TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Professur Psychologie digitaler Lernmedien
Institut für Medienforschung Philosophische Fakultät



Statistik I

Multiple Lineare Regression

Moneyball (2011). Sony Pictures Releasing.

Prof. Dr. Günter Daniel Rey

Überblick

- Multiple Regression
- Verfahren zur Auswahl unabhängiger Variablen
- Interaktionseffekte in der multiplen Regression
- Inkrement und Dekrement in der multiplen Regression
- Suppressorvariablen
- Prognosegüte und Kreuzvalidierung
- Indikatorcodierung
- Signifikanzprüfung zur multiplen Regression
- Konfidenzbänder
- Inferenzstatistische Voraussetzungen

Multiple lineare Regression

- Definition: Statistisches Verfahren zur Vorhersage einer Kriteriumsvariable durch mehrere Prädiktorvar. mittels Linearkombination
- Regressionsgleichung zur Regressions(hyper-)ebene:

$$\hat{y} = b_0 + x_1 \cdot b_1 + \ldots + x_m \cdot b_m$$

- Bestimmung der Regressionsgewichte (Beta-Gewichte) wieder mittels Methode der kleinsten Quadrate
- Unterschied zur linearen bivariaten Regression: Berechnung mit Matrizen statt mit Zahlen

$\hat{y}$	Vorhergesagte Kriteriumsvariable y
b_0	Achsenabschnitt der Regressionsgraden
x_1	Erste Prädiktorvariable
b_1	Steigung zur ersten Prädiktorvariablen
x_m	m-te Prädiktorvariable
b_m	Steigung zur m-ten Prädiktorvariablen

Verfahren zur Auswahl unabhängiger Variablen (Eid, Gollwitzer & Schmitt, 2017)

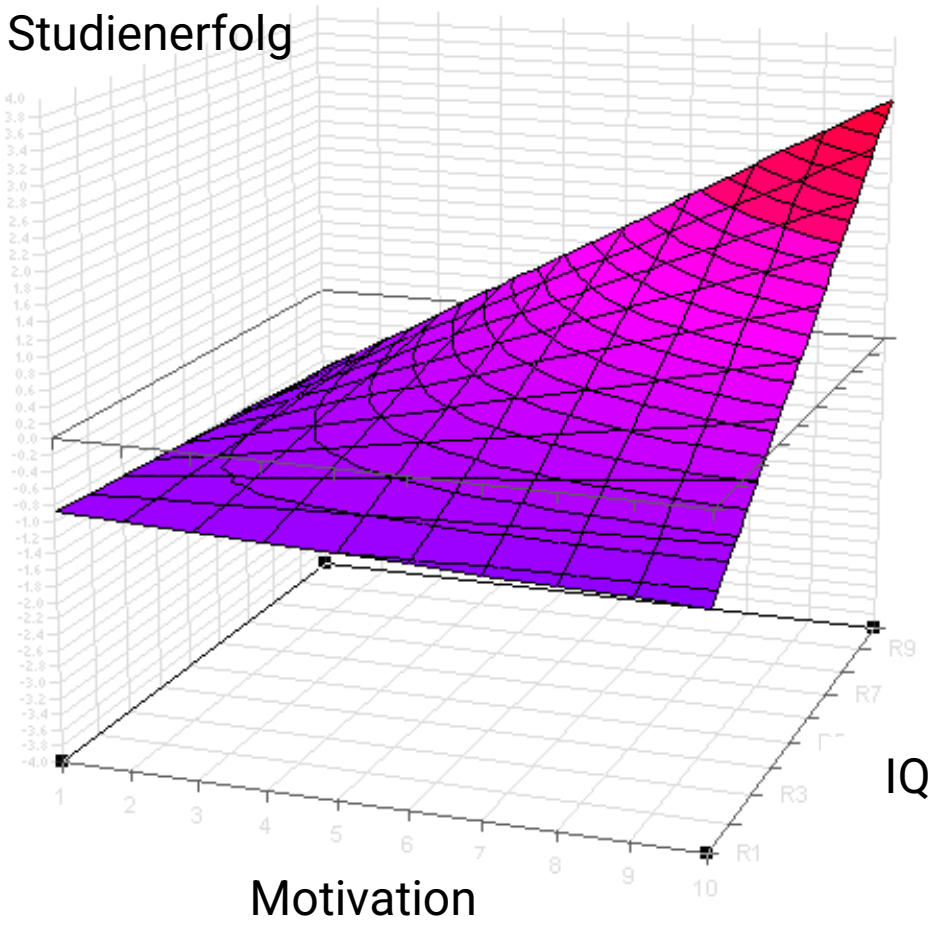
- Zwei generelle Strategien zur Auswahl unabhängiger Variablen (= Prädiktorvariablen) für ein Regressionsmodell
  - Auswahl aufgrund theoretischer Überlegungen
  - Datengesteuerte Auswahl zur Maximierung der Varianzaufklärung
- Strategien bei der datengesteuerten Auswahl
  - Vorwärtsselektion
  - Rückwärtselimination
  - Schrittweise Regression

