TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Professur Psychologie digitaler Lernmedien

Institut für Medienforschung

Philosophische Fakultät

Statistik I

Regression

Matrix (1999). Warner Bros. Pictures.

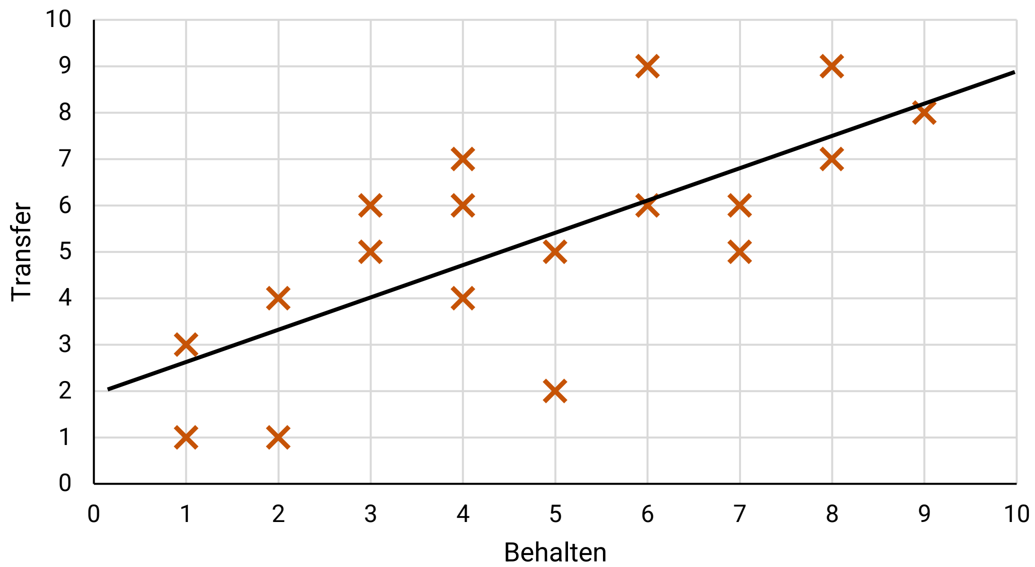
Prof. Dr. Günter Daniel Rey

Überblick

- Lineare bivariate Regression
- Methode der kleinsten Quadrate
- Nichtlineare Zusammenhänge
- Multiple Regression
- Indikatorcodierung
- Inferenzstatistische Voraussetzungen

Lineare bivariate Regression
(z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- **Lineare bivariate Regression:** Statistisches Verfahren zur Vorhersage einer Kriteriumsvariable durch eine Prädiktorvariable mittels linearer Funktion
- **Fiktives Beispiel:** Zusammenhang zwischen Behaltens- und Transferleistungen



