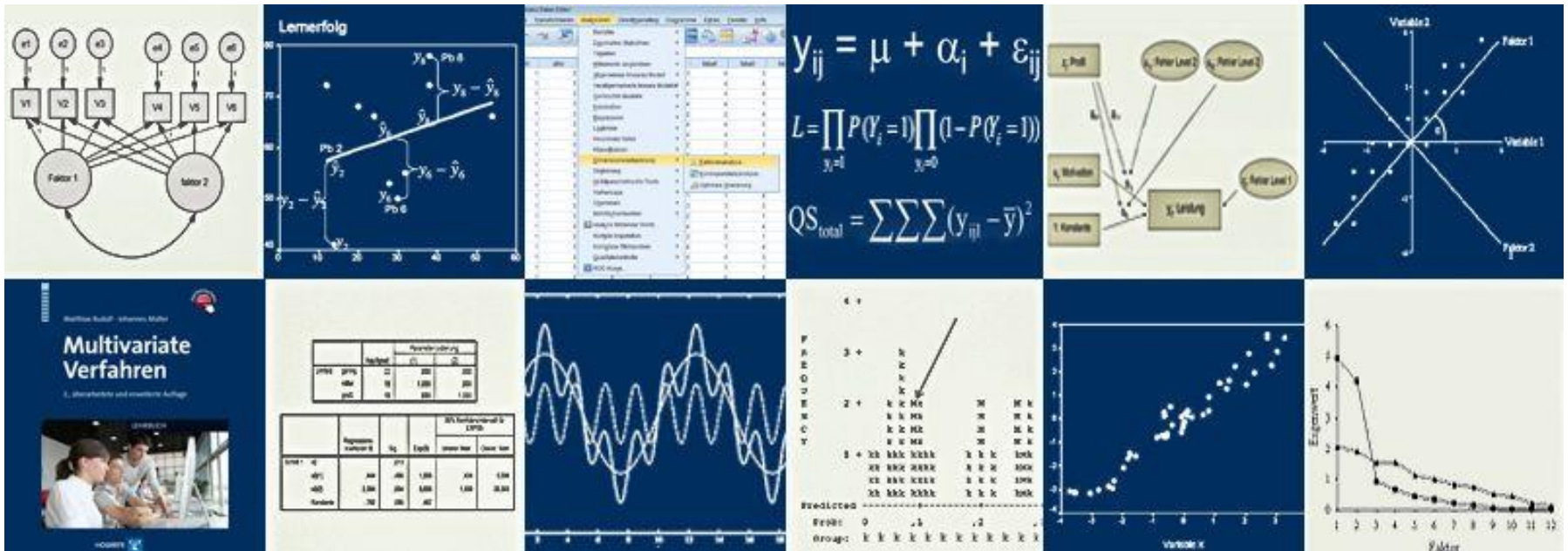


➤ Multivariate Statistik – Überblick

Dr. rer.nat. Matthias Rudolf

Diplom-Mathematiker

Professur Methoden der Psychologie



Multivariate Statistik – Überblick

Dr. rer.nat. Matthias Rudolf

Büro BZW A 317

Telefon: 463 36166

Matthias.Rudolf@TU-Dresden.de

<https://tu-dresden.de/mn/psychologie/methpsy/die-professur/beschaefigte/matthias-rudolf/index>

Allgemeine Sprechstunde: Jeden Montag 12.30 – 13.00 Uhr

➤ **Multivariate Statistik – Überblick**

**Spezielle M3-Sprechstunde im Wintersemester:
jeden Montag 11.45-12.30 Uhr**

(Raum BZW A317, ohne Voranmeldung)

- Klärung offener Fragen aus Vorlesung bzw. Computerseminar, Hilfe bei PVL-Aufgaben
- Diskussion von PVL-Aufgaben bzw. -Lösungen sowie von Aufgaben zur Prüfungsvorbereitung
- Besprechung sonstiger Probleme im Zusammenhang mit M3

➤ **Multivariate Statistik – Überblick**

Vorlesungen (ASB 28): Donnerstag, 3. DS

Computerseminare (SE2 / 101):

Gruppe 1: Montag, 2. DS

Gruppe 2: Montag, 5. DS

Gruppe 3: Montag, 6. DS

Gruppe 4: Dienstag, 2. DS

Gruppe 5: Mittwoch, 2. DS

Individuelles Tauschen von Terminen in Gruppen ist problemlos möglich – Tauschpartner erforderlich!

➤ Multivariate Statistik – Überblick

E	Einführung
E-SPSS	Einführung in die Arbeit mit SPSS
ERA	Einfache lineare Regressionsanalyse
MRA	Multiple lineare Regressionsanalyse und partielle Korrelationsanalyse
LogRA	Logistische Regressionsanalyse
FA	Faktorenanalyse
UVA	Mehrfaktorielle univariate Varianzanalyse
MVA	Multivariate Varianzanalyse, Varianzanalyse mit Messwiederholungen, Kovarianzanalyse
DA	Diskriminanzanalyse
CA	Clusteranalyse
ZRA	Zeitreihenanalyse
HLM	Ausblick auf komplexe multivariate Verfahren: Hierarchische lineare Modelle
LSM	Ausblick auf komplexe multivariate Verfahren: Lineare Strukturgleichungsmodelle
KA	Komplexe Anwendungsaufgabe

➤ Multivariate Statistik – Überblick

	Seminar Gr. 2 Montag, 5. DS, SE2/101	Seminar Gr. 4 Dienstag, 2. DS, SE2/101	Seminar Gr.5 Mittwoch, 2. DS, SE2/101	Seminar Gr. 1 Montag, 2. DS, SE2/101	Seminar Gr. 3 Montag, 6. DS, SE2/101	Vorlesung Donnerstag, 3. DS, ASB 28
14.10.- 18.10.	---	---	---	---	---	E/ERA
21.10.- 25.10.	E-SPSS	E-SPSS	E-SPSS	E-SPSS	E-SPSS	MRA
28.10.- 01.11.	ERA	ERA	ERA	ERA	ERA	Feiertag
04.11.- 08.11.	---	---	---	---	---	MRA
11.11.- 15.11.	MRA I	MRA I	MRA I	MRA I	MRA I	LogRA
18.11.- 22.11.	MRA II	MRA II	Feiertag	MRA II	MRA II	LogRA
25.11.- 29.11.	LogRA	LogRA	MRA II	LogRA	LogRA	FA Abgabe PVL 1

02.12.- 06.12.	FA I	FA I	LogRA	FA I	FA I	FA
09.12.- 13.12.	FA II	FA II	FA I	FA II	FA II	CA
16.12.- 20.12.	CA	CA	FA II	CA	CA	UVA
06.01.- 10.01.	---	---	CA	---	---	MVA
13.01.- 17.01.	UVA	UVA	UVA	UVA	UVA	MVA Abgabe PVL 2
20.01.- 24.01.	MVA I	MVA I	MVA I	MVA I	MVA I	HLM
27.01.- 31.01.	MVA II	MVA II	MVA II	MVA II	MVA II	LSM
03.02.- 07.02.	KA	KA	KA	KA	KA	ZRA / DA

➤ Multivariate Statistik – Überblick

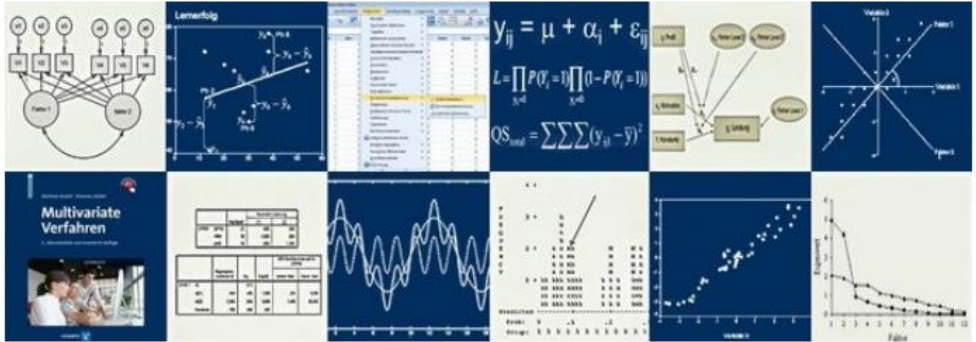
- Ablaufplan
- Inhaltsübersicht
- Literaturangaben
- **Probeklausur**
- Prüfungsrelevanz
- **Folien der Vorlesung**
- **Aufgaben, Lösungshilfen, Daten der Computerseminare**
- **PVL-Aufgaben**
- **Aufgaben zur Prüfungsvorbereitung**

STRUKTUR UND MODULE > MODUL M3

Modul M3 - Multivariate Statistik

Bachelorstudiengang Psychologie

MODULVERANTWORTLICHER:
> Dr. Matthias Rudolf



The grid contains 12 small images:

- 1. A network diagram with nodes and connections.
- 2. A scatter plot with a regression line and axes labeled x_1 and x_2 .
- 3. A screenshot of a software interface showing a list of data points.
- 4. A regression equation: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.
- 5. A diagram showing a flow from 'Lernaktivität' to 'Lernleistung' and 'Lernzeit'.
- 6. A scatter plot with axes labeled 'Wahrheit' and 'Fehler'.
- 7. A book cover titled 'Multivariate Verfahren'.
- 8. A table with columns for 'Faktor', 'Bedingung', and 'Ergebnis'.
- 9. A time series plot showing a fluctuating signal.
- 10. A table with columns for 'Faktor', 'Bedingung', and 'Ergebnis'.
- 11. A scatter plot showing a positive correlation between two variables.
- 12. A line graph showing a decreasing trend over time.

