

M24 Statistik 1: Wintersemester 2024 / 2025

# Seminar 05: Regression

MSc Albert Anoschin & Prof. Matthias Guggenmos  
Health and Medical University Potsdam



# Wozu dient eine „Regressionsgerade“?

- Mit der Korrelation können wir die *Stärke des linearen Zusammenhangs* zweier Variablen quantifizieren.
- Wir können aber auch Werte einer Variable (z.B. Depression) aufgrund von Werten auf einer anderen Variable (z.B. Social Media Nutzung) *vorhersagen*.

## Beispiel:

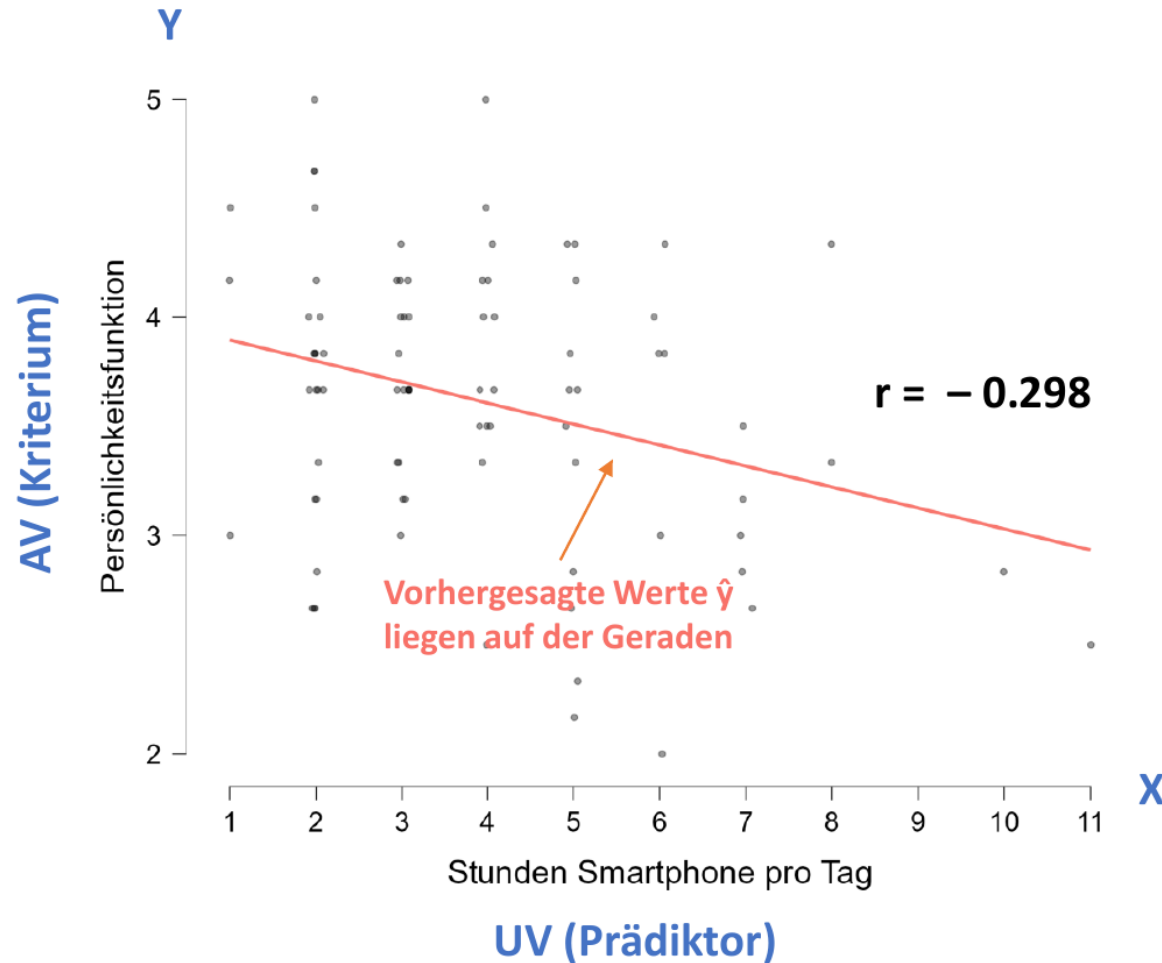
Welchen Wert (Depression) würden Sie für eine Person vorhersagen, über die Sie keine Informationen haben? Sie kennen die Verteilung von Depression in der Bevölkerung.

