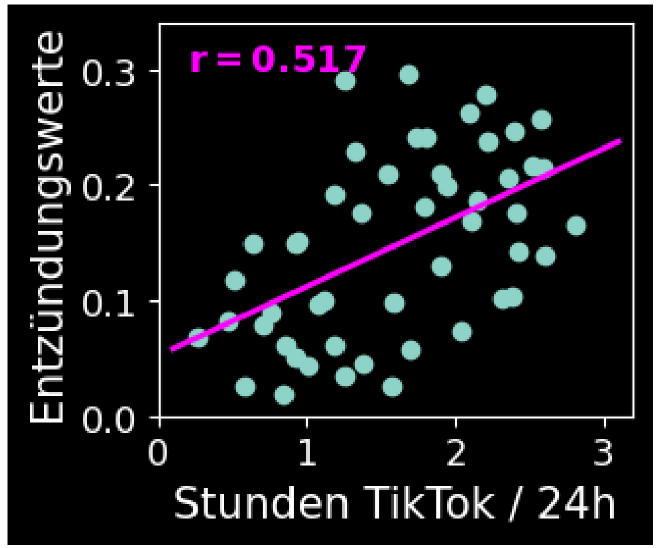
M24 Statistik 1: Wintersemester 23/24

Vorlesung 06: Regression

Prof. Matthias Guggenmos

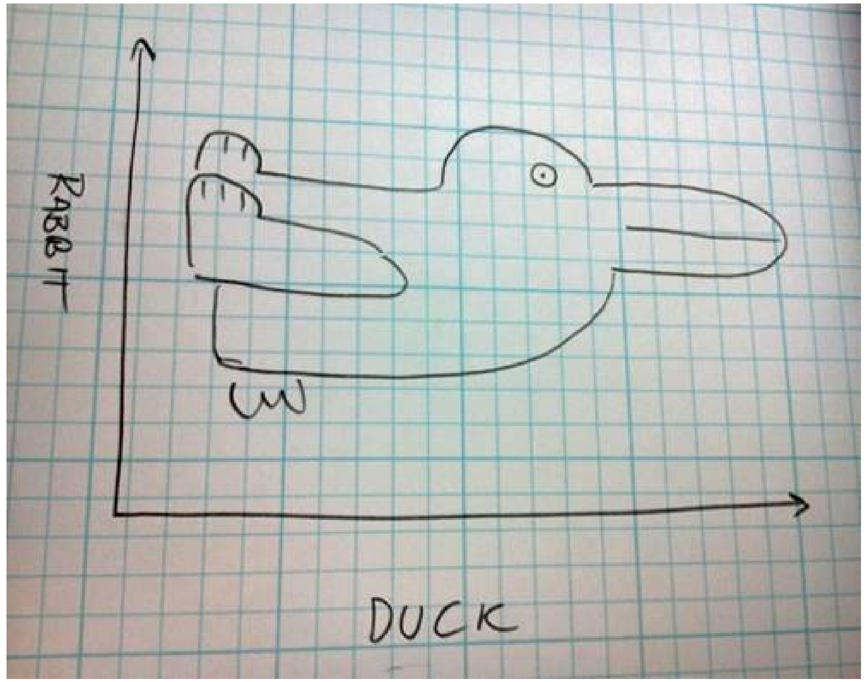
Health and Medical University Potsdam

Kurze Erinnerung: beim letzten Mal fanden wir einen Zusammenhang von TikTok-Online-Zeit und Entzündungsparametern:



Bei der Interpretation stellt sich einerseits die Kausalitätsfrage, andererseits, wie stark der Zusammenhang tatsächlich ist. Da die Pearson-Korrelation lediglich den **Grad der Linearität** beurteilt, fragen Sie sich: um wie viel erhöhen sich die Entzündungsparameter pro Stunde zusätzliche Zeit auf TikTok? Oder umgekehrt: um wie viel erhöht sich die Zeit auf TikTok, wenn die Entzündungswerte um einen Wert x ansteigen?

