

# Multivariate Verfahren

---

## Einheit 1: Wiederholung und Einführung in multivariate Verfahren

Wintersemester 2025 | Prof. Dr. Stephan Goerigk

# Kontakt

Prof. Dr. phil. Stephan Goerigk

Psychologische Methodenlehre

Infanteriestraße 11a · 80797 München ·

[stephan.goerigk@charlotte-fresenius-uni.de](mailto:stephan.goerigk@charlotte-fresenius-uni.de)

Zoom Sprechstunde (bitte per Email anmelden):

- Meeting-ID: 284 567 8838
- Kenncode: 807174

Commitment to Research Transparency

<http://www.researchtransparency.org>



# Übersicht Lehrveranstaltung

## Termine:

- Mittwoch 14:45 - 16:15
- Raum: HS P2 001

## Begleitende Übung zur Vorlesung (Blockunterricht):

- Dozentin: Janika Saretzki ([janika.saretzki@charlotte-fresenius-uni.de](mailto:janika.saretzki@charlotte-fresenius-uni.de))

## Materialien:

- werden auf **Studynet** ([https://studynet.hs-fresenius.de/ilias.php?baseClass=ilrepositorygui&ref\\_id=49744](https://studynet.hs-fresenius.de/ilias.php?baseClass=ilrepositorygui&ref_id=49744)) bereitgestellt

## Prüfungsleistung:

- Portfolio (Abgabetermin: März 2026)

# Material (bitte mitbringen)

Es werden Berechnungen mit R durchgeführt

- Installation R und RStudio (idealerweise) auf eigenem Laptop
- Material auf Lernplattform Studynet
- Foliensätze
- Wiederholung: R Einführung (Skriptum, Bachelor)
- Installationsskript für relevante R Pakete
- Markdown Code zur Bearbeitung der Prüfungsleistung (Portfolio)

# Prüfungsleistung - Portfolio

- Im Verlauf des Semesters angefertigtes Lernportfolio
- Besteht aus mehreren Aufgaben zur Überprüfung der im Modul vermittelten Kompetenzen
- Arbeiten mit Anwendungsbezug (z.B. Studienplanung)
- Analysen und Analyseprotokolle
- grafische Aufbereitungen
- Thesenpapiere und Reflexionen
- Die Aufgaben werden in einem RMarkdown Codebook bearbeitet.
- Aufgaben werden während des Semesters bearbeitet, Codebook wird am Semesterende in studynet abgegeben (Uploadbereich).

# Termine

Einheit	Datum	Thema
1	17.10.2024	Wiederholung und Einführung in multivariate Verfahren
2	24.10.2024	Mehrebenenanalyse - Lineare gemischte Modelle (1)
3	31.10.2024	Mehrebenenanalyse - Lineare gemischte Modelle (2)
4	07.11.2024	Mehrebenenanalyse - Lineare gemischte Modelle (3)
5	14.11.2024	Spezielle multivariate Hypothesen - Moderation
6	21.11.2024	Gruppieren: Clusteranalyse
7	28.11.2024	Gruppieren: Faktorenanalyse
8	12.12.2024	Multivariate (latente) Modellierung (1)
9	19.12.2024	Multivariate (latente) Modellierung (2)
10	02.01.2025	Multivariate (latente) Modellierung (3)
11	09.01.2025	Spezielle multivariate Hypothesen - Mediation
12	16.01.2025	Zeitreihenanalyse
13	23.01.2025	Netzwerkmodelle
14	30.01.2025	Multilevel-Vektor-Autoregression
15	06.02.2025	Freie Spitze

# Einführung in multivariate Verfahren

## Was sind multivariate Verfahren?

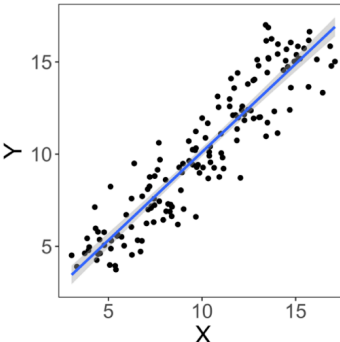
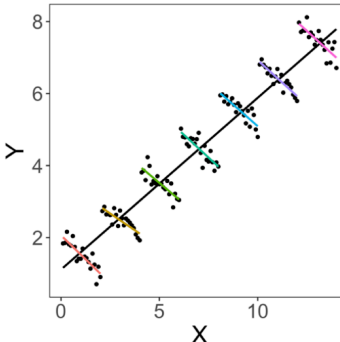
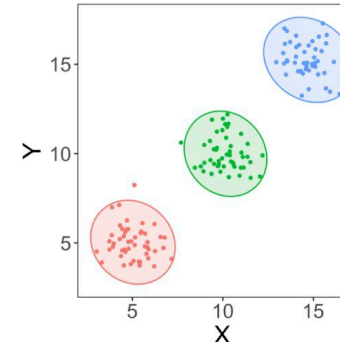
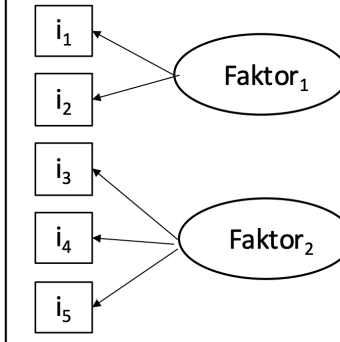
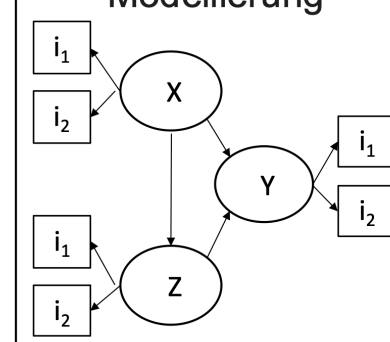
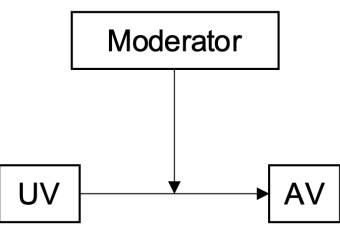
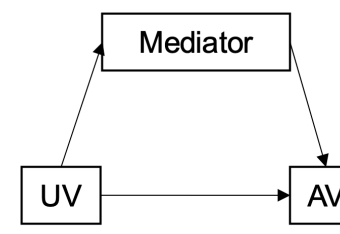
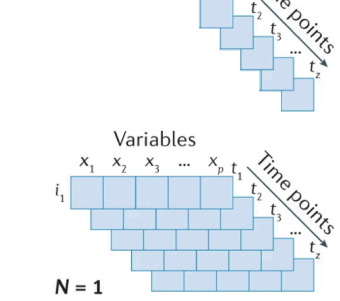
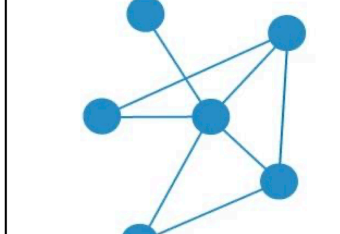
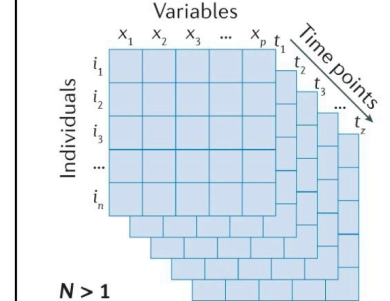
- Statistische Methoden zur gleichzeitigen Analyse mehrerer Variablen.
- Erfassen komplexer Zusammenhänge zwischen verschiedenen Einflussgrößen.
- Nutzen sowohl in der Hypothesengenerierung (strukturentdeckende Verfahren) als auch in der Hypothesenprüfung (strukturprüfende Verfahren).

## Rolle in der Psychologie

- Evaluation der Wirksamkeit von Interventionen auf unterschiedliche Maße (Outcomes).
- Vorhersage von Verhaltensweisen oder psychischen Störungen auf Basis mehrerer Prädiktoren.
- Gruppierung von Personen mit ähnlichen Merkmalen oder Symptomen (z.B. in der klinischen Psychologie).
- Evaluation von Theorien durch Analyse der Beziehungen zwischen mehreren Variablen gleichzeitig.

# Einführung in multivariate Verfahren

## Multivariate Verfahren - Semesterfahrplan

<p>Lineare Modelle</p> 	<p>Mehrebenenmodelle</p> 	<p>Clusteranalyse</p> 	<p>Faktorenanalyse</p> 	<p>Multivariate (latente) Modellierung</p> 
<p>Moderation</p> 	<p>Mediation</p> 	<p>Zeitreihenanalyse</p> 	<p>Netzwerkmodelle</p> 	<p>Multilevel-Vektor-Autoregression</p> 



# Wiederholung: Lineare Regressionsmodelle

- Regression ist eines der flexibelsten Modelle der Statistik - Logik des Regressionsmodells:

***Zu Erklärender Teil (AV) = Erklärender Teil (UVs)***

$$Y = X + \varepsilon$$

***Daten = Modell + Fehler***

***Daten = Erklärte Varianz + Unerklärte Varianz***

***Daten = Systematischer Anteil + Unsystematischer Anteil***

# Wiederholung: Lineare Regressionsmodelle

- Regression ist eines der flexibelsten Modelle der Statistik - Logik des Regressionsmodells:

$$Y = X + \varepsilon$$

## Y / Kriterium / Abhängige Variable

Alle Variablenarten erlaubt - Regressionstyp ändert sich

Allgemeines lineares Modell (LM):

- kontinuierlich numerisch (normalverteilte Fehler)

Verallgemeinertes lineares Modell (GLM):

- nominale Variablen / Kategorien (logistische Regression)
- Anzahl Events (z.B. Poisson-Regression, negativ-bionomiale Regression)
- Prozente (Probit-Regression, Beta-Regression)
- Zeit bis Event (Survivalanalyse)

## X / Prädiktor / Unabhängige Variable

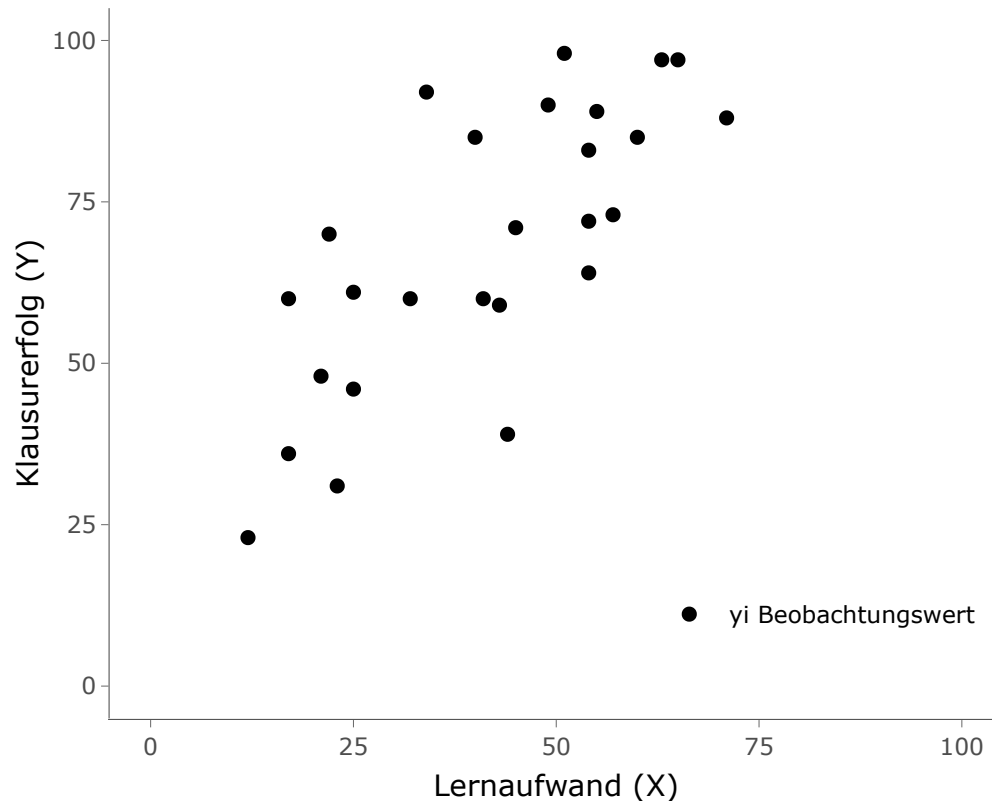
Alle Variablenarten erlaubt - Regressionstyp ändert sich nicht

