

Teil III - Multivariate Statistik

Tarek Carls

21. Oktober 2023

- 1 2. Varianzanalyse
 - 2.1 Einfaktorielle ANOVA
 - 2.2.2 Mehrfaktorielle ANOVA
 - 2.3 ANCOVA

- 2 3. Regressionsanalyse
 - 3.1 Lineare Regressionsanalyse
 - 3.2 - Logistische Regression

Was ist die Einfaktorielle ANOVA?

- ANOVA steht für "Analysis of Variance".
- Einfaktorielle ANOVA ist eine statistische Methode, um die Unterschiede zwischen den Mittelwerten von drei oder mehr unabhängigen Gruppen zu testen.
- Sie prüft, ob es signifikante Unterschiede zwischen den Gruppen gibt.

Annahmen der Einfaktoriellen ANOVA

- Unabhängigkeit der Beobachtungen
- Normalverteilung der Daten in jeder Gruppe
- Homogenität der Varianzen

Modellgleichung

Die Modellgleichung für die Einfaktorielle ANOVA ist:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

wobei:

- Y_{ij} der beobachtete Wert der j-ten Beobachtung in der i-ten Gruppe ist.
- μ der Gesamtdurchschnitt aller Beobachtungen ist.
- α_i der Effekt der i-ten Gruppe ist.
- ϵ_{ij} der zufällige Fehler der j-ten Beobachtung in der i-ten Gruppe ist.

Interpretation der Modellgleichung

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- μ gibt den Gesamtdurchschnitt aller Beobachtungen an.
- α_i gibt an, wie sehr sich die i-te Gruppe vom Gesamtdurchschnitt unterscheidet.
- ϵ_{ij} repräsentiert den zufälligen Fehler oder die natürliche Variation innerhalb der Gruppen.

Beispiel: Datenset für Einfaktorielle ANOVA

Gruppe	Beobachtung
A	10
A	12
A	11
B	20
B	19
B	21
C	15
C	14
C	16

Tabelle: Beispiel eines Datensets für eine Einfaktorielle ANOVA

Schritt 1: Daten eingeben

- Öffnen Sie SPSS und gehen Sie zur *Datenansicht*.
- Geben Sie Ihre Daten in zwei Spalten ein: Eine für die Gruppenvariable und eine für die Beobachtungsvariable.
- Beispiel:

Gruppe	Beobachtung
A	10
A	12
B	20

Schritt 2: ANOVA durchführen

- Wählen Sie im Hauptmenü *Analysieren*.
- Gehen Sie zu *Allgemeines lineares Modell* und dann zu *Univariat*.
- Verschieben Sie die Beobachtungsvariable in das Feld *Abhängige Variable*.
- Verschieben Sie die Gruppenvariable in das Feld *Feste Faktoren*.
- Klicken Sie auf *OK*, um die Analyse zu starten.

2.2.1 Einfaktorielle ANOVA - Durchführung in SPSS

- Auswahl der Variablen im Hauptfenster der allgemeinen Varianzanalyse.
- Bestimmung der Post-Hoc-Mehrfachvergleiche.
- Ergebnisse werden in drei Abschnitten angezeigt: Univariat, Post-Hoc-Tests, Mittelwert-Diagramme.

Zusätzliche Optionen in SPSS für ANOVA

• Diagrammoptionen:

- *Boxplots* - Zeigt die Verteilung der Daten für jede Gruppe.
- *Profilplots* - Zeigt die Mittelwerte jeder Gruppe.
- *Histogramme* - Zeigt die Verteilung der Residuen.
- *Streuungsdiagramme* - Zeigt Beziehungen zwischen zwei Variablen.
- *Normalitätsplots* - Zeigt, ob die Residuen normal verteilt sind.

