ZUSAMMENHÄNGE ZWISCHEN METRISCHEN VARIABLEN II

Martin-Luther-Universität Halle-Wittenberg
Institut für Soziologie
Übung Einführung in die deskriptive Statistik

Agenda

- Einführung in die **bivariate** lineare Regression
  - mathematisches Grundmodell
  - Regressionskonstanten und Regressionsgewichte
  - Vorhersagewerte bestimmen
  - Determinationskoeffizienten R^2

Medikament und Reaktionszeit (fiktiv)

Eine Forscherin interessiert sich dafür, ob die Dosierung eines Medikaments einen Einfluss auf die Reaktionszeit ausübt.

Für 6 Versuchspersonen erhält sie folgende Werte:

Person ID	1	2	3	4	5	6
Dosierung in mg	0	5	3	8	2	0
Reaktionszeit in ms	0	5	4	6	3	2

Quiz: Welches Skalenniveau liegt vor? Handelt es sich um eine symmetrischen oder asymmetrische Fragestellung?

Analyse

asymmetrische Fragestellung: Dosierung(X)→ Reaktionszeit(Y)

zwei metrische Variablen

Eine Forscherin interessiert sich dafür, ob die Dosierung eines Medikaments einen Einfluss auf die Reaktionszeit ausübt.

Person ID	1	2	3	4	5	6
Dosierung in mg	0	5	3	8	2	0
Reaktionszeit in ms	0	6	4	6	3	2

Aufgabe 1.1: Medikament und Reaktionszeit

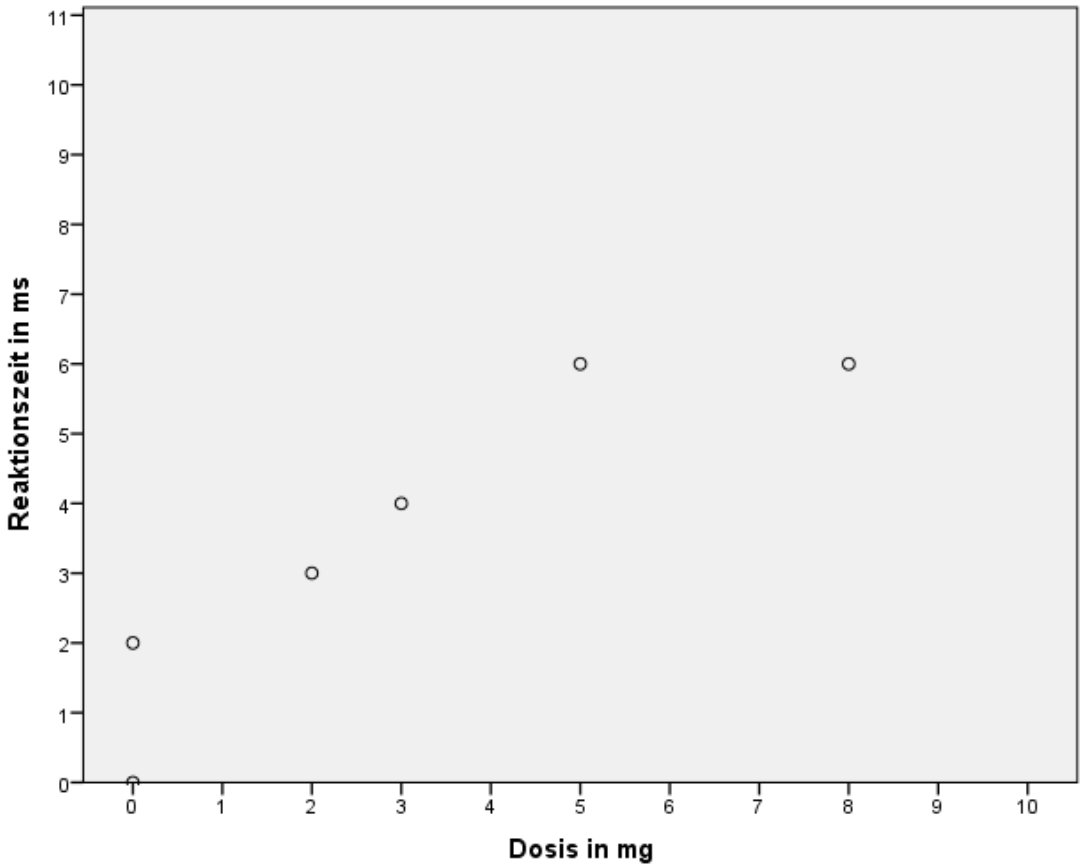
Eine Forscherin interessiert sich dafür, ob die Dosierung eines Medikaments einen Einfluss auf die Reaktionszeit ausübt.

Für 6 Versuchspersonen erhält sie folgende Werte:

Person ID	1	2	3	4	5	6
Dosierung in mg	0	5	3	8	2	0
Reaktionszeit in ms	0	6	4	6	3	2

Zeichnen Sie ein entsprechendes Streudiagramm!

Aufgabe 1.1: Lösung

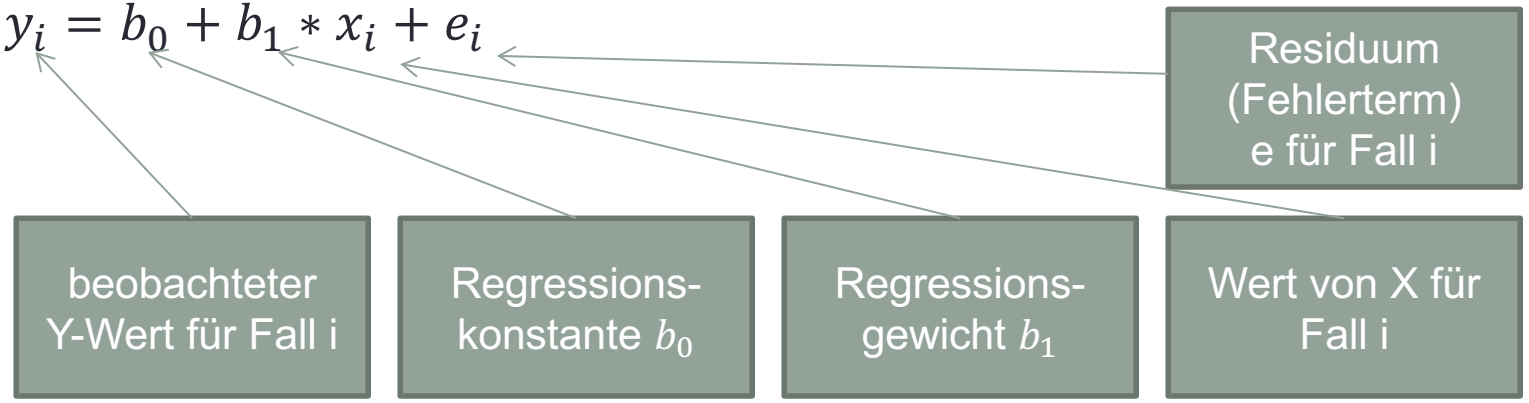


Quiz:

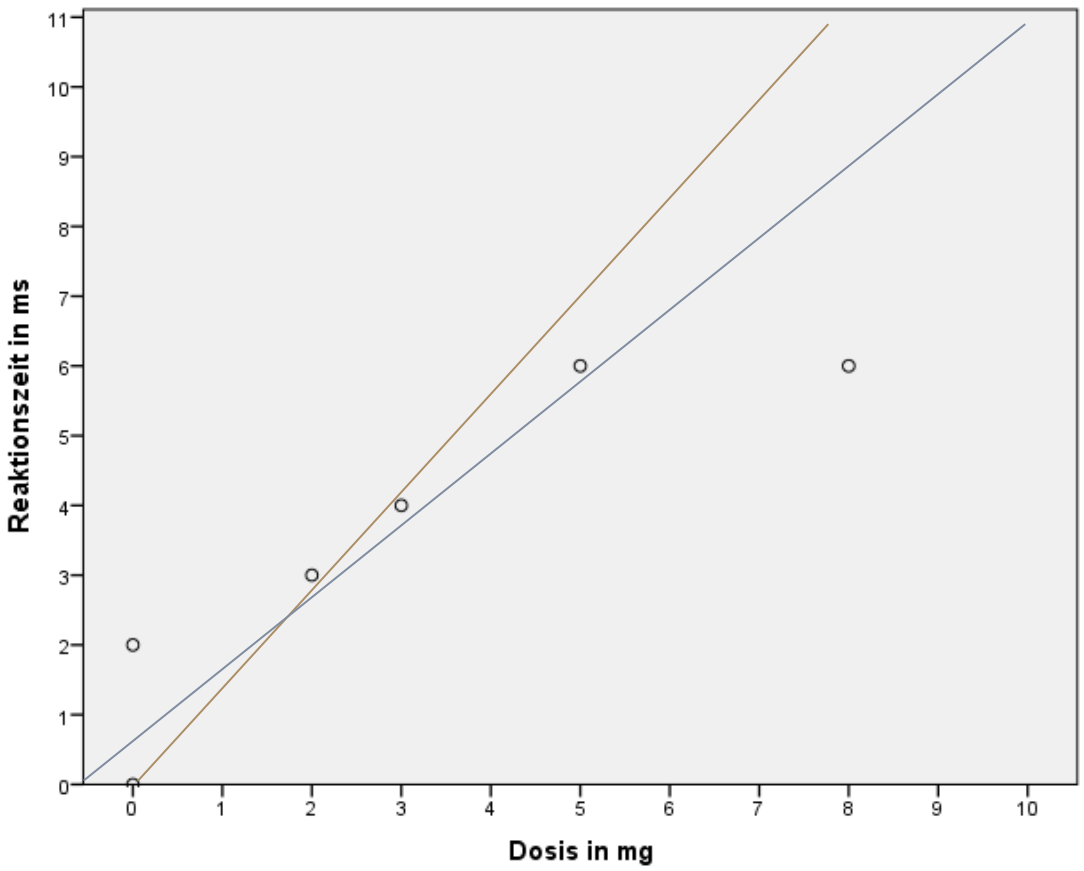
Welchen Zusammenhang zwischen den Variablen würden wir hier vermuten?

Lineare Regression I

- Idee:
  - Einfluss der unabhängigen Variablen (x) auf die abhängige Variable (y) soll mithilfe einer Geraden beschrieben werden
  - $y_i = b_0 + b_1 * x_i + e_i$



Lineare Regression II



Problem: Wie finden wir die ideale Gerade für unsere Fragestellung?

Lösung: OLS-Verfahren (Kleinstequadratemethode)

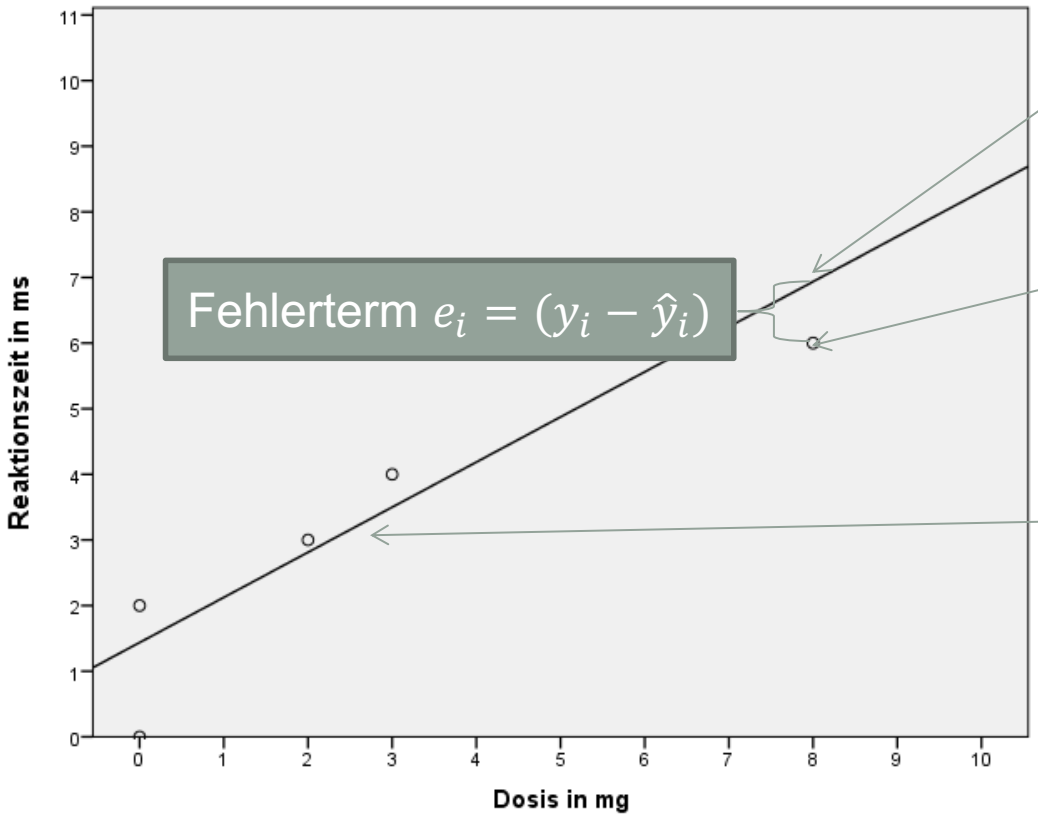
$$y_i = b_0 + b_1 * x_i + e_i$$

Lineare Regression III

- Ordinary-Least-Square-Verfahren:
  - Abweichung der Summe der Quadrate des Fehlerterms soll minimiert werden
  - $\sum e_i^2 \rightarrow min$

- Bestimmung der Regressionsgraden:
  - mithilfe des OLS-Verfahrens eindeutig möglich
  - Regressionsgewicht $b_1 = \frac{s_{X,Y}}{s_X^2} = \frac{SP_{X,Y}}{SAQ_X}$
  - Regressionskonstante $b_0 = \bar{y} - b_1 * \bar{x}$

Lineare Regression IV



Fehlerterm $e_i = (y_i - \hat{y}_i)$

vorhergesagter Wert $\hat{y}_i$ für Fall i, bei bekanntem x_i

tatsächlicher Wert y_i für Fall i bei bekanntem x_i

Regressionsgerade

Aufgabe 1.2: Medikament und Reaktionszeit

Eine Forscherin interessiert sich dafür, ob die Dosierung eines Medikaments einen Einfluss auf die Reaktionszeit ausübt.

Für 6 Versuchspersonen erhält sie folgende Werte:

Person ID	1	2	3	4	5	6
Dosierung in mg	0	5	3	8	2	0
Reaktionszeit in ms	0	6	4	6	3	2

Bestimmen Sie das Regressionsgewicht $b_1 = \frac{SP_{X,Y}}{SAQ_X}$ und die Regressionskonstante $b_0 = \bar{y} - b_1 * \bar{x}$!