Regression



Bildnachweis[1]

Der Forschungsprozess

Auf Basis einer neuen/aktualisierten Theorie ergeben sich häufig neue Fragen und Hypothesen

Wundern, Raten, Fragen
(best case: bereits eine Theorie)

↓

Hypothese

↓

Operationalisierung

↓

Experiment / Studie

↓

Daten

↓

Auswertung / Ergebnis

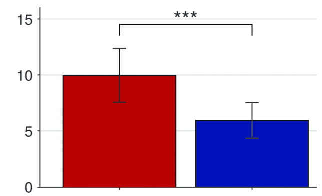
↓

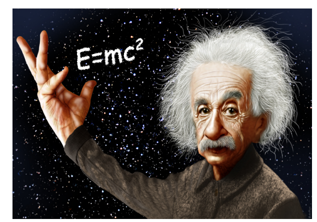
Schlussfolgerung / Interpretation

↓

Aufstellen oder Update einer Theorie

Player	Minutes	Points	Rebounds
A	41	20	6
B	30	29	7
C	22	7	7
D	26	3	3
E	20	19	8





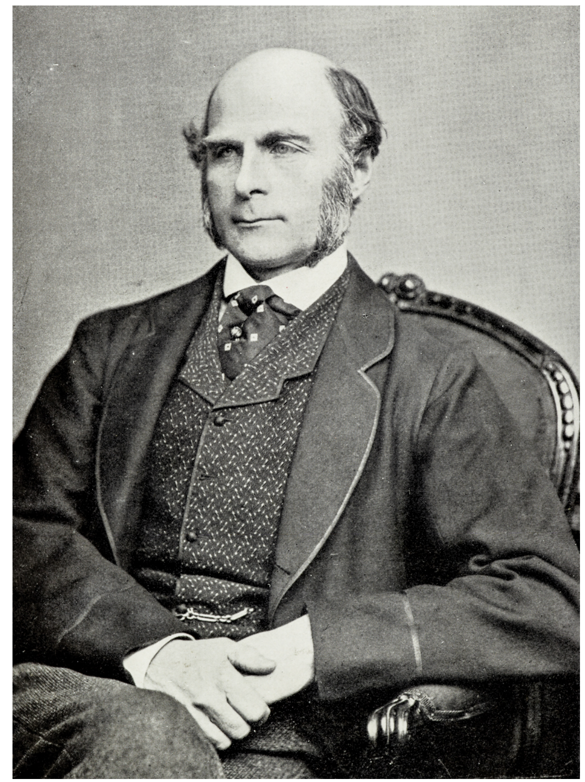
Woher kommt der Ausdruck “Regression”?

- Lateinisch »regredi« = „umkehren, zurückgehen“
- Psychoanalyse: Regression = Zurückfallen in kindliche Verhaltensmuster

> Wir heißen es **Regression**, wenn sich im Traum die Vorstellung in das sinnliche Bild zurückverwandelt, aus dem sie irgend einmal hervorgegangen ist.

Sigmund Freud (1900). “Traumdeutung”.

- In die Statistik wird der Ausdruck “Regression” klassischerweise auf **Francis Galton** (Cousin von Charles Darwin) attribuiert, der bereits 1885 ein Phänomen beschrieb, das er *regression toward mediocrity* (**Regression zur Mitte**) taufte
- Das Phänomen bestand darin, dass Nachfahren großer Eltern dazu tendieren, selbst nur durchschnittlich groß zu werden
- Neuere Forschung zeigt allerdings, dass sich Galton selbst wohl noch nicht des statistischen Ursprungs dieses Phänomens bewusst war und eine biologische Erklärung favorisierte[2].



Regression

- Dem Wortsinn nach ist Ziel der **Regression** eine abhängige Variable auf eine oder mehrere unabhängige Variablen zurückzuführen (auf diese zu *regredieren*)
- Eingängiger ist aber die umgekehrte Formulierung: Ziel der Regression ist es, auf Basis der unabhängigen Variablen die eine abhängige Variable **vorherzusagen** oder **zu erklären**
  - **Unabhängige Variable(n) = vorher*sagende* oder erklär*ende* Variable(n)** ("Ursache")
  - **Abhängige Variable = vorher*gesagte* oder erklär*ete* Variable** ("Auswirkung")

Schnittpunkt der Geraden mit der y-Achse (Ordinatenabschnitt, Intercept)

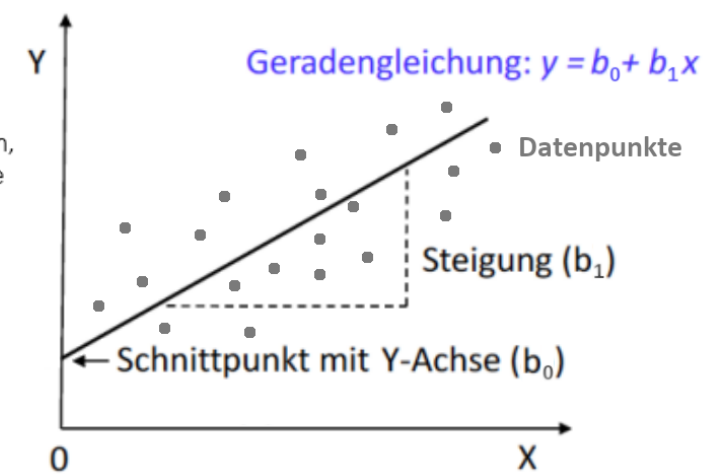
vorhergesagter Wert (Schätzung, d.h. kann vom echten Wert abweichen)