© MR

Kennwort:

<https://tu-dresden.de/mn/psychologie/methpsy/studium/struktur-und-module/modul-m3/index>

➤ Multivariate Statistik – Überblick

Verbindliche Prüfungsliteratur :

Rudolf, M. & Buse, J. (2019). Multivariate Verfahren. Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS (3. Aufl.). Göttingen: Hogrefe.

oder

Rudolf, M. & Müller, J. (2012). Multivariate Verfahren. Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS (2. Aufl.). Göttingen: Hogrefe.



➤ Multivariate Statistik – Überblick

Was wollen wir im Modul M3 erreichen?

Sie sollen ...

- ... die Prinzipien und Vorgehensweisen der behandelten Verfahren verstanden haben
- ... die Methoden zur Lösung fachlicher Fragestellungen sachgerecht auswählen und anwenden können
- ... die Ergebnisse der Verfahren (Outputs) interpretieren und auf fachwissenschaftliche Fragen übertragen können
- ... fit sein im Umgang mit SPSS (für Ihr Bachelorstudium und als sehr wichtige Grundlage für Ihr späteres Masterstudium)

➤ **Multivariate Statistik – Überblick**

➤ **Prüfungsleistung (PL)**

- Theoretische Fragen zu den Methoden
- Anwendung der Verfahren
- Interpretation von SPSS-Outputs

➤ **Prüfungsvorleistung (PVL)**

- Praktischer Umgang mit SPSS
- Interpretation der Ergebnisse

➤ **Aufgaben zur langfristigen Prüfungsvorbereitung**

- Langfristige Vorbereitung auf die PL

➤ Multivariate Statistik – Überblick

Prüfungsleistung (PL): Klausur

→ Dauer: 90 Minuten

→ Benutzbare Unterlagen:

Geschriebenes oder gedrucktes Material
(z.B. Mitschriften, Bücher)

→ Nicht benutzbar:

Rechner, jegliche „Außenverbindungen“ (z.B. Internet)

→ Mitzubringen:

leere A4-Blätter, Stifte

Personalausweis + Studentenausweis

➤ **Multivariate Statistik – Überblick**

Prüfungsleistung (PL): Klausur

Inhalte der Klausur:

- Inhalte der Vorlesung
- Inhalte der Computerseminare
- Prüfungsliteratur (wird zu jedem Kapitel angegeben)

Kompetenzen aus den Modulen M1 und M2 werden vorausgesetzt!

➤ Multivariate Statistik – Überblick

Im Internet: Klausur aus dem Wintersemester 2011/12

Aufgaben dieser Klausur und deren Lösungen werden z.B. im Rahmen der Aufgaben zur Prüfungsvorbereitung besprochen.

Aufgabe 1 (6 Punkte):

Im Ergebnis einer multiplen linearen Regression mit dem Kriterium Y und den Prädiktoren X_1 , X_2 und X_3 ergab sich das Bestimmtheitsmaß R^2_X .

In einer weiteren multiplen linearen Regression mit dem Kriterium Y und den Prädiktoren Z_1 , Z_2 und Z_3 ergab sich das Bestimmtheitsmaß R^2_Z .

In einer multiplen linearen Regression mit dem Kriterium Y und mit allen sechs Prädiktoren X_1 , X_2 , X_3 , Z_1 , Z_2 und Z_3 ergab sich das Bestimmtheitsmaß R^2_{XZ} .

- a) Kann dabei die Beziehung $R^2_{XZ} < R^2_X + R^2_Z$ gelten? Begründen Sie Ihre Antwort!
- b) Kann dabei die Beziehung $R^2_{XZ} = R^2_X + R^2_Z$ gelten? Begründen Sie Ihre Antwort!
- c) Kann dabei die Beziehung $R^2_{XZ} > R^2_X + R^2_Z$ gelten? Begründen Sie Ihre Antwort!

Aufgabe 2 (4 Punkte):

In einer Untersuchung soll geprüft werden, welchen Einfluss die tägliche Kalorienaufnahme (erfasst in kcal pro Tag) auf den Gesundheitszustand (erfasst auf einer metrischen Skala) von Erwachsenen hat. Dabei ist zu beachten, dass sich dieser Einfluss mit zunehmendem Alter (erfasst in Jahren) verändern (verstärken oder verringern) könnte.

- a) Welches der folgenden Modelle ist für diese Situation zutreffend: Partielle Korrelation, Moderation oder Mediation?
- b) Stellen Sie das entsprechende Modell für das gegebene Beispiel schematisch dar.
- c) Beschreiben Sie das Prinzip der statistischen Datenanalyse für das entsprechende Modell.

➤ Multivariate Statistik – Überblick

Prüfungsvorleistung (PVL): Lösungen von SPSS-Aufgaben

(individuelle Lösungen auf Basis der Matrikelnummer)

→ Erreichbare Punkte: 40

→ Notwendige Punktzahl: 20

→ Anzahl der Aufgabenblätter: 2 (je 20 Punkte)

→ Angebot der Aufgabenblätter: Internet

→ **Abgabe der Auswertungsblätter (keine Abgabe per mail möglich!): vor der Vorlesung am 28.11. bzw. 16.01.**

→ **Punktekontrolle:** im Internet ab 22.01.

→ **Nachfragen und jederzeit Hilfe:** in der M3-Sprechstunde

Multivariate Statistik – Überblick

Dr. Matthias Rudolf
Modul M3: Multivariate Statistik

01.09.2018

PVL Teil 1: Aufgaben 1-9

Abgabetermin: 15.11.2018 (Lösungsblätter in Papierform)

Im Internet finden Sie Daten-Datei **PVL_1_Daten_1.sav** zur Lösung der Aufgaben 1-4. Gegeben sind 40 Datensätze, wobei 2 Werte fehlen.

Proband	Praedikor_1	Praedikor_2	Praedikor_3	Kriterium	Geschlecht
1	51	100	43	.	.
2	62	118	80	1232,0	0
3	53	90	50	986,0	0
4	62	80	83	1069,0	0

- Geben Sie bei Proband 1 in der Variable Geschlecht den für Sie zutreffenden Wert ein (0= männlich, 1= weiblich).
- Tragen Sie bei Proband 1 in der Variablen Kriterium eine Zahl ein, die aus den letzten 3 Ziffern Ihrer Matrikelnummer besteht. (Beispiel: Falls Ihre Matrikelnummer z.B. 4128645 ist, tragen Sie bitte den Wert 645 ein).

Bearbeiten Sie alle folgenden Aufgaben mit der so modifizierten Datei. Geben Sie alle Ergebnisse mit mindestens 3 Nachkommastellen an.

Aufgabe 1 (2 Punkte)

Berechnen Sie Mittelwert und Standardabweichung der Variable Kriterium.

Multivariate Statistik – Überblick

Modul M3: Lösungsblatt PVL Teil 1; Abgabetermin 15.11.

Bitte drucken Sie dieses Lösungsblatt aus und tragen Sie die Ihre Lösungen ein. Falls der Platz nicht reichen sollte, nutzen Sie bitte zusätzlich die Rückseite des Blattes.

Das Blatt können Sie während der Vorlesungen abgeben oder bis zum Abgabetermin in den Briefkasten von Dr. Rudolf in der 3. Etage des BZW werfen. Elektronisch per mail verschickte Lösungen werden nicht berücksichtigt.

NAME:

VORNAME:

MATRIKELNUMMER:

Aufgabe 1 (2 Punkte):

Mittelwert =

Standardabweichung =

Aufgabe 2 (2 Punkt):

t =

p =

Multivariate Statistik – Überblick

Aufgaben zur langfristigen Prüfungsvorbereitung: inhaltliche Fragen zur selbständigen (!) Bearbeitung

Dr. Matthias Rudolf
Modul M3: Multivariate Statistik

01.09.2018

Aufgaben zur Prüfungsvorbereitung: Aufgaben 1-10

Lösungsbesprechung am 18.10. während der Vorlesung

Hinweis: Bei jeder Aufgabe können eine oder mehrere der angegebenen Lösungsalternativen korrekt sein.