Aufgabe 1.2: Lösung Hilfstabelle

ID	Dosier-ung (x_i)	Reaktions-zeit (y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$		$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	0					
2	5	6					
3	3	4					
4	8	6					
5	2	3					
6	0	2					
	$\bar{x} =$	$\bar{y} =$		$SAQ_X =$			$SP_{X,Y} =$

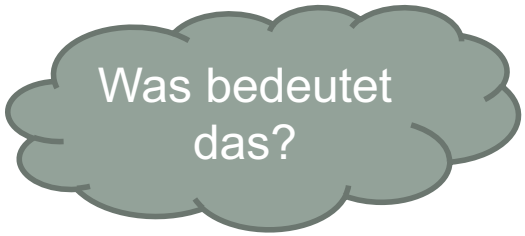
Aufgabe 1.2: Lösung Hilfstabelle

ID	Dosierung (x_i)	Reaktions-zeit (y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$		$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	0	-3	9	-3,5		10,5
2	5	6	2	4	2,5		5
3	3	4	0	0	0,5		0
4	8	6	5	25	2,5		12,5
5	2	3	-1	1	-0,5		0,5
6	0	2	-3	9	-1,5		4,5
	$\bar{x} = 3$	$\bar{y} = 3{,}5$		$SAQ_X = 48$			$SP_{X,Y} = 33$

Aufgabe 1.2: Lösung II

- Bestimmung Regressionsgewicht:
  - $b_1 = \frac{SP_{X,Y}}{SAQ_X}$
  - $b_1 = \frac{33}{48}$
  - $b_1 = 0{,}6875$
- Bestimmung Regressionskonstante:
  - $b_0 = \bar{y} - b_1 * \bar{x}$
  - $b_0 = 3{,}5 - 0{,}6875 * 3$
  - $b_0 = 1{,}4375$
- Bestimmung Regressionsgrade:
  - $y_i = b_0 + b_1 * x_i + e_i$
  - $y_i = 1{,}4375 + 0{,}6875 * x_i + e_i$

- $\bar{x} = 3$
- $\bar{y} = 3{,}5$
- $SAQ_X = 48$
- $SP_{X,Y} = 33$

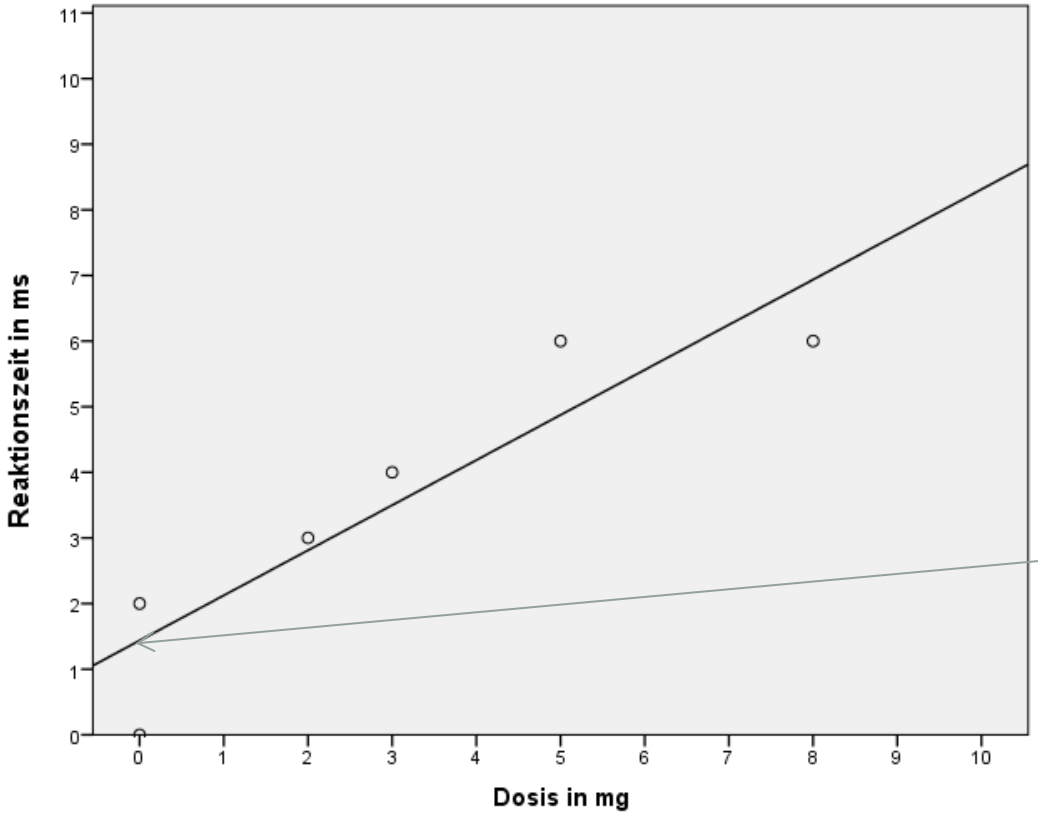


Lineare Regression V: Interpretation der Regressionskoeffizienten

- Regressionskonstante b_0:
  - geometrisch Schnittpunkt mit y-Achse
  - inhaltlich Wert von y der vorhergesagt würde, wenn $x = 0$ wäre
  - oft inhaltlich nicht sinnvoll interpretierbar → Zentrierung

- Regressionsgewicht b_1:
  - geometrisch Steigungskoeffizient
  - Vorzeichen gibt Richtung des Zusammenhangs an
  - bei Anstieg von x um eine Einheit verändert sich y um b_1 Einheiten

Medikamente und Reaktionszeit: Interpretation Regressionskonstante

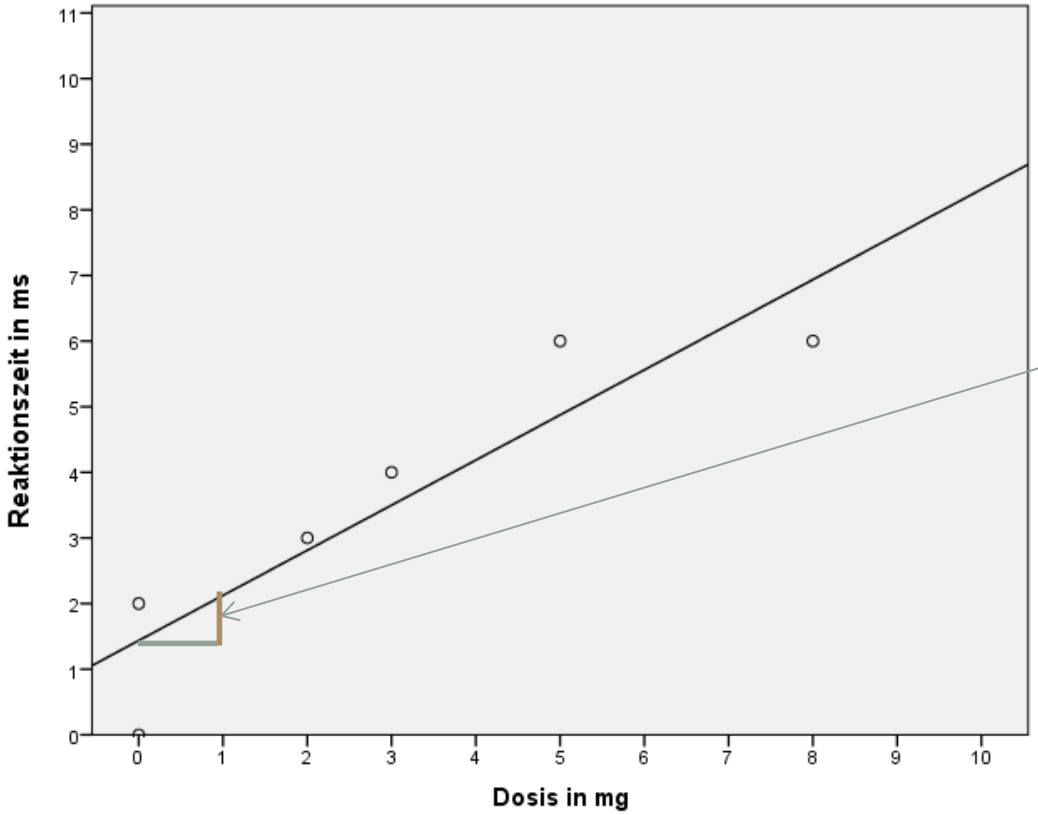


$$y_i = 1{,}4375 + 0{,}6875 * x_i + e_i$$

Quiz: Wie lässt sich die Regressionskonstante hier interpretieren?

