Nomenklatur - Übersicht

	Multivariate Regression	Hauptkomponenten-analyse	Korrespondenz-analyse	Clusteranalyse	Diskriminanz-analyse
Name der synthetischen Variable	Regressand	Hauptkomponente	-	Clusterzugehörigkeit	Diskriminanzfunktion
Wert der synthetischen Variable	Schätzwert für reale Variable	Faktorwert	Wert in 1./ 2. Dimension	(Clusterzugehörigkeit)	Diskriminanzwert
durch synth. Variable **erklärte Gesamt-Streuung**	erklärte Varianz	Anteil der Eigenwerte	Anteil der Eigenwerte (der Trägheit, der Streuung, der Inertia)	(Anteil der Fehlerquadratsumme)	Anteil der Eigenwerte (Diskrimanzanteil)
durch synth. Variable **erkl. Streuung der einzelnen Variablen**	-	Kommunalität	Anteil der Eigenwerte (der Trägheit, der Streuung, der Inertia)	-	-
Korrelation zwischen realer und synthetischer Variable	partielle Korrelation	Ladung	-	-	Ladung

Multivariate Analyse: Take-Home Message

1. Die meisten Standardmethoden basieren auf ***linearen Zusammenhängen*** im Datensatz.
2. In der Regel gibt es verschiedene ***Maßzahlen für die Güte des Verfahrens***, die unbedingt beachtet werden sollten (ein großes r^2 alleine sagt noch nicht viel aus!).
3. Multivariate Verfahren werden überwiegend eingesetzt, um
   - einzelne Werte ***vorherzusagen*** (Regression)
   - die ***Dimension*** des Datensatzes zu ***reduzieren*** (Prozessanalyse, Visualisierung)
   - ***Gruppen*** zu identifizieren (klassifizieren).
4. Die Verfahren verlangen i.d.R. ***±subjektive Entscheidungen*** des Anwenders, die zu begründen sind.
5. Die Ergebnisse der hier vorgestellten ***Verfahren sind nicht unabhängig*** voneinander.

Versuch / Experiment

Ein Versuch ist eine

(1) systematische Beobachtung der (***A***bhängige ***V***ariable ***AV***)

(2) Auswirkungen einer planmäßigen Veränderung (***U***nabhängige ***V***ariable ***UV***)

(3) unter weitestgehender Ausschaltung oder Kontrolle von Störfaktoren. (***S***tör***v***ariable ***SV)***

=> Forderungen (***MaxKonMin***-Prinzip; *Krelinger 1973*):

zu (2): ***Max***imale Variation der postulierten Einflussgrößen (***Primärvarianz***)

zu (3): ***Kon***trolle der Randbedingungen, Minimierung ihrer Varianz (***Sekundärvarianz***)

zu (1): ***Min***imierung der Beobachtungsfehler (***Fehlervarianz***)

Grundbegriffe der Versuchsplanung

Definitionen:

- Kausalität
- Hypothesen: $\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_3 = ...$; $\mathbf{H_1}$: $\mu_i \neq \mu_j$ = für mindestens zwei μ_i, μ_j
- Validität = Zulässigkeit der Schlussfolgerungen aus dem Experiment
  - intern: Ergebnisse der Untersuchung sind logisch eindeutig interpretierbar
  - extern: Ergebnisse der Untersuchung sind generalisierbar
- ***n***-faktoriell: ***n*** Einflussfaktoren (UV), ***p***-fach gestuft
- unabhängige Variable UV = Einflussgröße = Behandlung = Treatment: beliebig skaliert
- abhängige Variable AV: mind. intervallskaliert (=> Mittelwerte und Varianzen interpretierbar)

Minimierung der Fehlervarianz

- Richtigkeit und Präzision der Messung
- Ausreißer
- Fehlende Werte

Maximierung der Primärvarianz

Wenn die Beziehung zwischen **UV** und **AV**

- **linear** ist: Wahl von **extremen** Werten der **UV**
- **kurvilinear** ist: Wahl von **optimalen** Werten der **UV**
- **unbekannt** ist: Unterteilung in möglichst **viele Stufen** der **UV** (möglichst kleine Abstufungen)

UV *= unabhängige Variable*

AV *= abhängige Variable*

Minimierung der Sekundärvarianz

1. **Eliminierung** der Störvariablen
2. **Konstanthaltung** (Annahme einer linearen Beziehung zwischen **SV** und **AV**)
3. **Umwandlung** von Störvariablen in unabhängige Variablen (**SV** → **UV**)
4. **Parallelisierung** (Einzelmessungen werden in eine Rangreihe der Werte bzgl. der Störvariablen gebracht, dann nacheinander den Versuchsbedingungen zugeordnet)
5. **Wiederholungsmessung** (dieselbe Gruppe wird unter den verschiedenen Versuchsbedingungen getestet)
6. **Blockbildung** (Blöcke = Gruppen homogener Untereinheiten, auf die die einzelnen Versuchs-Varianten verteilt werden)
7. **Randomisierung** (zufällige Verteilung der **SV** auf die einzelnen Gruppen)