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

Wert der Variablen, die zur Vorhersage genutzt wird

Steigung der Geraden (Regressionskoeffizient, Regressionsgewicht)

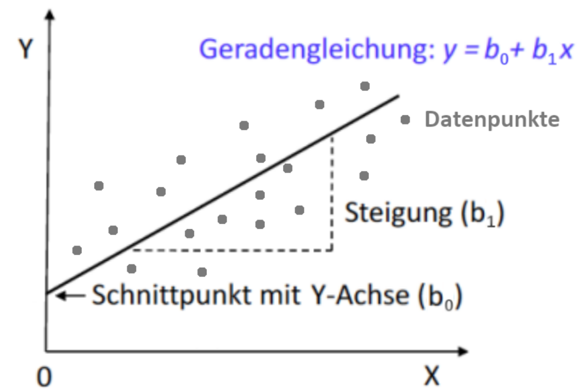
Das „Dach" (^) kennzeichnet vorhergesagte Werte

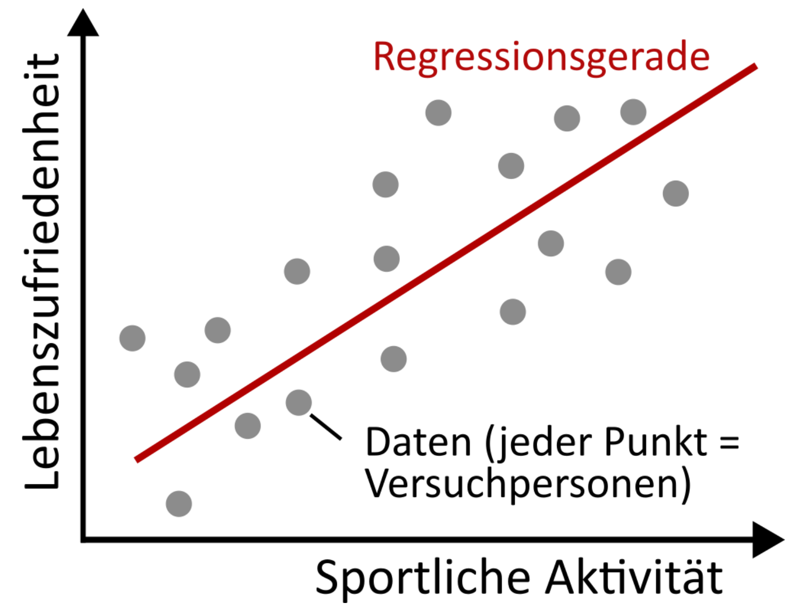


- Beispiel: Studie untersucht Zusammenhang von Lebenszufriedenheit und sportlicher Aktivität
  - **Lebenszufriedenheit:** unabhängige/vorhersagende/erklärende Variable
  - **Sportliche Aktivität:** abhängige/vorhergesagte/erklärte Variable

Regression

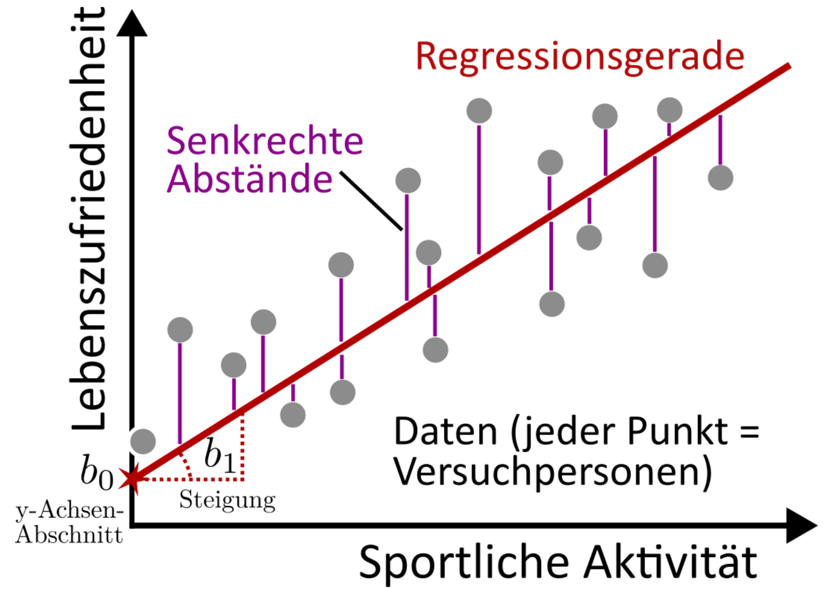
- Im Gegensatz zur Korrelation bestimmt die Regression nicht die Linearität des Zusammenhangs (vielmehr wird dies vorausgesetzt), sondern die **Steigung** des Zusammenhangs
- Aus diesem Grund ist die Regression (wieder im Gegensatz zur Korrelation) nicht symmetrisch – die Steigung ist abhängig davon welche Variable als abhängig und unabhängig deklariert wird.
  - Wie wir noch sehen werden, ist es auch nicht gestattet, die Regressionsgleichung zu invertieren ($x_i = \frac{1}{b_1}\hat{y}_i - \frac{b_0}{b_1}$) — im Allgemeinen ist $\frac{1}{b_1}$ *nicht* die Steigung, wenn die Rollen von X und Y vertauscht werden.
- Die Vorhersage/Erklärung von X durch Y geschieht durch eine Gleichung – die **Regressionsgleichung** – die im Streudiagramm als Gerade eingezeichnet werden kann.