Regressionskonstante:
- $b_0 = 1{,}4375$
- Schnittpunkt mit y-Achse liegt bei $1{,}4375$
- Bei einer Dosierung von 0 mg wird eine Reaktionszeit von etwa 1,4 ms vorhergesagt.

Fallbeispiel Medikamente und Reaktionszeit: Interpretation Regressionsgewicht



$$y_i = 1{,}4375 + 0{,}6875 * x_i + e_i$$

Quiz: Wie lässt sich das Regressionsgewicht hier interpretieren?

Regressionsgewicht:

- $b_1 = 0{,}6875$
- Wenn x um eine Einheit steigt, erhöht sich y im Durchschnitt um b_1 Einheiten
- Wenn die Dosis um 1 mg steigt, erhöht sich die Reaktionszeit im Durchschnitt um 0,7 ms

Lineare Regression VI: SPSS

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1,438			202	,092
	Dosis in mg	,687			342	,012

Regressionskonstante b_0

Regressionsgewicht b_1

Lineare Regression VII: Vorhersagewerte

- mathematisches Grundmodell:
  - $y_i = b_0 + b_1 * x_i + e_i$

y_i: realisierter Wert von y an Stelle x_i

- Vorhersagegleichung:
  - $\hat{y}_i = b_0 + b_1 * x_i$

$\hat{y}_i$: vorhergesagter Wert von y an Stelle x_i

Aufgabe 1.3: Vorhersagewert

Mithilfe von SPSS wurde für den Einfluss der Dosierung eines Medikaments in mg auf die Reaktionszeit in Millisekunden folgendes Regressionsmodell aufgestellt:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1,438	,653		2,202	,092
	Dosis in mg	,687	,158	,908	4,342	,012

Welche Reaktionszeit in Millisekunden würde hier für eine Dosis von 6 mg des Medikaments vorhergesagt?

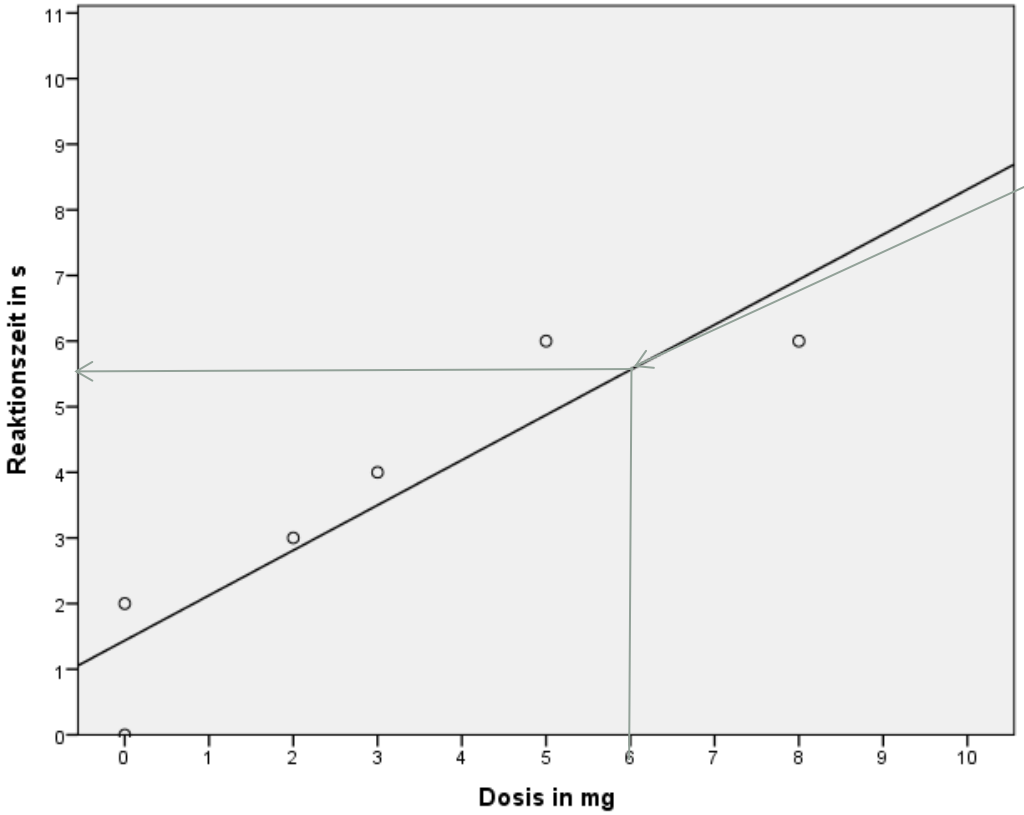
Hinweis: $\hat{y}_i = b_0 + b_1 * x_i$

Aufgabe 1.3: Lösung

- gegeben:
  - $b_0 = 1{,}438$
  - $b_1 = 0{,}687$
  - $x_i = 6$

- Berechnung:
  - $\hat{y}_i = 1{,}438 + 0{,}687 * 6$
  - $\hat{y}_i = 5{,}56$

Aufgabe 1.3: Lösung II



$\hat{y}_i = 5{,}56$

Für eine Dosis von 6 mg würde eine Reaktions-zeit von 5,6 ms vorhergesagt

Lineare Regression VIII: Bestimmung des Determinationskoeffizienten R²

- Fragestellung:
  - Wie gut ist unser Regressionsmodell?
  - Um wie viel Prozent verbessert sich die Prognose unserer abhängigen Variablen Y unter Hinzunahme der unabhängigen X-Variablen?
- Lösung:
  - PRE-Maß bestimmen
  - Determinationskoeffizient R²
- Berechnung:
  - $R^2 = PRE = \frac{E_0 - E_1}{E_0}$

Lineare Regression IX: E_0 –Fehler

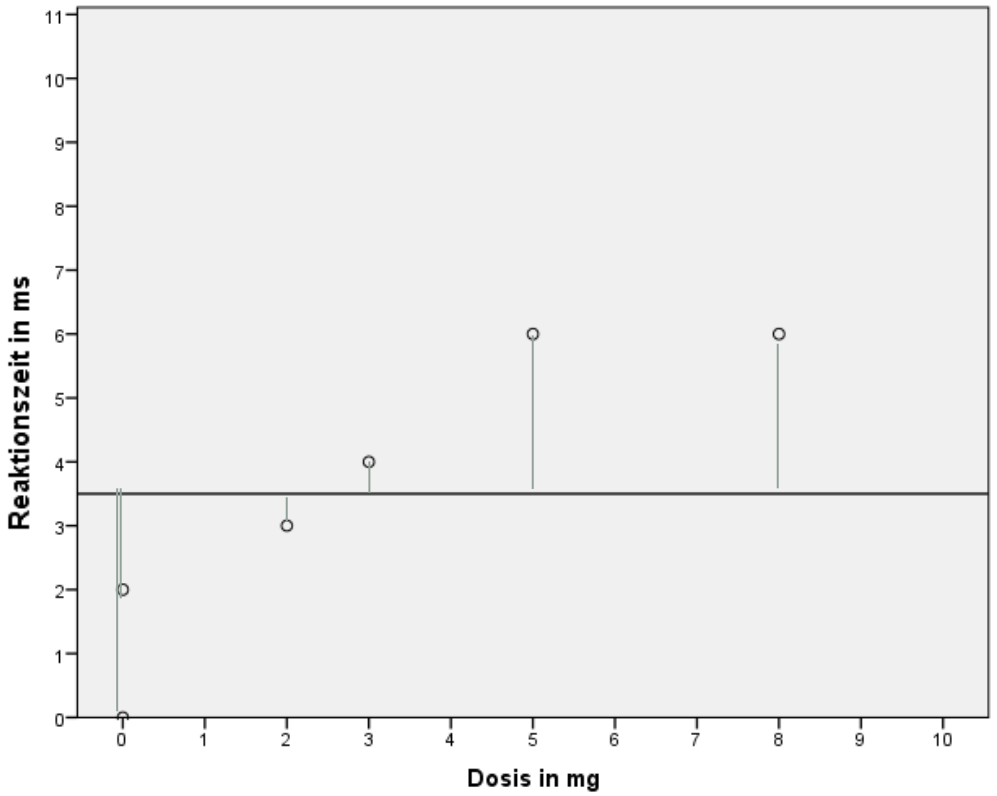
- E_0 –Fehler: Fehler ohne Kenntnis der unabhängigen Variablen

ID	Reaktionszeit (y_i)
1	0
2	6
3	4
4	6
5	3
6	2
	$\bar{y} = 3{,}5$

Quiz: Was wäre hier der beste Tipp, wenn wir möglichst wenig daneben liegen wollen?