Lateinisches Quadrat

(Latin Square)

Ziel:	Minimierung des Einflusses zweier Störvariabler.
Methode:	Jede Variante ist genau einmal in jeder Zeile und in jeder Spalte vertreten.
Einschränkungen:	Wechselwirkungen zwischen AV, SV1 und SV 2 können nicht untersucht werden.

n-dimensionale Erweiterung: ***Latin Hypercube***

Zunahme von **SV 1** →

Zunahme von **SV 2** ↓

A	D	C	B
C	B	A	D
D	A	B	C
B	C	D	A

Versuch / Experiment

Ein Versuch ist eine

(1) systematische Beobachtung der (***A***bhängige ***V***ariable ***AV***)

(2) Auswirkungen einer planmäßigen Veränderung (***U***nabhängige ***V***ariable ***UV***)

(3) unter weitestgehender Ausschaltung oder Kontrolle von Störfaktoren. (***S***tör***v***ariable ***SV)***

=> Forderungen (***MaxKonMin***-Prinzip; *Krelinger 1973*):

zu (2): ***Max***imale Variation der postulierten Einflussgrößen (***Primärvarianz***)

zu (3): ***Kon***trolle der Randbedingungen, Minimierung ihrer Varianz (***Sekundärvarianz***)

zu (1): ***Min***imierung der Beobachtungsfehler (***Fehlervarianz***)

Varianzanalyse = ANOVA

(Analysis of Variance)

Ziel: Bestimmung des Anteils verschiedener Einflussfaktoren (UV) an der beobachteten Varianz der AV

=> Untersuchung der ***Signifikanz von Mittelwertdifferenzen***

Varianz: = mittlere quadrierte Abweichung

= Summe der quadrierten Abweichungen, geteilt durch die Anzahl der Freiheitsgrade

beachte: ***alle*** *der bisher vorgestellten multivariaten Verfahren führen eine Zerlegung der Varianz durch, der Begriff der* ***ANOVA*** *ist jedoch für dieses Verfahren reserviert!*

Quadratsummenzerlegung

für einen einfaktoriellen, ***p***-fach gestufter Versuch mit jeweils ***n*** Wiederholungen

generell: Varianz $(\hat{\sigma}^2) = \frac{\text{Quadratsumme } (QS)}{\text{Freiheitsgrade } (df)}$

$$\boldsymbol{QS_{tot} = QS_{treat} + QS_{Fehler}} \text{ und } \boldsymbol{df_{tot} = df_{treat} + df_{Fehler}}$$

- **Gesamt-Varianz** (Stichprobenvarianz): $\hat{\sigma}^2_{tot} = \frac{QS_{tot}}{df_{tot}} = \frac{\sum_{i=1}^{n \cdot p} (x_i - \bar{x})^2}{(n \cdot p) - 1}$
- **Treatment-Varianz**:
  $\bar{x}_p$: *Mittelwert der Merkmalsausprägungen für die einzelnen Stufen der Behandlung*
  $\hat{\sigma}^2_{treat} = \frac{QS_{treat}}{df_{treat}} = \frac{n \cdot \sum_{I=1}^{p} (\bar{x}_i - \bar{x})^2}{p - 1}$
- **Fehler-Varianz**: $\hat{\sigma}^2_{Fehler} = \frac{QS_{Fehler}}{df_{Fehler}} = \frac{\sum_{i=1}^{p} \sum_{m=1}^{n} (x_m - \bar{x}_i)^2}{p \cdot (n-1)}$

Prüfgröße: F-Wert

- Wenn die $\mathbf{H_0}$ gilt ($\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_3 = ...$), dann stellt die Treatmentvarianz eine erwartungsgetreue Schätzung der Fehlervarianz dar: $\hat{\sigma}^2_{treat} = \hat{\sigma}^2_{Fehler}$
- **<u>Prüfgröße</u>** $\boldsymbol{F}$: $F = \hat{\sigma}^2_{treat} / \hat{\sigma}^2_{Fehler}$
- zu vergleichen mit tabellierten Werten für
  $df_{treat} = p\text{-}1$ Zählerfreiheitsgrade und
  $df_{Fehler} = p \cdot (n\text{-}1)$ Nennerfreiheitsgrade
- **<u>Interpretation</u>**: wird der tabellierte F-Wert überschritten, so unterscheiden sich **mindestens zwei** der p Stufen der Behandlung signifikant voneinander