Bestimmung der Regressionsgerade: Methode der kleinsten Quadrate

- Ziel der Regression ist es, die Gerade zu finden, die die Datenpunkte möglichst gut abbildet — es gibt jedoch verschiedene Definitionen dessen, was "möglichst gut" heißt
- Die häufigste Variante ist die **Methode der kleinsten Quadrate**, bei der die Gerade so gewählt wird, dass die Summe der quadrierten **senkrechten Abstände** jedes Datenpunktes zur Geraden minimal ist
  - Engl. *ordinary least square*

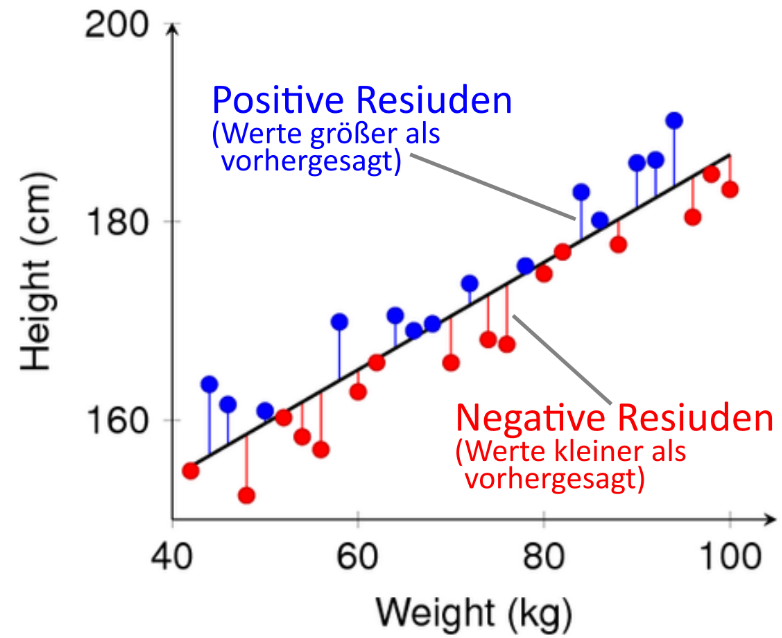


- Die **einfache Regression** mit nur einer unabhängigen Variablen hat zwei freie Parameter, um die Gerade an die Datenpunkte anzupassen (zu "fitten"):
  - **y-Achsenabschnitt** b_0 (engl. *intercept*)
  - **Steigung** b_1 (engl. *slope*)
- Die senkrechten Abstände der Datenpunkte von der gefitteten Geraden werden **Residuen** genannt.
- Exakt 0 wären die senkrechten Abstände nur, wenn alle Punkte auf einer perfekten Gerade liegen.

Warum weichen die Datenpunkte überhaupt von einer Geraden ab?