⇒ **Mittelwert**

Wie könnten Sie die Vorhersage der Depression einer Person verbessern? Sie wissen, dass Social Media Nutzung positiv mit Depression korreliert.

⇒ **Hinzunahme zusätzlicher Information (Stunden Social Media Nutzung als *Prädiktor*)**

# Lineare Regression

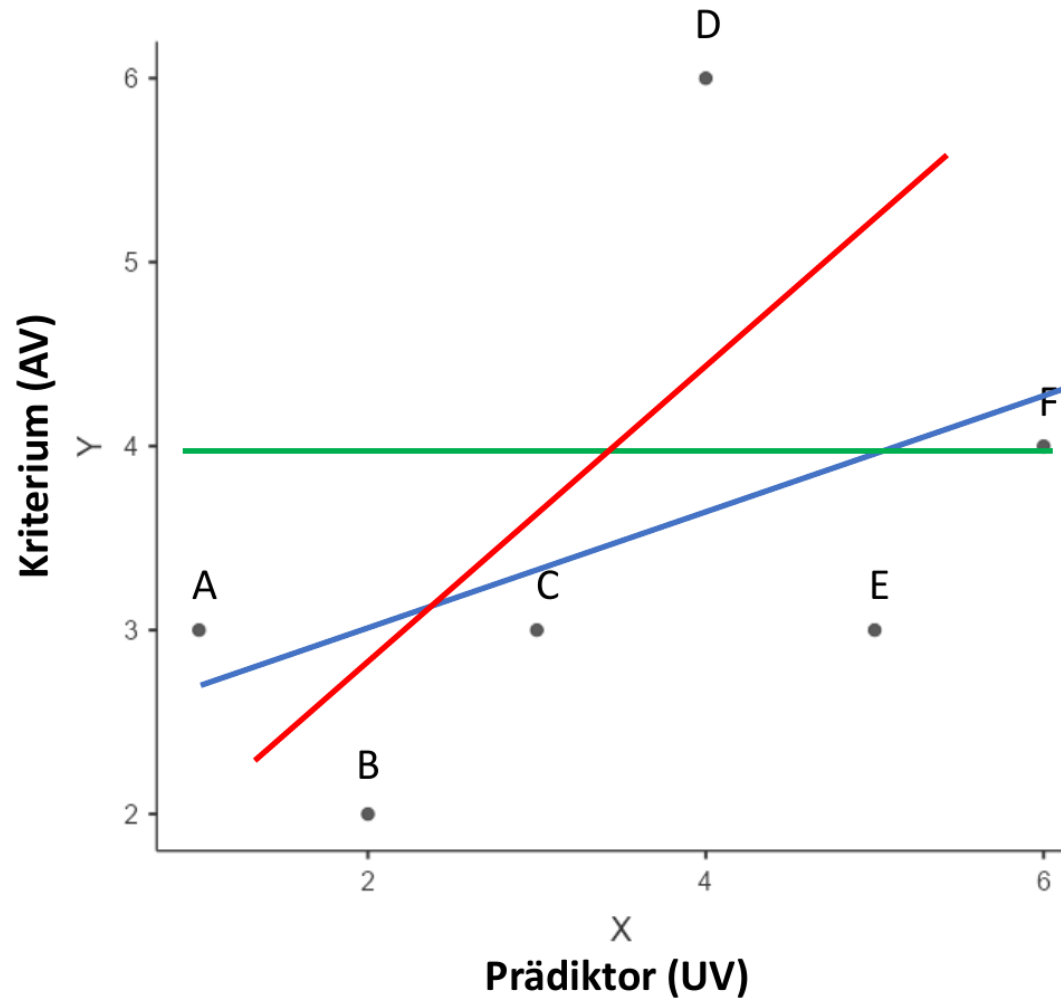


## Ziel der Regression:

Vorhersage von Werten auf der abhängigen Variablen (AV) aufgrund von Werten auf der unabhängigen Variable (UV)

**Beispiel:** Wie gut ist die Persönlichkeitsfunktion (psychosoziale Anpassung) einer Person, die neun Stunden am Tag vor ihrem Smartphone verbringt?