- 1 Prädiktor = Einfache Regression
- > 1 Prädiktoren = Multiple Regression
  - ohne Interaktion
  - mit Interaktion
- Für numerische Prädiktoren wird 1 Steigung geschätzt
- Kategoriale Prädiktoren werden dummy-codiert
  - es werden k-1 Steigungen geschätzt
  - Jede Steigung quantifiziert der Unterschied von Kategorie  $k_i$  zur Referenzkategorie  $k_0$

# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte

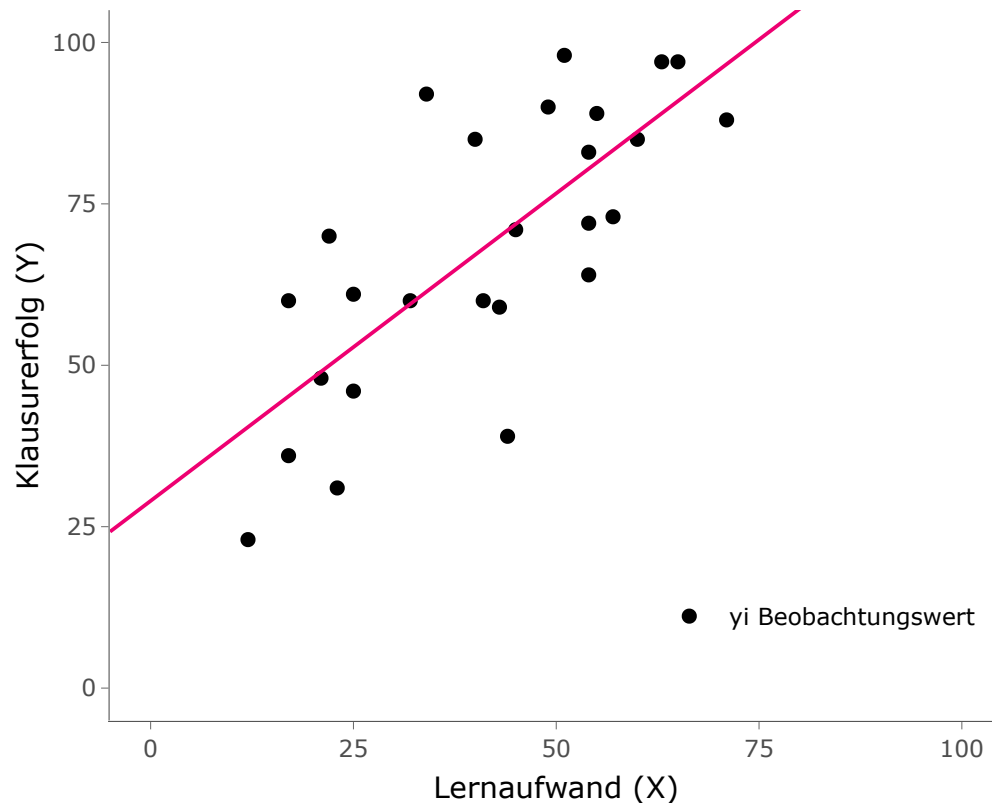


- Jeder Punkt repräsentiert eine Kombination aus  $X$  und  $Y$  Werten
- Wir könnten also sagen, jeder Punkt ist eine Person aus unserem Beispiel
- Es gilt, in der Regression eine Funktion zu finden, die diese Daten möglichst genau widerspiegelt

# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte

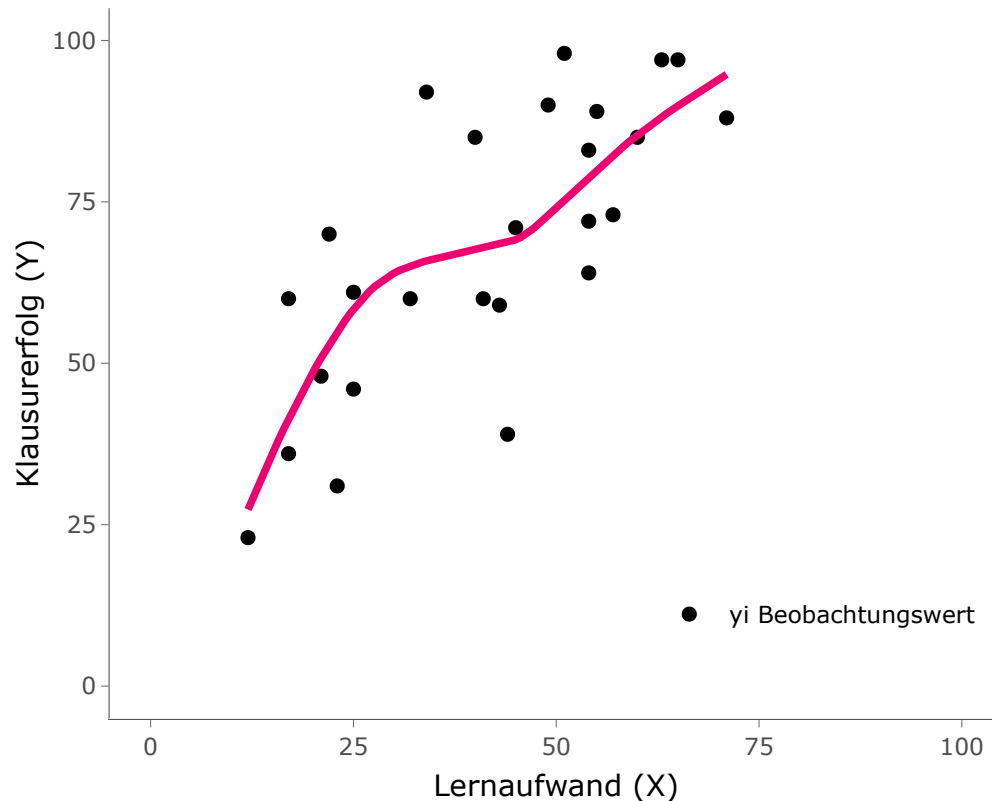


- Im Falle der *linearen* Regression wird unterstellt, dass diese Funktion linear, also eine Gerade ist

# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte

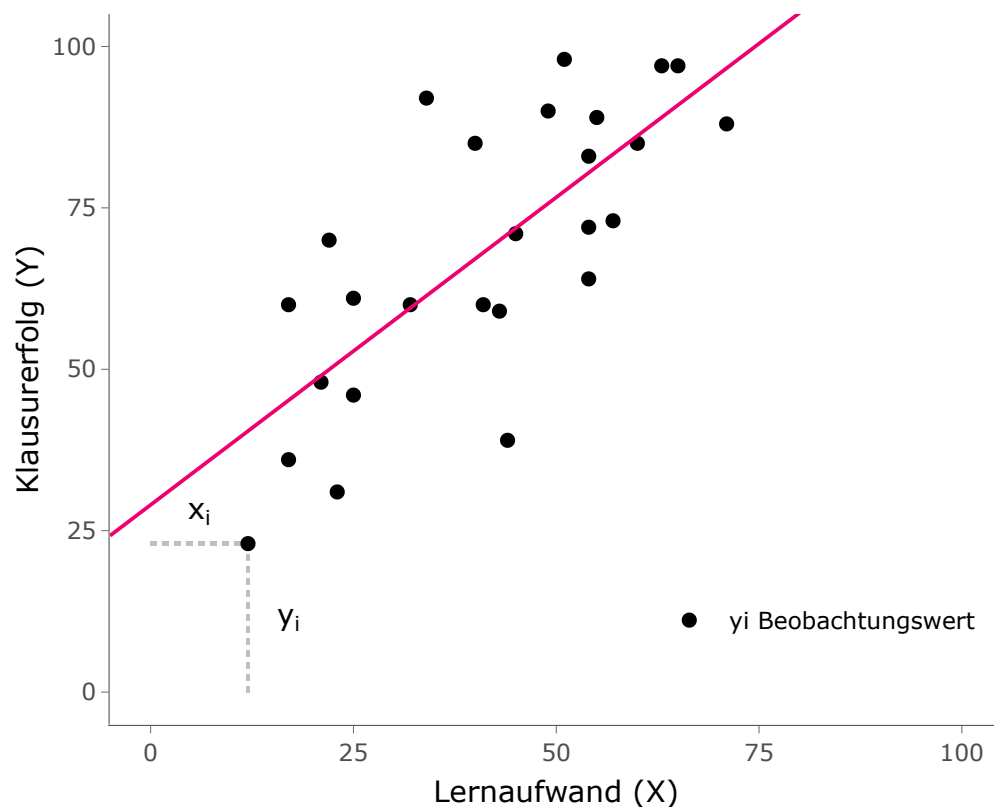


- Theoretisch wären allerdings auch andere Funktionen denkbar.
- Diese beschreiben die vorliegenden Daten ggf. besser, sind aber nicht so leicht interpretierbar/generalisierbar.

# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte



- Jeder Beobachtungspunkt hat für den  $X$  Wert einen entsprechenden  $Y$  Wert.
- Er ist somit eindeutig für die beiden Variablen definiert.

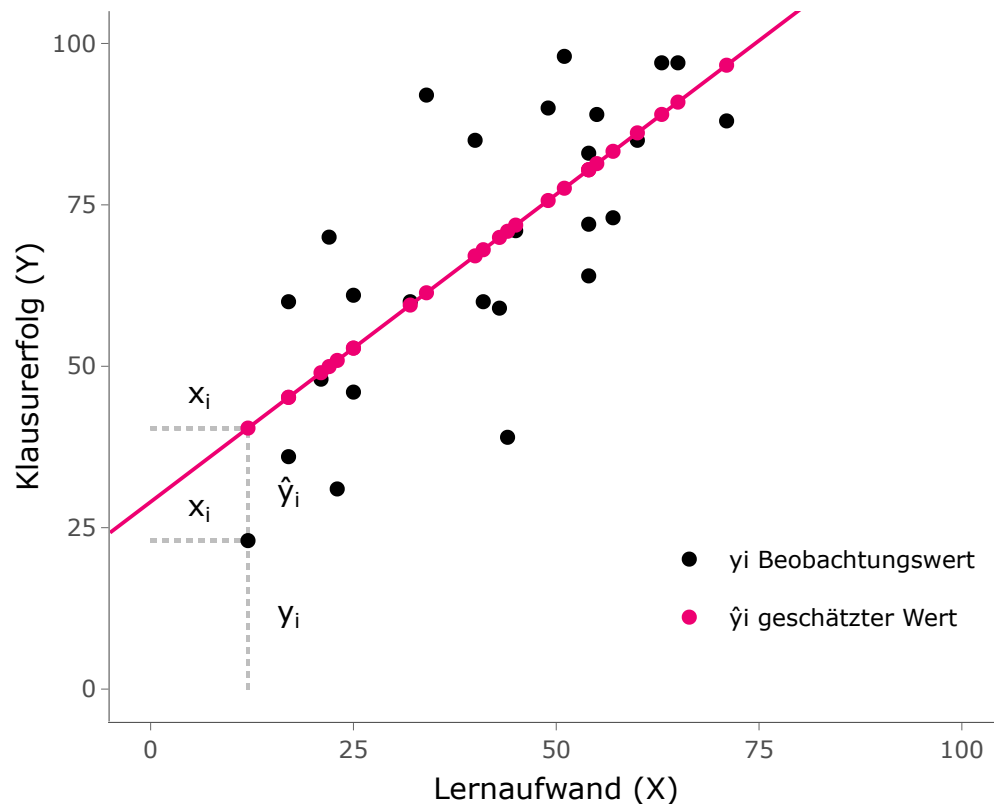
ABER:

- Für jeden gegebenen  $X$  Wert lässt sich ein Punkt auf der Geraden finden, der einen anderen  $Y$  Wert hat