Lineare bivariate Regression

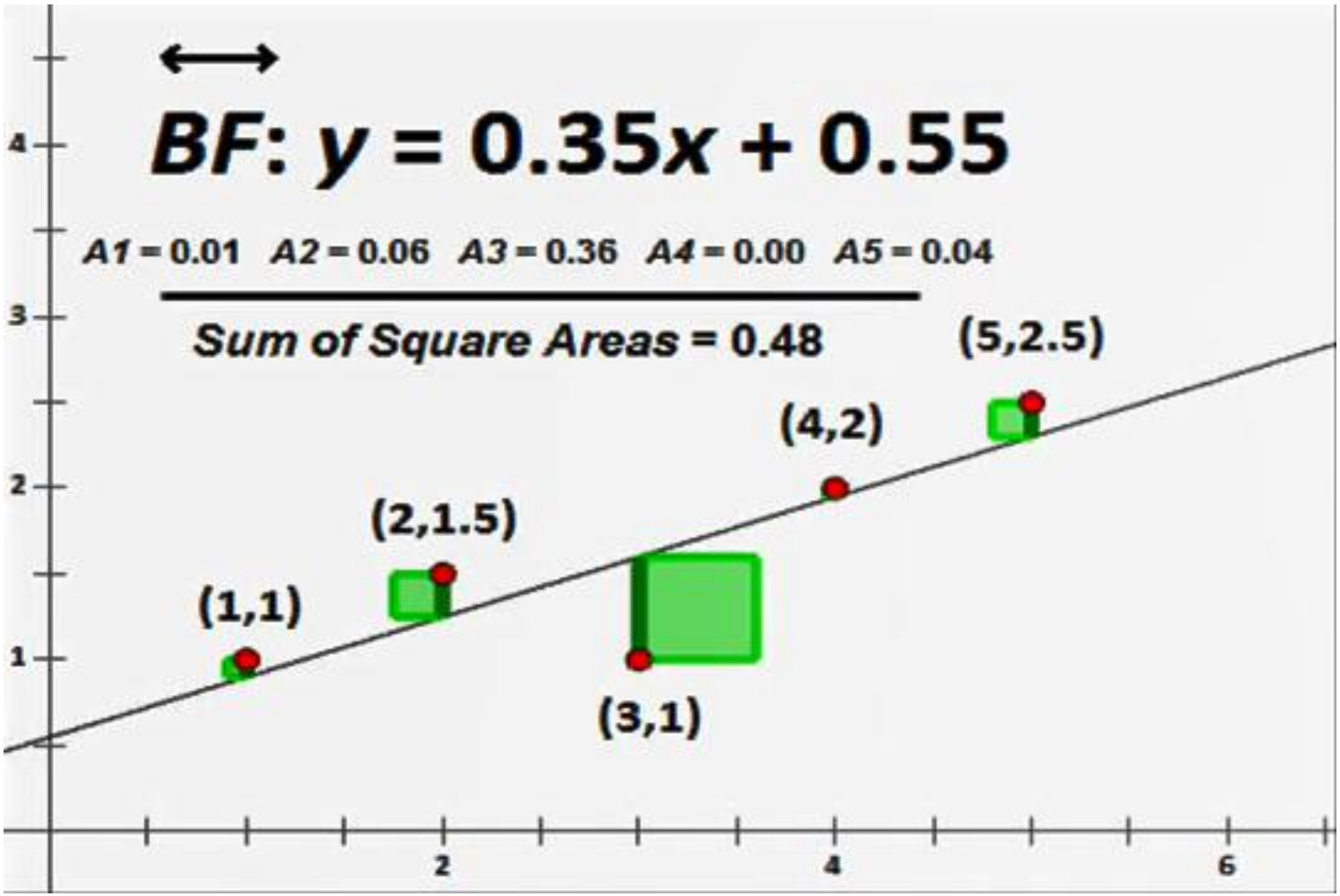
- Regressionsgrade soll den Gesamttrend der Einzelwerte bestmöglich wiedergeben
- Regressionsgleichung zur Regressionsgraden:

$$\boxed{\hat{y} = m \cdot x + b}$$

ŷ	= Vorhergesagte Kriteriumsvariable y
m	= Steigung der Regressionsgraden
x	= Prädiktorvariable x
b	= Achsenabschnitt der Regressionsgraden

- Berechnung der Regressionsgewichte m und b mittels Methode der kleinsten Quadrate