Verfahren zur Auswahl unabhängiger Variablen (Eid, Gollwitzer & Schmitt, 2017)

- Theoriegeleitete Vorgehensweise
  - Auswahl der Prädiktoren nach inhaltlicher Relevanz bzw. Hypothesen
  - alle (oder in Blöcken) gleichzeitig ins Modell aufnehmen
  - Interpretation erfolgt hypothesengeleitet, nicht datengesteuert
- Hierarchische Regression
  - Prädiktoren werden blockweise eingefügt (z. B. zuerst demografische, dann psychologische Variablen)
  - Vergleich von ΔR^2 zwischen Blöcken → zeigt inkrementelle Varianzaufklärung
- Empfehlung:
  - theoretische bzw. hierarchische Verfahren sind interpretativ stabiler; datengesteuerte Verfahren nur explorativ einsetzen

Interaktionseffekte in der multiplen Regression

- Interaktionseffekt (bzw. Moderatoreffekt bzw. Wechselwirkungseffekt)
- Fiktives Beispiel: Studienerfolg nur dann hoch, wenn IQ (x_1) und Motivation (x_2) hoch sind
- $\hat{y} = 1 \cdot b_0 + x_1 \cdot b_1 + x_1 \cdot x_2 \cdot b_3 + x_2 \cdot b_2$



Inkrement und Dekrement in der multiplen Regression

- Beitrag zur Varianzaufklärung: Für jede einzelne Prädiktorvariable lässt sich ein solcher Beitrag bestimmen
- Unterscheidung zwischen Inkrement und Dekrement
  - Inkrement (R_I^2): Zuwachs an aufgeklärter Varianz durch Hinzunahme weiterer Prädiktorvariablen
  - Dekrement (R_D^2): Abnahme an aufgeklärter Varianz durch Verzicht auf bestimmte Prädiktorvariablen

Inkrement und Dekrement in der multiplen Regression

- Orthogonaler Fall (sämtliche Prädiktorvariablen sind unkorreliert): Addition der Einzelkorrelationen zur Berechnung von R^2 ; R_I^2 (bzw. R_D^2) = $r^2_{x_j,y}$
- Kollinearer Fall (Prädiktoren sind korreliert)
  - R^2 kleiner als Summe der Einzelkorrelationen durch Informationsüberschneidungen (häufiger Fall)
  - R^2 größer als Summe der Einzelkorrelationen: Suppressoreffekte durch Informationspräzisierung (seltener Fall)

$$R^2 = \sum_{j=1}^{m} r^2_{x_j,y}$$

$$R^2 < \sum_{j=1}^{m} r^2_{x_j,y}$$

$$R^2 > \sum_{j=1}^{m} r^2_{x_j,y}$$

Suppressorvariablen in der multiplen Regression

- Suppressorvariablen erhöhen die aufgeklärte Varianz durch Unterdrückung irrelevanter Varianzen anderer Variablen
- Bedingungen für eine Suppressorvariable
  - Keine oder geringe Korrelation mit der Kriteriumsvariable
  - Deutliche Korrelation mit mindestens einer Prädiktorvariable
  - Inkrement bzw. Dekrement der Variable ist (deutlich) größer als einfacher Determinationskoeffizient (R^2) der Suppressorvariable
- Beispiel: Berufserfolg (AV) wird durch Abschlussnote im Studium (UV_1) und Prüfungsangst (UV_2) vorhergesagt
- Prüfungsangst könnte als mögliche Suppressorvariable irrelevante Varianz in der Abschlussnote unterdrücken

Prognosegüte und Kreuzvalidierung (Eid, Gollwitzer & Schmitt, 2017)