# Einfache lineare Regression

## Wiederholung: Lineare Regressionsmodelle

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte

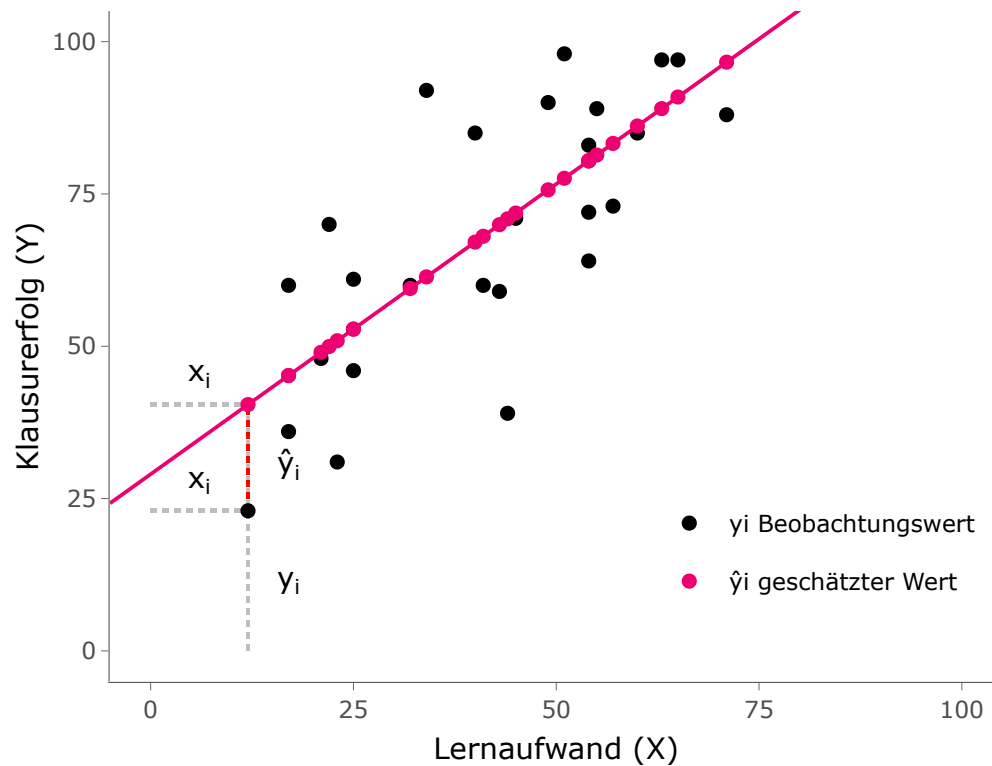


- Der pinke Punkt ist der gemäß der linearen Funktion geschätzte  $\hat{Y}$  Wert für den Punkt  $X$
- Es ist also der Wert, den man unter Annahme eines linearen Zusammenhangs **erwarten** würde
- Diese Punkte haben den  $X$  Wert gemeinsam aber sind unterschiedlich im  $Y$  Wert.

# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte



- Wie wir aber sehen, gibt es hier einen Unterschied in den beiden  $Y$  Werten
- Dieser Unterschied ist unser sogenannter Vorhersagefehler oder auch **Residuum**
- Differenz zwischen Beobachtungswert und vorhergesagtem Wert
- Das Residuum wird mit  $\varepsilon_i$  bezeichnet

Formel für das Residuum:

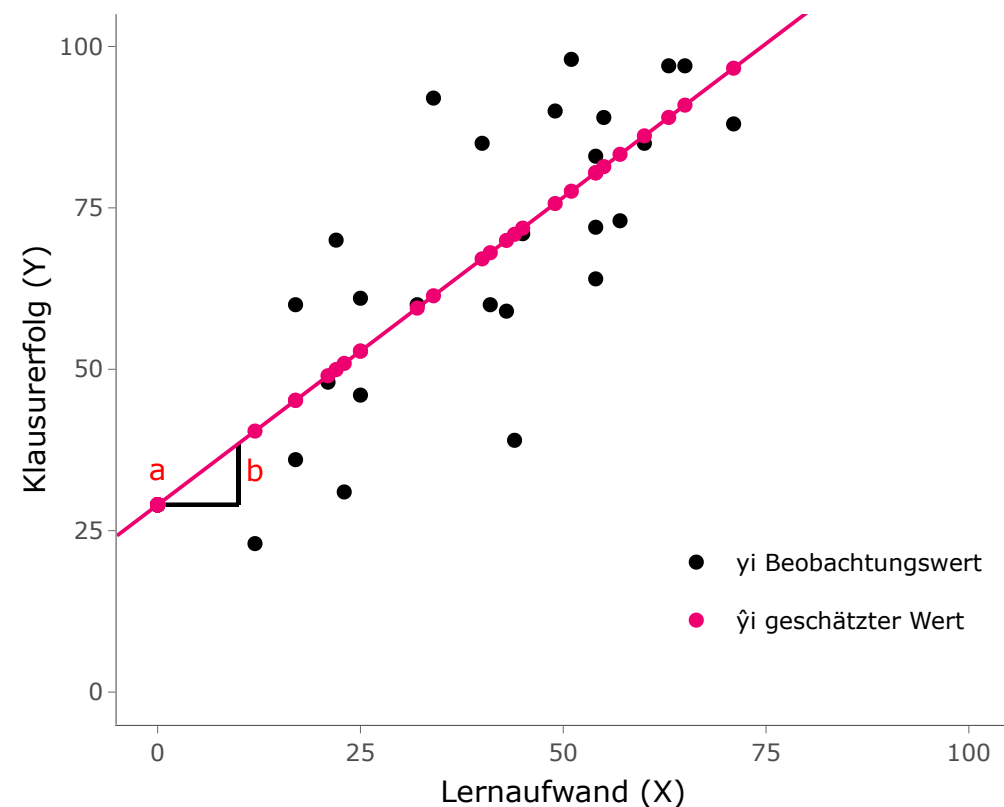
$$\varepsilon_i = \hat{y}_i - y_i$$



# Wiederholung: Lineare Regressionsmodelle

## Lineare Regressionsfunktion

$X \rightarrow Y$  Regressionsfunktion und Beobachtungswerte



.pull-right[

$$\hat{y}_i = a + b \cdot x_i + \varepsilon_i$$

$a$  : Y-Achsenabschnitt

$b$  : Steigungsparameter

### Interpretation:

$a$  : Wert, den  $Y$  hat, wenn  $X = 0$  ist

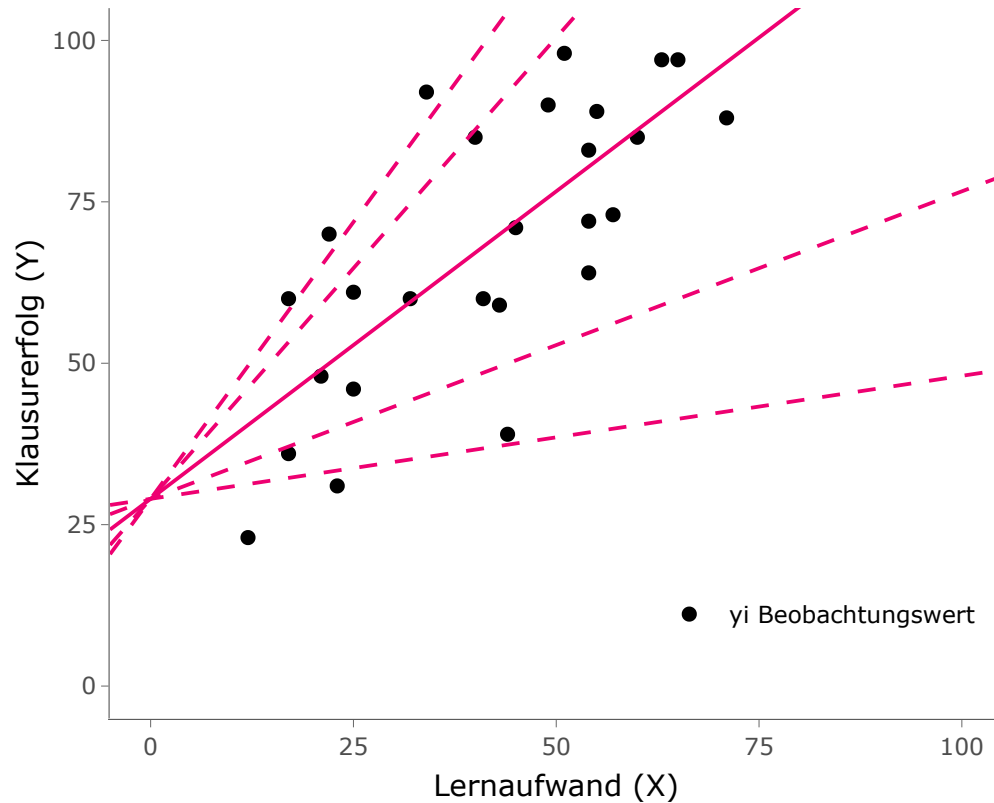
$b$  : Veränderung von  $Y$  bei Zunahme von  $X$  um 1 Einheit

In dieser Vorlesung nutzen wir oft folgende Notation:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

# Wiederholung: Lineare Regressionsmodelle

## Residuen und Zielfunktion



$$\hat{y}_i = a + b \cdot x_i + \varepsilon_i$$

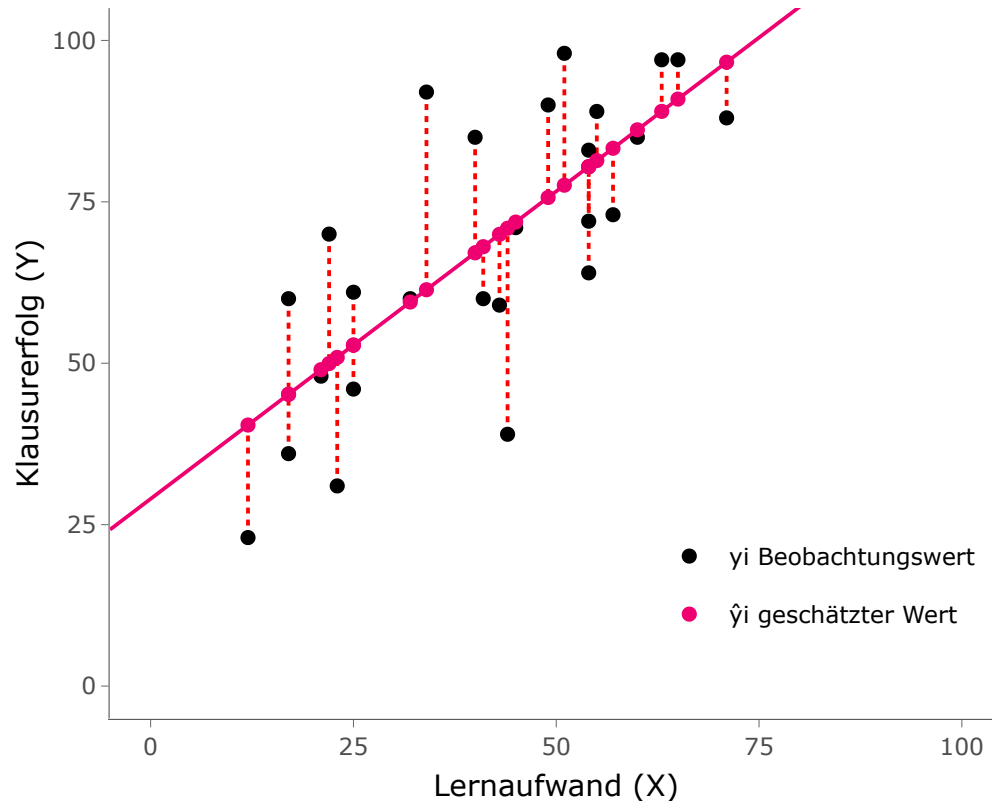
$a$  : Y-Achsenabschnitt

$b$  : Steigungsparameter

- Theoretisch sind endlos viele Geraden denkbar, die die Punktwolke alle an unterschiedlichen Stellen durchschneiden
- Wir wollen aber genau die Gerade finden, welche die Daten am allerbesten beschreibt.

