BA KW | Vorlesung

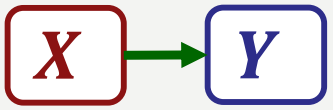
Einführung in die Statistik

Regression

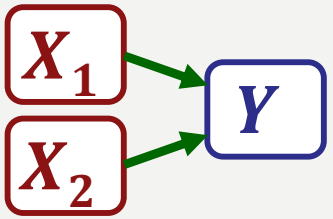
Prof. Thomas Hanitzsch

Lineare Regression: Grundlagen

- Regression im statistischen Sinne meint das „Zurückführen“ einer abhängigen Variable Y auf eine oder mehrere unabhängige Variablen X_1, X_2 etc.
  - Kriterium: vorherzusagende abhängige Variable Y
  - Prädiktor: beeinflussende unabhängige Variable X
- Damit postuliert die Regression einen Zusammenhang, der in eine bestimmte Richtung verläuft
  - Einfache bzw. bivariate Regression: ein Prädiktor



  - Multiple Regression: zwei und mehr Prädiktoren



Einfache lineare Regression

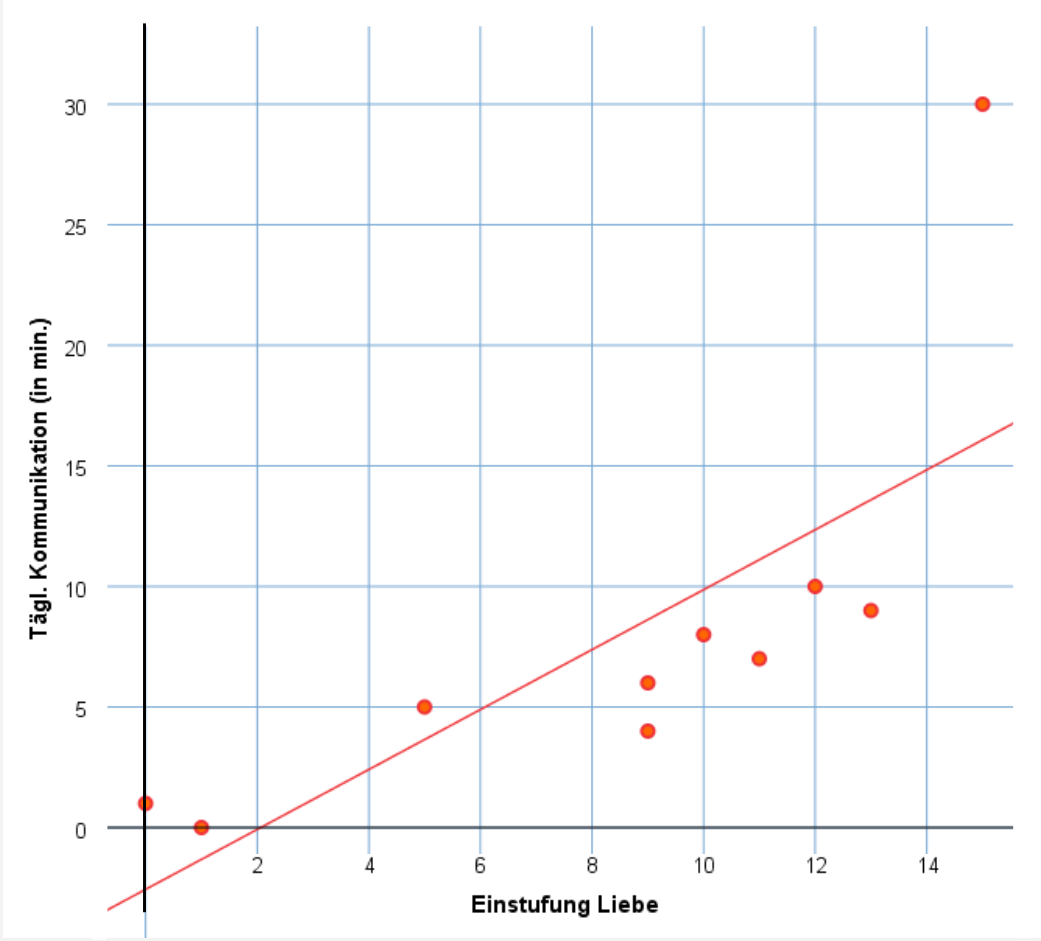
- Ausgangssituation:
  - Zwei Variablen sind metrisch skaliert
  - Das Streudiagramm legt einen linearen Zusammenhang nahe
- Ziel:
  - Auffinden einer linearen Funktion, die die Verteilung der Wertepaare (x_i; y_i → „Punktewolke“) möglichst gut abbildet (Regressionsgerade)