• Weitere Optionen:

- *Kontraste* - Zum Vergleich spezifischer Gruppenkombinationen.
- *Homogenitätstests* - Tests auf Gleichheit der Varianzen (z.B. Levene-Test).
- *Geschätzte Randmittel* - Gibt Mittelwerte, Standardabweichungen und Konfidenzintervalle für die einzelnen Gruppen an.
- *Deskriptive Statistiken* - Zeigt Mittelwerte, Standardabweichungen und andere deskriptive Statistiken für jede Gruppe.

Mehrfaktorielle ANOVA

- Die Mehrfaktorielle ANOVA, auch als Zweiweg-ANOVA oder n-Weg-ANOVA bekannt, untersucht, wie zwei oder mehr unabhängige Variablen gleichzeitig den Mittelwert einer abhängigen Variablen beeinflussen.
- Sie ermöglicht die Untersuchung von Haupteffekten und Interaktionseffekten.

Haupteffekte und Interaktionen

- **Haupteffekt:** Der Einfluss einer einzelnen unabhängigen Variablen auf die abhängige Variable, ohne Berücksichtigung anderer unabhängiger Variablen.
- **Interaktionseffekt:** Wenn der Effekt einer unabhängigen Variablen auf die abhängige Variable von dem Niveau einer anderen unabhängigen Variablen abhängt.

Modellgleichung für Mehrfaktorielle ANOVA

Die Modellgleichung für eine Zweiweg-ANOVA ist:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

wobei:

- Y_{ijk} der beobachtete Wert für die Kombination der i-ten Stufe des ersten Faktors und der j-ten Stufe des zweiten Faktors ist.
- μ der Gesamtdurchschnitt aller Beobachtungen ist.
- α_i der Effekt der i-ten Stufe des ersten Faktors ist.
- β_j der Effekt der j-ten Stufe des zweiten Faktors ist.
- $(\alpha\beta)_{ij}$ der Interaktionseffekt zwischen den beiden Faktoren ist.
- ϵ_{ijk} der zufällige Fehler ist.

Beispiel: Datenset für Mehrfaktorielle ANOVA

Faktor 1	Faktor 2	Beobachtung
Männlich	A	10
Männlich	B	12
Männlich	A	11
Weiblich	A	20
Weiblich	B	19
Weiblich	A	21

Tabelle: Beispiel eines Datensets für eine Mehrfaktorielle ANOVA

Durchführung in SPSS

- Wählen Sie im Hauptmenü *Analysieren*.
- Gehen Sie zu *Allgemeines lineares Modell* und dann zu *Univariat*.
- Verschieben Sie die abhängige Variable in das Feld *Abhängige Variable*.
- Verschieben Sie die unabhängigen Variablen in das Feld *Feste Faktoren*.
- Unter *Plots* können Sie Interaktionsdiagramme hinzufügen.
- Klicken Sie auf *OK*, um die Analyse zu starten.

Interpretation der Ergebnisse

- SPSS gibt eine ANOVA-Tabelle aus, die Haupteffekte und Interaktionseffekte zeigt.
- Signifikante Haupteffekte deuten darauf hin, dass es Unterschiede zwischen den Stufen der unabhängigen Variablen gibt.
- Ein signifikanter Interaktionseffekt deutet darauf hin, dass der Effekt einer unabhängigen Variablen von dem Niveau einer anderen unabhängigen Variablen abhängt.

Zusätzliche Optionen in SPSS

• Diagrammoptionen:

- *Boxplots* - Zeigt die Verteilung der Daten für jede Gruppe.
- *Profilplots* - Zeigt die Mittelwerte jeder Gruppe.
- *Histogramme* - Zeigt die Verteilung der Residuen.
- *Streuungsdiagramme* - Zeigt Beziehungen zwischen zwei Variablen.
- *Normalitätsplots* - Zeigt, ob die Residuen normal verteilt sind.