Methode der kleinsten Quadrate

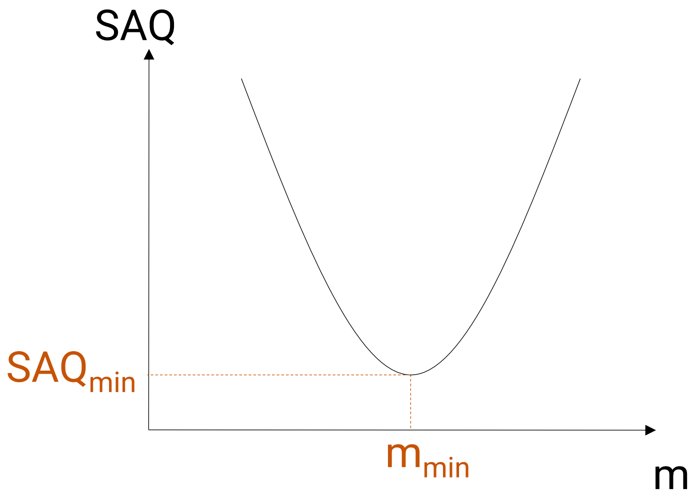


Quelle: http://www.youtube.com/watch?v=jEEJNz0RK4Q

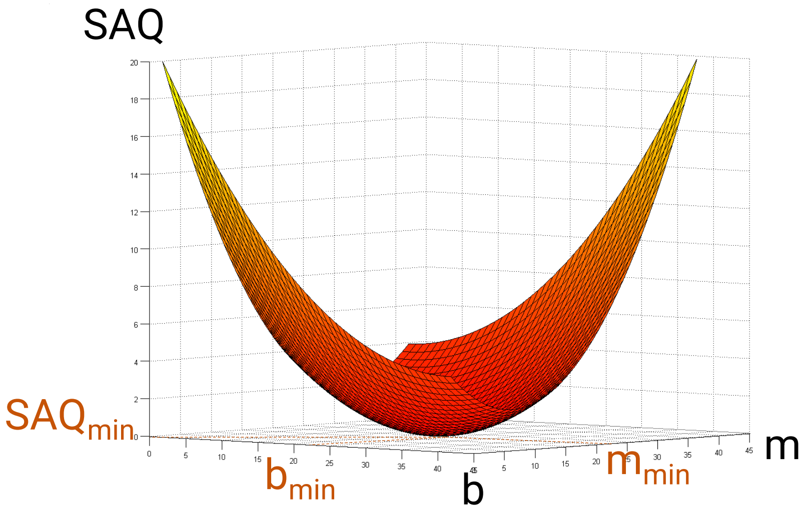
Methode der kleinsten Quadrate

- Summe der Abweichungsquadrate (SAQ) soll ein Minimum ergeben
- Ein Gewicht: Parabel (→ Regressionsgerade durch Achsenursprung)
- Zwei Gewichte: Paraboloid (→ Regressionsgrade)

Für ein Gewicht



Für zwei Gewichte



Methode der kleinsten Quadrate

- Summe der Abweichungsquadrate (SAQ) soll ein Minimum ergeben
- Formel:

$$SAQ = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} [y_i - (m \cdot x_i + b)]^2 = \min$$

- Erste Ableitung bilden und auf Null setzen ergibt für m und b:

$$m_{yx} = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}$$

$$b_{yx} = \bar{y} - m_{yx} \cdot \bar{x}$$

y	= Beobachtete Werte der Variablen y
ŷ	= Vorhergesagte Kriteriumsvariable y
m	= Steigung der Regressionsgraden
x	= Prädiktorvariable x
b	= Achsenabschnitt der Regressionsgraden
i	= Person i

Lineare bivariate Regression

- Beispiel: Berechnung von b und m zu dem rechts dargestellten Datensatz:

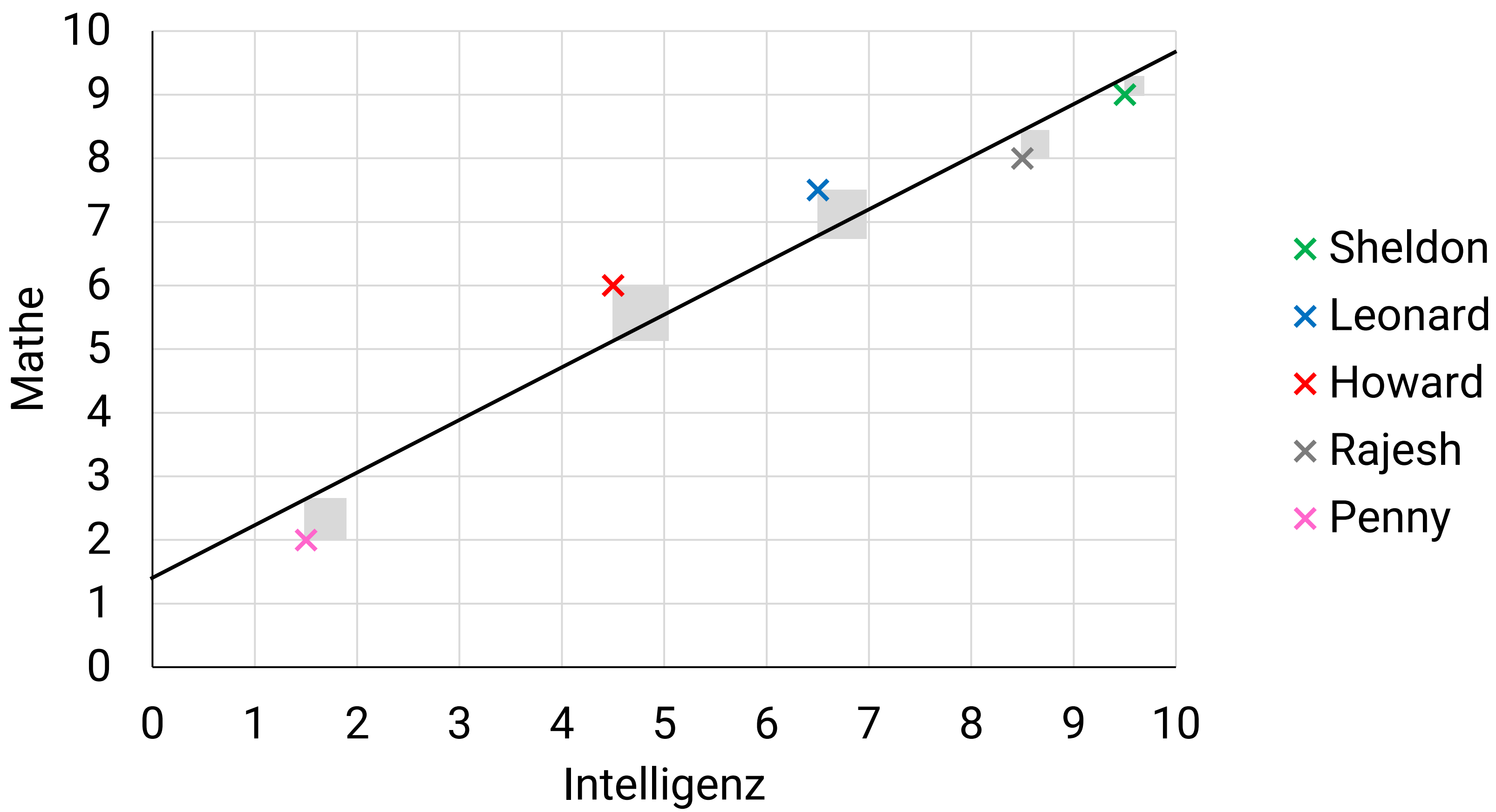
$$m_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} \approx \frac{8.5}{3.21^2} \approx 0.82$$

$$b_{yx} = \bar{y} - m_{yx} \cdot \bar{x} \approx 6.5 - 0.82 \cdot 6.1 = 1.47$$

VPN	IQ	Mathe
Sheldon	9.5	9.0
Leonard	6.5	7.5
Howard	4.5	6.0
Rajesh	8.5	8.0
Penny	1.5	2.0
M	6.1	6.5
SD	3.21	2.74

Lineare bivariate Regression

- Beispiel: Regressionsgrade mit b = 1.47 und m = 0.82:



Lineare bivariate Regression

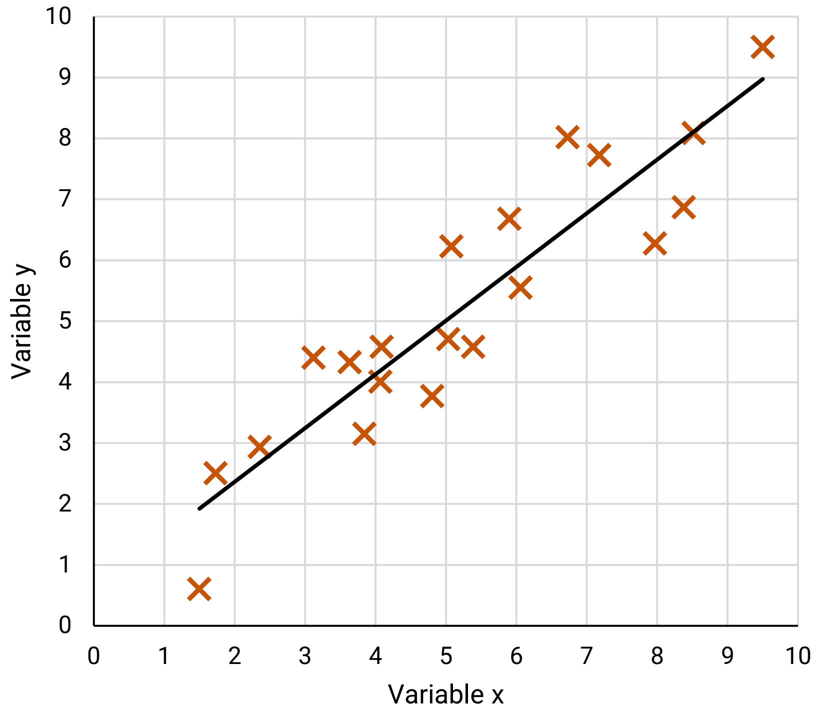
Wie hoch ist die Intelligenz laut Regressionsgleichung für das Beispiel auf der vorherigen Folie bei einer Person mit einem Mathewert von 4?