→ Lineare Regressionsfunktion: $y_i = a + b \cdot x_i + e_i$

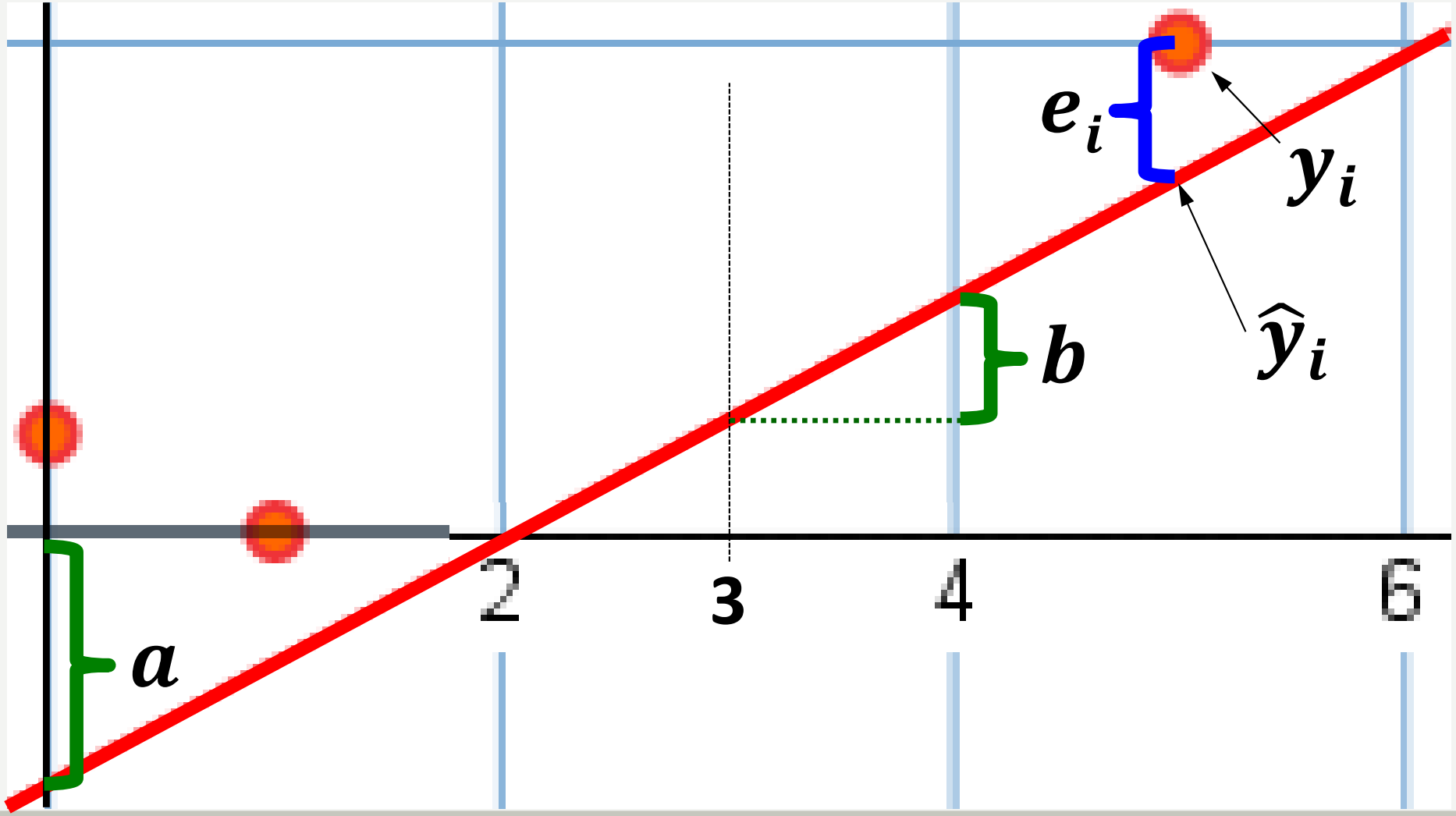
y_i = beobachteter Y-Wert für den X-Wert der i-ten Beobachtung
a = Schnittpunkt mit der Y-Achse (Achsenabschnitt bzw. „Konstante“)
b = Steigung der Regressionsgeraden
e_i = Residuum bzw. Abweichung des beobachteten Werts y_i vom vorhergesagten Wert $\hat{y}_i$ (e=„error“; Fehler)