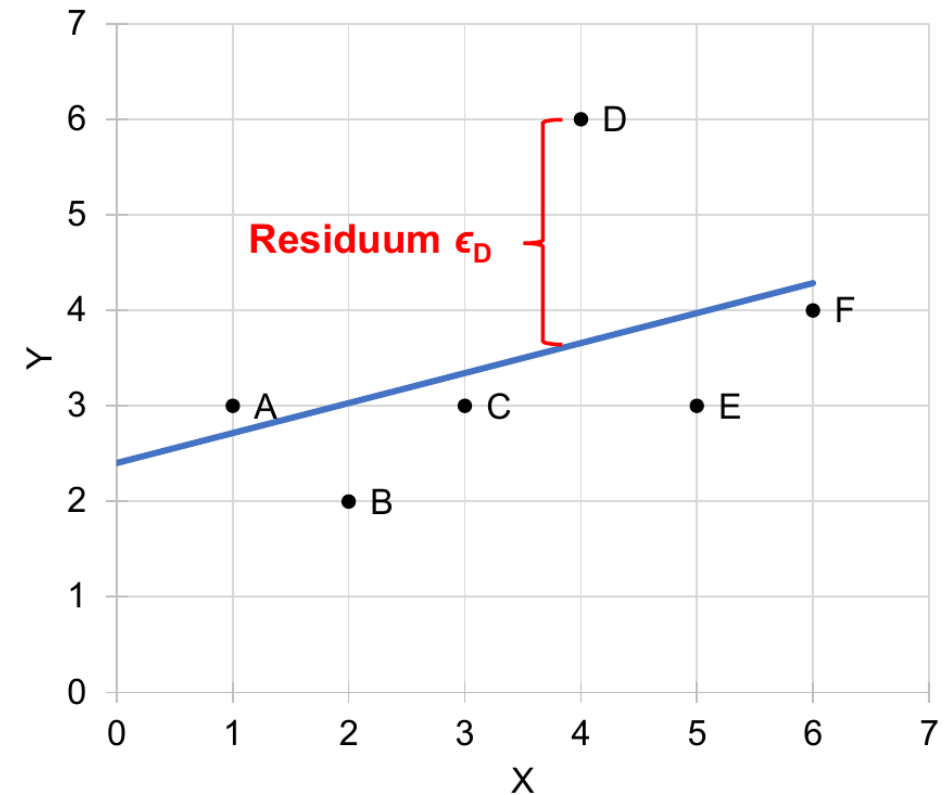
# Bestimmung der Regressionsgeraden: Welche Linie passt am besten?