Verschiedene Gründe:

- Variablen korrelieren überhaupt nicht
- Einfluss von Störvariablen
- Messungenauigkeit



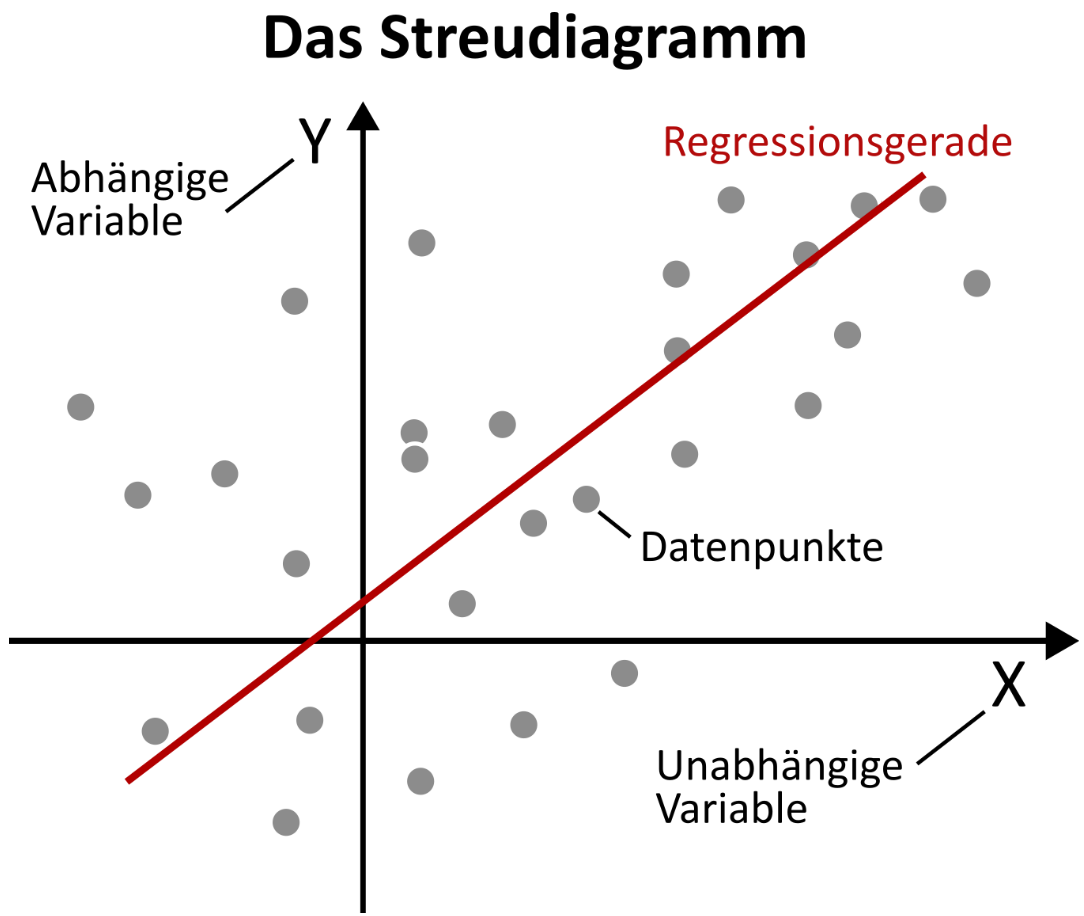
In der Psychologie gibt es (bis auf triviale Fälle) keine perfekten linearen Zusammenhänge, d.h. es verbleiben immer **Residuen** $\Delta\hat{y}_i$:

$$\text{Residuum:}\quad \Delta\hat{y}_i = \hat{\epsilon}_i = \hat{y}_i - y_i$$

Residuum = Differenz von vorhergesagtem Wert $\hat{y}_i$ und tatsächlichem Wert y_i

Streudiagramm bei Regression

- in der Regel UV auf der x-Achse und AV auf der y-Achse



https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_all.html?locale=de

Totale, erklärte und Residuenquadratsumme

- Die Methode der kleinsten Quadrate minimiert die **Residuenquadratsumme (SQR)**:

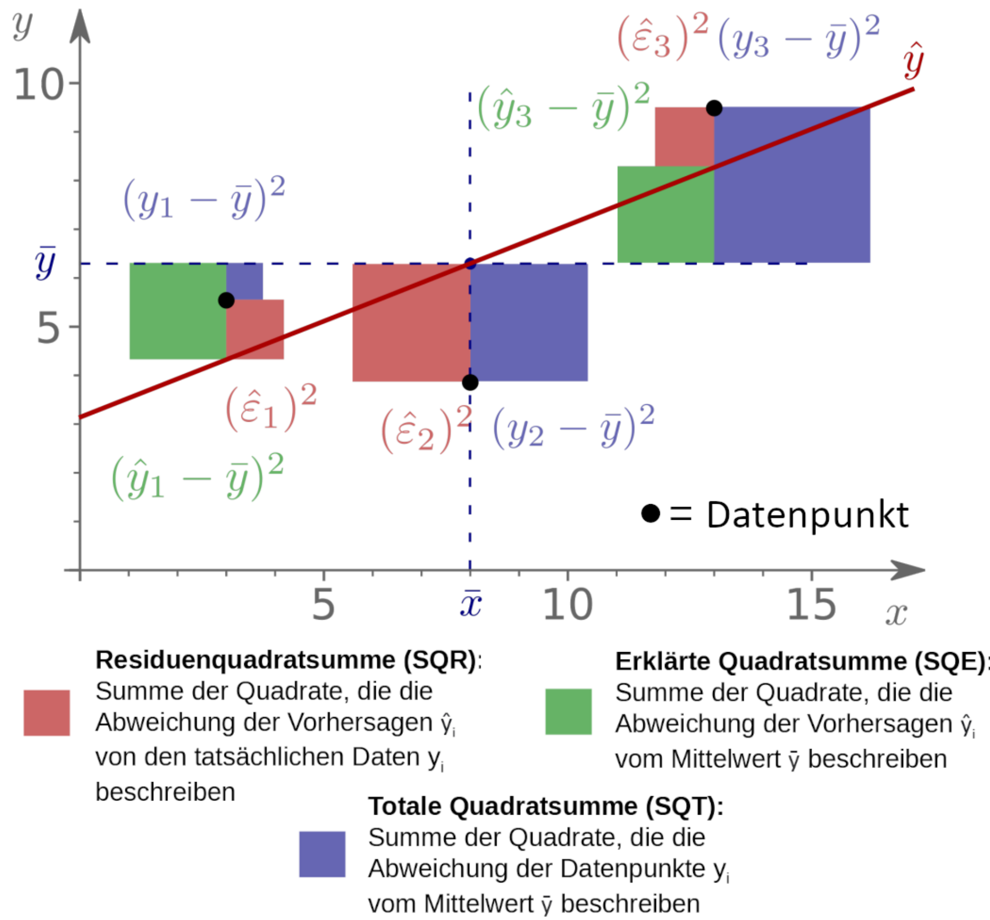
$$SQR = \sum (\hat{y}_i - y_i)^2 = \sum \hat{\epsilon}_i^2$$

- Diese wiederum lässt sich in Bezug setzen zur **totalen Quadratsumme (SQT)** und zur **erklärten Quadratsumme (SQE)**:

$$SQR = SQT - SQE \qquad \text{bzw.}$$
$$SQT = SQT + SQR$$

- Mit

$$SQE = \sum (\hat{y}_i - \bar{y})^2$$
$$SQT = \sum (y_i - \bar{y})^2$$



- Falls alle Punkte exakt auf der Regressionsgeraden liegen ($\hat{y}_i = y_i$) ist die erklärte Quadratsumme identisch der totalen Quadratsumme ($SQE = SQT$) und die Residuenquadratsumme ist 0

Bestimmtheitsmaß

- Das **Bestimmtheitsmaß** R^2 gibt an, wie gut die Datenpunkte durch die Regressionsgerade gefittet werden ("Anpassungsgüte")
- Es gibt an, welcher Anteil der Datenvarianz $Var(Y)$ durch die Varianz der Vorhersage $Var(\hat{Y})$ erklärt wird..

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SQE}{SQT}$$

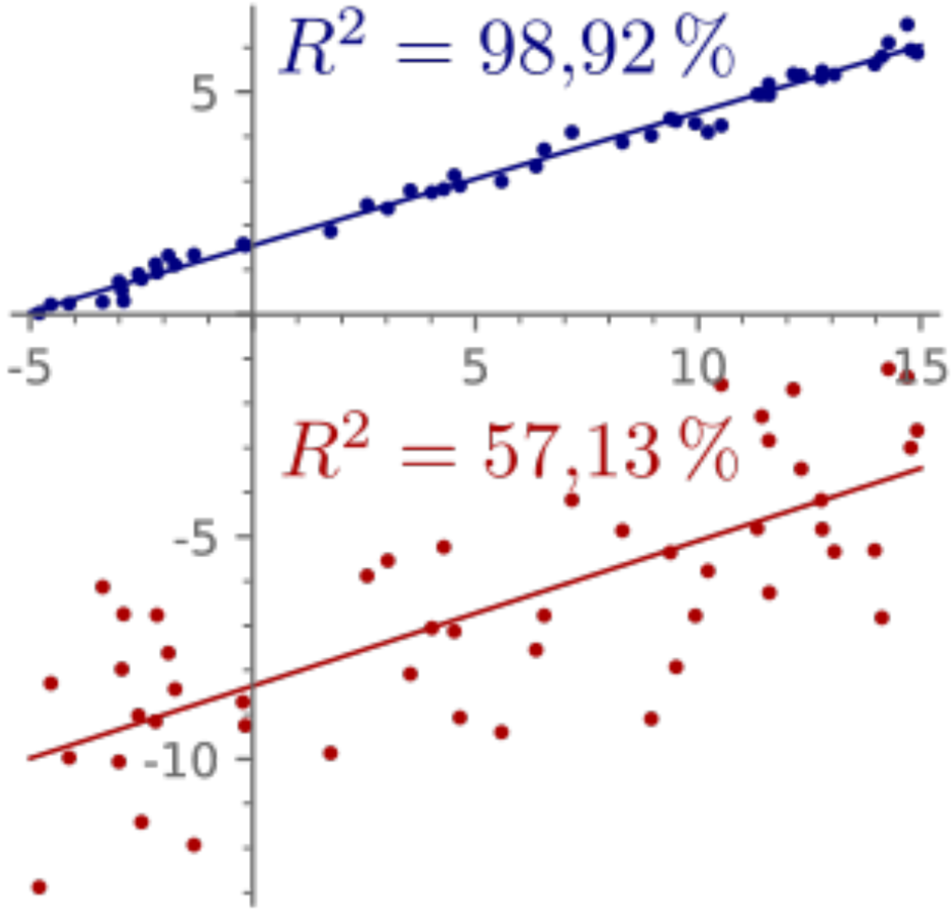
- .. oder äquivalent, den Anteil der erklärten Quadratsumme an der totalen Quadratsumme.