Einfache lineare Regression: Beispiel

Proband (i)	Punkte auf „Liebe“-Skala (x_i)	Tägliche Kommunikation in Minuten (y_i)
1	9	4
2	15	30
3	13	9
4	10	8
5	0	1
6	9	6
7	5	5
8	12	10
9	1	0
10	11	7
$Mittelw.$	8,5	8,0
s	4,99	8,38



Einfache lineare Regression: Parameter



Methode der kleinsten Quadrate (OLS)

- Regressionsgerade wird so berechnet, dass die Summe der quadrierten Residuen minimiert wird:

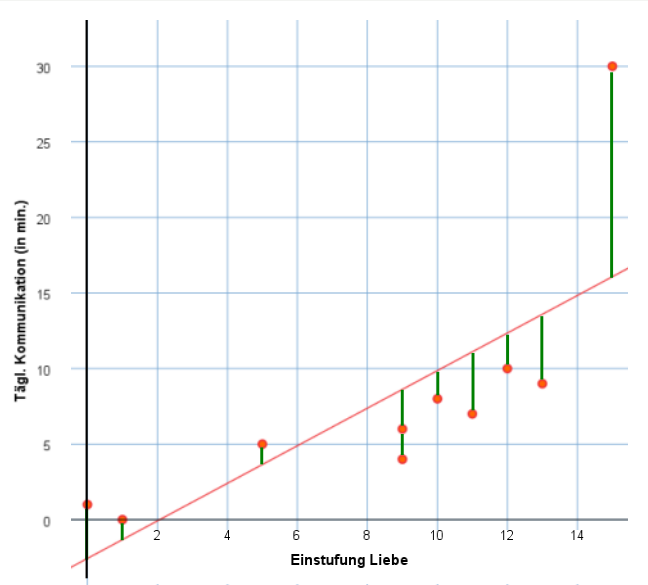
$$\sum_{i=1}^{N} e_i^2 \qquad \text{mit: } e_i = y_i - \hat{y}_i$$

 optimale Anpassung der Geraden an die beobachten Werte

- Berechnung der Parameter:

$$b = \frac{s_{XY}}{s_X^2}$$

$$a = \overline{y} - b \cdot \overline{x}$$



Einfache lineare Regression: Beispiel

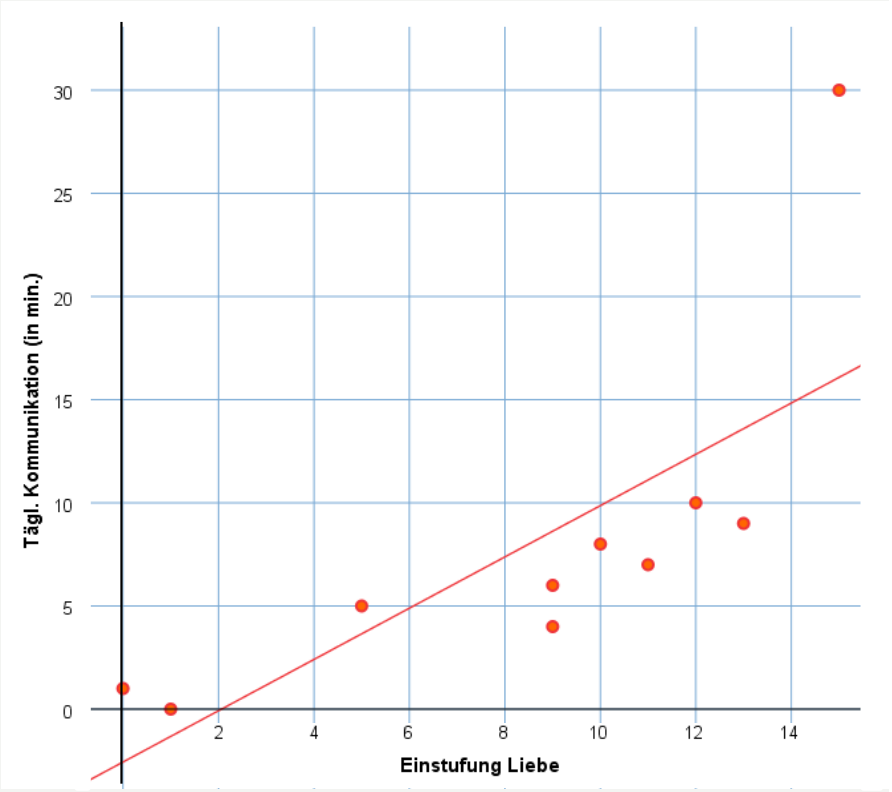
$$s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{279}{9} = 31$$