Ungleiche Stichprobenumfänge

n_i: Stichprobenumfang für die Treatmentstufe ***i***

N: Summe aller Untersuchungseinheiten $N = \sum_{p} n_i$

für ***gleiche*** Stichprobenumfänge:

$$QS_{treat} = \sum_{I=1}^{p} (\bar{x}_i - \bar{x})^2$$

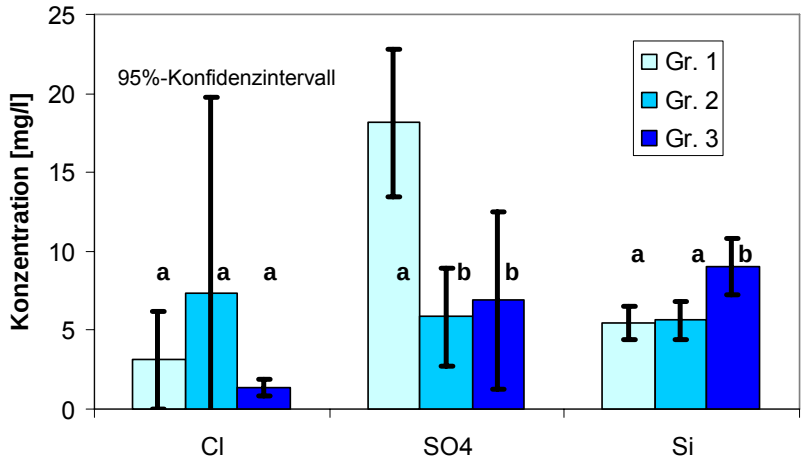
$$df_{tot} = (n \cdot p) - 1 = N - 1$$
$$df_{treat} = p - 1$$
$$df_{Fehler} = p \cdot (n-1) = N - p$$

für ***ungleiche*** Stichprobenumfänge:

$$QS_{treat} = \sum_{I=1}^{p} (\bar{x}_i - \bar{x})^2 \cdot n_i$$

$$df_{tot} = N - 1$$
$$df_{treat} = p - 1$$
$$df_{Fehler} = N - p$$

Beispiel



Mehrere t-Tests oder eine ANOVA?

- Für den **einfaktoriellen**, **zweifach gestuften** Ansatz entspricht die ANOVA einem t-Test
- Für den **mehrfaktoriellen** und/oder **mehrfach gestuften** Ansatz wären
    1. viele einzelne t-Tests erforderlich => steigende Wahrscheinlichkeit, dass einzelne Tests fälschlicherweise signifikante Unterschiede aufweisen
    2. Untersuchungen der ***Wechselwirkungen*** zwischen verschiedenen UV mittels t-Test nicht möglich

Wechselwirkungen zwischen UVs

= der Anteil des Gesamteffekts verschiedener Faktoren, der von der addierten Wirkung (Superposition) der Einzeleffekte abweicht

Bsp.: Erhöhung des Weizenertrags durch

- N-Düngung um 25%,
- Applikation eines Fungizids um 20%,
- Kombination aus N-Düngung und Fungizid-Applikation um 30% (*statt um 50%*).

Voraussetzungen der ANOVA

1. **Quadratsummenzerlegung** → Voraussetzungs-frei

2. **F-Test** → Voraussetzungen:
   - Normalverteilung der Fehlerkomponenten (= Abweichungen der Messwerte vom jeweiligen Stichprobenmittel) (*selten überprüft*)
   - gleiche Varianzen der Fehlerkomponenten (*Bartlett-Test, Levene-Test*)
   - Unabhängigkeit der Fehlerkomponenten innerhalb und zwischen den Stichproben (*s. Randomisierung*)

   kritisch: **kleine, ungleichgroße** Stichproben und **heterogene Varianzen**

   => Ausweichen auf verteilungsfreie Verfahren **(Kruskal-Wallis)**

Aufgabe

1. Testen Sie jeweils mittels t-Test und einfaktorieller ANOVA die beiden Gruppen auf signifikante Unterschiede der Mittelwerte für verschiedene Variablen.
2. Stellen Sie zum Vergleich die Mittelwerte und 95%-Konfidenzintervalle ***aller Variablen*** sowie der ***Hauptkomponentenwerte*** der beiden Gruppen grafisch dar. Kennzeichnen Sie signifikante Unterschiede mittels unterschiedlicher Buchstaben.
3. Führen Sie zum Vergleich für die drei Gruppen eine Diskriminanzanalyse mit den Werten der ***Variablen*** durch.