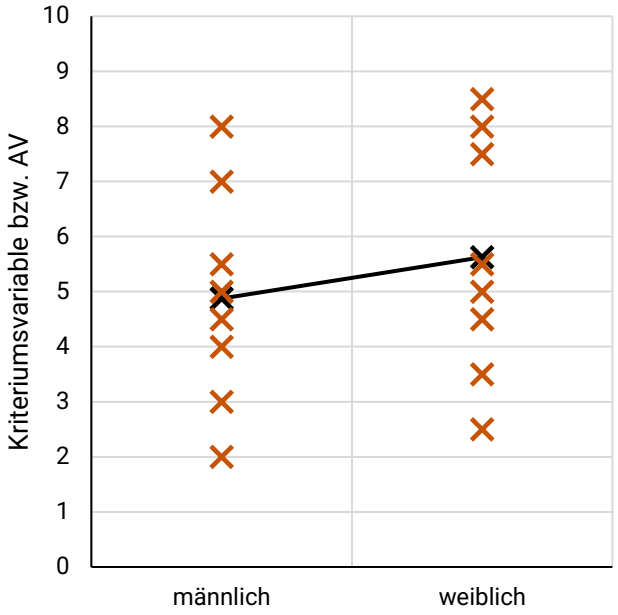
- Prognosegüte: Wie gut kann das Regressionsmodell zukünftige Fälle vorhersagen?
- Überprüfung der Prognosegüte
  - Weitere Stichprobenziehung
  - Kreuzvalidierung
- Kreuzvalidierung (engl. cross-validation): Teilung des Datensatzes in Trainings-Substichprobe und Test-Substichprobe
- Ziel: Overfitting vermeiden

Modellgüte und Residualdiagnostik (z. B. Field, 2018)

- Residualanalyse:
  - Streudiagramm der Residuen gegen vorhergesagte Werte zur Prüfung von Linearität und Homoskedastizität
  - Q-Q-Plot der Residuen zur Prüfung der Normalverteilung
- Überprüfung des Einflusses einzelner Fälle auf Modellschätzung
  - Leverage/Hebelwert: sagt aus, wie stark ein Datenpunkt im Raum der Prädiktorvariablen vom Zentrum entfernt ist und damit, wie stark dieser Punkt die Form der Regressionsgerade "hebeln" kann
  - Mahalanobis-Distanz: misst wie der Hebelwert die Distanz eines Punktes vom Zentrum der Daten im Prädiktorraum, berücksichtigt aber die Korrelationen zwischen den unabhängigen Variablen
  - Cook-Distanz: misst als wichtigster Wert den tatsächlichen Einfluss eines Datenpunkts auf die geschätzten Regressionskoeffizienten; sie kombiniert die Informationen aus dem Hebelwert (Distanz im Prädiktorraum) und dem Residuum (Distanz im Kriteriumraum)

Indikatorcodierung

- Regressionsanalyse mittels Indikatorcodierung auch bei fehlendem Intervallskalenniveau der Prädiktorvariable(n) möglich
- Indikatorcodierung: Umrechnung von nominal- oder ordinalskalierten Prädiktorvariablen in künstliche, intervallskalierte Prädiktorvariablen
- Beispiel: Umrechnung der Variable Geschlecht in eine Indikatorvariable (z. B. ♂ = 0 und ♀ = 1)
- Äquidistanz: Diese Indikatorvariable enthält nur ein Intervall, welches zu sich selbst äquidistant ist und somit Intervallskalenniveau besitzt
- Wichtig: Durch Indikatorcodierung und das Allgemeine Lineare Modell gilt mathematisch: Varianzanalyse = Regressionsanalyse



Signifikanzprüfung zur multiplen Regression (z. B. Eid, Gollwitzer & Schmitt, 2017)

- Signifikanztest für die multiple Regression mittels *F*-Test
- Berechnung des empirischen *F*-Wertes für die multiple Regression:

$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{(1-R^2)}$$

n = Stichprobenumfang
k = Anzahl an Prädiktorvariablen
R^2 = Determinationskoeffizient

- Formel für die Freiheitsgrade:
  - Zählerfreiheitsgrade: $df_Z = k$
  - Nennerfreiheitsgrade: $df_N = n - k - 1$
- Inferenzstatistische Entscheidung: Vergleich empirischer *F*-Wert mit kritischem *F*-Wert → Ergebnis signifikant ($F_{emp} \geq F_{krit}$) oder nicht signifikant ($F_{emp} < F_{krit}$)

Signifikanzprüfung zur multiplen Regression (z. B. Eid, Gollwitzer & Schmitt, 2017)

- Beispiel: Berechnung des *F*-Wertes zu einer multiplen Regression mit zwei Prädiktorvariablen
- Weitere Angaben: R^2 = .338; *N* = 237
- Berechnung des F-Wertes:

$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{(1-R^2)} = \frac{237-2-1}{2} \cdot \frac{0.338}{(1-0.338)} = \frac{234}{2} \cdot \frac{0.338}{0.662} \cong 59.737$$