Aufgabe 1:

In einem Forschungsprojekt soll die Überlegenheit einer neu entwickelten Therapie gegenüber einer Standardtherapie nachgewiesen werden (Verringerung der Dauer der Krankheit).

Als statistische Hypothesen bezeichnet man die statistische Nullhypothese H_0 und die statistische Alternativhypothese H_1 .

Welche dieser Hypothesen ergibt sich unmittelbar aus der inhaltlichen Hypothese?

- H_0
- H_1
- beide
- keine von beiden

➤ **Multivariate Statistik – Überblick**

➤ **Prüfungsleistung (PL)**

- Theoretische Fragen zu den Methoden
- Anwendung der Verfahren
- Interpretation von SPSS-Outputs

➤ **Prüfungsvorleistung (PVL)**

- Praktischer Umgang mit SPSS
- Interpretation der Ergebnisse
- **Unterstützung in der M3-Sprechstunde**

➤ **Aufgaben zur langfristigen Prüfungsvorbereitung**

- Langfristige Vorbereitung auf die PL
- **Unterstützung in der M3-Sprechstunde**

➤ Multivariate Statistik – Überblick

Statistik-Software

- I Programmpakete
- SPSS (incl. AMOS)
 - SAS
 - BMDP
 - STATISTICA
 -

- II Spezielle Software
- HLM
 - Mplus
 - StatXact
 - STATA
 - EquivTest
 -

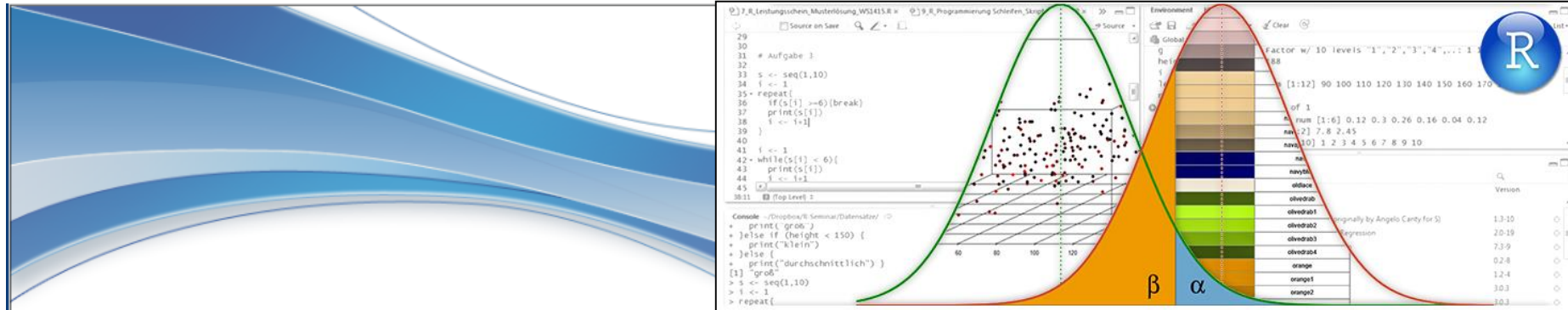
- III freie Software im Internet
- R
 - Stichprobenumfangsplanung
 -

Normal Power Calculations

Normal Distribution 2-Sample Equal Variances

μ_1 The Mean of Population 1	<input type="text"/>
μ_2 The Mean of Population 2	<input type="text"/>
Sigma Common Standard Deviations for both Populations	<input type="text"/>
Number of Sides Specifies Alternative Hypothesis. One sided and $\mu_1 > \mu_2 \Rightarrow H_1 : \mu_1 > \mu_2$ One sided and $\mu_1 < \mu_2 \Rightarrow H_1 : \mu_1 < \mu_2$ Two sided $\Rightarrow H_1 : \mu_1 \text{ not equal } \mu_2$	<input checked="" type="radio"/> 1 Side <input type="radio"/> 2 Sides
Significance Level The Significance Level of the test or Prob(reject null hypothesis ($H_0 : \mu_1 = \mu_2$) given it is true)	<input type="text"/>
Power The Power desired for the test or Prob(reject H_0 given that H_a is true)	<input type="text"/>
<input type="button" value="Submit Query"/>	

➤ Multivariate Statistik – Überblick



IBM® SPSS® Statistics
Version 21

Stata/IC 12.1 - [Results]
File Edit Data Graphics Statistics User Window Help

The HLM for Windows software interface is shown. The title bar reads "HLM for Windows". The menu bar includes "File", "Basic Settings", "Other Settings", "Run Analysis", and "Help". The main window displays the following text: "Stephen Raudenbush", "Anthony Bryk", "Richard Congdon", and a large "HLM" logo. Below the logo, it says "Hierarchical Linear and Nonlinear Modeling". The bottom right corner of the window has a "Mixed" dropdown menu.

The "Unnamed project : Group number 1 : Input" software interface is shown. The menu bar includes "File", "Edit", "View", "Diagram", "Analyze", "Tools", "Plugins", and "Help". The interface features a toolbar with various icons for data manipulation and analysis. The main area is divided into sections: "Group number 1" (a large empty text area), "Default model" (a smaller empty text area), and "Unstandardized estimates" / "Standardized estimates" (a section for displaying results).

➤ Multivariate Statistik – Überblick

SPSS – in der Psychologie mit Abstand am weitesten verbreitet

The screenshot shows the IBM SPSS Statistics 'Daten-Editor' window for a file named 'FABA.sav'. The 'Analysieren' menu is open, displaying various statistical analysis options. The 'Dimensionsreduzierung' option is highlighted, and its sub-menu is also open, showing 'Faktorenanalyse...', 'Korrespondenzanalyse...', and 'Optimale Skalierung...'. In the background, a data table is visible with columns 'gruppe', 'geschl', and 'alter', and rows numbered 1 to 21.

	gruppe	geschl	alter
1	1	1	3
2	1	1	3
3	1	1	3
4	1	1	3
5	1	1	3
6	1	1	3
7	1	1	3
8	1	1	3
9	1	1	3
10	1	1	3
11	1	1	3
12	1	1	3
13	1	1	3
14	1	1	3
15	1	1	3
16	1	1	3
17	1	1	3
18	1	1	3
19	1	1	3
20	1	1	3
21	1	1	3

➤ **Multivariate Statistik – Überblick**

Nutzungsmöglichkeiten von SPSS:

Campuslizenz – Rechenzentrum der TUD

In vielen PC-Pools der TU installiert

Test-Version für Abschlussarbeiten (über SPSS)

Studenten-Version (Buchhandel)

Demo-Versionen

In der Fachrichtung Psychologie:

Voll-Versionen in den PC-Pools

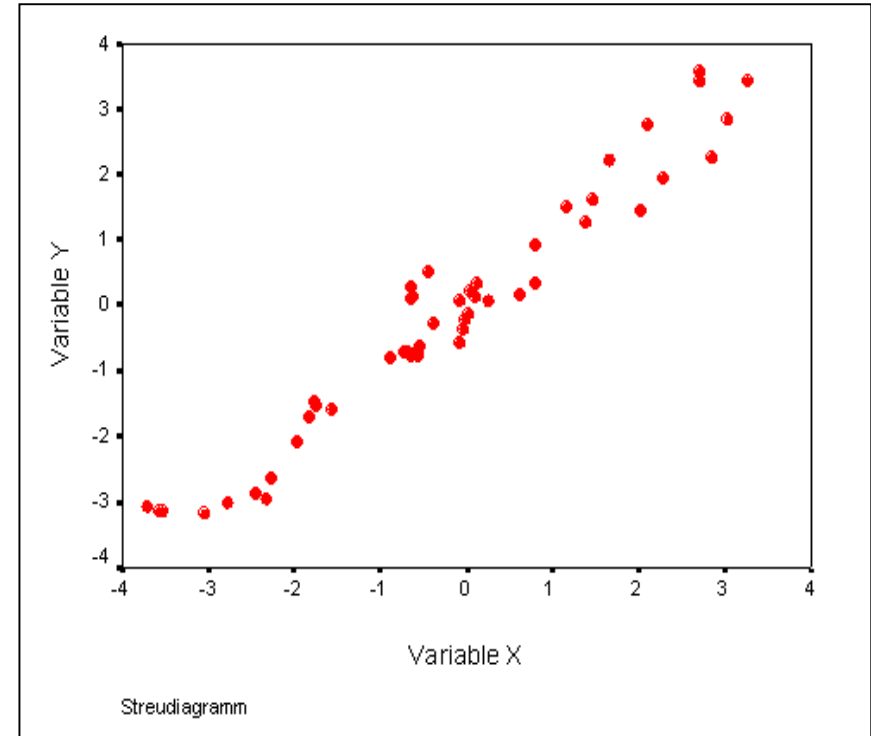
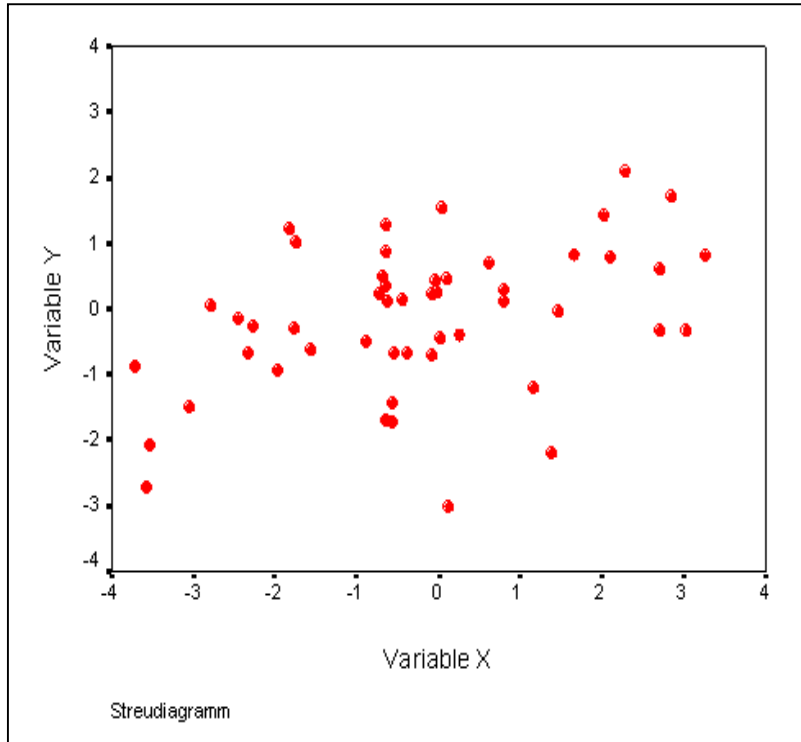
Voll-Versionen in den Instituten

