#    (Intercept) grouptrt1 grouptrt2
# 1            1         0         0
# 11           1         1         0
# 21           1         0         1
coef(m)
# (Intercept)  grouptrt1  grouptrt2
#       5.032     -0.371      0.494
anova(m)
#       Df  Sum Sq Mean Sq F value
# group  2   3.766   1.883   4.846
# Resid 27  10.492   0.389
```

Effektkodierung

```
options(contrasts = c("contr.sum",
                      "contr.poly"))
m <- lm(weight ~ group, PlantGrowth)

model.matrix(m) |> unique()
#    (Intercept) group1 group2
# 1            1      1      0
# 11           1      0      1
# 21           1     -1     -1
coef(m)
# (Intercept)    group1    group2
#       5.073    -0.041    -0.412
anova(m)
#       Df  Sum Sq Mean Sq F value
# group  2   3.766   1.883   4.846
# Resid 27  10.492   0.389
```

Multivariate Datenanalyseverfahren

Einführung

Was sind multivariate statistische Verfahren?

- Im weiteren Sinne: Alle statistischen Verfahren, in denen der Zusammenhang von mehr als zwei Variablen untersucht wird (unabhängig von deren Klassifikation als unabhängige bzw. abhängige Variable)

  Beispiele: Zweifaktorielle Varianzanalyse (2 UV, 1 AV), multiple Regression

- Im engeren Sinne: Alle statistischen Verfahren, in denen mehr als eine abhängige Variable untersucht wird

  Beispiele: Einfaktorielle multivariate Varianzanalyse (z. B. 1 UV, 2 AV), nicht aber zweifaktorielle Varianzanalyse (2 UV, 1 AV)

Einführung

Warum werden multivariate statistische Verfahren verwendet?

- Viele psychologische Konstrukte sind nur multivariat zu operationalisieren, z. B. Intelligenz, Persönlichkeit ...
- Mit der Durchführung mehrerer univariater statistischer Verfahren ist die Problematik der α-Fehler-Kumulierung verbunden, außerdem bleiben Effekte der Kombination verschiedener Variablen unberücksichtigt
- Ist in einer psychologischen Untersuchung keine ausreichende Kontrolle von Störvariablen möglich, so möchte man deren Effekt zumindest statistisch abschätzen können
- Identifikation von Scheineffekten, optimale Vorhersage eines Kriteriums, Datenreduktion und strukturelle Vereinfachung, geeignete Klassifikationsverfahren (z. B. in der Diagnostik, Überprüfung komplexer Zusammenhangsmodelle, ...)

Literatur

Bortz, J. (1999). Statistik für Sozialwissenschaftler. Berlin: Springer.

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Erdfelder, E., Mausfeld, R., Meiser, T. & Rudinger, G. (Eds.). (1996). Handbuch Quantitative Methoden. Weinheim: Psychologie Verlags Union.

Fahrmeir, L., Hamerle, A., & Tutz, G. (Eds.) (1996). Multivariate statistische Verfahren. Berlin: De Gruyter.

Hays, W.L. (1988). Statistics for the social sciences (4th ed.). New York: Holt, Rinehart & Winston.

Johnson, R.A. & Wichern, D.W. (1988). Applied multivariate analysis. London: Prentice Hall.