# Wiederholung: Lineare Regressionsmodelle

## Residuen und Zielfunktion



### Ziel:

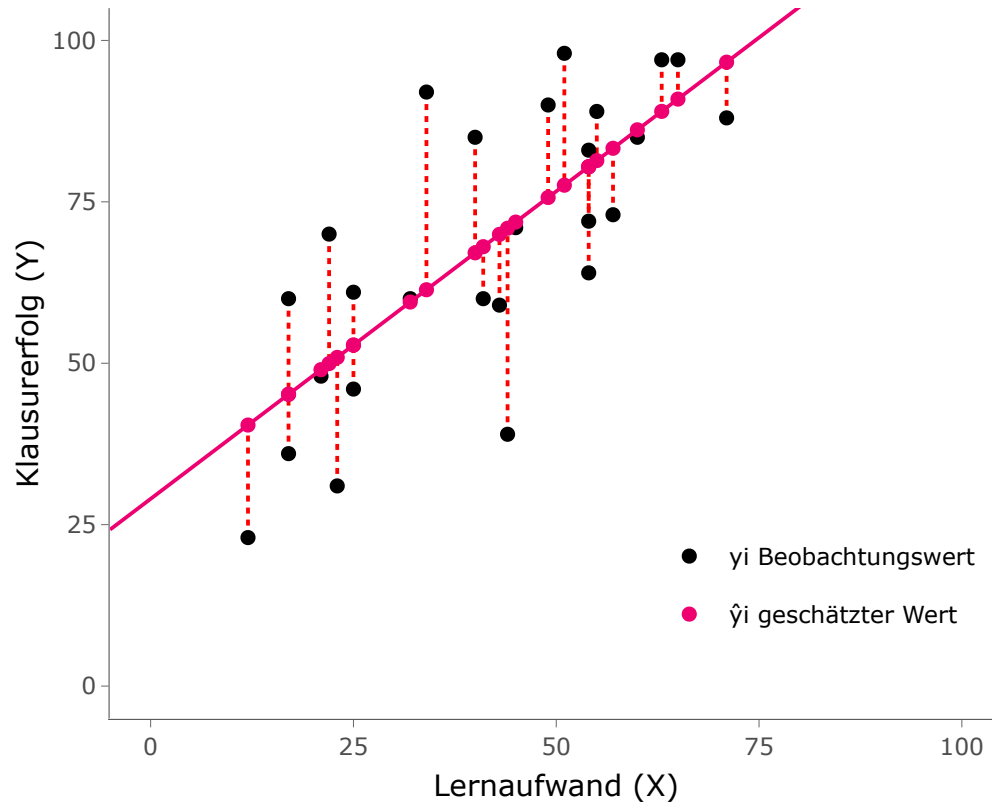
- Y-Achsenabschnitt und Steigung so wählen, dass die lineare Funktion die Punkte möglichst gut widerspiegelt
- gut widerspiegeln = Abstand zwischen dem Beobachtungswert und dem gemäß linearer Funktion geschätzten Wert möglichst klein halten

### Bildliche Vorstellung:

Wenn ich die Residuen aller Beobachtungswerte zu einer Schnur aneinanderhänge, soll diese Schnur möglichst kurz sein

# Wiederholung: Lineare Regressionsmodelle

## Residuen und Zielfunktion



Es liegt ein Optimierungsproblem vor:

- Die Summe der quadrierten Residuen wird über alle Beobachtungswerte minimiert
- So werden die optimalen Werte für a und b gefunden
- Quadrierung verhindert, dass sich negative und positive Werte ausgleichen

$$\sum_{i=1}^n \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 \dots + \varepsilon_n^2 \rightarrow \min_{a,b}$$

# Wiederholung: Lineare Regressionsmodelle

## Kategoriale Prädiktoren: Dummy und Effektkodierung

- kategorialer Prädiktor lässt sich mathematisch integrieren, indem Kategorien numerisch kodiert werden
- Bei 0 und 1 spricht man von einer **Dummy-Kodierung** (z.B.: gesund = 0, erkrankt = 1)
- Bei -1 und 1 spricht man von einer **Effekt-Kodierung** (z.B.: gesund = -1, erkrankt = 1)
- Zahlen sind arbiträr (0 und 4 wäre auch möglich), aber Kodierung mit 1 ist leichter zu interpretieren.

UV: Gruppe (nominal dichotom)	UV: Gruppe (Dummy-kodiert)	UV: Gruppe (Effekt-kodiert)	AV: Sorgen (skaliert von 1-12)
Gesund	0	-1	3.44
Gesund	0	-1	3.77
Gesund	0	-1	5.56
Gesund	0	-1	4.07
Gesund	0	-1	4.13
Gesund	0	-1	5.72
Gesund	0	-1	4.46
Gesund	0	-1	2.73
GAD	1	1	8.31
GAD	1	1	8.55
GAD	1	1	10.22
GAD	1	1	9.36
GAD	1	1	9.40
GAD	1	1	9.11
GAD	1	1	8.44
GAD	1	1	10.79

# Wiederholung: Lineare Regressionsmodelle

## Kategoriale Prädiktoren: Dummy und Effektkodierung

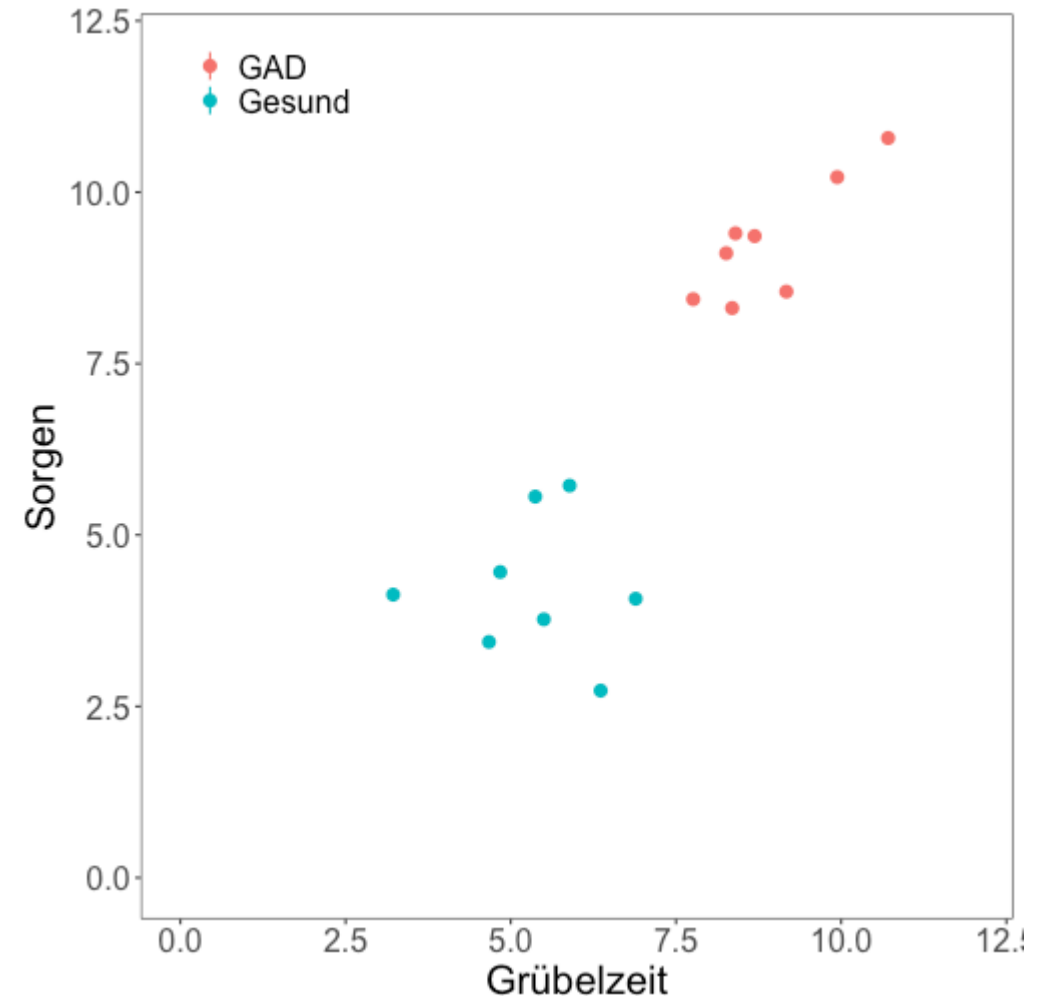
- Art der Kodierung ist wichtig für Interpretation von Modellkoeffizienten
- Auswirkung auf Y-Achsenschnittpunkt (Intercept)
- Auswirkung auf Steigungskoeffizient anderer Prädiktoren

UV: Gruppe (nominal dichotom)	UV: Gruppe (Dummy-kodiert)	UV: Gruppe (Effekt-kodiert)	AV: Sorgen (skaliert von 1-12)	UV: Grübelzeit
Gesund	0	-1	3.44	4.67
Gesund	0	-1	3.77	5.50
Gesund	0	-1	5.56	5.37
Gesund	0	-1	4.07	6.89
Gesund	0	-1	4.13	3.22
Gesund	0	-1	5.72	5.89
Gesund	0	-1	4.46	4.84
Gesund	0	-1	2.73	6.36
GAD	1	1	8.31	8.35
GAD	1	1	8.55	9.17
GAD	1	1	10.22	9.94
GAD	1	1	9.36	8.69
GAD	1	1	9.40	8.40
GAD	1	1	9.11	8.26
GAD	1	1	8.44	7.76
GAD	1	1	10.79	10.71

# Wiederholung: Lineare Regressionsmodelle

## Kategoriale Prädiktoren: Dummy und Effektkodierung

- Art der Kodierung ist wichtig für Interpretation von Modellkoeffizienten
- Auswirkung auf Y-Achsenschnittpunkt (Intercept)
- Auswirkung auf Steigungskoeffizient anderer Prädiktoren



# Wiederholung: Lineare Regressionsmodelle

## Kategoriale Prädiktoren: Dummy und Effektkodierung

**Dummy-Kodierung** (gesund = 0, erkrankt = 1)

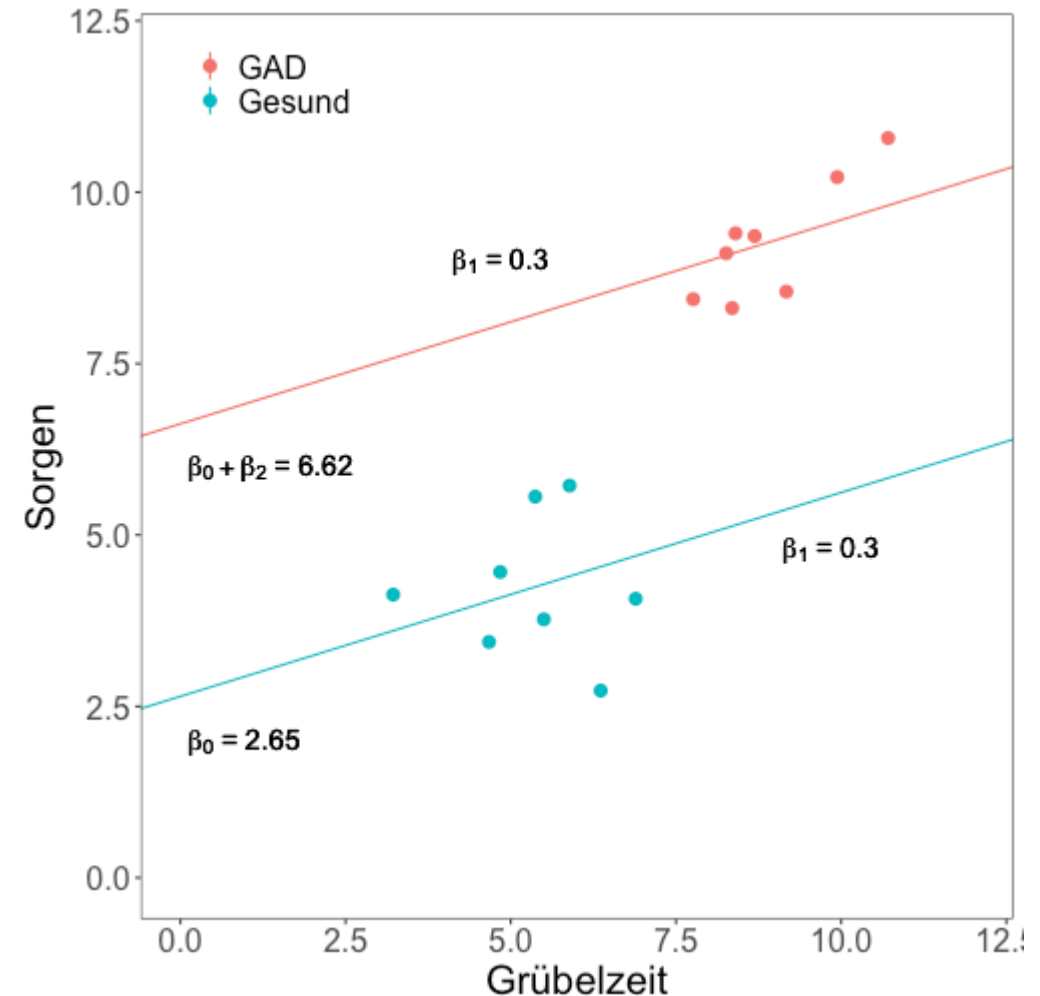
Notation:

$$Sorgen_i = \beta_0 + \beta_1 Gr\ddot{u}belzeit_i + \beta_2 Gruppe_i + \varepsilon_i$$