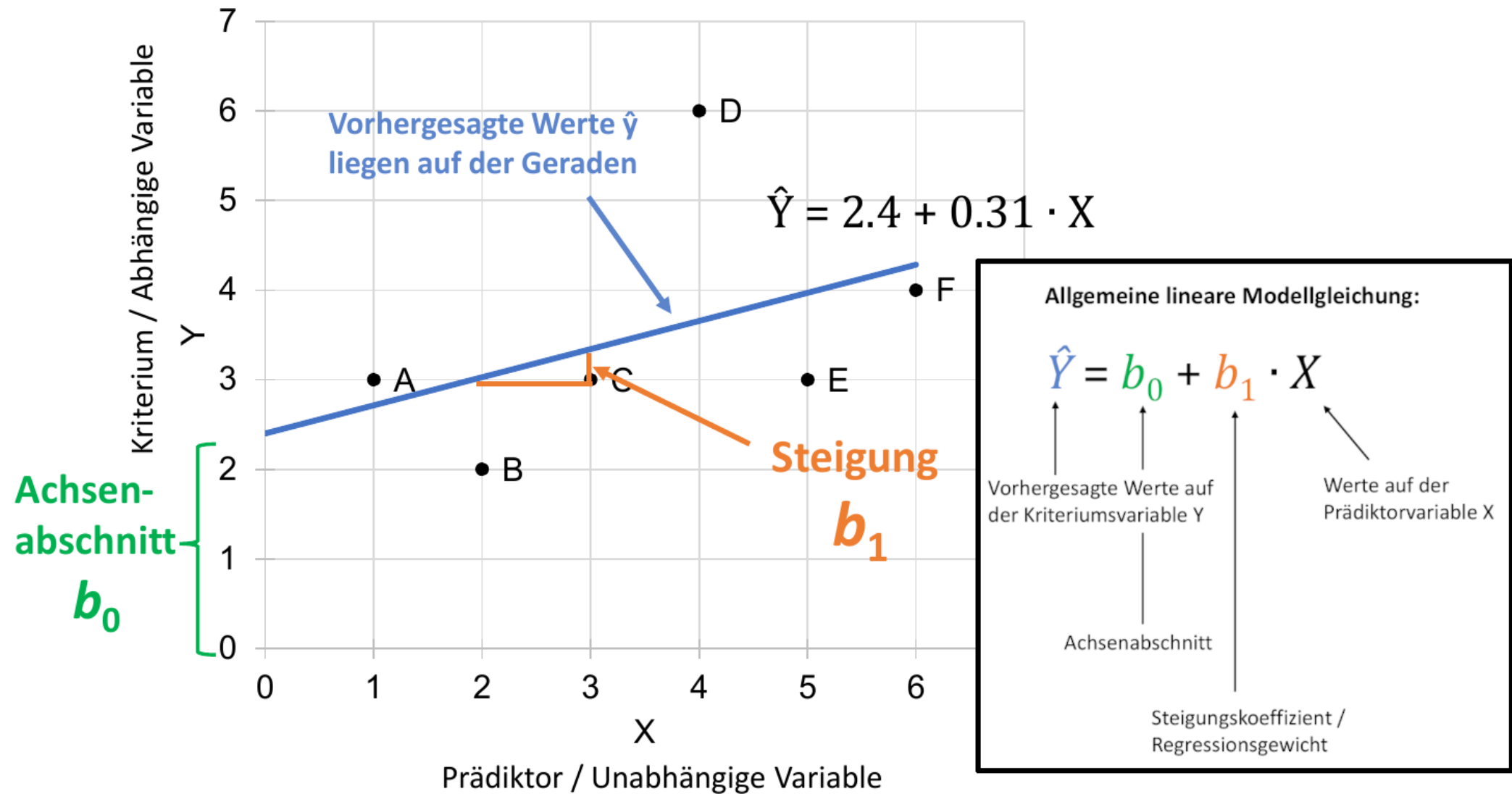
**Regressionsgerade:** Gerade mit dem kleinsten Abstand zu den beobachteten Werten (Methode kleinster Quadrate).

# Residuen

- Ein Regressionsmodell kann in der Praxis nie alle beobachteten Werte perfekt vorhersagen.
- D.h. die beobachteten Werte ( $y_A, y_B, y_C, \dots$ ) weichen fast immer von den vorhergesagten Werten ( $\hat{y}_A, \hat{y}_B, \hat{y}_C, \dots$ ) ab.
- Diese Abweichung nennt man Residuum.
- Die Summe aller Regressionsresiduen ist gleich 0.
- Die Summe aller quadrierten Residuen ist minimal  
→ **Methode der kleinsten Quadrate**