Korrelations- und Regressionsanalyse

- 1 Korrelations- und Regressionsanalyse
 - 1.1 Grundlagen und allgemeiner Überblick
 - 1.2 Einfache lineare Regression
 - 1.2.1 Modell
 - 1.2.2 Methode der kleinsten Quadrate
 - 1.2.3 Voraussetzungen
 - 1.2.4 Varianzzerlegung und Bestimmtheitsmaß
 - 1.2.5 Tests und Vorhersage
 - 1.3 Partielle Korrelation
 - 1.4 Multiple lineare Regression
 - 1.4.1 Modell
 - 1.4.2 Prinzipielle Vorgehensweise
 - 1.4.3 Multikollinearität
 - 1.4.4 Merkmalsselektionsverfahren
 - 1.4.5 Hierarchische Regression
 - 1.5. Moderation und Mediation
 - 1.5.1 Analyse von Moderatoreffekten
 - 1.5.2 Analyse von Mediatoreffekten
 - 1.6. Nichtlineare Regression

Prüfungsliteratur: Rudolf & Müller S. 37-94, 342

➤ Korrelations- und Regressionsanalyse



Korrelationsanalyse: Untersuchung von Existenz und Stärke des Zusammenhanges von (zwei oder mehr) Variablen

Regressionsanalyse: Untersuchung der Art des Zusammenhanges von (zwei oder mehr) Variablen; Beurteilung der Vorhersagemöglichkeiten der Kriteriumsvariablen aus der (den) Prädiktorvariablen; Anpassung einer Regressionsfunktion

Korrelations- und Regressionsanalyse

Verfahren der Korrelationsanalyse

- Bivariate Korrelationsanalyse bei unterschiedlichen Datenniveaus; → Modul M2
- Multiple Korrelation → Abschnitt 1.4
- Kanonische Korrelation (Zusammenhangsanalyse von 2 Merkmalsmengen (X_1, X_2, \dots, X_p) und (Y_1, Y_2, \dots, Y_q) → Spezialliteratur
- Partielle Korrelation → Abschnitt 1.3
-

Verfahren der Regressionsanalyse

- Einfache lineare Regression, z.B. $y = b_1 \cdot x + b_0$ → Abschnitt 1.2
- Multiple lineare Regression, z.B. $y = b_2 \cdot x_2 + b_1 \cdot x_1 + b_0$ → Abschnitt 1.4
- Nichtlineare Regression, z.B. $y = b_1 \cdot x^2 + b_0$ → Abschnitt 1.6
- (Binäre) logistische Regression (Kriterium dichotom) → Abschnitt 2
- Multinomiale logistische Regression; Probitanalyse; Regression mit gewichteten Daten, → Spezialliteratur

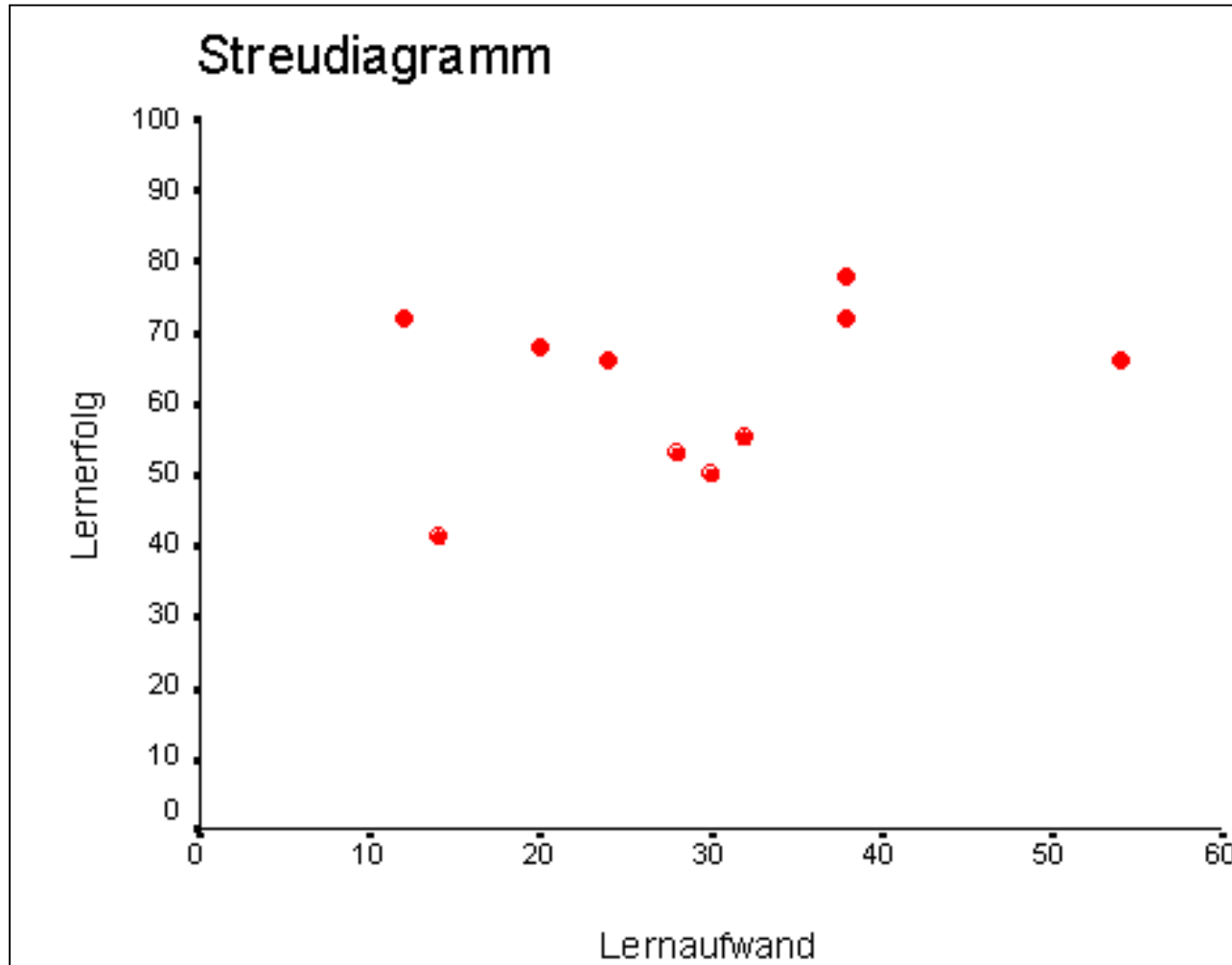
➤ Einfache lineare Regression

Beispiel: Untersuchung der Abhängigkeit des Lernerfolgs von Schülern, gemessen als Punktzahl in einer Klausur (Kriterium Y) von der Zahl der in der Vorbereitung aufgewendeten Übungsstunden (Prädiktor X).