Stevens, J. (2001). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Von der univariaten zur multivariaten Statistik

Univariater statistischer Hypothesentest

Zur Illustration des Übergangs von der univariaten zur multivariaten Statistik betrachten wir den Test des Erwartungswerts μ einer Normalverteilung bei unbekannter Varianz σ^2

- Hypothesen
  Nullhypothese $H_0\colon \mu = \mu_0$ und Alternativhypothese $H_1\colon \mu \neq \mu_0$
- Sind die Zufallsvariablen $X_1, \ldots, X_n$ stochastisch unabhängig und identisch normalverteilt, dann ist die Testfunktion

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$$

t-verteilt mit $n-1$ Freiheitsgraden (S^2 ist die Stichprobenvarianz)

Univariater statistischer Hypothesentest

- Testenscheidung
  Nullhypothese H_0 ablehnen, wenn

$$\left| \frac{\bar{x} - \mu_0}{s} \sqrt{n} \right| > t_{1-\alpha/2}(n-1)$$

  mit $t_{1-\alpha/2}(n-1)$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden
- Bei der multivariaten Verallgemeinerung des statistischen Tests wird das Quadrat

$$T^2 = n \, \frac{(\bar{X} - \mu_0)^2}{S^2}$$

  der Prüfgröße betrachtet, wobei T^2 dann χ^2-verteilt ist mit $n-1$ Freiheitsgraden

Multivariater statistischer Hypothesentest

Der Test des Erwartungswerts im multivariaten Fall geht von folgender Situation aus

- Die Zufallsvariablen $X_1, \ldots, X_p$ repräsentieren p verschiedene Messwerte (z. B. Ergebnisse in verschiedenen psychologischen Tests), die simultan betrachtet werden sollen
- Es wird angenommen, dass die $X_1, \ldots, X_p$ gemeinsam normalverteilt sind, mit Erwartungswerten $\mu_1 = E(X_1), \ldots, \mu_p = E(X_p)$
- Im Rahmen einer Stichprobenziehung vom Umfang n werden dann für sämtliche Zufallsvariablen $X_1, \ldots, X_p$ jeweils n Realisierungen bestimmt

Bestandteile des multivariaten Tests

Eine Liste $\bar{X}_1, \ldots, \bar{X}_p$ von arithmetischen Mittelwerten, die wir mit $\bar{\mathbf{X}}$ bezeichnen

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix}$$

Eine Liste $\mu_1, \ldots, \mu_p$ der Erwartungswerte der Zufallsvariablen $X_1, \ldots, X_p$, die wir mit $\boldsymbol{\mu}$ bezeichnen

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

Bestandteile des multivariaten Tests

Für die Zufallsvariablen $X_1, \ldots, X_p$ sind für den multivariaten Test neben den jeweiligen Stichprobenvarianzen $(S^2(X_1), \ldots)$ auch die (Stichproben-)Kovarianzen $(S(X_1, X_2), S(X_1, X_3), \ldots)$ zu berücksichtigen, die wir in folgendem Schema $\mathbf{S}$ zusammenfassen

$$\mathbf{S} = \begin{pmatrix} S^2(X_1) & S(X_1, X_2) & \ldots & S(X_1, X_p) \\ S(X_2, X_1) & S^2(X_2) & \ldots & S(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ S(X_p, X_1) & S(X_p, X_2) & \ldots & S^2(X_p) \end{pmatrix}$$

der *Kovarianzmatrix*

Nullhypothese

Damit ist die Nullhypothese

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}$$

bzw. in Kurzschreibweise

$$\boldsymbol{\mu} = \boldsymbol{\mu_0}$$

Formulierung des Tests

Bei der univariaten Variante des Tests wurde die χ^2-verteilte Prüfgröße

$$T^2 = n \, \frac{(\bar{X} - \mu_0)^2}{S^2}$$

betrachtet. Diese Gleichung lässt sich auch folgendermaßen schreiben

$$T^2 = n \cdot (\bar{X} - \mu_0) \cdot (S^2)^{-1} \cdot (\bar{X} - \mu_0)$$

Im Rahmen der multivariaten Betrachtung wird in Analogie zum univariaten Fall die Prüfgröße