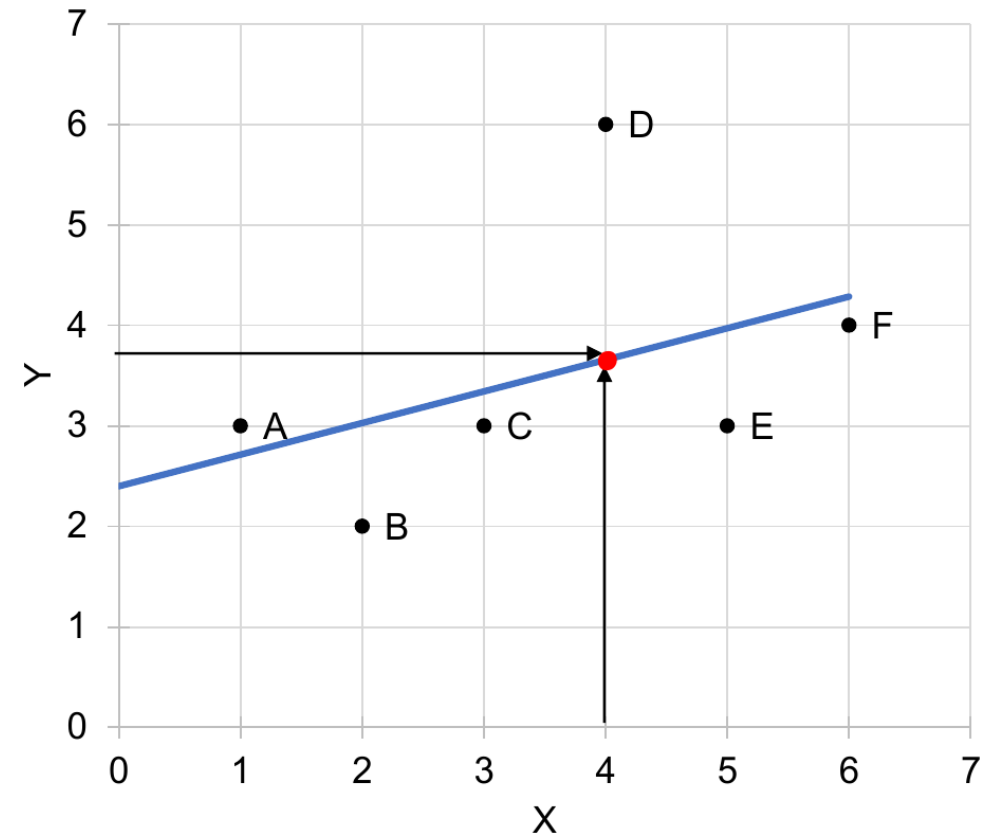
# Regressionsparameter



# Modellvorhersagen

- Marvin hat den Wert 4 auf der X-Variable. Welchen Wert würden wir für ihn auf der Y-Variable erwarten?

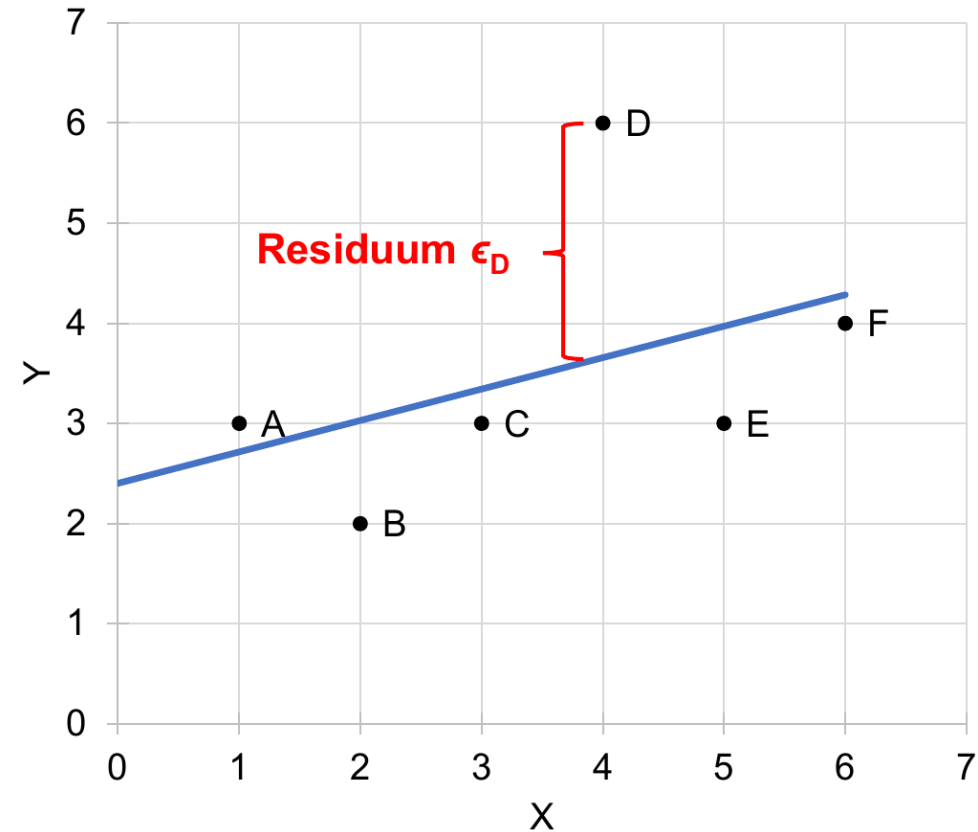
$$\hat{Y} = 2.4 + 0.31 \cdot X$$
$$\hat{Y}_{\text{Marvin}} = 2.4 + 0.31 \cdot 4 = 3.64$$