Erhobene Daten:

Schüler	X: Aufwand	Y: Erfolg	Schüler	X: Aufwand	Y: Erfolg
1	12	72	6	30	50
2	14	41	7	32	55
3	20	68	8	38	78
4	24	66	9	38	72
5	28	53	10	54	66

➤ Einfache lineare Regression



➤ Modell der ERA

Modellgleichung:

$$y_i = b_0 + b_1 \cdot x_i + e_i (i = 1, \dots, n)$$

y_i : Wert der Kriteriumsvariablen Y des i-ten Probanden

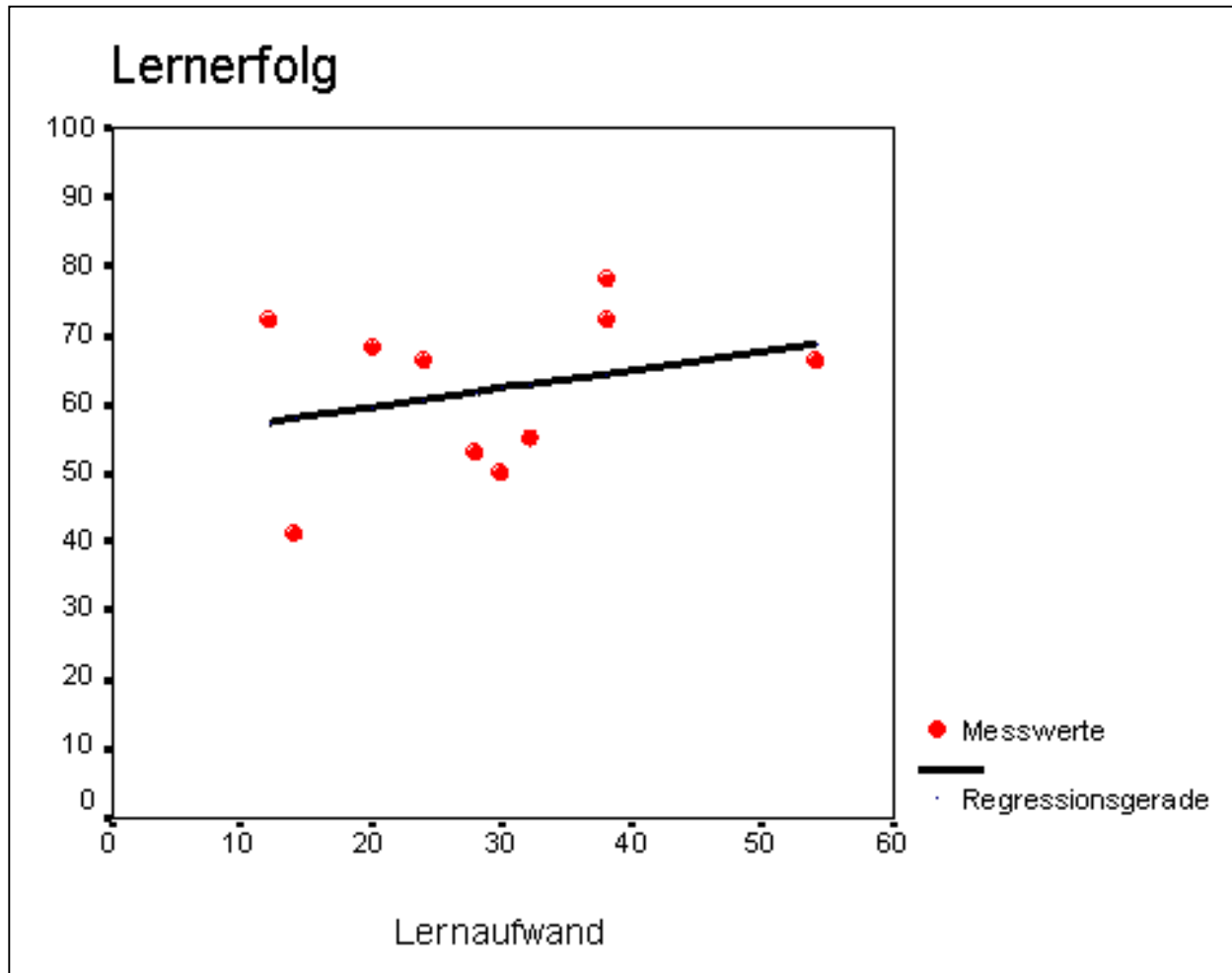
x_i : Wert der Prädiktorvariablen X des i-ten Probanden