- ☐ A: 3
- ☐ B: 3.09 (gerundet)
- ☐ C: 4
- ☐ D: 4.75
- ☐ E: Wert kann nicht berechnet werden

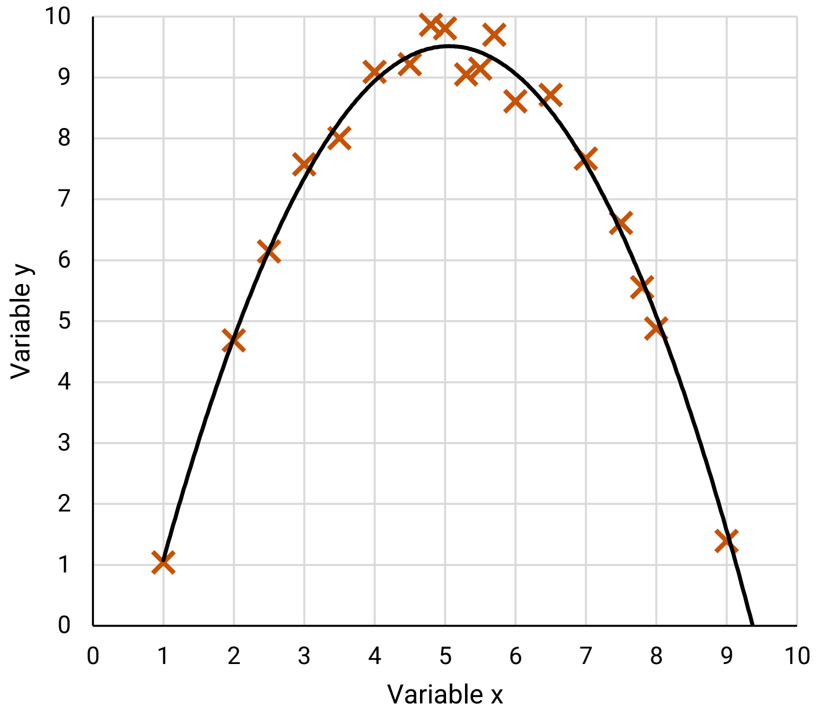
Nichtlineare Zusammenhänge (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Beispiele für lineare und nonlineare Zusammenhänge