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{224{,}5}{9} = 24{,}94$$

$$b = \frac{s_{XY}}{s_X^2} = \frac{31}{24{,}94} = \mathbf{1{,}24}$$

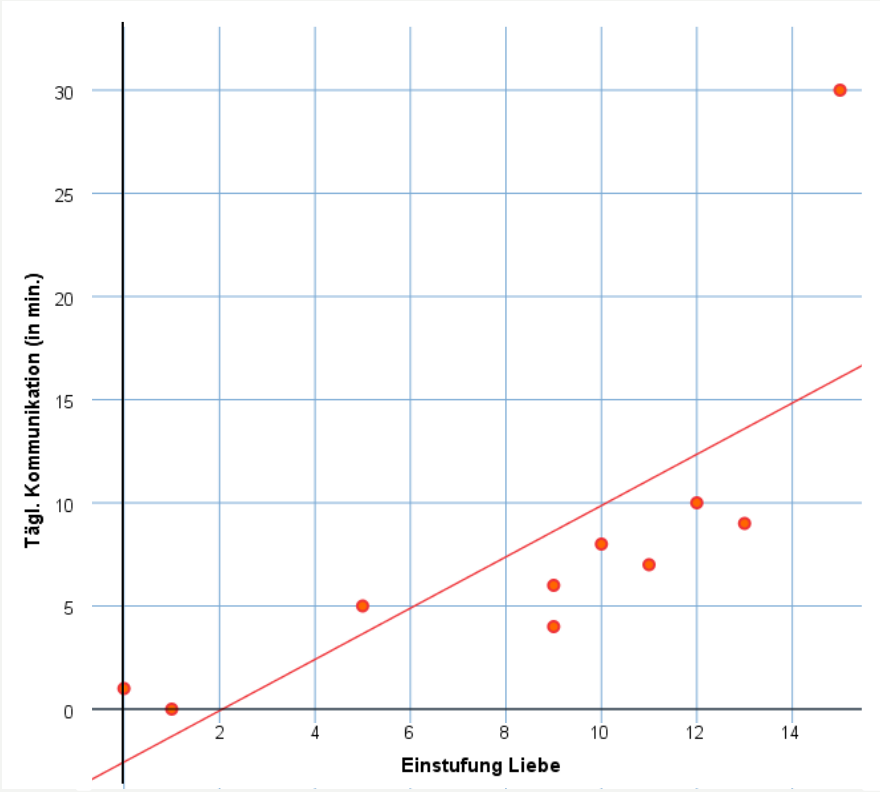
$$a = \bar{y} - b \cdot \bar{x} = 8{,}0 - 1{,}24 \cdot 8{,}5 = \mathbf{-2{,}56}$$

$$\boldsymbol{\hat{y}_i = -2{,}56 + 1{,}24 \cdot x_i}$$



Einfache lineare Regression: Beispiel

- Interpretation: $y_i = -2{,}56 + 1{,}24 \cdot x_i + e_i$
  - Für eine Person mit 0 Punkten auf der „Liebesskala“ wird vorhergesagt, dass sie -2,56 Minuten täglich mit dem/der Partner/in kommuniziert macht hier wenig Sinn…
  - Mit jedem zusätzlichen Punkt $(x_i + 1)$ auf der „Liebesskala“ werden 1,24 Minuten täglich mehr mit dem/der Partner/in kommuniziert



Standardisierung von Regressionskoeffizienten

- Unstandardisierte Regressionskoeffizienten b sind abhängig vom Wertebereich von X und Y
- Wenn Regressionskoeffizienten miteinander verglichen werden sollen (z.B. bei der multiplen Regression), dann empfiehlt sich eine Standardisierung der Koeffizienten
- Berechnung:

$$\beta = b \cdot \frac{s_X}{s_Y}$$

Im Beispiel:

$$\beta = 1{,}24 \cdot \frac{4{,}99}{8{,}38} = \mathbf{0{,}74}$$

(zum Vergleich: $r_{XY} = 0{,}74$)

Varianzzerlegung

Das Prinzip der Varianzzerlegung:

$$y_i = a + b \cdot x_i + e_i$$

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$s_Y^2 \quad = \quad s_{\hat{Y}}^2 \quad + \quad s_e^2$$

s_Y^2 = Varianz der beobachteten Werte des Kriteriums Y

$s_{\hat{Y}}^2$ = Varianz der vorhergesagten Werte $(\hat{y})$, d.h. die durch X „erklärte“ Varianz

s_e^2 = Restvarianz bzw. „unerklärte“ Varianz

Anpassungsgüte (R^2)