• Weitere Optionen:

- *Kontraste* - Zum Vergleich spezifischer Gruppenkombinationen.
- *Homogenitätstests* - Tests auf Gleichheit der Varianzen (z.B. Levene-Test).
- *Geschätzte Randmittel* - Gibt Mittelwerte, Standardabweichungen und Konfidenzintervalle für die einzelnen Gruppen an.
- *Deskriptive Statistiken* - Zeigt Mittelwerte, Standardabweichungen und andere deskriptive Statistiken für jede Gruppe.

ANCOVA (Analyse der Kovarianz)

- ANCOVA kombiniert die Techniken der ANOVA und Regression.
- Sie untersucht den Einfluss kategorialer unabhängiger Variablen auf eine abhängige Variable, während sie gleichzeitig den Einfluss einer oder mehrerer kontinuierlicher Kovariaten kontrolliert.
- Ziel ist es, den Effekt der unabhängigen Variablen auf die abhängige Variable zu bestimmen, nachdem der Effekt der Kovariaten herausgenommen wurde.
- Kovariablen sind metrische Variablen, bei denen ein Einfluss auf die Zielgröße vermutet wird, der jedoch nicht primär von Interesse ist.
- Durch Berücksichtigung von Kovariablen kann die Reststreuung (auch Residuen genannt) reduziert werden.
- Dies führt zu einer höheren Power, d.h. einer höheren Wahrscheinlichkeit, einen vorhandenen Effekt zu entdecken.

Warum ANCOVA verwenden?

- Um die Varianz der abhängigen Variable zu reduzieren, die durch die Kovariate erklärt wird.
- Um Unterschiede zwischen Gruppen zu untersuchen, wenn es vorher bestehende Unterschiede zwischen den Gruppen in der Kovariate gibt.
- Um die Effektgröße eines kategorialen Prädiktors zu erhöhen, indem die Streuung der abhängigen Variable reduziert wird.

Modellgleichung für ANCOVA

Die Modellgleichung für ANCOVA ist:

$$Y_i = \mu + \alpha_i + \beta(X_i - \bar{X}) + \epsilon_i$$

wobei:

- Y_i der beobachtete Wert der abhängigen Variablen für die i-te Beobachtung ist.
- μ der Gesamtdurchschnitt aller Beobachtungen ist.
- α_i der Effekt der i-ten Stufe der unabhängigen Variablen ist.
- β die Regressionssteigung der Kovariate ist.
- X_i der Wert der Kovariate für die i-te Beobachtung ist.
- \bar{X} der Durchschnittswert der Kovariate über alle Beobachtungen ist.
- ϵ_i der zufällige Fehler ist.

Beispiel: Datenset für ANCOVA

Gruppe (Faktor)	Kovariate	Beobachtung
A	25	10
A	30	12
B	28	20
B	27	19
C	29	15
C	31	14

Tabelle: Beispiel eines Datensets für eine ANCOVA

Durchführung in SPSS

- Wählen Sie im Hauptmenü *Analysieren*.
- Gehen Sie zu *Allgemeines lineares Modell* und dann zu *Univariat*.
- Verschieben Sie die abhängige Variable in das Feld *Abhängige Variable*.
- Verschieben Sie die unabhängige Variable in das Feld *Feste Faktoren*.
- Verschieben Sie die Kovariate in das Feld *Kovariaten*.
- Klicken Sie auf *OK*, um die Analyse zu starten.

Interpretation der Ergebnisse

- SPSS gibt eine ANCOVA-Tabelle aus, die den Haupteffekt der unabhängigen Variablen und den Effekt der Kovariate zeigt.
- Ein signifikanter Effekt der unabhängigen Variablen deutet darauf hin, dass es Unterschiede zwischen den Gruppen gibt, nachdem der Effekt der Kovariate kontrolliert wurde.
- Ein signifikanter Effekt der Kovariate deutet darauf hin, dass es einen linearen Zusammenhang zwischen der Kovariate und der abhängigen Variablen gibt.