Bei einer einfachen Regression gilt: $R^2 = r^2$
Das Bestimmtheitsmaß ist bei einer einfachen Regression also identisch dem quadrierten Korrelationskoeffizienten zwischen den Variablen X und Y!

Beispiel

Lebenszufriedenheit und sportliche Aktivität haben eine Korrelation von $r = 0{,}8$.
$\Rightarrow$ Sportliche Aktivität erklärt $r^2 = 0{,}64 \mathrel{\widehat{=}} 64\%$ der Varianz von Lebenszufriedenheit (und umgekehrt).

Bestimmtheitsmaß



Beispiele für zwei Regressionen mit Bestimmtheitsmaß $R^2 = 98{,}92\%$ und $R^2 = 57{,}13\%$. Selbst das schwächere Beispiel mit $57{,}13$ wäre für typische Effekte in der Psychologie noch ein außerordentlich hoher Wert.[3]

- Das Bestimmtheitsmaß R^2 gibt an, wie gut sich die Variable Y mit einer linearen Gleichung basierend auf X vorhersagen lässt.
- Der Maximal wert von R^2 ist 1. In diesem Fall erklärt die lineare Gleichung in X die Daten Y perfekt.
- Da das Bestimmtheitsmaß R^2 angibt, welcher Anteil der Varianz in den Daten durch die lineare Gleichung erklärt wird, wird es manchmal in Prozent ausgedrückt (d.h. mit 100 multipliziert; wie im Bild links). Der Maximalwert von R^2 ist dann 100%.

Analytische Form der Regressionskoeffizienten (einfache Regression)

Die optimalen **Regressionskoeffizienten** b_0 (Achsenabschnitt) und b_1 (Steigung) lassen sich analytisch herleiten:

Achsen-abschnitt → $b_0 = \bar{Y} - b_1 \bar{X}$

Regressions-koeffizient

Mittelwert des Prädiktors (x)

Mittelwert des Kriteriums (y)

Regressions-koeffizient → $b_1 = \frac{Cov(X,Y)}{Var(X)}$

Kovarianz der Variablen x und y

Varianz des Prädiktors (x)

- Aus den Mittelwerten und der Kovarianz von X und Y, sowie der Varianz von X, lassen sich also die Regressionskoeffizienten vollständig bestimmen.
- Auch hier zeigt sich wieder die Assymmetrie der Regression: während bei der Formel für die Pearson-Korrelation $Var(X)Var(Y)$ im Nenner steht, ist es beim Regressionskoeffizienten lediglich die Varianz der unabhäbgigen Variable $Var(X)$
- Wäre stattdessen Y die unabhängige Variable, stünde $Var(Y)$ im Nenner, und der Regressionskoeffizient hätte i.d.R. einen anderen Wert.
    - Dies ist auch der Grund, weshalb die Regressionsgleichung nicht einfach invertiert werden darf:

$$\cancel{x_i = \frac{1}{b_1}\hat{y}_i - \frac{b_0}{b_1}}$$

$$\left(\text{.. und } \frac{1}{b_1} \text{ im Allgemeinen } \textbf{nicht} \text{ der Regressionskoeffizient für } Y \text{ als unabhängige Variable ist.}\right)$$