Welches Maß können wir verwenden, um zu messen, wie sehr wir daneben liegen?

Lineare Regression X: E_0 –Fehler



bester Tipp ohne Kenntnis von X ist $\bar{y}$

E0-Fehler ergibt sich aus der Summe der quadrierten Abweichungen des arithmetischen Mittels von den realisierten Werten.

$$E_0 = SAQ_Y = \sum(y_i - \bar{y})^2$$

Bestimmen Sie den E0-Fehler!

Lösung Hilfstabelle

ID	Dosierung (x_i)	Reaktionszeit (y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	0	-3	9	-3,5		10,5
2	5	6	2	4	2,5		5
3	3	4	0	0	0,5		0
4	8	6	5	25	2,5		12,5
5	2	3	-1	1	-0,5		0,5
6	0	2	-3	9	-1,5		4,5
	$\bar{x} = 3$	$\bar{y} = 3{,}5$		$SAQ_X = 48$		$SAQ_Y =$	$SP_{X,Y} = 33$

Lösung Hilfstabelle II

ID	Dosierung (x_i)	Reaktionszeit (y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	0	-3	9	-3,5	12,25	10,5
2	5	6	2	4	2,5	6,25	5
3	3	4	0	0	0,5	0,25	0
4	8	6	5	25	2,5	6,25	12,5
5	2	3	-1	1	-0,5	0,25	0,5
6	0	2	-3	9	-1,5	2,25	4,5
	$\bar{x} = 3$	$\bar{y} = 3{,}5$		$SAQ_X = 48$		$SAQ_Y = 27{,}5$	$SP_{X,Y} = 33$

$E_0 = SAQ_Y = 27{,}5$

Lineare Regression X: E_1 –Fehler

- E_1 –Fehler: Fehler unter Kenntnis der unabhängigen X-Variablen und des damit von uns aufgestellten Prognosegleichung

ID	Dosierung (x_i)	Reaktions-zeit (y_i)
1	0	0
2	5	6
3	3	4
4	8	6
5	2	3
6	0	2
	$\bar{x} = 3$	$\bar{y} = 3{,}5$

$$\hat{y}_i = 1{,}438 + 0{,}687 * x_i$$

Grundidee:
Für jeden Wert x_i bestimmen wir den vorhergesagten Wert von $\hat{y}_i$

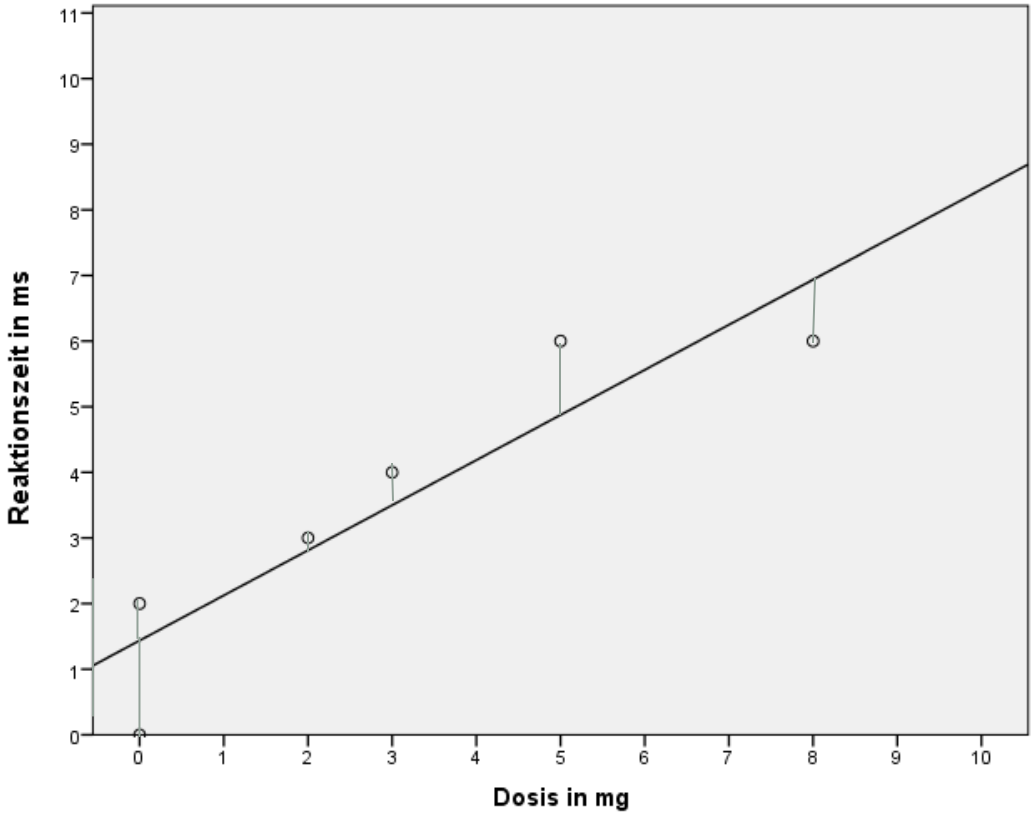
Lineare Regression X: E_1 –Fehler

- E_1 –Fehler: Fehler unter Kenntnis der unabhängigen X-Variablen und des damit von uns aufgestellten Prognosegleichung

ID	Dosierung (x_i)	tatsächliche Reaktionszeit (y_i)	vorhergesagte Reaktionszeit $(\hat{y}_i)$
1	0	0	1,438
2	5	6	4,873
3	3	4	3,499
4	8	6	6,934
5	2	3	2,812
6	0	2	1,438
	$\bar{x} = 3$	$\bar{y} = 3{,}5$	

$$\hat{y}_i = 1{,}438 + 0{,}687 * x_i$$

Lineare Regression XI: E1-Fehler



bester Tipp mit Kenntnis von X ist Vorhersagewert $\hat{y}_i$

E1-Fehler ergibt sich aus der quadrierten Abweichung der realisierten Werten von den vorhergesagten Werten

$$E_1 = SAQ_E = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2$$

Lineare Regression XII: E1-Fehler

ID	Dosierung (x_i)	tatsächliche Reaktionszeit (y_i)	vorhergesagte Reaktionszeit $(\hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	0	0	1,438	2,068
2	5	6	4,873	1,270
3	3	4	3,499	0,251
4	8	6	6,934	0,872
5	2	3	2,812	0,035
6	0	2	1,438	0,316
	$\bar{x} = 3$	$\bar{y} = 3{,}5$		$\boldsymbol{E_1 = SAQ_E = 4,813}$

Die Berechnung des E1-Fehler ist so relativ aufwändig… können wir uns die Bestimmung von R² irgendwie erleichtern?