Zusätzliche Optionen in SPSS

- **Diagrammoptionen:**

- *Boxplots*
- *Profilplots*
- *Histogramme*
- *Streuungsdiagramme*
- *Normalitätsplots*

- **Weitere Optionen:**

- *Kontraste*
- *Homogenitätstests*
- *Modellterme speichern*
- *Deskriptive Statistiken*

3. Einleitung

- Kapitel 3 konzentriert sich auf die Regressionsanalyse.
- Die Regressionsanalyse ist anwendbar, wenn die Zielgröße metrisch ist.
- Es gibt verschiedene Formen der Regressionsanalyse, abhängig von der Skalierung der Zielgröße.
- Für metrische Zielgrößen wird die lineare Regressionsanalyse verwendet.
- Für kategoriale Zielgrößen wird die logistische Regressionsanalyse verwendet.

3. Prinzip der Regressionsanalyse

- Eine mathematische Gleichung wird aufgestellt, die den Zusammenhang zwischen zwei oder mehreren Variablen optimal widerspiegelt.
- Die Regressionskoeffizienten in der Regressionsgleichung quantifizieren den Zusammenhang.
- Bei nur einer Einflussgröße spricht man von der Einfachregression.
- Bei mehreren Einflussgrößen von der multiplen Regression.

3. Beispiele

- Einfluss von sportlicher Betätigung, Alter und BMI auf die Gewichtsveränderung.
- Einfluss verschiedener Variablen wie Geschlecht, Alter und sportliche Betätigung auf die Zufriedenheit mit der Reha-Maßnahme.

Lineare Regressionsanalyse

- Die lineare Regressionsanalyse ist eine statistische Methode zur Modellierung der Beziehung zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen Variablen.
- Das Hauptziel ist es, die beste Linie (Regressionslinie) zu finden, die die Datenpunkte am besten beschreibt.
- Es können Vorhersagen für neue Datenpunkte gemacht werden.

Annahmen der linearen Regression

- Linearität: Die Beziehung zwischen den Variablen sollte linear sein.
- Unabhängigkeit: Die Beobachtungen sollten unabhängig voneinander sein.
- Homoskedastizität: Die Varianz der Fehlerterme sollte konstant sein.
- Normalität: Die Fehlerterme sollten normalverteilt sein.
- Keine Multikollinearität: Bei mehrfacher linearer Regression sollten die unabhängigen Variablen nicht hoch korreliert sein.

Streudiagramm in SPSS erstellen

- 1 Starten Sie SPSS und öffnen Sie Ihren Datensatz.
- 2 Wählen Sie im Hauptmenü *Grafiken*.
- 3 Gehen Sie zu *Diagrammerstellung (Legacy)* und wählen Sie *Streudiagramm*.
- 4 Wählen Sie *Einfach* und klicken Sie auf *Definieren*.
- 5 Verschieben Sie die abhängige Variable in das Feld *Y-Achse* und die unabhängige Variable in das Feld *X-Achse*.
- 6 Klicken Sie auf *OK*, um das Streudiagramm zu erstellen.

Hinweis: Ein Blick auf das Streudiagramm kann Ihnen helfen, die Art des Zusammenhangs zwischen den Variablen zu erkennen und ob lineare Regression geeignet ist.

Grundlegende Gleichung

Die grundlegende Gleichung für eine einfache lineare Regression ist:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

wobei:

- Y die abhängige Variable ist.
- X die unabhängige Variable ist.
- β_0 der Y-Achsenabschnitt ist.
- β_1 die Steigung der Regressionslinie ist.
- ϵ der Fehlerterm ist.

Gleichung für Multiple Lineare Regression

Die Gleichung für eine multiple lineare Regression mit p Prädiktoren und Kovariaten ist:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

wobei:

- Y die abhängige Variable ist.
- X_1, X_2, \dots, X_p die unabhängigen Variablen (Prädiktoren und Kovariaten) sind.
- β_0 der Y-Achsenabschnitt ist.
- $\beta_1, \beta_2, \dots, \beta_p$ die Regressionskoeffizienten für die jeweiligen unabhängigen Variablen sind.
- ϵ der Fehlerterm ist.