$$T^2 = n \cdot (\bar{\mathbf{X}} - \boldsymbol{\mu_0})' \cdot \mathbf{S}^{-1} \cdot (\bar{\mathbf{X}} - \boldsymbol{\mu_0})$$

betrachtet, die als Hotellings T^2 bezeichnet wird (Mit $'$ wird die Transposition bezeichnet, d. h. das Vertauschen von Zeilen und Spalten)

Multivariate Varianzanalyse

Grundlagen

Die multivariate Varianzanalyse (MANOVA) ist ein multivariates Verfahren im engeren Sinn und untersucht den Einfluss kategorialer unabhängiger Variablen auf *mehrere* abhängige Variablen

Vorteile einer multivariaten Varianzanalyse

- Bei Durchführung mehrerer univariater Varianzanalysen besteht das Problem der Kumulierung des Fehlers erster Art
- Die Kovarianz zwischen den abhängigen Variablen wird ignoriert
- Auch wenn bei einzelnen Variablen keine signifikanten Effekte auftreten, kann deren Zusammenwirken bei multivariater Betrachtung zu einem signifikanten Effekt führen (Konsequenz der größeren Power)

Annahmen

In Verallgemeinerung des univariaten Falles werden die Vektoren, die die individuellen Fehlerterme beschreiben, als stochastisch unabhängig und identisch multinormalverteilt mit Erwartungswert $\mathbf{0}$ und Kovarianzmatrix $\mathbf{\Sigma}$ angenommen ($\mathbf{\Sigma}$ ist eine $(q \times q)$-Matrix)

Der Annahme gleicher Varianzen (Varianzhomogenität) im univariaten Fall entspricht hier die Annahme gleicher Kovarianzmatrizen $\mathbf{\Sigma}$

- Diese Annahme ist vergleichsweise restriktiv
- In allen Fällen sollen hier nicht nur die Varianzen der betrachteten Variablen, sondern auch deren paarweise Kovarianzen über alle Beobachtungen hinweg konstant sein

Die Annahme der Multinormalverteilung lässt sich statistisch bzw. grafisch nicht in zufriedenstellender Weise testen

Hypothesentests

Die Hypothesen lassen sich unmittelbar ausgehend vom univariaten Fall verallgemeinern

Für eine einfaktorielle MANOVA lautet die Nullhypothese der Gleichheit sämtlicher p gruppenspezifischer Erwartungswerte beispielsweise

$$H_0: \quad \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_p$$

Zur Überprüfung der Hypothesen berechnet man die multivariaten Analoga der Gesamtvariation SS_T (total sum of squares) und der Variationen SS_B (sum of squares between groups) bzw. SS_W (sum of squares within groups) zwischen bzw. innerhalb der Gruppen, die SSCP-Matrizen $\mathbf{T}$, $\mathbf{B}$ und $\mathbf{W}$

Die verwendeten Teststatistiken beziehen sich auf verschiedene Kennwerte dieser SSCP-Matrizen oder die in den Matrizen $\mathbf{W}^{-1}\mathbf{B}$ bzw. $\mathbf{T}^{-1}\mathbf{B}$ beschriebene Variation

Wie viele abhängige Variablen?

Stevens (2001, S. 245) referiert eine Reihe von Gründen, die Anzahl der abhängigen Variablen gering zu halten

- Die Verwendung weniger abhängiger Variablen erleichtert die Interpretation der Ergebnisse
- Kleine oder vernachlässigbare Unterschiede bei den meisten Variablen können die bei einigen Variablen tatsächlich vorhandenen Effekte überdecken
- Die Power multivariater Tests nimmt allgemein mit zunehmender Anzahl abhängiger Variablen ab, ebenso die Robustheit (d. h. die MANOVA reagiert zunehmend sensitiver bei Verletzungen der grundlegenden Annahmen)

Festlegung der Stichprobengröße

Folgende Tabelle gibt eine Empfehlung für die Stichprobengröße pro Gruppe bei einer Teststärke ≈ 0.7, einem Signifikanzniveau von $\alpha = 0.05$ und für 3-6 abhängige Variablen (Stevens, 2001, S. 247; siehe auch Appendix A)