- Inferenzstatistische Entscheidung: Vergleich empirischer *F*-Wert mit kritischem *F*-Wert

Signifikanzprüfung zur multiplen Regression (z. B. Eid, Gollwitzer & Schmitt, 2017)

- „Varianzanalytische Darstellung“: *F*-Wert für die multiple Regression auch als Quotient aus Quadratsummen darstellbar:

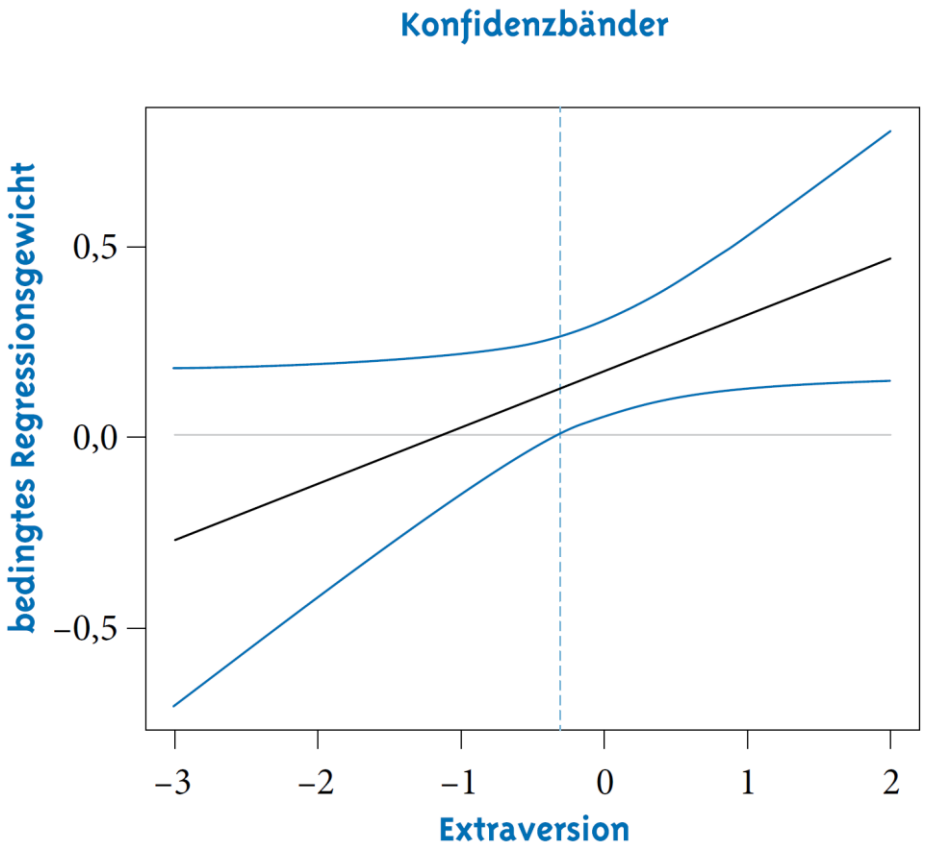
$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{(1-R^2)} \quad \Rightarrow \quad F = \frac{MQSR}{MQSE}$$

MQSR = Mittlere Quadratsumme der Regression
MQSE = Mittlere Quadratsumme der Residuen

- Weiterer Beleg dafür, dass mathematisch gilt:
- Regressionsanalyse = Varianzanalyse

Konfidenzbänder (Eid, Gollwitzer & Schmitt, 2017)

- Konfidenzintervall (= Vertrauensintervall): Beschreibt den Bereich um den geschätzten Populationsparameter, für den gilt, dass er mit einer Wahrscheinlichkeit von 1−α (z. B. 95%) den „wahren“ Populationsparameter beinhaltet
- Konfidenzbänder: Vgl. Konfidenzintervalle
- Gekrümmte Form der Konfidenzbänder



Quelle: Eid, Gollwitzer und Schmitt (2017)

Inferenzstatistische Voraussetzungen (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Linearität zwischen den Prädiktoren und dem Kriterium
- Intervallskalenniveau der Kriteriumsvariable
- Normalverteilung der Residuen
- Unabhängigkeit der einzelnen Messwerte verschiedener Personen
- Homoskedastizität: Homogenität der Streuungen der zu einem x-Wert gehörenden y-Werte über den gesamten Wertebereich von x (vgl. inferenzstatistische Voraussetzungen der MANOVA ohne MW; Vorlesung „Statistik II“)
- Keine Multikollinearität zwischen den Prädiktoren (keine zu hohe Korrelation)
- Keine einflussreichen Ausreißer, die das Modell unverhältnismäßig stark beeinflussen

Beispiele für die multiple lineare Regression in Fachzeitschriften

We then used the attributes to predict perceived warmth and competence. Two separate linear multiple regressions were conducted with perceived warmth and competence as the dependent variables. We used multiple dummy codes for the gender displayed, developmental categories, and locomotion.