Durchführung in SPSS

- Wählen Sie im Hauptmenü *Analysieren*.
- Gehen Sie zu *Regression* und dann zu *Linear*.
- Verschieben Sie die abhängige Variable in das Feld *Abhängige Variable*.
- Verschieben Sie die unabhängige(n) Variable(n) in das Feld *Unabhängige Variable(n)*.
- Klicken Sie auf *OK*, um die Analyse zu starten.

Interpretation der Ergebnisse

- SPSS gibt eine Ausgabe mit verschiedenen Statistiken, darunter:
 - Regressionskoeffizienten (β_0 und β_1).
 - Bestimmtheitsmaß R^2 - gibt den Anteil der Varianz in der abhängigen Variable an, der durch die unabhängige Variable erklärt wird.
 - t-Tests für die Koeffizienten - testen, ob die Koeffizienten signifikant von null verschieden sind.
 - F-Test - testet die Gesamtbedeutung des Modells.
- Ein signifikanter t-Test für β_1 deutet darauf hin, dass es einen signifikanten linearen Zusammenhang zwischen der unabhängigen und der abhängigen Variable gibt.
- Zusätzliche Ausgaben:
 - Über das Feld *Statistiken* können nützliche Zusatzausgaben ausgewählt werden.
 - Schätzer und die dazugehörigen 95%-Konfidenzintervalle sollten in der Ausgabe enthalten sein.
 - Anpassungsgüte des Modells, d.h. das Bestimmtheitsmaß, sollte ebenfalls ausgegeben werden.

Logistische Regressionsanalyse

- Die logistische Regressionsanalyse ist eine statistische Methode zur Modellierung der Beziehung zwischen einer binären abhängigen Variablen und einer oder mehreren unabhängigen Variablen.
- Anstatt den tatsächlichen Wert der abhängigen Variablen vorherzusagen (wie bei der linearen Regression), schätzt sie die Wahrscheinlichkeit, dass ein bestimmtes Ereignis eintritt.

Grundlegende Gleichung

Die grundlegende Gleichung für die logistische Regression ist:

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

wobei:

- p die Wahrscheinlichkeit des Ereignisses ist.
- X_1, X_2, \dots, X_p die unabhängigen Variablen sind.
- $\beta_0, \beta_1, \dots, \beta_p$ die Regressionskoeffizienten sind.

Durchführung in SPSS

- Wählen Sie im Hauptmenü *Analysieren*.
- Gehen Sie zu *Regression* und dann zu *Binär logistisch*.
- Verschieben Sie die binäre abhängige Variable in das Feld *Abhängige Variable*.
- Verschieben Sie die unabhängige(n) Variable(n) in das Feld *Unabhängige Variable(n)*.
- Klicken Sie auf *OK*, um die Analyse zu starten.

Interpretation der Ergebnisse

- SPSS gibt eine Ausgabe mit verschiedenen Statistiken, darunter:
 - Regressionskoeffizienten (β_0, β_1, \dots).
 - Odds Ratios - zeigen, wie sich eine Einheit Veränderung in der unabhängigen Variablen auf die Odds des Ereignisses auswirkt.
 - Wald-Tests - testen die Signifikanz der Koeffizienten.
- Ein signifikanter Wald-Test für einen Koeffizienten deutet darauf hin, dass die entsprechende unabhängige Variable einen signifikanten Einfluss auf die abhängige Variable hat.

Annahmen der logistischen Regression

- Die abhängige Variable ist binär.
- Es gibt keine hohen Interkorrelationen (Multikollinearität) zwischen den Prädiktoren.
- Es gibt eine lineare Beziehung zwischen den logit-transformierten Chancen der abhängigen Variablen und den unabhängigen Variablen.
- Es werden ausreichend Fälle für jede Kombination von unabhängigen Variablen benötigt.