AVn		3	4	5	6
Effekt-größe	sehr groß	12-16	14-18	15-19	16-21
	groß	25-32	28-36	31-40	33-44
	mittel	42-54	48-62	54-70	58-70
	klein	92-120	105-140	120-155	130-170

Hauptkomponentenanalyse

Zielsetzung

Die Hauptkomponentenanalyse wird eingesetzt zur Reduktion einer größeren Menge p von untereinander korrelierten Variablen auf einige wenige, die einen möglichst großen Anteil der Gesamtvarianz auf sich vereinigen

Die beobachteten Variablen werden linear transformiert zu neuen Variablen, den untereinander unkorrelierten *Hauptkomponenten*, die nach abnehmender Varianzaufklärung angeordnet werden

Abgrenzung von Hauptkomponentenanalyse und Faktorenanalyse

- Die Hauptkomponentenanalyse ist im Wesentlichen ein deskriptives Verfahren, für das kein explizites Modell formuliert wird
- Der Faktorenanalyse liegt ein explizites Modell zugrunde, mit dem durch die Einführung latenter Variablen (*Faktoren*) versucht wird, die Zusammenhänge zwischen den beobachteten Variablen aufzuklären (Weiterentwicklung durch lineare Strukturgleichungsmodelle)

Grundlagen

Für die Zufallsvariablen $X_1, \ldots, X_p$ sollen jeweils n Beobachtungen vorliegen, also n-dimensionale Spaltenvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_p$, die zu einer Datenmatrix $\mathbf{X}$ zusammengefasst seien

Den Ausgangspunkt der Berechnungen im Rahmen der Hauptkomponentenanalyse bildet alternativ

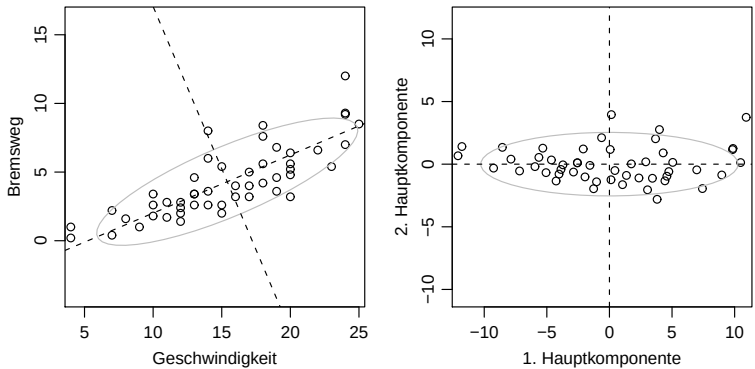
- die Kovarianzmatrix $\mathbf{S}$ zur Datenmatrix $\mathbf{X}$
- die Korrelationsmatrix $\mathbf{R}$ zur Datenmatrix $\mathbf{X}$

Da die Ergebnisse der Hauptkomponentenanalyse nicht invariant sind bezüglich linearer Transformationen der einzelnen Variablen, ergeben sich für $\mathbf{S}$ andere Lösungen als für $\mathbf{R}$

Hauptachsentransformation

Im Zweidimensionalen lässt sich die Kovarianz als Ellipse darstellen; diese ist durch die Lage des Mittelpunkts und durch Länge und Ausrichtung der orthogonalen *Hauptachsen* charakterisiert

Die Hauptachsen können als neues Koordinatensystem für die Beobachtungen aufgefasst werden



Die Bestimmung der Hauptachsen aus der Kovarianz- bzw. Korrelationsmatrix stellt ein *Eigenwertproblem* dar

Dimensionsreduktion

Bezüglich der Hauptkomponenten $Y_1, \ldots, Y_p$ erhält man eine vollständige Beschreibung des Datensatzes

Häufig will man eine sparsame Beschreibung der Daten durch $k < p$ Variablen erreichen, mit der ein möglichst geringer "Informationsverlust" verbunden ist

Mit der Auswahl der ersten k Hauptkomponenten $Y_1, \ldots, Y_k$ erreicht man mit wachsendem k eine zunehmend bessere näherungsweise Rekonstruktion der Daten

- Kriterien zur Bestimmung der Anzahl der beizubehaltenden Hauptkomponenten: Kaiser-Guttman-Kriterium, Scree-Test

Interpretation der Hauptkomponenten

Zur Beschreibung der Beiträge, die die Variablen $X_1, \ldots, X_p$ für die einzelnen Hauptkomponenten leisten, betrachtet man die bivariaten Korrelationen dieser Variablen mit den Hauptkomponenten

- Diese *Faktorladungen* werden üblicherweise in einer *Faktor-Struktur-Matrix* angegeben

Zeigen sich wesentliche Korrelationen für viele der Variablen $X_1, \ldots, X_p$, so ist für die entsprechende Hauptkomponente eine inhaltliche Interpretation schwierig