Lineare Regression XI: Bestimmung des Determinationskoeffizienten R²

- Sonderfall bivariate Regression:
  - $R^2 = (r_{X,Y})^2$
  - $r_{X,Y} = \frac{s_{X,Y}}{s_X * s_Y}$

Kovarianz (X,Y)

Standardabweichung von X bzw. Y

  - $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}}$

Kovariation(X,Y)

Variation von X bzw. Y

Die letztere Formel ist für uns am einfachsten, da uns hier nur noch SAQ_Y fehlt ☺

Aufgabe 1.4: Medikament und Reaktionszeit

Eine Forscherin interessiert sich dafür, ob die Dosierung eines Medikaments einen Einfluss auf die Reaktionszeit ausübt.

Für 6 Versuchspersonen erhält sie folgende Werte:

Person ID	1	2	3	4	5	6
Dosierung in mg	0	5	3	8	2	0
Reaktionszeit in ms	0	6	4	6	3	2

Wie gut eignet sich die Dosierung des Medikaments zur Prognose der Reaktionszeit? Berechnen Sie den Determinationskoeffizienten R^2 und interpretieren Sie Ihr Ergebnis?

Aufgabe 1.4: Lösung II

ID	Dosierung (x_i)	Reaktionszeit (y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	0	-3	9	-3,5	12,25	10,5
2	5	6	2	4	2,5	6,25	5
3	3	4	0	0	0,5	0,25	0
4	8	6	5	25	2,5	6,25	12,5
5	2	3	-1	1	-0,5	0,25	0,5
6	0	2	-3	9	-1,5	2,25	4,5
	$\bar{x} = 3$	$\bar{y} = 3{,}5$		$SAQ_X = 48$		$SAQ_Y = 27{,}5$	$SP_{X,Y} = 33$

- $R^2 = (r_{X,Y})^2$
- $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}}$

Aufgabe 1.4: Lösung III

- Berechnung:
  - $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}} = \frac{33}{\sqrt{48*27,5}}$
  - $r_{X,Y} = 0,9083$
  - $R^2 = \left(r_{X,Y}\right)^2 = 0,9083^2$
  - $R^2 = 0,825$
- Interpretation:
  - Durch Kenntnis der Dosis verbessert sich die Prognose der Reaktionszeit um 82,5 Prozent.

Lineare Regression XII: SPSS II

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,908[a]	,825	,781	1,097

a. Prädiktoren: (Konstante), Dosis in mg

Determinations-koeffizient R^2

ANOVA[a]

Modell		Quadratsumme	df
1	Regression	22,688	1
	Residuum	4,813	4
	Gesamtsumme	27,500	5

$E_0 - E_1 = SAQ_{Regression}$

$E_1 = SAQ_{Residuum} = SAQ_E$

$E_0 = SAQ_{Gesamt} = SAQ_Y$

$$R^2 = \frac{E_0 - E_1}{E_0} = \frac{27{,}5 - 4{,}813}{27{,}5} = 0{,}825$$

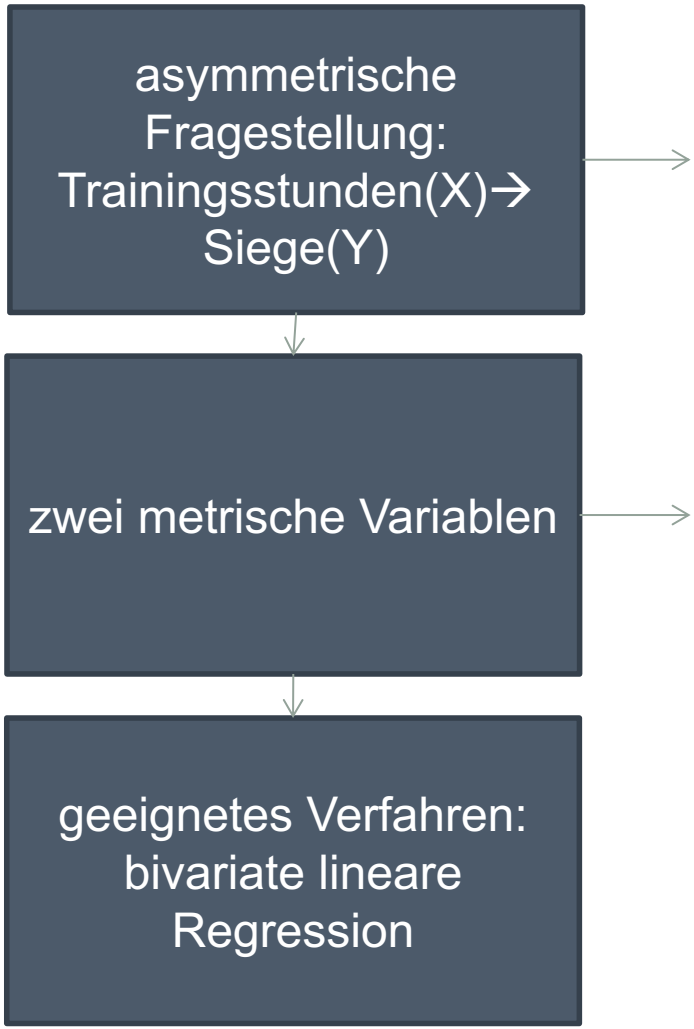
Aufgabe 2: Tischfußballturnier

Ein Sportsoziologe interessiert sich für die Fragestellung, ob die Anzahl der Siege bei einem Tischfußballturnier von der Anzahl der Trainingsstunden abhängt. Für 9 Spieler wurde folgende Tabelle erzielt:

Spieler	Trainingsstunden (X)	Siege (Y)
1	1	2
2	2	0
3	3	3
4	4	5
5	5	1
6	6	2
7	7	6
8	8	3
9	9	5

a) Mit welchem Verfahren lässt sich der Zusammenhang untersuchen? Wie lautet das allgemeine Grundmodell (Gleichung)?
b) Wenden Sie das Verfahren an und interpretieren Sie ihr Ergebnis!
c) Wie gut eignet sich das Modell zur Prognose der Siege?

Aufgabe 2: Analyse



Ein Sportsoziologe interessiert sich für die Fragestellung, ob die Anzahl der Siege bei einem Tischfußballturnier von der Anzahl der Trainingsstunden abhängt.

Spieler	Trainingsstunden (X)	Siege (Y)
1	1	2
2	2	0
3	3	3
4	4	5
5	5	1
6	6	2
7	7	6
8	8	3
9	9	5

Aufgabe 2: Analyse II

a) Mit welchem Verfahren lässt sich der Zusammenhang untersuchen? Wie lautet das allgemeine Grundmodell?	→ Verfahren benennen und allgemeine Gleichung aufschreiben
b) Wenden Sie das Verfahren an und interpretieren Sie ihr Ergebnis!	→ Regressionsgerade (über Hilfstabelle) bestimmen und Regressionskonstante und -gewicht interpretieren.
c) Wie gut eignet sich das Modell zur Prognose der Siege?	→ R^2 als geeignetes PRE-Maß bestimmen.