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y}$$

Die Anpassungsgüte:

- Wird oft auch als „Bestimmtheitsmaß“ bezeichnet
- Wertebereich: zwischen 0 und 1 bzw. zwischen 0 und 100%
- Bemisst den Anteil der Varianz von Y, der durch einen linearen Zusammenhang zwischen X und Y erklärt werden kann

$\bar{x} = 8{,}5$
$\bar{y} = 8{,}0$
$a = -2{,}56$
$b = 1{,}24$
$s_Y = 8{,}38$

Anpassungsgüte: Beispiel

Proband	Punkte auf „Liebe“-Skala (x_i)	Tägl. Kommunikation in Minuten (y_i)	Tägl. Kommunikation vorhergesagt ($\hat{y}_i = a + b \cdot x_i$)	Abweichung der vorhergesagten Werte $(\hat{y}_i - \bar{y})^2$	Varianz der vorhergesagten Werte ($s^2_{\hat{Y}}$)
1	9	4	8,60	0,36	
2	15	30	16,04	64,64	
3	13	9	13,56	30,91	
4	10	8	9,84	3,39	
5	0	1	-2,56	111,51	
6	9	6	8,60	0,36	
7	5	5	3,64	19,01	
8	12	10	12,32	18,66	
9	1	0	-1,32	86,86	
10	11	7	11,08	9,49	
s^2		**70,22**		**Σ=345,20**	**38,36**

Vorhergesagte Werte: $\hat{y}_i = -2{,}56 + 1{,}24 \cdot x_i$

Anpassungsgüte: $R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y} = \frac{38{,}36}{70{,}22} = \mathbf{0,546}$

Anpassungsgüte (R^2)

Vereinfachte Berechnung:

- Bei der einfachen linearen Regression (und nur dort!!) lässt sich R^2 auch einfacher berechnen
  - über die Formel der **Korrelation**:

$$R^2 = r_{XY}^2 = \left(\frac{s_{XY}}{s_X \cdot s_Y}\right)^2 = \left(\frac{31{,}0}{4{,}99 \cdot 8{,}38}\right)^2 = 0{,}74^2 = 0{,}549$$

  - über den **standardisierten Regressionskoeffizienten**:

$$R^2 = \beta^2 = \left(b \cdot \frac{s_X}{s_Y}\right)^2 = \left(1{,}24 \cdot \frac{4{,}99}{8{,}38}\right)^2 = 0{,}74^2 = 0{,}549$$

Regressionkoeffizient b: Signifikanztest

- **Funktion:**
  - Prüft, ob das empirische Regressionsgewicht b aus einer Grundgesamtheit stammt, in der das wahre Regressionsgewicht gleich Null ist $(b = 0)$.
    ⍰ gilt für zweiseitige Testung!
- **Formulierung der Hypothesen:**
  - Zweiseitiges Problem:

    $H_0: b = 0$ $\qquad$ $H_1: b \neq 0$
  - Regressionskoeffizienten lassen sich bei Vorliegen starker Annahmen auch einseitig testen

Regressionkoeffizient *b*: Signifikanztest

$$t = \frac{b}{SE_b} \quad \text{mit} \quad df = N - 2$$

$$SE_b = \frac{s_e}{\sqrt{QS_X}}$$

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}$$

$$QS_X = \sum_{i=1}^{N} (x_i - \overline{x})^2$$

- Testentscheidung nach festgelegtem Signifikanzniveau:
  - kritischer Wert t_{krit} kann aus der t-Tabelle ausgelesen werden (unter Berücksichtigung von α und df)
  - H_0 wird abgelehnt, wenn $|t| > t_{krit}$