e_i : Residuum (Realisierung des Modellfehlers E) des i-ten Pb

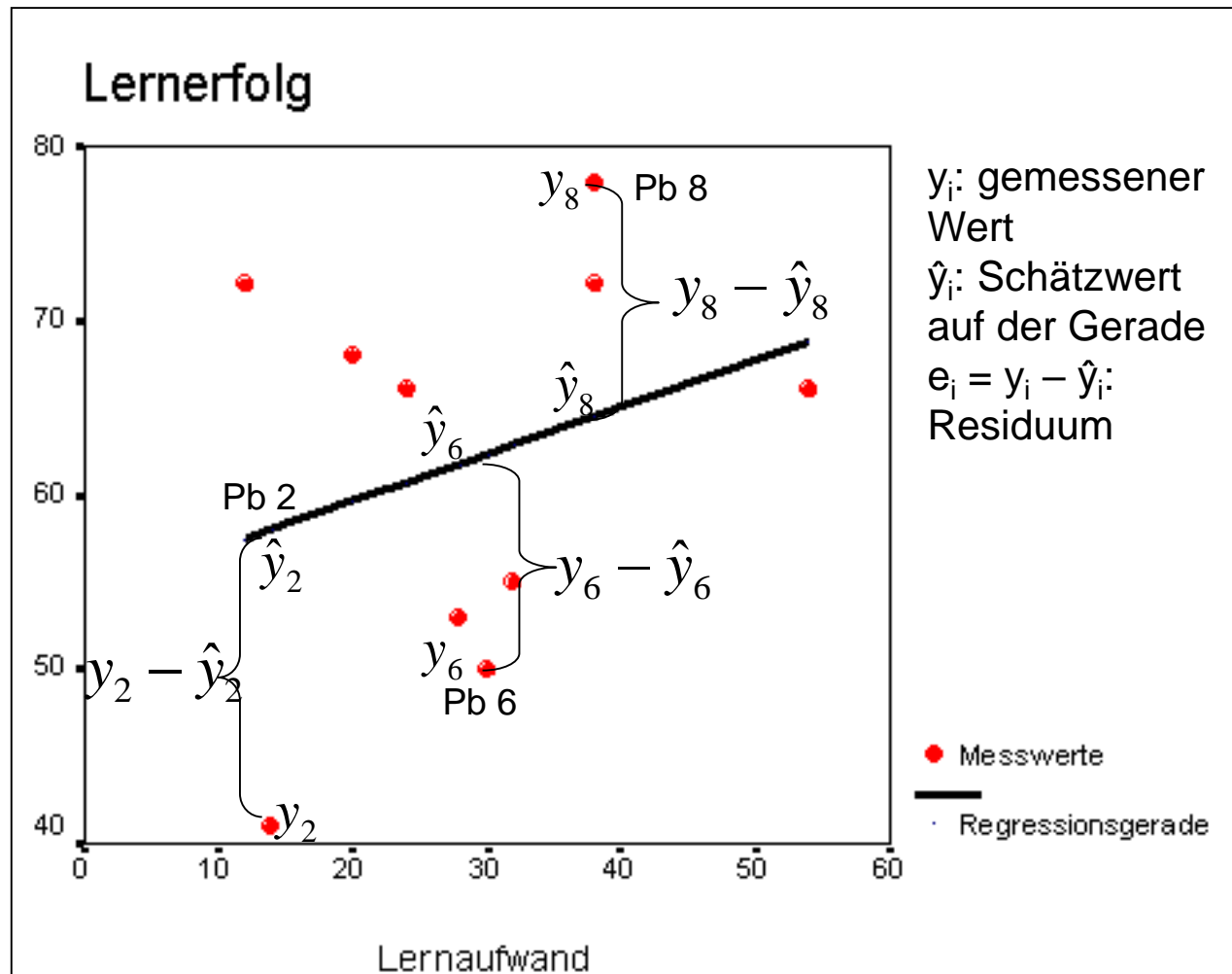
b_0, b_1 : Regressionskoeffizienten

n : Anzahl der Probanden

➤ Modell der ERA

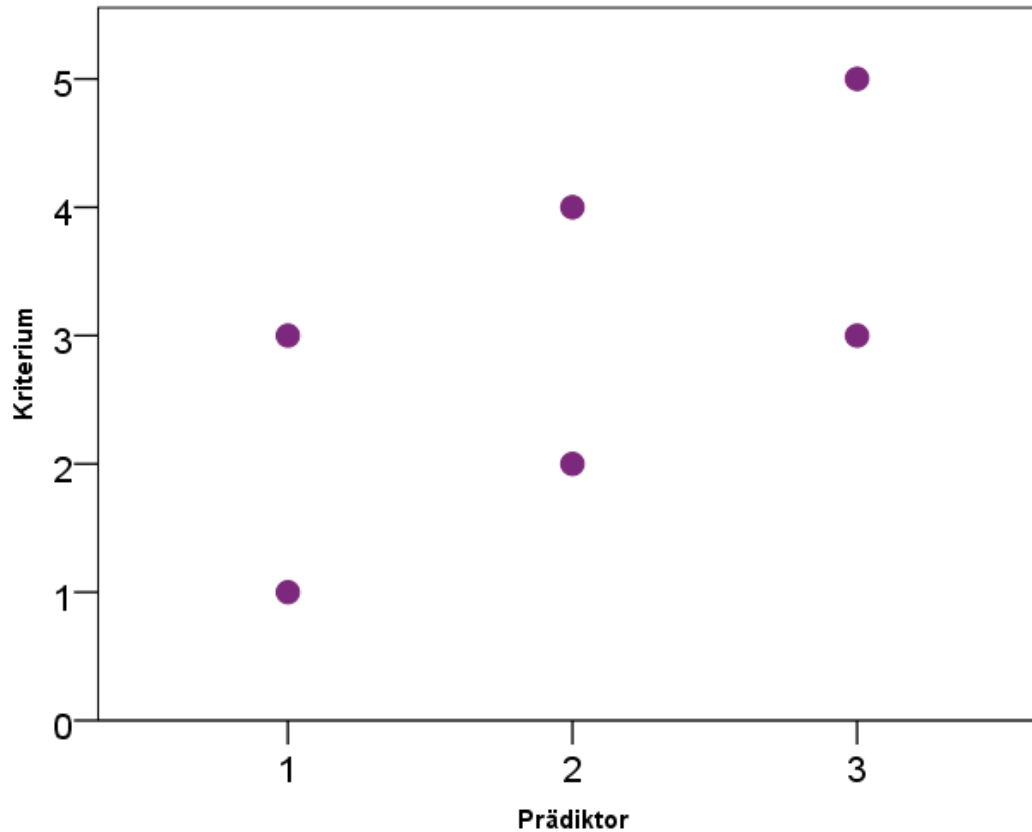


➤ Methode der kleinsten Quadrate



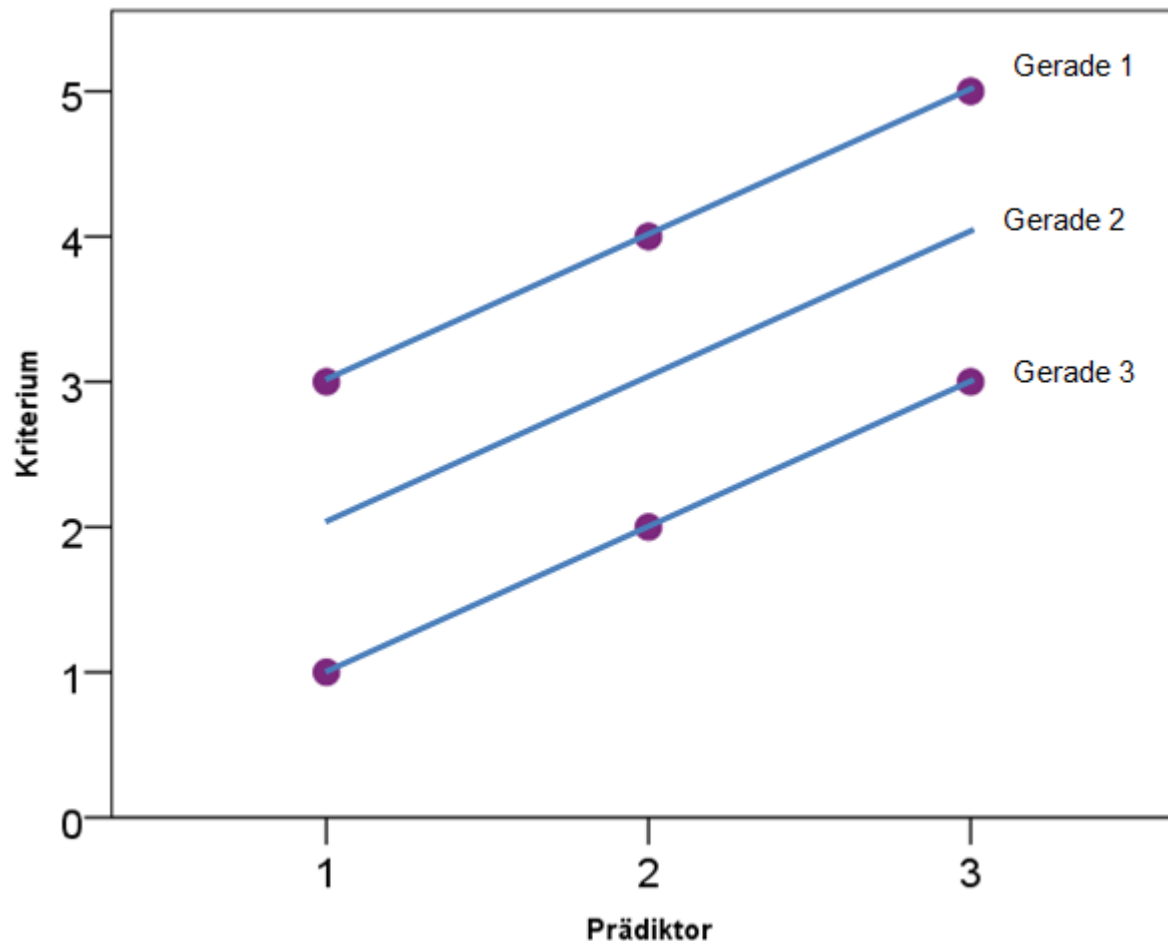
➤ Methode der kleinsten Quadrate

Warum Methode der kleinsten Quadrate (MkQ)?



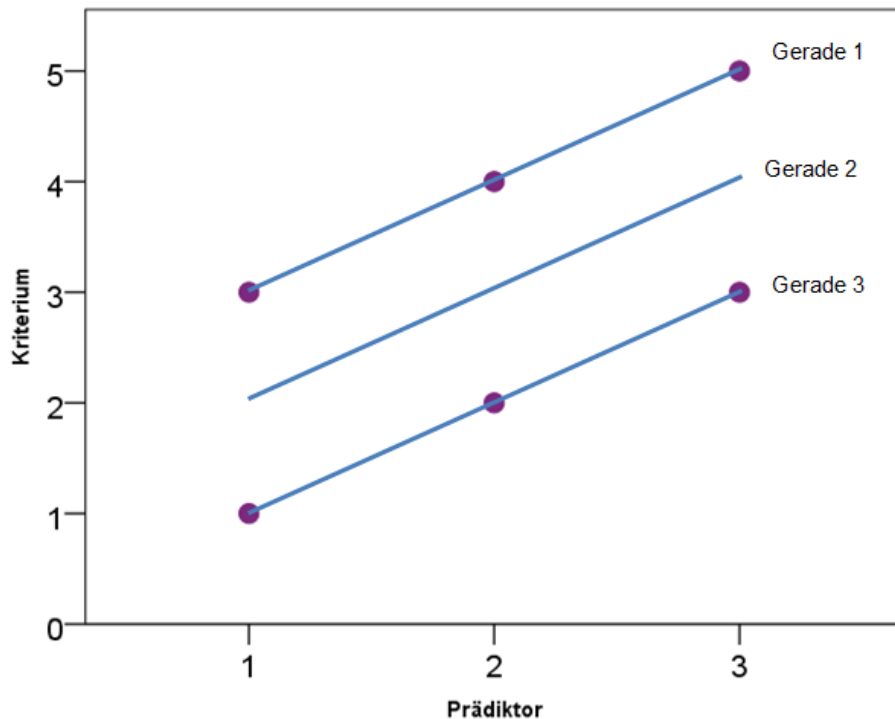
➤ Methode der kleinsten Quadrate

Warum Methode der kleinsten Quadrate (MkQ)?



➤ Methode der kleinsten Quadrate

Warum Methode der kleinsten Quadrate (MkQ)?