Aufgabe 2a: Lösung

- Benennung des Verfahrens:
  - bivariate lineare Regression
- allgemeines mathematisches Grundmodell:
  - $y_i = b_0 + b_1 * x_i + e_i$

Aufgabe 2b: Lösung Hilfstabelle I

Spieler	Trainingsstunden (X)	Siege (Y)	$x_i - \bar{x}$	$(x_i-\bar{x})^2$	$y_i - \bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x}) * (y_i-\bar{y})$
1	1	2					
2	2	0					
3	3	3					
4	4	5					
5	5	1					
6	6	2					
7	7	6					
8	8	3					
9	9	5					
	$\bar{x} =$	$\bar{y} =$		$SAQ_X =$		$SAQ_Y =$	$SP_{XY} =$

Aufgabe 2b: Lösung Hilfstabelle II

Spieler	Trainingsstunden (X)	Siege (Y)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	1	2	-4	16	-1	1	4
2	2	0	-3	9	-3	9	9
3	3	3	-2	4	0	0	0
4	4	5	-1	1	2	4	-2
5	5	1	0	0	-2	4	0
6	6	2	1	1	-1	1	-1
7	7	6	2	4	3	9	6
8	8	3	3	9	0	0	0
9	9	5	4	16	2	4	8
	$\bar{x} = 5$	$\bar{y} = 3$		$SAQ_X = 60$		$SAQ_Y = 32$	$SP_{XY} = 24$

Aufgabe 2b: Lösung III

- Bestimmung Regressionsgewicht:
  - $b_1 = \frac{SP_{X,Y}}{SAQ_X}$
  - $b_1 = \frac{24}{60}$
  - $b_1 = 0{,}4$
- Bestimmung Regressionskonstante:
  - $b_0 = \bar{y} - b_1 * \bar{x}$
  - $b_0 = 3 - 0{,}4 * 5$
  - $b_0 = 1$
- Bestimmung mathematisches Modell:
  - $y_i = b_0 + b_1 * x_i + e_i$
  - $y_i = 1 + 0{,}4 * x_i + e_i$

- $\bar{x} = 5$
- $\bar{y} = 3$
- $SAQ_X = 60$
- $SAQ_Y = 32$
- $SP_{XY} = 24$

Aufgabe 2b: Lösung IV

- Interpretation Regressionskonstante:
  - $b_0 = 1$
  - Der Schnittpunkt mit der y-Achsen liegt bei 1. Das bedeutet, dass für jemanden der nicht trainiert hat (0 Trainingsstunden hatte) ein einziger Sieg vorhergesagt wird.
- Interpretation Regressionsgewicht:
  - $b_1 = 0{,}4$
  - Wenn x um eine Einheit steigt, dann steigt y im Durchschnitt um 0,4 Einheiten. Für jede zusätzliche Trainingsstunde werden also im Durchschnitt 0,4 Siege mehr vorhergesagt.

Aufgabe 2c: Lösung

- Güte des Modells → R^2
  - $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}} = \frac{24}{\sqrt{60*32}}$
  - $r_{X,Y} = 0{,}5477$
  - $R^2 = \left(r_{X,Y}\right)^2 = 0{,}5477^2$
  - $R^2 = 0{,}3$
- Interpretation:
  - Durch Kenntnis der Anzahl der Trainingsstunden lässt sich die Prognose der Siege um 30 % verbessern.
  - 30 % der Streuung der Siege lässt sich mithilfe der Anzahl der Trainingsstunden erklären.

- $\bar{x} = 5$
- $\bar{y} = 3$
- $SAQ_X = 60$
- $SAQ_Y = 32$
- $SP_{XY} = 24$

Übungsaufgabe 1: Einkommensungleichheit

Hat der prozentuale Anteil der öffentlichen Ausgaben am Bruttoinlandsprodukt in % (publicspending) einen Einfluss auf den Gini-Koeffizienten (in Punkten)?

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	,030	1	,030	13,868	,001[b]
	Residuum	,060	28	,002		
	Gesamtsumme	,090	29			

a. Abhängige Variable: Ginikoeffizient Mitte 2000

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	,427	,032		13,154	,000
	Öffentliche Sozialausgaben in % vom BIP	-,006	,002	-,576	-3,724	,001

a. Abhängige Variable: Ginikoeffizient Mitte 2000

a) Bestimmen Sie mithilfe der vorliegenden Tabelle die Regressionsgleichung und interpretieren Sie die Regressionskonstante sowie das -gewicht.

b) Wie gut eignet sich das Modell zur Vorhersage des Gini-Koeffizienten?

Übungsaufgabe 1a: Lösung

- Interpretation Regressionskonstante:
  - $b_0 = 0{,}427$
  - Der Schnittpunkt mit der y-Achse liegt bei 0,427. Für ein Land mit einem prozentualen Anteil der öffentlichen Aufgaben von 0 Prozent am BIP würde ein Wert von 0,427 Punkten vorhergesagt.
- Interpretation Regressionsgewicht:
  - $b_1 = -0{,}006$
  - Wenn sich x um eine Einheit erhöht, verringert sich y um 0,006 Einheiten. Mit jedem zusätzlichen Prozent öffentlicher Ausgaben am BIP sinkt der Gini-Koeffizient um 0,006Punkte.

Übungsaufgabe 1b: Lösung

- Berechnung Determinationskoeffizienten:
  - $R^2 = \frac{E_0 - E_1}{E_0} = \frac{SAQ_{Regresion}}{SAQ_{Gesamt}}$
  - $R^2 = \frac{0{,}03}{0{,}09}$
  - $R^2 = 0{,}333$
- Interpretation Determinationskoeffizient:
  - Durch Kenntnis des Anteils der öffentlichen Ausgaben am GDP verbessert sich die Prognose des GINI-Koeffizienten um 33,1%.

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	,030	1	,030	13,868	,001[b]
	Residuum	,060	28	,002		
	Gesamtsumme	,090	29			

a. Abhängige Variable: Ginikoeffizient Mitte 2000

b. Prädiktoren: (Konstante), Öffentliche Sozialausgaben in % vom BIP

Übungsaufgabe 2: Wasserqualität und Kindersterblichkeit

- Eine Forscherin der OECD nimmt an, dass die Kindersterblichkeit pro 1.000 Lebendgeburten u.a. von der (wahrgenommenen) Wasserqualität abhängt. Für die folgende Aufgabe wurde eine Zufallsstichprobe mit 10 Fällen gezogen.

a) Welches Verfahren sollte zum Einsatz kommen? Spezifizieren Sie das mathematische Modell! (2)

b) Erstellen Sie eine Arbeitstabelle und berechnen Sie das entsprechende Modell. (6)

c) Interpretieren Sie Ihr Ergebnis statistisch und inhaltlich. (2)

d) Bitte berechnen Sie die Güte des Modells und interpretieren Sie Ihr Resultat! (3)

Land	Zufriedenheit mit Wasser in %	Kindersterblichkeit auf 1.000 Lebendgeburten
Belgien	84,7	3,4
Kanada	91,3	5,1
Chile	84,5	7
Finnland	95	2,6
Israel	64,3	3,8
Italien	80,6	3,7
Spanien	81,6	3,5
Schweiz	96,7	4
Türkei	64,1	17
UK	94,8	4,7

Übungsaufgabe 2a: Analyse

Land	Zufriedenheit mit Wasser in %	Kindersterblichkeit auf 1.000 Lebendgeburten
Belgien	84,7	3,4
Kanada	91,3	5,1
Chile	84,5	7
Finnland	95	2,6
Israel	64,3	3,8
Italien	80,6	3,7
Spanien	81,6	3,5
Schweiz	96,7	4
Türkei	64,1	17
UK	94,8	4,7

metrische Variable

metrische Variable

Eine Forscherin der OECD nimmt an, dass die Kindersterblichkeit pro 1.000 Lebendgeburten u.a. von der Wasserqualität **abhängt**.

geeignetes Verfahren:
bivariate lineare Regression
$y_i = b_0 + b_1 * x_i + e_i$
geeignetes Maß: R^2

asymmetrischer Zusammenhang:
Wasserqualität (X)
→Kindersterblichkeit (Y)