In R:

```
mod1 = lm(Sorgen ~ Gr  belzeit + Gruppe_dummy, data = df)
coef(mod1)
```

(Intercept)	Gr��belzeit	Gruppe_dummy
2.6454950	0.2975208	3.9760945





# Wiederholung: Lineare Regressionsmodelle

## Kategoriale Prädiktoren: Dummy und Effektkodierung

**Effekt-Kodierung** (gesund = -1, erkrankt = 1)

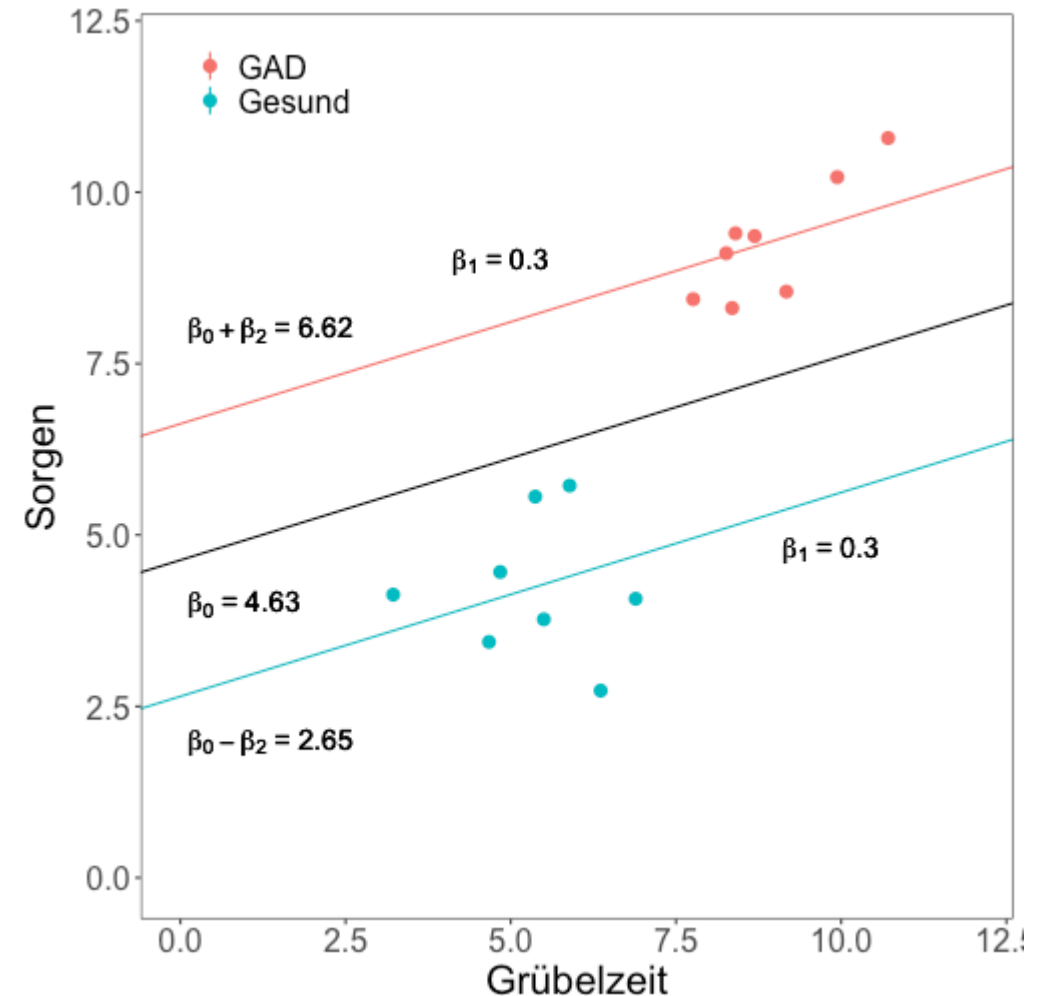
Notation:

$$Sorgen_i = \beta_0 + \beta_1 Gr\ddot{u}belzeit_i + \beta_2 Gruppe_i + \varepsilon_i$$

In R:

```
mod2 = lm(Sorgen ~ Gr  belzeit + Gruppe_effekt, data = df)
coef(mod2)
```

(Intercept)	Gr��belzeit	Gruppe_effekt
4.6335423	0.2975208	1.9880472



# Wiederholung: Lineare Regressionsmodelle

## Inferenz - Regression und Hypothesentests

Hypothesentest	AV	UVs	Fragestellung	Teststatistik
Ein-Stichproben z-Test	Intervallskaliert (Populationsvarianz bekannt)	Keine UV, nur Referenzwert	Unterschied zwischen Stichprobenmittelwert und Referenzwert?	z-Wert
Unabhängiger z-Test	Intervallskaliert (Populationsvarianz bekannt)	1 kategoriale UV, 2 Stufen	Unterschied zwischen 2 Gruppen?	z-Wert
Abhängiger z-Test	Intervallskaliert (Populationsvarianz bekannt)	1 UV Messwiederholung, 2 Messungen	Unterschied zwischen 2 Messzeitpunkten?	z-Wert
Ein-Stichproben t-Test	Intervallskaliert	Keine UV, nur Referenzwert	Unterschied zwischen Stichprobenmittelwert und Referenzwert?	t-Wert
Unabhängiger t-Test	Intervallskaliert	1 kategoriale UV, 2 Stufen	Unterschied zwischen 2 Gruppen?	t-Wert
Abhängiger t-Test	Intervallskaliert	1 UV Messwiederholung, 2 Messungen	Unterschied zwischen 2 Messzeitpunkten?	t-Wert
Einfaktorielle ANOVA	Intervallskaliert	1 kategoriale UV, $\geq 2$ Stufen	Unterschied zwischen $\geq 2$ Gruppen?	F-Wert
ANOVA mit Messwiederholung	Intervallskaliert	1 UV Messwiederholung, $\geq 2$ Messungen	Unterschied zwischen $\geq 2$ Messzeitpunkten?	F-Wert
Einfache Regression	Intervallskaliert	1 kategoriale UV oder 1 stetige UV	Kann UV die AV vorhersagen?	t-Wert (Steigung) oder F-Wert (Omnibus)
Mehrfaktorielle ANOVA	Intervallskaliert	2 kategoriale UVs	Unterschiede zwischen den Stufen der Faktoren? Besteht Interaktion?	F-Wert
Multiple Regression	Intervallskaliert	2 kategoriale oder stetige UVs	Können UVs die AV vorhersagen? Besteht Interaktion?	t-Wert (Steigung) oder F-Wert (Omnibus)
Mixed ANOVA	Intervallskaliert	2 UVs, davon 1 kategoriale UV und eine Messwiederholung	Unterschiede zwischen Stufen und Zeitpunkten? Besteht Interaktion?	F-Wert
$\chi^2$ -Test	Nominalskaliert (dichotom)	1 kategoriale UV, 2 oder mehr Stufen	Unterschiede in Verteilungen/Häufigkeiten?	$\chi^2$ -Wert

# Wiederholung: Lineare Regressionsmodelle

## Inferenz - Regression und Hypothesentests

```
summary(mod1) # Dummy-codiert
```

Call:

```
lm(formula = Sorgen ~ Grübelzeit + Gruppe_dummy, data = df)
```

Residuals:

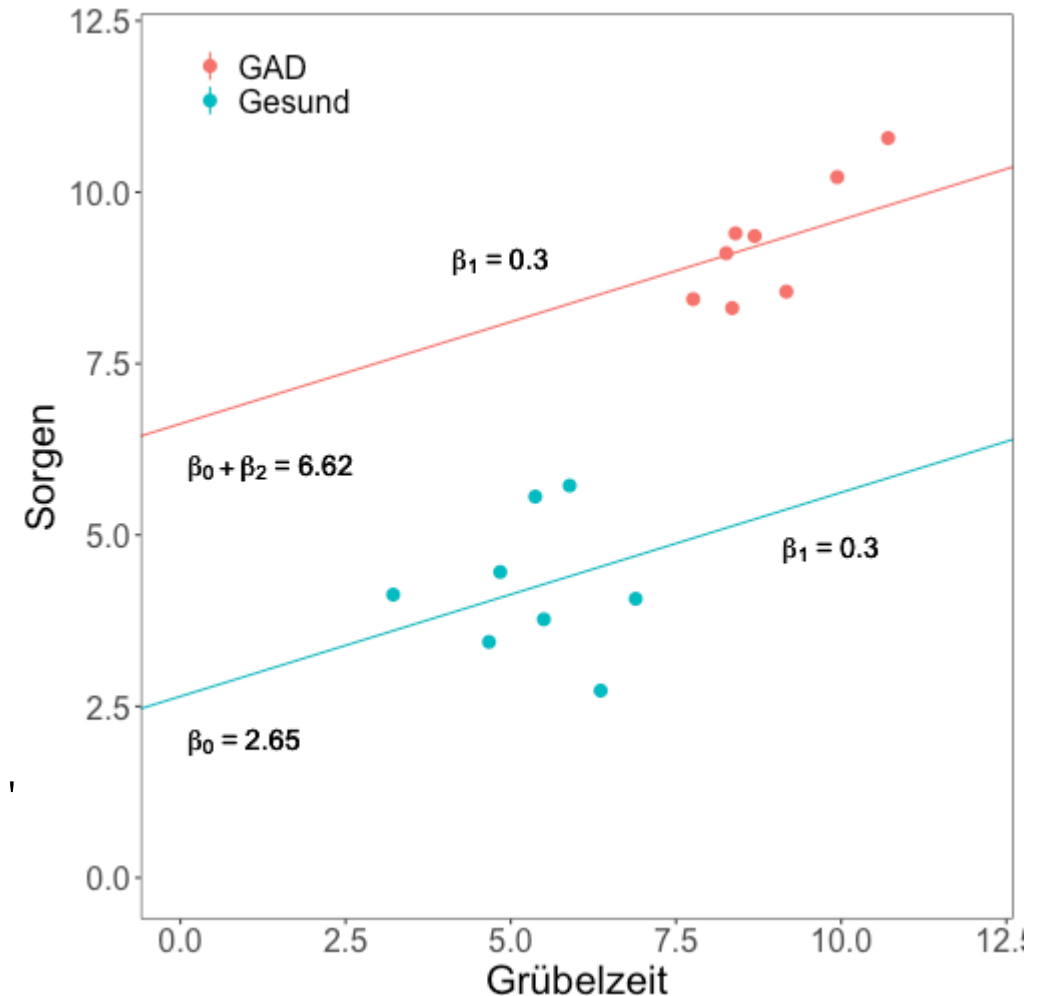
Min	1Q	Median	3Q	Max
-1.80773	-0.60254	0.09192	0.55513	1.32211