Signifikanztest: Beispiel

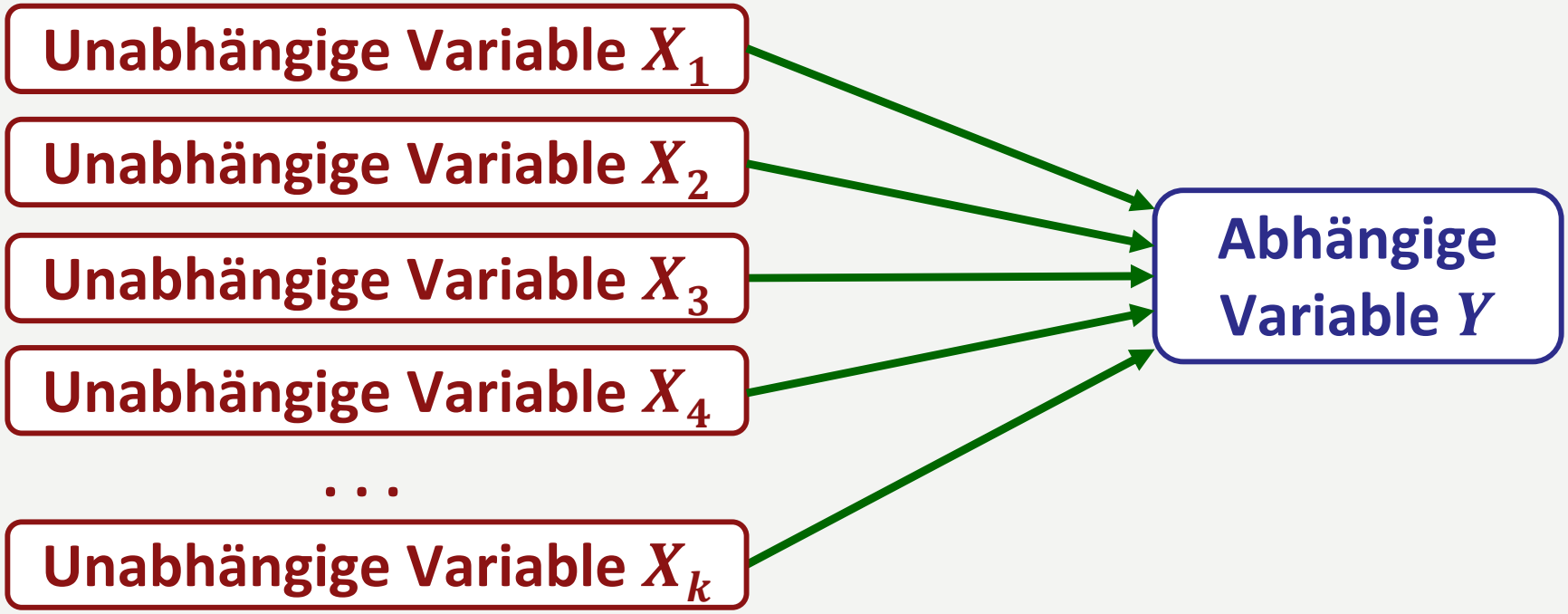
$$SE_b = 0{,}40$$

$$t = \frac{b}{SE_b} = \frac{1{,}24}{0{,}40} = \mathbf{3{,}10} \qquad df = N - 2 = \mathbf{8}$$

- Kritischer Wert bei $\alpha = 0{,}05$:
  - für $df = 8$: $t_{\text{krit}} = 2{,}31$
- Testentscheidung:
  - $|t| = 3{,}10 > t_{\text{krit}}$ ⇒ H_0 wird abgelehnt

Exkurs: Multiple lineare Regression

- Eine abhängige Variable Y (Kriterium) wird auf mehrere unabhängige Variablen X_1, X_2, …, X_k (Prädiktoren) „zurückgeführt“



Multiple lineare Regression

- Ziel:
  - Auffinden einer linearen Funktion, die die Punktewolke möglichst gut abbildet (Regressionsgerade)

    Einfache lineare Regressionsfunktion:

    $$Y = a + b \cdot X + e$$

    → Lineare Regressionsfunktion für *k* Prädiktoren:

    $$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ \ldots + b_k \cdot X_k + e$$
  - Anhand der standardisierten Regressionskoeffizienten (β, „Betas“) lässt sich sagen, welche Prädiktoren besser und welche schlechter für eine Vorhersage geeignet sind

Multiple lineare Regression

- Gestufter Einschluss von Prädiktoren:
  - Hierarchisch:
    - einzelne Prädiktoren oder Blöcke von Prädiktoren gehen in einer vorab definierten Reihenfolge in die Berechnung ein
    - damit lässt sich prüfen, welchen zusätzlichen Gewinn man durch den Einschluss bestimmter Prädiktoren erzielt (über die Statistik „Änderung in R^2“)
  - Schrittweise:
    - anhand statistischer Kriterien werden Prädiktoren in die Regression aufgenommen oder entfernt, um die erklärte Varianz zu maximieren
    - Entscheidung über Inklusion und Exklusion von Prädiktoren wird damit der Software überlassen

Multiple lineare Regression: Beispiel

- Ein Forscherteam interessiert sich für die Faktoren, die das Vertrauen der deutschen Bevölkerung in die Medien beeinflussen. Das Team nimmt an, dass das Vertrauen maßgeblich von folgenden Faktoren abhängt:
  - Soziodemografische Merkmale: Alter, Bildung und Einkommen
  - Regionale Herkunft („Ost“ vs. „West“)
  - Soziales Vertrauen
  - Kulturelle Werte
  - Zufriedenheit mit der Demokratie in Deutschland
  - Politisches Vertrauen
- Mit den deutschen Daten des European Social Survey führt das Forscherteam eine hierarchische Regressionsanalyse durch (*N*=2034)