Summe der „Abweichungsbeträge“:

Gerade 1: 6

Gerade 2: 6

Gerade 3: 6

keine eindeutige Lösung

Summe der „Abweichungsquadrate“:

Gerade 1: 12

Gerade 2: 6

Gerade 3: 12

eindeutige Lösung

➤ Methode der kleinsten Quadrate

Parameterschätzungen: **Methode der kleinsten Quadrate**

$$QS_{\text{Rest}} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (b_1 \cdot x_i + b_0))^2 \xrightarrow{b_0, b_1} \text{Minimum}$$

QS_{Rest} : Fehler-Quadratsumme

y_i : Wert der Kriteriumsvariablen Y des i-ten Pb

\hat{y}_i : Schätzwert für den Wert der Variablen Y (i-ter Pb)

x_i : Wert der Prädiktorvariablen X des i-ten Pb

e_i : Residuum des i-ten Pb

b_0, b_1 : Regressionskoeffizienten

n : Anzahl der Pb (Probanden)

➤ Methode der kleinsten Quadrate

- **Herleitung der Regressionskoeffizienten (1):**

$$\begin{aligned} & \sum_i (y_i - (b_1 \cdot x_i + b_0))^2 \\ &= \sum_i (y_i^2 - 2 \cdot y_i \cdot (b_1 x_i + b_0) + (b_1 x_i + b_0)^2) \\ &= \sum_i (y_i^2 - 2y_i \cdot b_1 x_i - 2y_i \cdot b_0 + b_1^2 x_i^2 + 2 \cdot b_1 x_i \cdot b_0 + b_0^2) \\ &= \sum_i y_i^2 - 2b_1 \sum_i y_i x_i - 2b_0 \sum_i y_i + b_1^2 \sum_i x_i^2 + 2b_1 b_0 \sum_i x_i + n b_0^2 \end{aligned}$$

➤ Methode der kleinsten Quadrate

- **Herleitung der Regressionskoeffizienten (2):**

$$\frac{d}{db_0} \left(\sum_i y_i^2 - 2b_1 \sum_i y_i x_i - 2b_0 \sum_i y_i + b_1^2 \sum_i x_i^2 + 2b_1 b_0 \sum_i x_i + n b_0^2 \right)$$

$$= -2 \sum_i y_i + 2b_1 \sum_i x_i + 2n b_0 = 0$$

$$\frac{d}{db_1} \left(\sum_i y_i^2 - 2b_1 \sum_i y_i x_i - 2b_0 \sum_i y_i + b_1^2 \sum_i x_i^2 + 2b_1 b_0 \sum_i x_i + n b_0^2 \right)$$

$$= -2 \sum_i y_i x_i + 2b_1 \sum_i x_i^2 + 2b_0 \sum_i x_i = 0$$

➤ Methode der kleinsten Quadrate

- **Herleitung der Regressionskoeffizienten (3):**

$$-2 \sum_i y_i + 2b_1 \sum_i x_i + 2nb_0 = 0$$

$$\Rightarrow b_0 = \frac{2 \sum_i y_i - 2b_1 \sum_i x_i}{2n} = \bar{y} - b_1 \bar{x}$$

➤ Methode der kleinsten Quadrate

- **Herleitung der Regressionskoeffizienten (4):**

$$-2 \sum_i y_i x_i + 2b_1 \sum_i x_i^2 + 2b_0 \sum_i x_i = 0 \Rightarrow -2 \sum_i y_i x_i + 2b_1 \sum_i x_i^2 + 2(\bar{y} - b_1 \bar{x}) \sum_i x_i = 0$$

$$\Rightarrow -2 \sum_i y_i x_i + 2b_1 \sum_i x_i^2 + 2 \left(\frac{\sum_i y_i}{n} - b_1 \frac{\sum_i x_i}{n} \right) \sum_i x_i = 0$$

$$\Rightarrow 2b_1 \sum_i x_i^2 - 2b_1 \frac{\sum_i x_i}{n} \sum_i x_i = 2 \sum_i y_i x_i - 2 \frac{\sum_i y_i}{n} \sum_i x_i$$

$$\Rightarrow b_1 = \frac{2 \sum_i y_i x_i - 2 \frac{\sum_i y_i}{n} \sum_i x_i}{2 \sum_i x_i^2 - 2 \frac{\sum_i x_i}{n} \sum_i x_i} = \frac{n \sum_i y_i x_i - \sum_i y_i \sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

➤ Methode der kleinsten Quadrate

Parameterschätzungen der einfachen linearen Regression:

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \cdot \bar{x}$$

$$\hat{b}_1 = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

➤ Methode der kleinsten Quadrate

Parameterschätzungen (MkQ) im Beispiel:

$$\bar{x} = 29$$

$$\bar{y} = 62.1$$

$$\sum x_i = 290$$

$$\sum y_i = 621$$

$$\sum x_i y_i = 18390$$

$$\sum x_i^2 = 9828$$

$$\left(\sum x_i\right)^2 = 84100$$

$$\hat{b}_0 = 54.308$$

$$\hat{b}_1 = .269$$

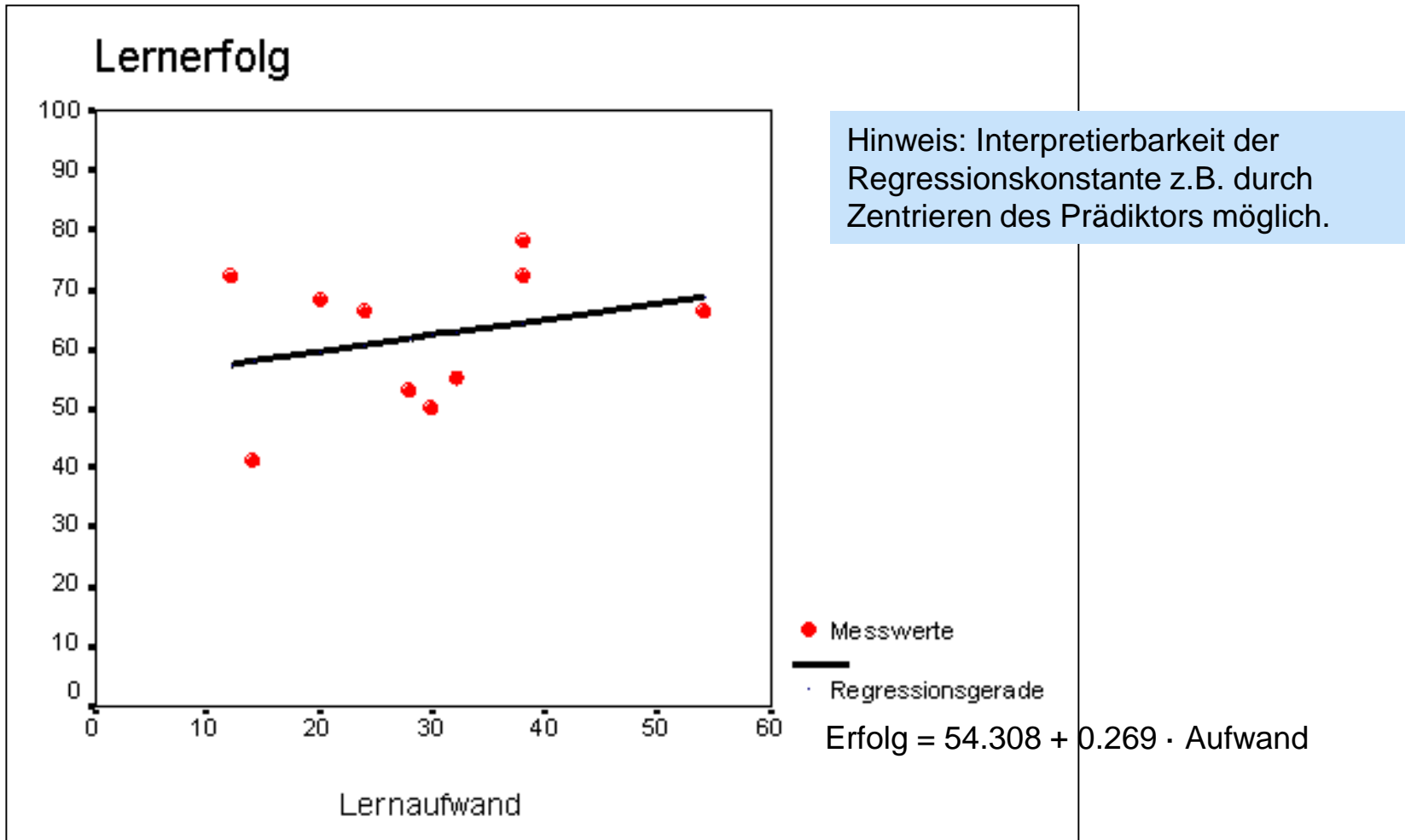
Regressionsgleichung (MkQ):

$$y = 54.308 + .269 \cdot x$$

bzw.