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.6455	1.2918	2.048	0.06133 .
Grübelzeit	0.2975	0.2339	1.272	0.22563
Gruppe_dummy	3.9761	0.9544	4.166	0.00111 **

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9268 on 13 degrees of freedom  
Multiple R-squared: 0.9021, Adjusted R-squared: 0.887  
F-statistic: 59.9 on 2 and 13 DF, p-value: 2.753e-07



# Wiederholung: Lineare Regressionsmodelle

## Äquivalente Modelle

- Verschiedene Arten der Zentrierung (bzw. allgemeiner lineare Transformationen) der Prädiktorvariablen resultieren in mathematisch äquivalenten Modellen.
- Die Regressionskoeffizienten unterscheiden sich, jedoch bleiben  $R^2$ , F-Statistik und Residuen unverändert!

```
summary(mod1) # Dummy-codiert
```

Call:  
lm(formula = Sorgen ~ Grübelzeit + Gruppe\_dummy, data = df)

Residuals:

Min	1Q	Median	3Q	Max
-1.80773	-0.60254	0.09192	0.55513	1.32211

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.6455	1.2918	2.048	0.06133 .
Grübelzeit	0.2975	0.2339	1.272	0.22563
Gruppe_dummy	3.9761	0.9544	4.166	0.00111 **

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9268 on 13 degrees of freedom  
Multiple R-squared: 0.9021, Adjusted R-squared: 0.887  
F-statistic: 59.9 on 2 and 13 DF, p-value: 2.753e-07

```
summary(mod2) # Effekt-codiert
```

Call:  
lm(formula = Sorgen ~ Grübelzeit + Gruppe\_effekt, data = df)

Residuals:

Min	1Q	Median	3Q	Max
-1.80773	-0.60254	0.09192	0.55513	1.32211

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.6335	1.6828	2.754	0.01643 *
Grübelzeit	0.2975	0.2339	1.272	0.22563
Gruppe_effekt	1.9880	0.4772	4.166	0.00111 **

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

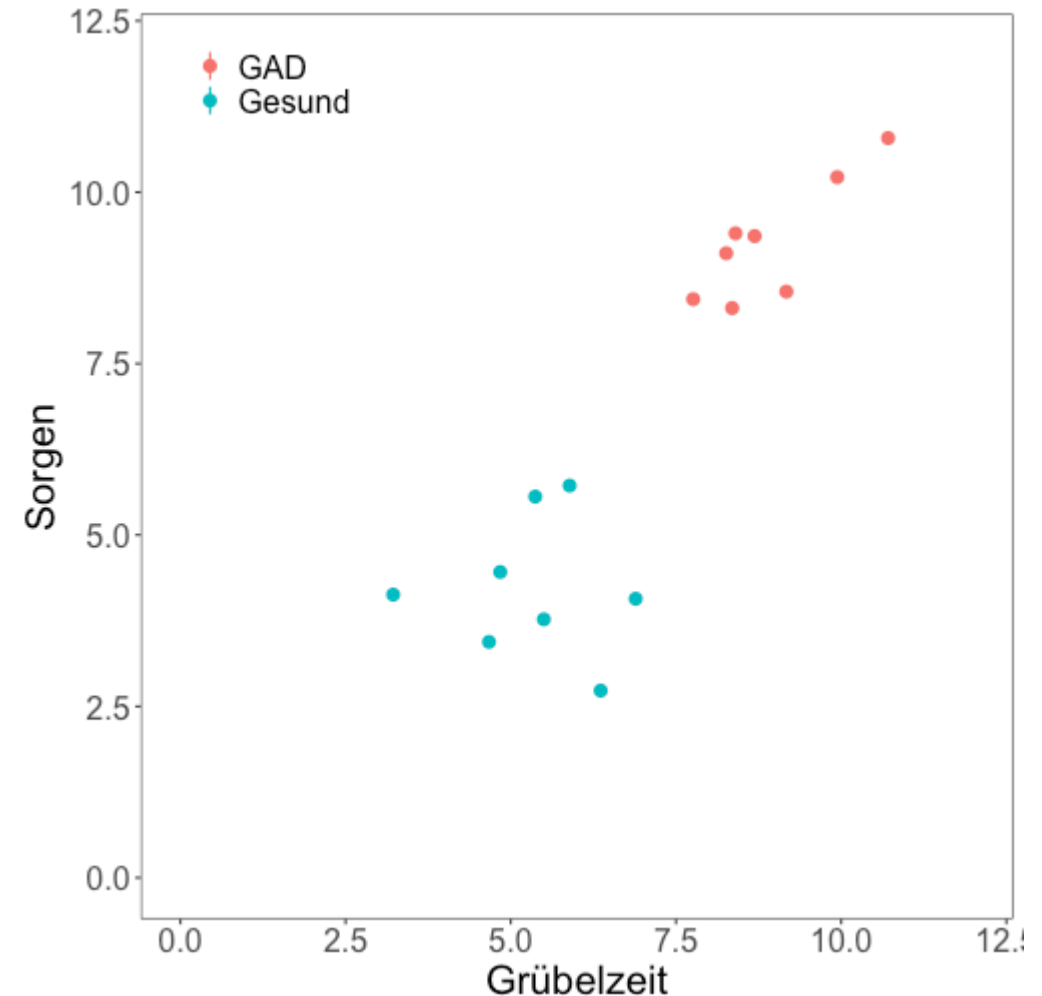
Residual standard error: 0.9268 on 13 degrees of freedom  
Multiple R-squared: 0.9021, Adjusted R-squared: 0.887  
F-statistic: 59.9 on 2 and 13 DF, p-value: 2.753e-07

# Wiederholung: Lineare Regressionsmodelle

## Interaktion zwischen Prädiktoren

- Wechselwirkung zwischen den Prädiktoren (Grübelzeit x Gruppe)
- Frage: Unterschiedlicher Effekt von Prädiktor 1 in den Stufen von Prädiktor 2?
- Statt Annahme einer einheitlichen Steigung wird eine Steigung pro Gruppe geschätzt.
- Unterschied zwischen Steigungen kann auf Signifikanz geprüft werden:

*"Ist der Effekt der Grübelzeit im Durchschnitt unterschiedliche, je nachdem in welcher Gruppe ich bin?"*



# Wiederholung: Lineare Regressionsmodelle

## Interaktion zwischen Prädiktoren

Interaktion zwischen Prädiktoren:

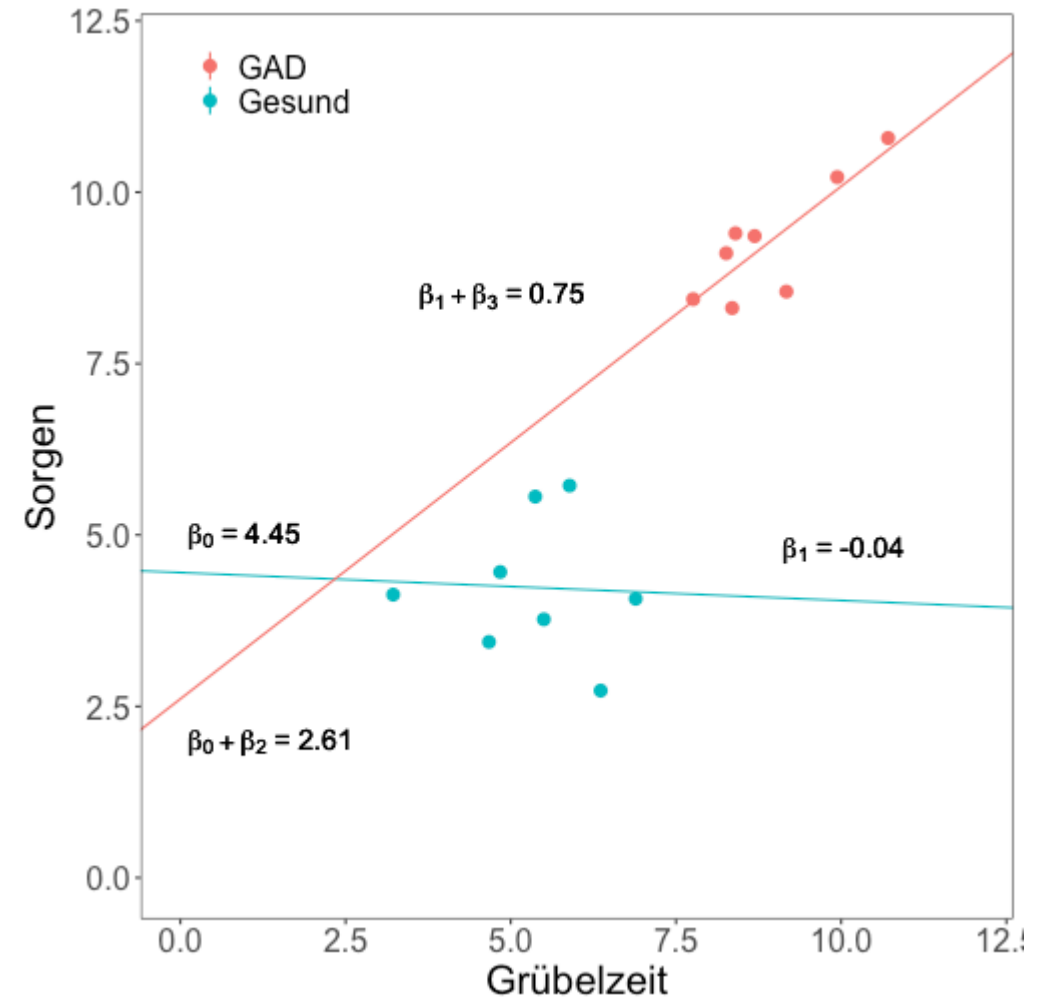
Notation:

$$Sorgen_i = \beta_0 + \beta_1 Gr\ddot{u}belzeit_i + \beta_2 Gruppe_i + \beta_3 Gr\ddot{u}belzeit_i \cdot Gruppe_i + \varepsilon_i$$

In R:

```
mod = lm(Sorgen ~ Gruebelzeit * Gruppe_dummy, data = df)
coef(mod)
```

(Intercept)	Gruebelzeit	Gruppe_dummy	Gruebelzeit:Gruppe_dummy
4.45310622	-0.04082475	-1.84405759	0.78868685



# Wiederholung: Lineare Regressionsmodelle

## Interaktion zwischen Prädiktoren

Interaktion zwischen Prädiktoren:

In R:

```
summary(mod)
```

Call:

```
lm(formula = Sorgen ~ Grübelzeit * Gruppe_dummy, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.46346	-0.47985	0.09944	0.26992	1.50735

Coefficients:

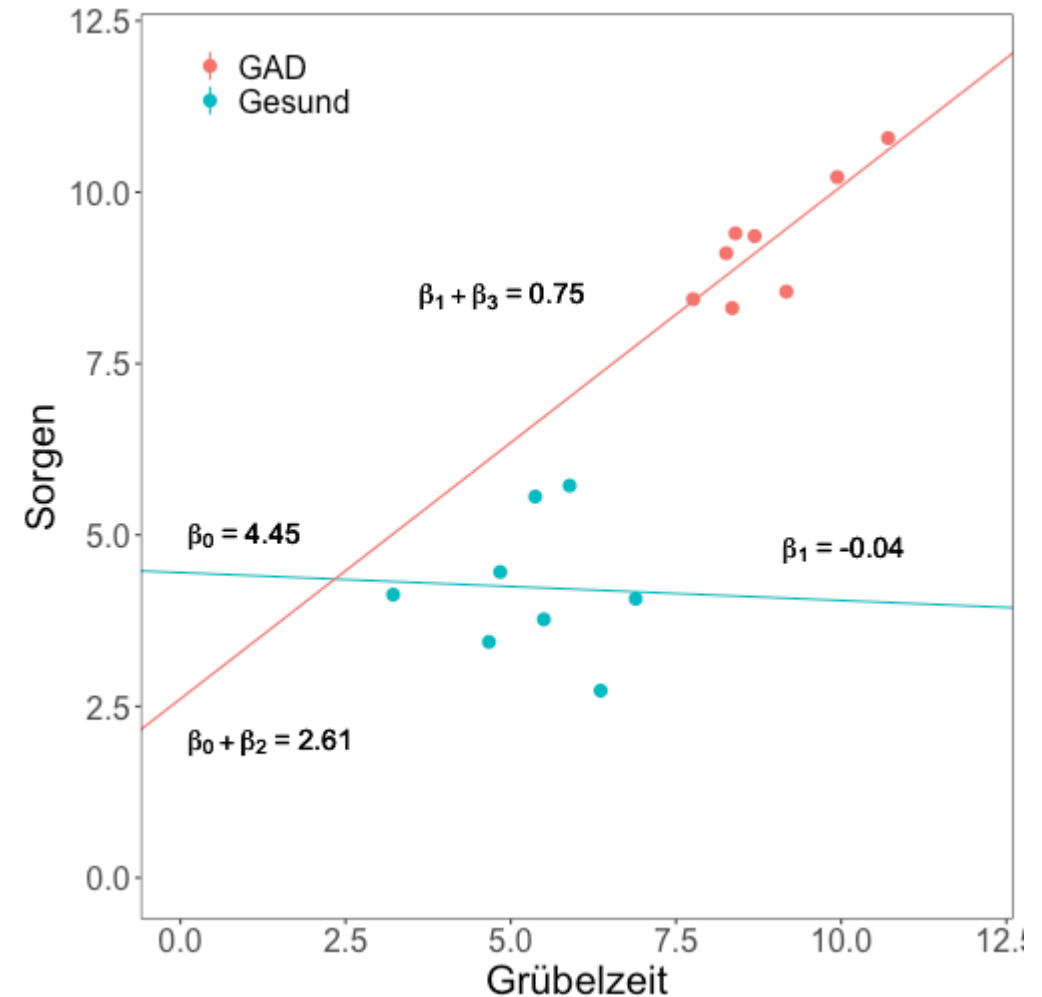
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.45311	1.55531	2.863	0.0143 *
Grübelzeit	-0.04082	0.28557	-0.143	0.8887
Gruppe_dummy	-1.84406	3.33577	-0.553	0.5905
Grübelzeit:Gruppe_dummy	0.78869	0.43600	1.809	0.0956 .

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.855 on 12 degrees of freedom

Multiple R-squared: 0.9231, Adjusted R-squared: 0.9039

F-statistic: 48 on 3 and 12 DF, p-value: 5.869e-07



# Wiederholung: Lineare Regressionsmodelle

## Annahmen der (multiplen) linearen Regression

1. Die Zufallsvariablen hängen linear zusammen
  2. Die Residuen ( $\varepsilon_i$ ) sind unabhängig voneinander
    - Verletzung dieser Annahme in hierarchischen Datenstrukturen
1.  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \rightarrow$  d.h. Die Residuen ( $\varepsilon_i$ )...
    - ...sind normalverteilt mit Erwartungswert 0
    - ...haben konstante Varianz (Homoskedastizität)



# Wiederholung: Lineare Regressionsmodelle

## Annahmen der (multiplen) linearen Regression

Exkurs zur Normalverteilung der Residuen:

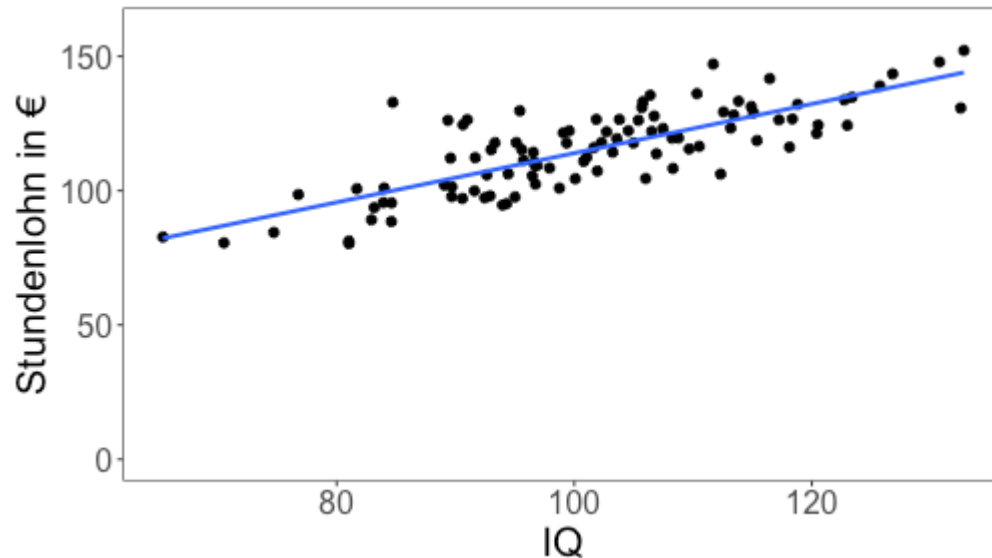
- Modellannahme ist Normalverteilung der Residuen → **nicht** der AV und der Prädiktoren
- CAVE: nicht-normalverteilte AVs/UVs können eher zu nicht-normalverteilten Residuen führen
- daher werden nicht-normalverteilte Variablen oft vor der Modellierung transformiert
- NV-Annahme ist für die Modellparameterschätzung i.d.R. irrelevant (Gelman & Hill, 2007)
- Verletzung der NV-Annahme ist für die Standardfehlerschätzung i.d.R. unkritisch

# Zentrierung von Prädiktoren

- Der Y-Achsenabschnitt beschreibt den durchschnittlichen Wert, wenn alle Prädiktorvariablen 0 sind
- Frage: Ist die 0 ein sinnvoller Wert?

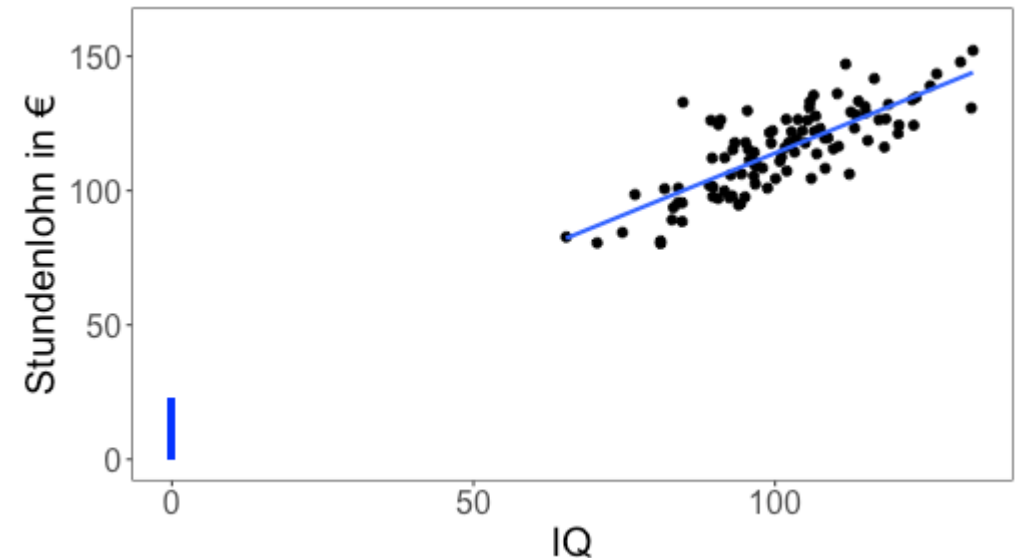
```
fit <- lm(Einkommen ~ IQ, data = data_iq)
coef(fit)
```

```
(Intercept)      IQ
22.4700771    0.9150189
```



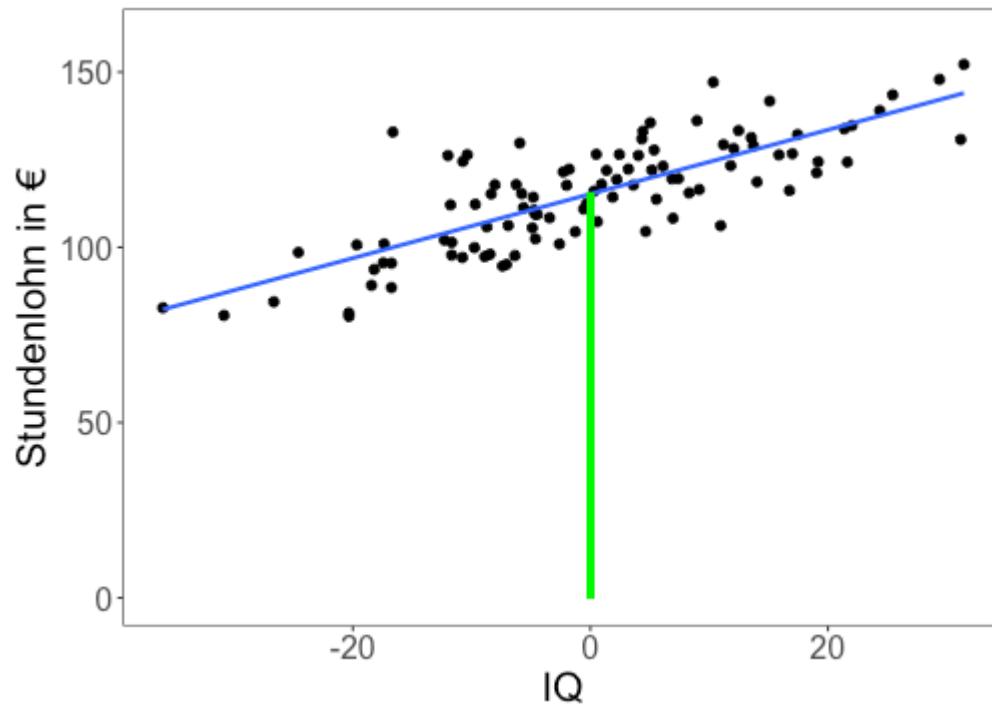
```
# Vorhergesagtes Einkommen bei IQ=0
# Einsetzen
22.4700771 + 0 * 0.9150189
```