# Gleichungen für beobachtete Werte

- Die Regressionsgleichung für einen beobachteten Werte  $x_i$  enthält das Residuum für diesen Wert  $\epsilon_i$ .
- Sie entspricht also der Gleichung für den vorhergesagten Wert mit Addition bzw. Subtraktion des Residuums:

$$y_i = b_0 + b_1 \cdot x_i + \epsilon_i$$





# Mathematische Bestimmung der Regressionsparameter

## Regressionsgewicht

$$b_1 = \frac{\text{Kovarianz von X und Y}}{\text{Varianz von X}} = r_{x,y} \frac{\text{Standardabweichung von y}}{\text{Standardabweichung von x}}$$

Diagramm zur Bestimmung des Regressionsgewichts  $b_1$ :

- Oben links: Kovarianz von X und Y (Pfeil nach unten zu  $\hat{\sigma}_{XY}$ )
- Oben rechts: Standardabweichung von y (Pfeil nach unten zu  $\sigma_y$ )
- Unten links: Varianz von X (Pfeil nach oben zu  $\hat{\sigma}_X^2$ )
- Unten rechts: Standardabweichung von x (Pfeil nach oben zu  $\sigma_x$ )

## Achsenabschnitt

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Diagramm zur Bestimmung des Achsenabschnitts  $b_0$ :

- Oben links: Mittelwert von Y (Pfeil nach unten zu  $\bar{y}$ )
- Oben rechts: Mittelwert von X (Pfeil nach unten zu  $\bar{x}$ )
- Unten: Regressionsgewicht  $b_1$  (Pfeil nach oben zu  $b_1$ )