$$\text{Erfolg} = 54.308 + .269 \cdot \text{Aufwand}$$

➤ Methode der kleinsten Quadrate

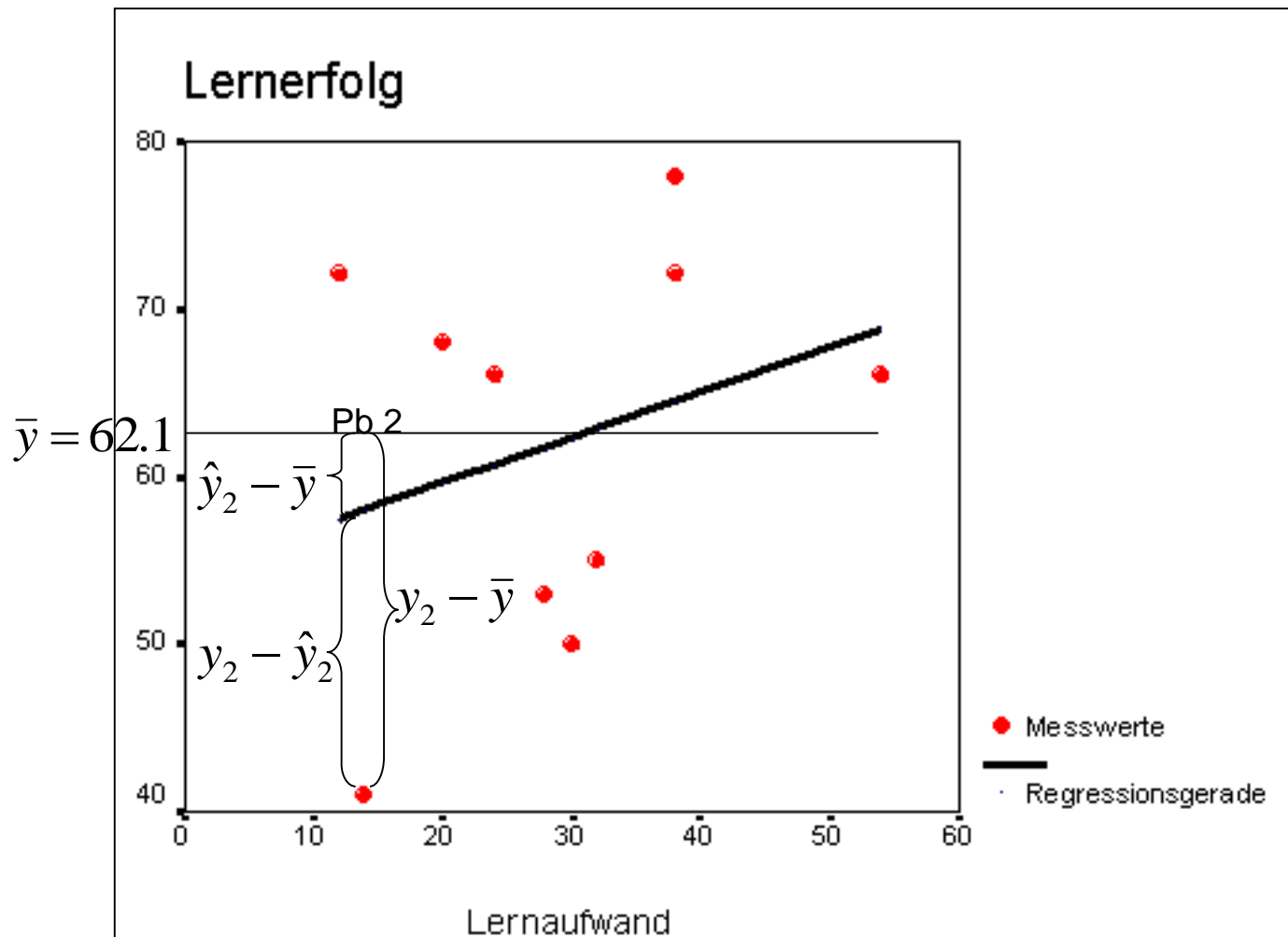


Voraussetzungen

1. Festlegen von Prädiktor und Kriterium (metrisch)
 2. Gültigkeit der Modellgleichung (lineares Modell)
 3. Statistische Unabhängigkeit der Modellfehler (Residuen)
 4. Modellfehler $N(0, \sigma^2)$ verteilt
- Voraussetzungen 1 – 3 für Parameterschätzungen
 - Tests relativ robust gegen geringfügige Verletzungen der Voraussetzung 4

➤ Bestimmtheitsmaß

Vorüberlegungen zur Varianzzerlegung:



➤ Bestimmtheitsmaß

Varianzzerlegung (1):

$$QS(y) = \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 = QS(\hat{y}) + QS(e)$$

**Totale Quadratsumme =
erklärte Quadratsumme + nicht erklärte Quadratsumme**

Messwerte: $y_1 = 72, y_2 = 41, \dots, y_{10} = 66$ $\bar{y} = 62.1$

Totale Quadratsumme:

$$\begin{aligned} QS(y) &= \sum (y_i - 62.1)^2 \\ &= (72 - 62.1)^2 + (41 - 62.1)^2 + \dots + (66 - 62.1)^2 = 1238.90 \end{aligned}$$

➤ Bestimmtheitsmaß

Varianzzerlegung (2):

$$QS(y) = \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 = QS(\hat{y}) + QS(e)$$

**Totale Quadratsumme =
erklärte Quadratsumme + nicht erklärte Quadratsumme**

Schätzwerte:

$$\hat{y}_1 = 54.308 + .269 \cdot 12 = 57.5, \dots, \hat{y}_{10} = 54.308 + .269 \cdot 54 = 68.6 \quad \bar{y} = 62.1$$

Erklärte Quadratsumme:

$$\begin{aligned} QS(\hat{y}) &= \sum (\hat{y}_i - \bar{y})^2 \\ &= (57.5 - 62.1)^2 + (58.1 - 62.1)^2 + \dots + (68.6 - 62.1)^2 = 102.37 \end{aligned}$$

➤ Bestimmtheitsmaß

Varianzzerlegung (3):

$$QS(y) = \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 = QS(\hat{y}) + QS(e)$$

**Totale Quadratsumme =
erklärte Quadratsumme + nicht erklärte Quadratsumme**

Schätzwerte: $\hat{y}_1 = 57.5, \dots, \hat{y}_{10} = 68.6$

Messwerte: $y_1 = 72, \dots, y_{10} = 66$

Nicht erklärte Quadratsumme:

$$QS(e) = \sum (y_i - \hat{y}_i)^2 =$$
$$(72 - 57.5)^2 + (41 - 58.1)^2 + \dots + (66 - 68.8)^2 = 1136.53$$

➤ Bestimmtheitsmaß

Varianzzerlegung (4):

$$QS(y) = \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 = QS(\hat{y}) + QS(e)$$

**Totale Quadratsumme =
erklärte Quadratsumme + nicht erklärte Quadratsumme**

$$QS(y) = 1238.90 = 102.37 + 1136.53 = QS(\hat{y}) + QS(e)$$

➤ Bestimmtheitsmaß

Bestimmtheitsmaß: Anteil der Varianz des Kriteriums, der mit Hilfe der Regression, d.h. durch den Prädiktor aufgeklärt werden kann

$$b = r^2 = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}} = \frac{QS(\hat{y})}{QS(y)} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Im Beispiel:

$$b = r^2 = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}} = \frac{QS(\hat{y})}{QS(y)} = \frac{102.37}{1238.9} = 0.083$$

d.h. 8.3% der Varianz des Kriteriums Lernerfolg kann durch den Prädiktor Lernaufwand erklärt werden.

➤ Tests und Vorhersage

A) Beurteilung der globalen Güte der Regression: Prüfung des Bestimmtheitsmaßes

$$H_1: b = r^2 > 0 \quad H_0: b = r^2 = 0$$

$$\text{F-Test: } F = \frac{\text{erklärte Varianz} / df_1}{\text{nicht erklärte Varianz} / df_2} = \frac{QS(\hat{y})}{QS(e)/n-2}$$

Im Beispiel:

$$F = \frac{QS(\hat{y})}{QS(e)/n-2} = \frac{102.37}{1136.53/8} = 0.72 \quad (p = 0.42 > 0.05)$$

➤ Tests und Vorhersage

B) Prüfung der Regressionskoeffizienten:

$$H_1: b_1 \neq 0 \quad H_1: b_1 = 0$$

(allgemein: $H_1: b_1 \neq b^*$ $H_1: b_1 = b^*$)

$$\text{t-Test: } t = \frac{b_1}{s_{b_1}} = \frac{\text{Schätzwert}}{\text{Standardfehler}}$$

Beispiele in den Computerseminaren;
Einzelheiten siehe Rudolf & Müller, S. 39-41

➤ Tests und Vorhersage

C) Vorhersage aus bekanntem Prädiktorwerte x_0 :

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 \cdot x_0 \quad (x_0 \in [x_{\min}, x_{\max}])$$

Vorhersage- und
Konfidenzintervalle berechenbar