```
[1] 22.47008
```



# Zentrierung von Prädiktoren

- Der Y-Achsenabschnitt beschreibt den durchschnittlichen Wert, wenn alle Prädiktorvariablen 0 sind
- Frage: Ist die 0 ein sinnvoller Wert?



- Zentrierung anhand Stichprobenmittelwert des IQ
  - Mittelwert wird berechnet und von jedem IQ-Wert abgezogen
  - neue 0 entspricht Mittelwert
  - Werte  $>0$  sind überdurchschnittlich
  - Werte  $<0$  sind unterdurchschnittlich
- vereinfachte und sinnvollere Interpretation

# Zentrierung von Prädiktoren

- Der Y-Achsenabschnitt beschreibt den durchschnittlichen Wert, wenn alle Prädiktorvariablen 0 sind
- Frage: Ist die 0 ein sinnvoller Wert?
- Beispiele für i.d.R. nicht sinnvolle 0 Werte:
- IQ
- Puls
- Konzentration roter Blutkörperchen
- Faustregel: Prädiktorvariablen immer so zentrieren, dass 0 einen sinnvollen Wert beschreibt.

## Gängige Zentrierungsstrategien:

Prädiktor	Zentrierung
Stetige Prädiktoren	Auf Stichprobenmittelwert zentrieren (0 entspricht Wert eines durchschnittlichen Probanden)
Wenn Normdaten vorliegen	Auf Populationsmittelwert zentrieren
Bei Messwiederholungen	Auf ersten Messzeitpunkt (Baseline) zentrieren
Bei Likert-Skalen	Auf den semantischen Mittelpunkt zentrieren: -3 = trifft gar nicht zu; 0 = unentschieden; +3 = trifft voll zu

# Over- & Underfitting

## Underfitting

- Manchmal repräsentiert das (lineare) Modell die Systematik der Daten unzureichend (in blau dargestellt).

- Beispiele:

- A:  $y = a \cdot b^{1/x}$

- B:  $y = a + b_1 \cdot x + b_2 \cdot x^2$

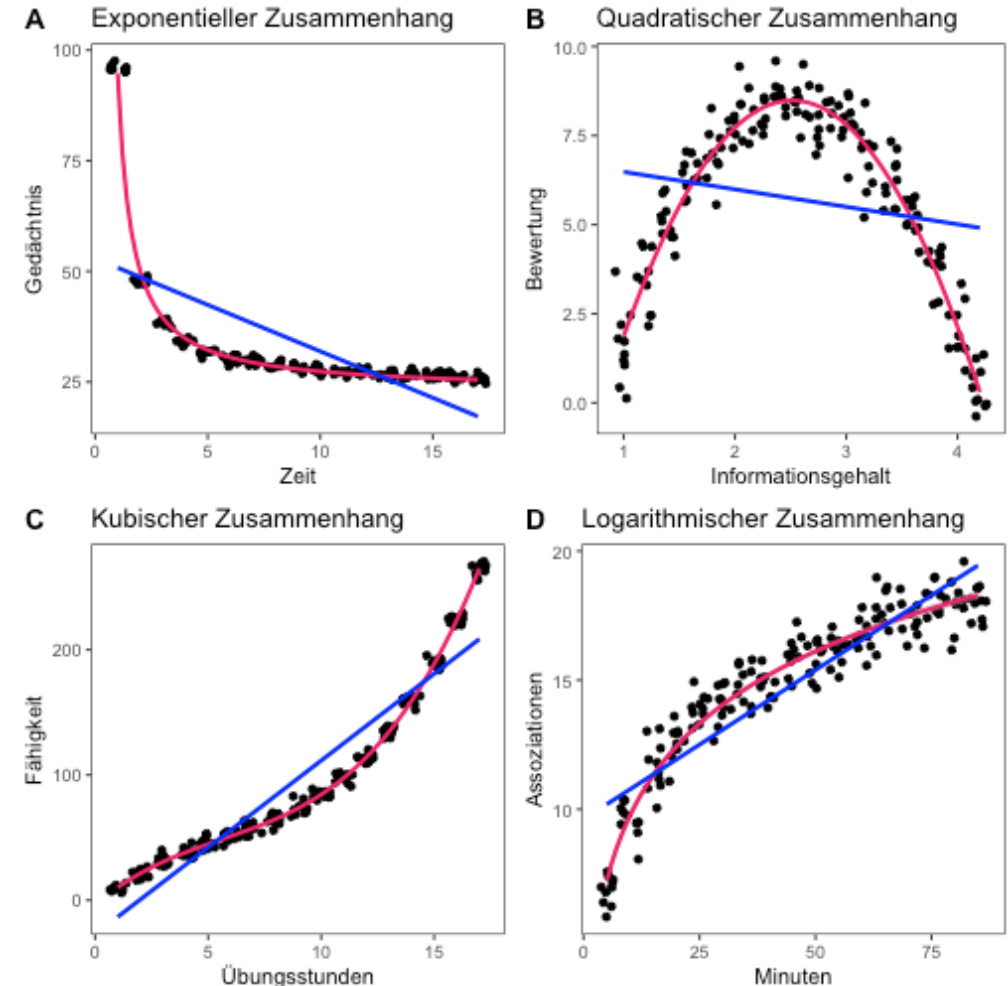
- C:  $y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$

- D:  $y = a + b \cdot \log(x)$

→ Bessere Passung durch Erhöhung der Modellkomplexität

→ Es gibt nach wie vor nur die Variablen  $X$  und  $Y$

→ Lediglich die angenommene (modellierte) Beziehung ändert sich



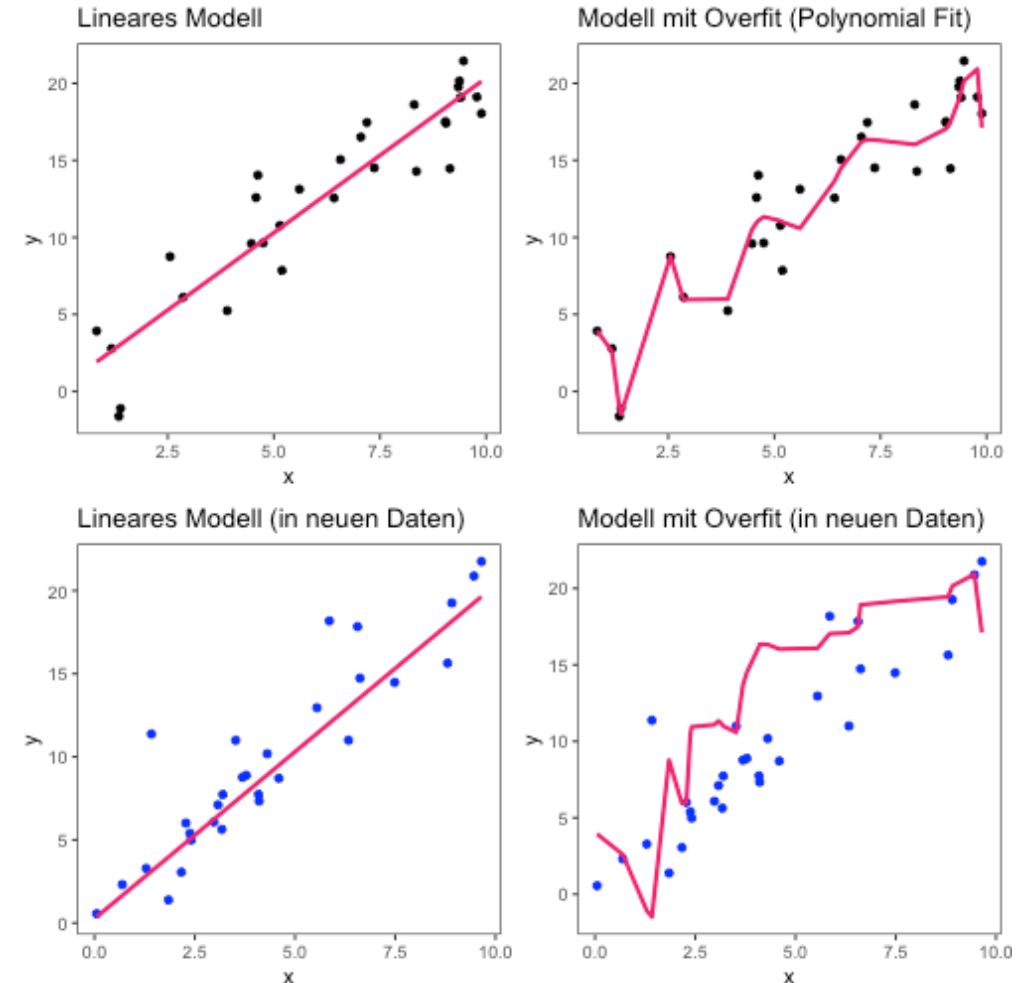
# Over- & Underfitting

## Overfitting

- Manchmal ist das Modell an die Daten "überangepasst".
- Vorteil: Modell bildet Stichprobendaten gut ab („in-sample prediction“) → guter Modellfit/hohes  $R^2$
- Nachteil: Modell lässt sich schlechter auf neue Daten (blau) übertragen (out-of-sample prediction)
- Neue Daten = Daten auf die das Modell nicht angepasst wurde

→ mangelnde Generalisierbarkeit des Modells

- Beispiel: Polynom 12. Grades:
- $$y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + b_4 \cdot x^4 + b_5 \cdot x^5 + b_6 \cdot x^6 + b_7 \cdot x^7 + b_8 \cdot x^8 + b_9 \cdot x^9 + b_{10} \cdot x^{10} + b_{11} \cdot x^{11} + b_{12} \cdot x^{12}$$

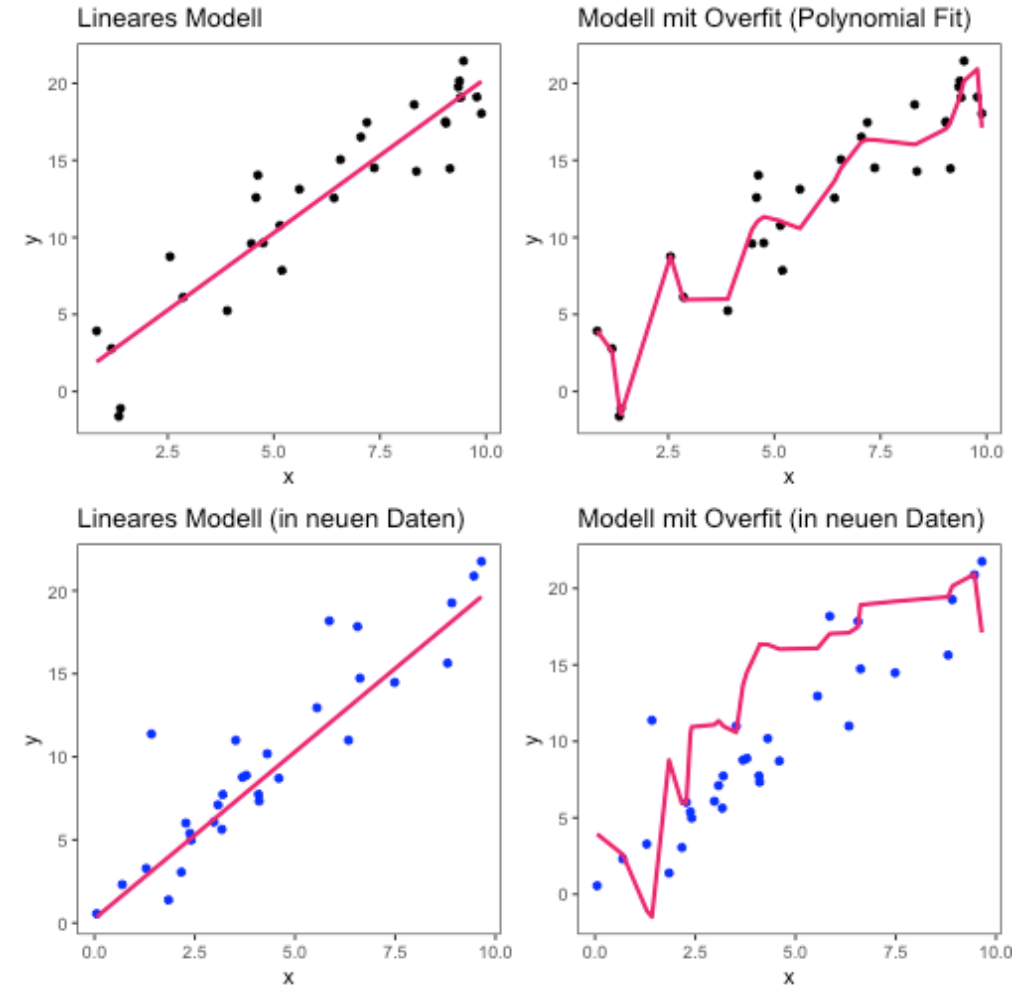


# Over- & Underfitting

## Overfitting

Wie passiert Overfitting?

- Ein komplexes Modell verwendet seine große Anzahl an Parametern, um sich an zufällige Details des Datensatzes anzupassen.
- Dabei wird nicht nur das eigentliche Signal erfasst (systematische Varianz), auch, das zufällige Rauschen
- Da das Rauschen in jedem neuen Datensatz unterschiedlich ist, kann sich die Vorhersagegenauigkeit komplexer Modelle im Vergleich zu einfachen Modellen verschlechtern.



- Allgemeines lineares Modell als flexibles Werkzeug um empirische Daten abzubilden.
- Kann ein Kriterium **verherzusagen** und **Hypothesen testen**.
- Flexibel **anpassbar** (Kriterium) und **erweiterbar** (Prädiktoren).
- Gute Modelle spiegeln Stichproben adäquat wieder (**Modellfit**) und sind gleichzeitig verallgemeinerbar (**Generalisierbarkeit**).
- Teststatistiken richten sich nach der Hypothese (z.B. ganzes Modell → F-Test; einzelner Koeffizient → t-Wert)
- Prädiktoren werden für bessere Interpretierbarkeit oft **zentriert**.
- Valide Interpretationen des Regressionsmodells erfordern, dass die **Voraussetzungen** gelten (z.B. Unabhängigkeit der Fehler).