A robot was perceived with higher warmth if its mechanics were invisible ($\beta = -0.13$, $SE = 0.05$, and $p < .05$), it had more degrees of freedom for movement ($\beta = 0.12$, $SE = 0.01$, $p < .05$), younger age ($\beta = -0.14$, $SE = 0.01$, $p < .05$), and displayed as a child ($\beta = -0.14$, $SE = 0.01$, $p < .05$). The regression was significant $F(25, 316) = 5.90$, $p < .001$ with the four factors explaining 26.4% of the variance in perceived warmth. The standardized beta coefficients are presented in Table 6.

Quelle: Reeves, Hancock und Liu (2020)

However, the three other cases (Fig. 1b–e) could not be covered with this multiple regression equation as in these cases both (suboptimal and optimized) conditions have non-linear progression lines. To solve this issue, it was decided to include two dummy variables in the regression equation, one indicating the optimized visual design (as dumOpt, see Eq. 1) and one indicating the suboptimal visual design (as dumSub, see Eq. 2):

$$\text{perf} = b_0 + b_1\text{abi} + b_{21}\text{dumOpt} + b_{22}\text{dumSub} + b_{31}\text{abi} \times \text{dumOpt} + b_{32}\text{abi} \times \text{dumSub} + b_{41}(\text{abi} \times \text{dumOpt})^2 + b_{42}(\text{abi} \times \text{dumSub})^2 \quad (3)$$

Quelle: Kühl, Fehringer und Münzer (2022)

Table 3 Multiple regression to predict students' grade in the first exam

	B	SE B	β	*p*
Constant	-3.93	0.59		<0.001
Cognitive strategy use	0.21	0.18	0.17	0.24
Metacognitive strategy use	0.00	0.14	0.00	0.99
Self-efficacy	0.23	0.11	0.24	0.04

Quelle: Gentner, Respondek und Seufert (2024)

Zusammenfassung I

- Multiple Regression zur Vorhersage einer Kriteriumsvariable durch mehrere Prädiktorvariablen mittels Linearkombination
- Auswahl unabhängiger Variablen aufgrund theoretischer Überlegungen oder datengesteuert zur Maximierung der Varianzaufklärung
- Interaktionseffekte in der multiplen Regression
- Inkrement/Dekrement als Zuwachs/Abnahme an aufgeklärter Varianz in der multiplen Regression durch Hinzunahme/Verzicht von Prädiktorvariablen
- Suppressorvariablen erhöhen die aufgeklärte Varianz durch Unterdrückung irrelevanter Varianzen anderer Variablen
- Überprüfung der Prognosegüte durch weitere Stichprobenziehung oder Kreuzvalidierung zur Overfitting-Vermeidung

Zusammenfassung II

- Indikatorcodierung als Umrechnung in künstliche, intervallskalierte Prädiktorvariablen
- Signifikanzprüfung zur multiplen Regression mittels *F*-Test (vgl. Varianzanalyse)
- Konfidenzbänder u. A. zur Abschätzung des Wertebereiches, in dem sich der „wahre“ Wert mit hoher Wahrscheinlichkeit befindet
- Inferenzstatistische Voraussetzungen: Intervallskalenniveau & Normalverteilung der Kriteriumsvariable, Unabhängigkeit der einzelnen Messwerte & Homoskedastizität

Prüfungsliteratur

- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
  - Multiple Regressionsanalyse (S. 629–726) → Ohne Rechenaufgaben

Weiterführende Literatur

- Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2021). *Quantitative Methoden 1: Einführung in die Statistik für Psychologie, Sozial- & Erziehungswissenschaften* (5. Aufl.). Heidelberg: Springer.
  - Merkmalszusammenhänge (S. 105–119)
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.
  - Partielle Korrelation und multiple lineare Regression (S. 339–361)
- Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. Sage.
  - The Lineare Model (Regression) (S. 369–436)
- Leonhart, R. (2022). *Lehrbuch Statistik. Einstieg und Vertiefung* (5. Auflage). Bern: Huber.
  - Multiple Korrelation und Multiple Regression (S. 355–400)