Linearer Zusammenhang



Nonlinearer Zusammenhang



Multiple univariate Regression

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

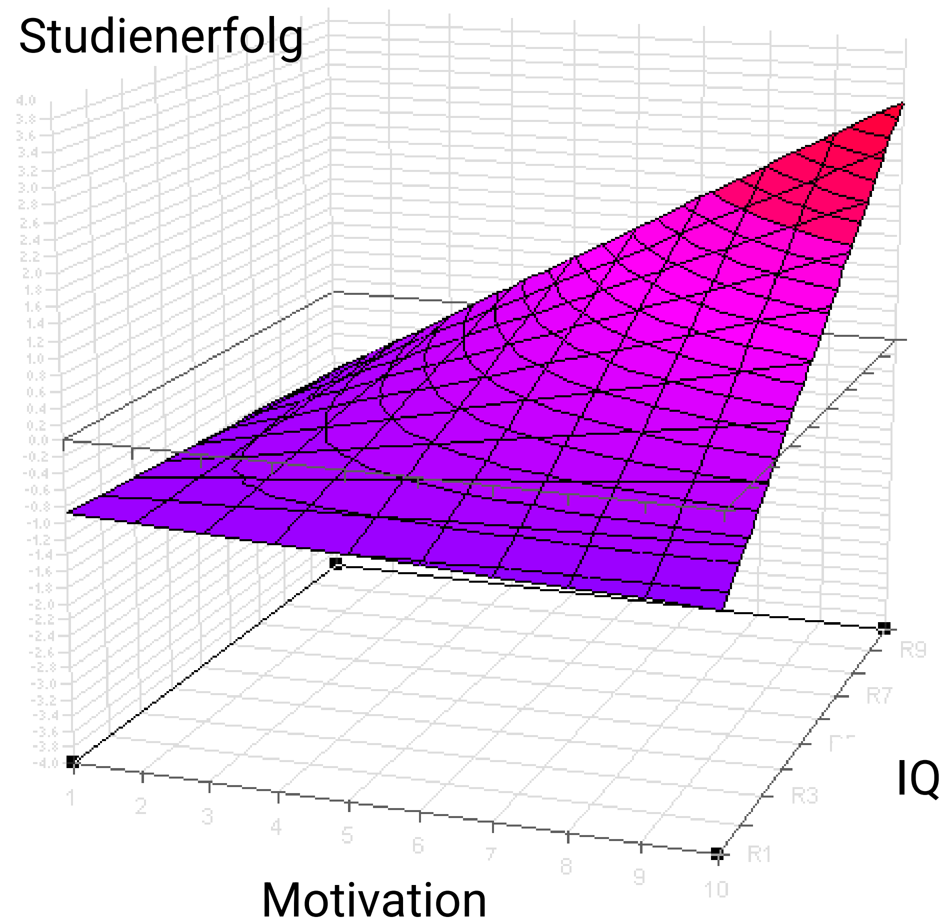
$$\boxed{\hat{y} = 1 \cdot b_0 + x_1 \cdot b_1 + \ldots + x_m \cdot b_m}$$

- 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

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)

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

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

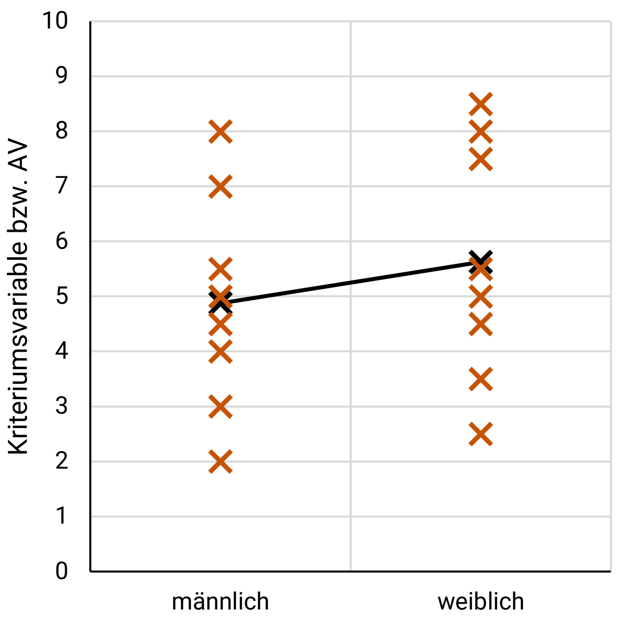
$$R^2 > \sum_{j=1}^{m} r^2_{x_j,y}$$

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

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



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

- Intervallskalenniveau der Kriteriumsvariable
- Normalverteilung der Kriteriumsvariable in der Population
- 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)

Beispiele für Korrelationen und Regressionen in Fachzeitschriften

The data were analyzed by means of a 2 × 2 ANOVA with the learner's gender and the speaker's gender as between-factors. For the analysis of problem-solving performance, two additional control variables were included, namely the "Abiturnote" (i.e., final high school grade point average) and intrinsic motivation, resulting in a 2 × 2 ANCOVA. Both covariates showed a significant correlation with problem-solving performance (Abiturnote: $r = -.36$, $P = .001$, whereby better school grades were associated with better learning outcomes; intrinsic motivation: $r = .26$, $P = .02$, whereby higher intrinsic motivation was associated with better learning outcomes), but were independent from each other ($r = -.01$, $P = .94$). The results of the experiment are shown in Table 1.