Multiple lineare Regression: Beispiel

MODELLZUSAMMENFASSUNG 1

```
Call:
lm(formula = Medienvertrauen ~ Alter + Bildung + Einkommen, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59088 -0.41414  0.02808  0.54688  1.77013

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1393047  0.0505882  22.521  < 2e-16 ***
Alter        0.0035679  0.0008081   4.415 1.06e-05 ***
Bildung     -0.0746876  0.0269063  -2.776  0.00556 **
Einkommen    0.0192074  0.0067453   2.848  0.00445 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6429 on 2030 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.01632,   Adjusted R-squared:  0.01487
F-statistic: 11.23 on 3 and 2030 DF,  p-value: 2.635e-07
```

MODELLZUSAMMENFASSUNG 2

```
lm(formula = Medienvertrauen ~ Alter + Bildung + Einkommen +
    Region, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61553 -0.41281  0.02502  0.55398  1.78677

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1643987  0.0526189  22.129  < 2e-16 ***
Alter        0.0036159  0.0008082   4.474  8.1e-06 ***
Bildung     -0.0741601  0.0268950  -2.757  0.00588 **
Einkommen    0.0185015  0.0067545   2.739  0.00621 **
Region      -0.0492218  0.0285639  -1.723  0.08500 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6426 on 2029 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.01776,   Adjusted R-squared:  0.01582
F-statistic: 9.171 on 4 and 2029 DF,  p-value: 2.425e-07
```

Multiple lineare Regression: Beispiel

MODELLZUSAMMENFASSUNG 3

```
Call:
lm(formula = Medienvertrauen ~ Alter + Bildung + Einkommen +
    Region + SozVertrauen + Postmaterial, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68134 -0.40903 -0.01946  0.51661  1.74631

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2993349  0.0601998  21.584  < 2e-16 ***
Alter        0.0035202  0.0007962   4.421 1.03e-05 ***
Bildung     -0.0588922  0.0266001  -2.214   0.0269 *
Einkommen    0.0138175  0.0066814   2.068   0.0388 *
Region      -0.0534930  0.0280768  -1.905   0.0569 .
SozVertrauen 0.1613458  0.0243368   6.630 4.30e-11 ***
Postmaterial -0.0649336  0.0104629  -6.206 6.57e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6308 on 2027 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.05421,   Adjusted R-squared:  0.05142
F-statistic: 19.37 on 6 and 2027 DF,  p-value: < 2.2e-16
```

MODELLZUSAMMENFASSUNG 4

```
Call:
lm(formula = Medienvertrauen ~ Alter + Bildung + Einkommen +
    Region + SozVertrauen + Postmaterial + Demokratie + PolitVertrauen,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.78205 -0.37485 -0.03353  0.38961  1.80014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7049552  0.0745151   9.461  < 2e-16 ***
Alter        0.0025274  0.0007615   3.319  0.00092 ***
Bildung     -0.0825037  0.0260077  -3.172  0.00154 **
Einkommen    0.0022204  0.0065154   0.341  0.73329
Region      -0.0235966  0.0269217  -0.876  0.38087
SozVertrauen 0.0982444  0.0234010   4.198 2.81e-05 ***
Postmaterial -0.0470819  0.0102108  -4.611 4.27e-06 ***
Demokratie   0.0382182  0.0073080   5.230 1.88e-07 ***
PolitVertrauen 0.3043215  0.0207727  14.650  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5867 on 1928 degrees of freedom
  (109 observations deleted due to missingness)
Multiple R-squared:  0.1925,    Adjusted R-squared:  0.1891
F-statistic: 57.44 on 8 and 1928 DF,  p-value: < 2.2e-16
```

Multiple lineare Regression: Beispiel

KOEFIZZIENTEN: MODELL 1

	Variable	B	StdErr	beta	t	p
1	(Intercept)	1.139304696	0.0505882342	*NA*	22.521140	1.927647e-100
2	Alter	0.003567902	0.0008081133	0.09734761	4.415101	1.062657e-05
3	Bildung	-0.074687558	0.0269062853	-0.06339520	-2.775841	5.556423e-03
4	Einkommen	0.019207400	0.0067453244	0.06511479	2.847513	4.450554e-03