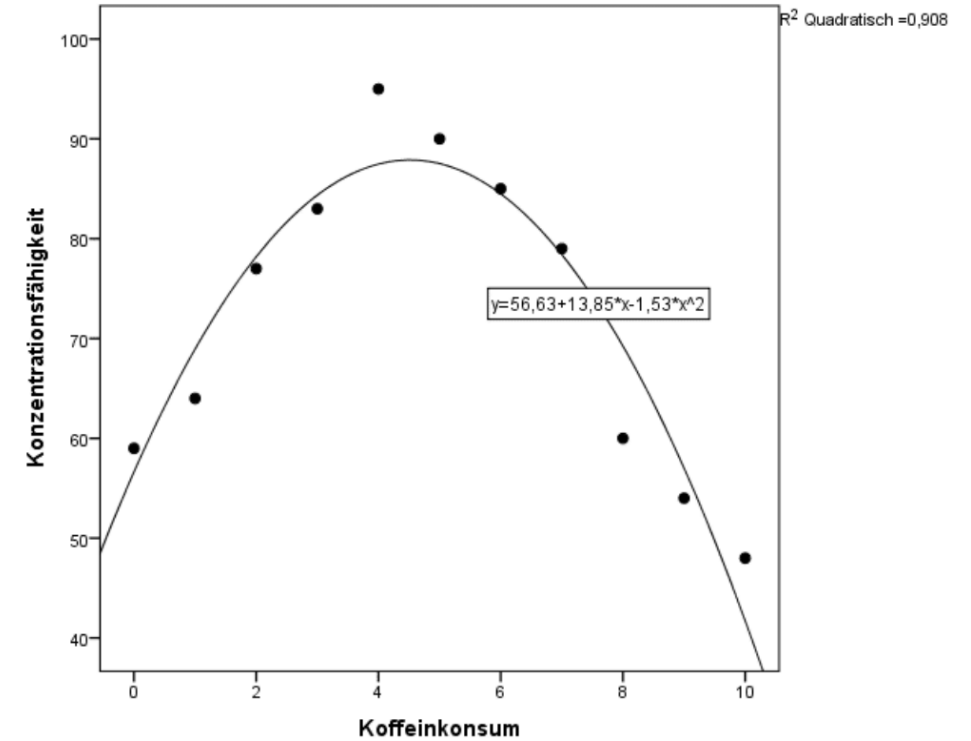
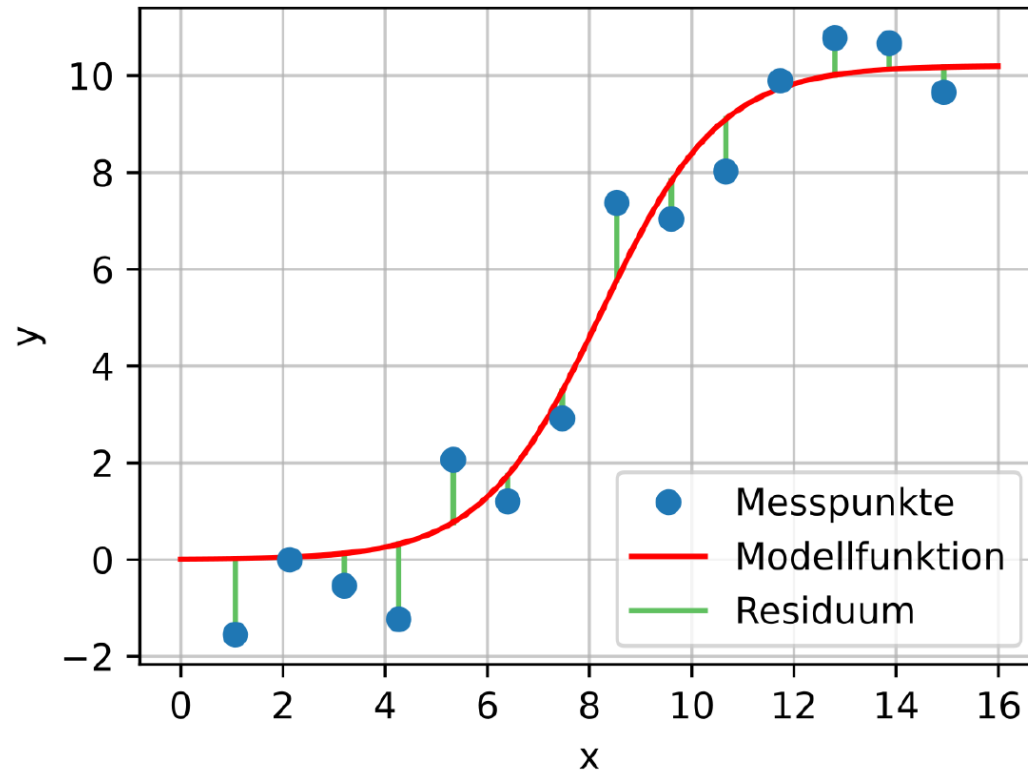
# Korrelation vs. Regression

- Das Regressionsgewicht  $b_1$  hängt von den Einheiten der Variablen ab. Es beantwortet die Frage „Um wie viel Einheiten erhöht sich die AV, wenn man die UV um 1 erhöht?“
- Der Korrelationskoeffizient  $r$  ist ein einheitsloses Maß für die Stärke des linearen Zusammenhangs
- Bei einer univariaten Regression (ein Prädiktor, eine abhängige Variable) lassen sich  $b_1$  und  $r$  ineinander überführen:

$$r_{X,Y} = b_1 \frac{\sigma_X}{\sigma_Y}$$

- Die Regression ist ein vielseitiges Verfahren. Sie ermöglicht u.a. die Untersuchung von Zusammenhängen zwischen mehreren Variablen (multivariate Regression) und von nicht-linearen Zusammenhängen.