Quelle: Linek, Gerjets und Scheiter (2010)

Table 6. Correlations between indices of game performance, pre-test and learning outcome

	2	3	4	5	6
(1) Level Reached	.99***	.49**	.41*	.43*	.18
(2) Unique Maths Tasks		.55**	.38*	.44*	.20
(3) All Maths Tasks			.-.23	.37*	.25
(4) Accuracy				.06	-.14
(5) Pre-test					-.01
(6) Gain					

Note. * = p<0.05, ** = p<0.01, *** = p<0.001 (two-tailed test of significance).

Quelle: Habgood und Ainsworth (2011)

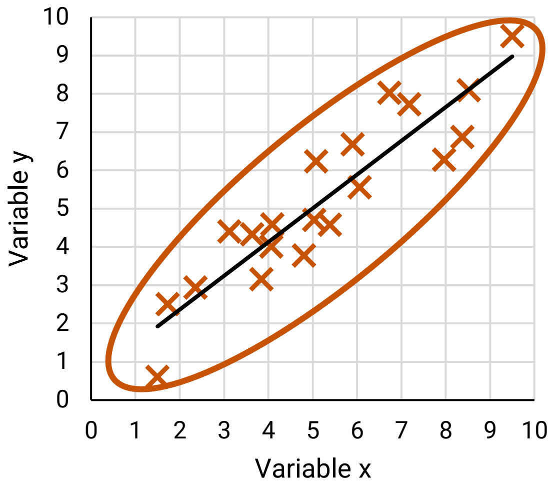
Table 2
Summary of regression analysis for variables predicting posttest scores.

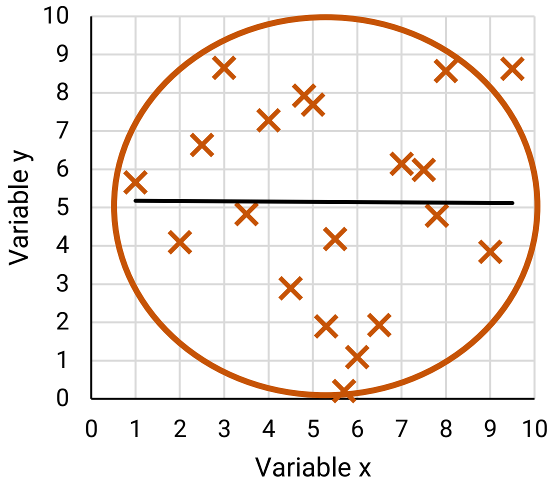
Predictor variable	*B*	*SE*	*β*
Unique simulation experiments	0.05	0.02	0.23
Relations aspect model quality	−0.12	0.04	−0.29

Quelle: Mulder, Lazonder und de Jong (2014)

Zusammenfassung

- Lineare bivariate Regression: Statistisches Verfahren zur Vorhersage einer Kriteriumsvariable durch eine Prädiktorvariable mittels linearer Funktion
- Methode der kleinsten Quadrate zur Berechnung der Regressionsgewichte
- Nichtlineare Zusammenhänge ebenfalls vorhersagbar
- Multiple Regression zur Vorhersage einer Kriteriumsvariable durch mehrere Prädiktorvariablen mittels Linearkombination
- Indikatorcodierung als Umrechnung in künstliche, intervallskalierte Prädiktorvariablen
- Inferenzstatistische Voraussetzungen: Intervallskalenniveau & Normalverteilung der Kriteriumsvariable, Unabhängigkeit der einzelnen Messwerte & Homoskedastizität





Prüfungsliteratur

- 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)

Weiterführende Literatur I

- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.
  - Einfache lineare Regression (S. 183–202)
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
  - Abhängigkeiten zwischen zwei Variablen: Einfache lineare Regression (S. 589–613)
- Leonhart, R. (2022). *Lehrbuch Statistik. Einstieg und Vertiefung* (5. Auflage). Bern: Huber.
  - Korrelation und Regression (S. 261–378)

Weiterführende Literatur II

- Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik: Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (3. Aufl.). München: Pearson.
  - Lineare Regression (S. 245–288)