KOEFIZZIENTEN: MODELL 2

	Variable	B	StdErr	beta	t	p
1	(Intercept)	1.164398734	0.0526189232	*NA*	22.128897	2.234662e-97
2	Alter	0.003615923	0.0008082022	0.09865783	4.474033	8.100614e-06
3	Bildung	-0.074160087	0.0268949848	-0.06294748	-2.757395	5.878595e-03
4	Einkommen	0.018501507	0.0067544878	0.06272175	2.739143	6.213840e-03
5	Region	-0.049221775	0.0285639478	-0.03800346	-1.723213	8.500231e-02

Multiple lineare Regression: Beispiel

KOEFIZZIENTEN: MODELL 3

	Variable	B	StdErr	beta	t	p
1	(Intercept)	1.299334860	0.0601998347	*NA*	21.583695	3.592915e-93
2	Alter	0.003520209	0.0007962177	0.09604635	4.421164	1.033622e-05
3	Bildung	-0.058892191	0.0266001447	-0.04998801	-2.213980	2.694107e-02
4	Einkommen	0.013817500	0.0066814143	0.04684255	2.068050	3.876190e-02
5	Region	-0.053493013	0.0280767674	-0.04130122	-1.905241	5.689011e-02
6	SozVertrauen	0.161345830	0.0243368315	0.14495229	6.629697	4.299746e-11
7	Postmaterial	-0.064933615	0.0104628809	-0.13591725	-6.206093	6.568300e-10

KOEFIZZIENTEN: MODELL 4

	Variable	B	StdErr	beta	t	p
1	(Intercept)	0.704955190	0.0745151084	*NA*	9.4605672	8.580998e-21
2	Alter	0.002527350	0.0007614663	0.068500363	3.3190575	9.200391e-04
3	Bildung	-0.082503743	0.0260077179	-0.067775756	-3.1722792	1.536284e-03
4	Einkommen	0.002220434	0.0065154079	0.007318957	0.3407975	7.332932e-01
5	Region	-0.023596640	0.0269217427	-0.018111753	-0.8764901	3.808729e-01
6	SozVertrauen	0.098244438	0.0234010265	0.087958129	4.1982961	2.811690e-05
7	Postmaterial	-0.047081903	0.0102107739	-0.096133468	-4.6110024	4.270371e-06
8	Demokratie	0.038218236	0.0073080264	0.116901859	5.2296249	1.882863e-07
9	PolitVertrauen	0.304321498	0.0207727168	0.326600953	14.6500576	3.699942e-46

Multiple lineare Regression

Voraussetzungen:

- Es liegt ein linearer Zusammenhang vor Prüfung der Streudiagramme
- Alle relevanten Variablen sind berücksichtigt
- Die Werte der Kriteriumsvariable sind unabhängig voneinander
- Die Residuen (d.h. die Fehler der Vorhersage) sind normalverteilt
- Die Residuen für die einzelnen Beobachtungen sind unkorreliert
- Die Varianz der Residuen ist unabhängig von den x-Werten (Homoskedastizität)
- Es liegt keine bzw. nur eine geringe Multikollinearität vor

- Zudem: Ausreißer können die Schätzung extrem verzerren!

Aufgabe 1

a) Pierre (siehe Übungsblatt 8, Aufgabe 1) nimmt nun an, dass die TV-Nutzungsdauer einen linearen Einfluss auf das Vertrauen in andere Menschen hat. Welchen Wert für das Vertrauen würden wir für Personen prognostizieren, die täglich 6 Stunden fernsehen?

ID	TV-Nutzung	Vertrauen
1	1	5
2	2	6
3	3	4

b) Wie gut kann das Regressionsmodell das interpersonale Vertrauen erklären?

Aufgabe 1

c) Können wir annehmen, dass ein Einfluss der TV-Nutzungsdauer auch in der Grundgesamtheit vorhanden ist? (Signifikanzniveau $\alpha = 0{,}05$. Die Residuen sind normalverteilt und unkorreliert, ihre Varianz ist unabhängig von X. Ausreißer liegen nicht vor.)

d) Pierres Cousine Pamela führt dieselbe Studie mit einer anderen Stichprobe durch, erfasst das Vertrauen aber anders als Pierre auf einer 7-stufigen Ratingskala. Sie ermittelt eine Kovarianz von $s_{XY} = -\ 0{,}3$, sowie Standardabweichungen für das Vertrauen von $s_Y = 0{,}75$ und für die TV-Nutzungsdauer erneut von $s_X = 1{,}0$. Hat Pamela oder hat Pierre einen stärkeren Einfluss der TV-Nutzungsdauer (für ihre Stichprobe!) ermittelt?