# Nicht-lineare Regressionen



# Ausblick auf Statistik II: Multivariate Regression

Wenn wir die Unterschiede zwischen Personen in der Persönlichkeitsfunktion (= Varianz der AV) besser erklären wollen, können wir weitere Prädiktoren in das Regressionsmodell aufnehmen:

**Modell mit 1 Prädiktor:**

$$\text{Persönlichkeitsfunktion} = b_0 + b_1 \cdot \text{Smartphonestunden}$$

**Modell mit 2 Prädiktoren:**

$$\text{Persönlichkeitsfunktion} = b_0 + b_1 \cdot \text{Smartphonestunden} + b_2 \cdot \text{Optimismus}$$

Betrachten wir also mehr als einen Prädiktor, können wir (möglicherweise) bessere Vorhersagen treffen (und  $R^2$  erhöhen).

# Lineare Regressionen in JASP

**Hypothese:** Gewissenhaftere Personen haben weniger Schwierigkeiten beim Stillsitzen.

Koeffizienten

Modell		Unstandardisiert	Standardfehler	Standardisiert	t	p
H <sub>0</sub>	(Konstante)	3.647	0.428		8.516	< .001
H <sub>1</sub>	(Konstante)	7.141	1.574		4.537	< .001
	Gewissenhaftigkeit	-1.042	0.455	-0.509	-2.288	0.037

Regressionsgleichung:  $\hat{Y} = b_0 + b_1 \cdot X$

$$\hat{Y} = 7.141 + 1.042 \cdot X$$

# Übung

**Sie wollen die Lebenszufriedenheit von Personen mittels Extraversion vorhersagen.**

1. Formulieren Sie eine gerichtete Zusammenhangshypothese.
2. Bestimmen Sie X (Prädiktor) und Y (Kriterium)
3. Berechnen Sie von Hand das Regressionsgewicht  $b_1$ . Lassen Sie sich dafür zunächst in JASP die Deskriptiven Statistiken und den Korrelationskoeffizienten  $r$  zwischen X und Y ausgeben.
4. Berechnen Sie von Hand den Achsenabschnitt  $b_0$
5. Überprüfen Sie Ihre Berechnungen, indem Sie sich in JASP das Regressionsmodell ausgeben lassen.

Menü: Regression -> Lineare Regression. Metrische Prädiktoren werden im JASP-Menü „Kovariaten“ genannt, nominale Prädiktoren (z.B. Gruppen) heißen „Faktoren“.

6. Lesen Sie die Regressionsparameter ab und notieren Sie die Regressionsgleichung.
7. Sagen Sie anhand der Regressionsgleichung den Wert der Lebenszufriedenheit für eine Person vorher, die einen Extraversionswert von 2 aufweist.

# Anpassungsgüte: $R^2$

- Als standardisiertes Maß der Anpassungsgüte eines Regressionsmodells verwendet man den *Determinationskoeffizienten*  $R^2$
- $R^2$  liegt zwischen 0 und 1.
- Ist über verschiedene Studien und Modelle hinweg vergleichbar
- Interpretation von  $R^2$ :
  - Anteil der Gesamtvarianz in der AV (z.B. Schwierigkeiten beim Stillsitzen), der durch systematische Varianz in der UV (z.B. Gewissenhaftigkeit) erklärt wird.

Modell-Zusammenfassung - Stillsitzen

Modell	R	$R^2$	Korrigiertes $R^2$	RMSE
$H_0$	0.000	0.000	0.000	1.766
$H_1$	0.509	0.259	0.209	1.570

Bei einer univariaten linearen Regression entspricht  $R^2 = r^2$