Die Interpretation wird erleichtert, wenn nur die Faktorladungen weniger Variablen betragsmäßig hoch sind

Interpretation der Hauptkomponenten

Um die Interpretierbarkeit der Hauptkomponenten zu erhöhen, hat man die Möglichkeit, die Hauptkomponenten zu rotieren, wobei sich das Muster der Faktorladungen verändert

Grundsätzlich existieren zwei verschiedene Arten von Rotationen

- Orthogonale Rotationen, bei denen die Unkorreliertheit der Komponenten erhalten bleibt (jedoch geht die Eigenschaft der Varianzmaximierung verloren)
- Oblique (schiefwinklige) Rotationen erhalten die Unkorreliertheit der Komponenten nicht

Loglineare Modelle

Mehrdimensionale Kontingenztabellen

Im Rahmen von zweidimensionalen Kontingenztabellen haben wir Verfahren kennen gelernt (χ^2-Tests), die den Zusammenhang zweier kategorialer Variablen über eine statistische Analyse der darin beschriebenen Häufigkeiten untersuchen

Eine entsprechende Analyse lässt sich auch auf der Grundlage generalisierter linearer Modelle (vgl. logistisches Regressionsmodell) durchführen

Diese Verfahren lassen sich auf Zusammenhänge von mehr als zwei Variablen verallgemeinern, wie sie in der Praxis häufig vorkommen

- Im Falle von drei kategorialen Variablen hat man anstatt einer zweidimensionalen Tabelle dann beispielsweise ein Schema, das man als Quader betrachten kann

Das nachfolgende Beispiel zeigt eine Kontingenztabelle für vier dichotome Variablen

Beispiel: Mehrdimensionale Kontingenztabellen

Die folgende Tabelle fasst Beobachtungen von 68694 PKW-Insassen zusammen, die 1991 im Bundesstaat Maine (USA) in Verkehrsunfälle verwickelt waren

		Sicherheitsgurt: ja		Sicherheitsgurt: nein	
		Verletzung		Verletzung	
Geschlecht	Ort	ja	nein	ja	nein
w	Stadt	759	11587	996	7287
	Land	757	6134	973	3246
m	Stadt	380	10969	812	10381
	Land	513	6693	1084	6123

Loglineare Modelle

Die Nullhypothese eines fehlenden Zusammenhangs (zwischen zwei oder mehr Variablen) bezieht sich stets auf die Annahme stochastischer Unabhängigkeit, die als Produkt der entsprechenden Randwahrscheinlichkeiten formalisiert wird

Dieser nichtlineare Zusammenhang geht nach Logarithmieren in ein additives (also lineares) Zusammenwirken über

- Die Verfahren werden daher als *loglineare Modelle* bezeichnet

Die Assoziationen zwischen zwei oder mehr Variablen in diesem Rahmen können durchaus als analog zur Analyse der Interaktionen bei einer mehrfaktoriellen Varianzanalyse betrachtet werden (bei den Kontingenztabellen muss jedoch keine der Variablen als abhängige Variable ausgezeichnet sein)

Schlussfolgerungen basierend auf statistischer Signifikanz

Der routinemäßige mißbräuchliche Einsatz von Signifikanztests außerhalb eines strikt konfirmatorischen Rahmens (vgl. Folie 262) begünstigt die Verbreitung von nicht replizierbaren Ergebnissen

In ihrem Lehrbuch *Regression and other stories* identifizieren Gelman, Hill und Vehtari (2020) folgende

Problems with the concept of statistical significance

1. Statistical significance is not the same as practical importance
2. Non-significance is not the same as zero
3. The difference between "significant" and "not significant" is not itself statistically significant
4. Researcher degrees of freedom, *p*-hacking, and forking paths
5. The statistical significance filter