Übungsaufgabe 2b: Lösung

Land	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}) * (y_i - \overline{y})$
Belgien	84,7	3,4	0,94	-2,08	0,8836	4,3264	-1,9552
Kanada	91,3	5,1	7,54	-0,38	56,8516	0,1444	-2,8652
Chile	84,5	7	0,74	1,52	0,5476	2,3104	1,1248
Finnland	95	2,6	11,24	-2,88	126,3376	8,2944	-32,3712
Israel	64,3	3,8	-19,46	-1,68	378,6916	2,8224	32,6928
Italien	80,6	3,7	-3,16	-1,78	9,9856	3,1684	5,6248
Spanien	81,6	3,5	-2,16	-1,98	4,6656	3,9204	4,2768
Schweiz	96,7	4	12,94	-1,48	167,4436	2,1904	-19,1512
Türkei	64,1	17	-19,66	11,52	386,5156	132,7104	-226,4832
UK	94,8	4,7	11,04	-0,78	121,8816	0,6084	-8,6112
	$\overline{x} = \mathbf{83,76}$	$\overline{y} = \mathbf{5,48}$			$SAQ_X = \mathbf{1253,804}$	$SAQ_Y = \mathbf{160,496}$	$SP_{XY} = \mathbf{-247,71}$

Übungsaufgabe 2b: Lösung II

- Berechnung des Modells:
  - $b_1 = \frac{SP_{XY}}{SAQ_X} = \frac{-247{,}718}{1253{,}804} = -0{,}198$
  - $b_0 = \bar{y} - b_1 * \bar{x} = 5{,}48 - (-0{,}198 * 83{,}75) = 22{,}027$
  - $y_i = 22{,}027 - 0{,}198 * x_i + e_i$

Übungsaufgabe 2c: Lösung I

- Die Regressionskonstante b_0 liegt bei 22,027.
- Dies bedeutet, dass der Schnittpunkt der Geraden mit der Y-Achsen bei diesem Wert liegt.
- In einem Land, wo die Zufriedenheit mit dem Wasser bei 0 liegen würde, würde ein eine Kindersterblichkeit von 22,027 pro 1.000 Lebendgeborenen vorhergesagt. (1)

Übungsaufgabe 2c: Lösung II

- Das Regressionsgewicht b_1 liegt bei -0,198. Wenn sich X um eine Einheit erhöht, so sinkt Y um 0,198 Einheiten. Dies bedeutet, dass für jede Einheit Zufriedenheit (hier Prozent) die Kindersterblichkeit um 0,198 pro 1.000 Lebendgeburten sinkt. (1)
- Wahrscheinlich ist hier die wahrgenommene Wasserqualität eher eine Proxyvariable für tatsächliche Umweltzustände.

Übungsaufgabe 2d: Lösung

- Berechnung von R^2 im bivariaten Fall:
  - $r_{XY} = \frac{SP_{XY}}{\sqrt{SAQ_X * SAQ_Y}} = \frac{-247{,}71}{\sqrt{1253{,}804 * 160{,}496}} = -0{,}552$ (1)
  - $R^2 = (r_{XY})^2 = 0{,}548^2 = 0{,}305$ (1)

- Interpretation:
  - R² ist als PRE-Maß zu interpretieren. Durch Kenntnis der Zufriedenheit mit dem Wasser lässt sich die Streuung der Kindersterblichkeit zu 30,5 % erklären. (1)

Übungsaufgabe 3: Einfluss Urbanisierung auf Alphabetisierungsrate

Basierend auf Daten der Weltbank aus dem Jahre 2009 möchte ein Wissenschaftler untersuchen, ob der Urbanisierungsgrad einen Einfluss auf die Alphabetisierungsrate hat. Leider liegen ihm die Originaldaten nicht mehr vor, so dass er nur noch auf folgende Tabelle zurückgreifen kann:

Variable	Mittelwert	Varianz	Kovarianz
Urbanisierungsgrad in Prozent	57,1	591,9	176,1
Alphabetisierungsrate über 15 Jahre in Prozent	83,8	317,2	

a) Mit welchem Verfahren können Sie die vermutete Beziehung untersuchen? Spezifizieren Sie das mathematische Modell. (3)
b) Berechnen Sie das in a) spezifizierte Modell. (2)
c) Interpretieren Sie Ihr Ergebnis. (2)
d) Wie gut ist das Modell zur Erklärung der Alphabetisierungsrate geeignet? (3)

Übungsaufgabe 3: Analyse

a) Mit welchem Verfahren können Sie die vermutete Beziehung untersuchen? Spezifizieren Sie das mathematische Modell. (3)
b) Berechnen Sie das in a) spezifizierte Modell. (2)
c) Interpretieren Sie Ihr Ergebnis. (2)
d) Wie gut ist das Modell zur Erklärung der Alphabetisierungsrate geeignet? (3)

Grundmodell der einfachen linearen Regression gesucht

Berechnung des Modells, also Regressionskonstante und Regressionsgewicht gesucht.

statistische und inhaltliche Interpretation von Regressionsgewicht und -konstante.

Erklärungskraft = Berechnung des PRE-Maßes R^2

Übungsaufgabe 3: Analyse II

Variable	Mittelwert	Varianz	Kovarianz
Urbanisierungsgrad in Prozent	57,1	591,9	176,1
Alphabetisierungsrate über 15 Jahre in Prozent	83,8	317,2	

arithmetisches Mittel $\bar{x}$

arithmetisches Mittel $\bar{y}$

Varianz s_x^2

Varianz s_y^2

Kovarianz s_{xy}

Achtung: genau überlegen, welche Variable die abhängige und welche die unabhängige ist. Ebenso sollte unbedingt darauf geachtet werden, ob die Varianz s_x^2 oder Variation SS_x angegeben ist. Das wird häufig verwechselt. Auch von Dozenten. ☺

Übungsaufgabe 3a/b: Lösung

- Verfahren:
  - bivariate lineare OLS-Regression (1)
  - Grundmodell lautet: $y_i = b_0 + b_1 * x_i + e_i$ (2)

- Berechnung Koeffizienten:
  - $b_1 = \frac{s_{XY}}{s_X^2} = \frac{176{,}1}{591{,}9} = 0{,}2975$
  - $b_0 = \bar{y} - b_1 * \bar{x}$
  - $b_0 = 83{,}8 - 0{,}2975 * 57{,}1$
  - $b_0 = 66{,}81275$

Übungsaufgabe 3c: Lösung

- Interpretation Regressionskonstante:
  - Die Regressionskonstante liegt bei 66,8.
  - Dies bedeutet geometrisch, dass der Schnittpunkt mit der y-Achse bei 66,8 liegt.
  - Ein Land mit einem Urbanisierungsgrad von 0% hätte nach Modell eine Alphabetisierungsrate von 66,8%.
- Interpretation Regressionsgewicht:
  - Das Regressionsgewicht liegt bei 0,2975.
  - Dies ist geometrisch gesehen der Anstieg, wenn X um eine Einheit ansteigt.
  - Dies bedeutet, dass mit jedem zusätzlichen Prozent Urbanisierungsgrad die Alphabetisierungsrate um 0,2975 % ansteigt

Übungsaufgabe 3d: Lösung

- Berechnung R^2 über r_{XY}:
  - $r_{XY} = \frac{SP_{XY}}{\sqrt{SAQ_X * SAQ_Y}} = \frac{s_{XY}}{\sqrt{s_X^2 * S_Y^2}} = \frac{176{,}1}{\sqrt{591{,}9 * 317{,}2}} = 0{,}4064 \ldots$
  - $R^2 = \left(r_{XY}\right)^2 = 0{,}165$
- Interpretation:
  - Der Determinationskoeffizient liegt bei 0,165.
  - Dies bedeutet, dass sich die Vorhersage der Streuung der Alphabetisierungsrate bei Kenntnis des Urbanisierungsgrades um 16,5 % verbessert.

Literaturhinweise

- Kerstin Völkl / Christoph Korb (2018): Deskriptive Statistik. Eine Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler. S. 217-230.
- Steffen-M. Kühnel / Dagmar Krebs (2012): Statistik für die Sozialwissenschaften. S. 456-469.
- Hans Benninghaus (2007): Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler. S. 192-227.