Zusammenhang Regression ↔ Korrelation

- Folgendener Zusammenhang gilt zwischen der Steigung b_1 und dem Korrelationskoeffizienten r:

$$b_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(X,Y)}{s_X^2} = \frac{s_Y}{s_Y}\frac{Cov(X,Y)}{s_X s_X} = \frac{s_Y}{s_X}\underbrace{\frac{Cov(X,Y)}{s_X s_Y}}_{r} = \frac{s_Y}{s_X}r$$

- Es gilt also

$$b_1 = \frac{s_Y}{s_X}r \qquad \text{bzw.} \qquad r = \frac{s_X}{s_Y}b_1$$

Sind die Standardabweichungen s_X und s_Y bekannt, kann aus der Steigung b_1 der Regression immer auch der Korrelationskoeffizient r bestimmt werden (und umgekehrt).

- Der Ausruck $\frac{s_X}{s_Y}b_1$ wird auch **standardisierter Regressionskoeffizient** β_1 genannt:

$$\beta_1 = \frac{s_X}{s_Y}b_1$$

- Bei der einfachen Regression ist der standardisierte Regressionskoeffizient identisch mit dem Korrelationskoeffizienten: $\beta_1 = r$

Standardisierter Regressionskoeffizient

- Wie gesehen erhält man den **standardisierten Regressionskoeffizienten** β durch die Transformation $\beta_1 = \frac{s_X}{s_Y} b_1$
- Im Gegensatz zu b_1 ist β_1 unabhängig von der Skalierung von X und Y (also z.B. ob die Einheit als cm oder m gewählt wurde) $\Rightarrow$ β-Koeffizienten lassen sich besser zwischen verschiedenen Regressionen vergleichen

---

Definition **Standardisierung einer Variable** X:= Variable X durch ihre Standardabweichung s_X teilen
Die Variable X hat nach der Standardisierung die Standardabweichung $s_X = 1$.

- Wurden sowohl X als auch Y vor der Regression standardisiert, also $s_X = s_Y = 1$, so sind die Regressionskoeffizienten automatisch standardisiert:

$$\beta_1 = \frac{s_X}{s_Y} b_1 = \frac{1}{1} b_1 = b_1$$

Interpretation im Kontext der Regressionsgleichungen $\hat{Y} = b_0 + b_1 X$ bzw. $\hat{Y} = \beta_0 + \beta_1 X$	
b_1	Um welchen Wert ändert sich Y bei einer Änderung von X um den Wert 1?
β_1	Um welchen Wert ändert sich Y bei einer Änderung von X um *eine* Standardabweichung s_X?

Herleitung der Regressionskoeffizienten

- Die Methode der kleinsten Quadrate entspricht der Minimierung der quadratischen Residuen:

$$SQR = \sum \left(\hat{y}_i - y_i\right)^2 = \sum \left(b_0 + b_1 x_i - y_i\right)^2 \stackrel{!}{=} \min$$

- Um das Minimum von SQR in Abhängigkeit von b_0 und b_1 zu finden, setzen wir die Ableitungen von SQR nach den Parametern gleich Null (Infinitesimalrechnung@Schule 😉)
- Zunächst leiten wir SQR nach b_0 ab (Kettenregel):

$$\frac{dSQR}{db_0} = \sum 2\left(b_0 + b_1 x_i - y_i\right) = 2nb_0 + 2\sum \left(b_1 x_i - y_i\right) = 0$$

$$\boxed{\longrightarrow b_0 = \frac{1}{n}\sum \left(y_i - b_1 x_i\right) = \frac{1}{n}\sum y_i - \frac{b_1}{n}\sum x_i = \bar{y} - b_1 \bar{x}}$$

- … jetzt benötigen wir noch b_1

KEIN KLAUSUR STOFF

Herleitung der Regressionskoeffizienten

- SQR nach b_1 ableiten und gleich Null setzen:

$$\frac{dSQR}{db_1} = \sum 2(b_0 + b_1 x_i - y_i)x_i = 2b_0 \sum x_i + 2b_1 \sum x_i^2 - 2 \sum x_i y_i = 0$$

$$\longrightarrow b_1 = \frac{\sum x_i y_i}{\sum x_i^2} - \frac{b_0 \sum x_i}{\sum x_i^2} \overset{(b_0 \text{ einsetzen})}{=} \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\bar{y} \sum x_i}{\sum x_i^2} + b_1 \frac{\bar{x} \sum x_i}{\sum x_i^2}$$

- Alle b_1-Terme auf die linke Seite bringen und einige Umformungen vornehmen:

$$b_1 - b_1 \frac{\bar{x} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\bar{y} \sum x_i}{\sum x_i^2}$$

$$b_1 \left(1 - \frac{\bar{x} \sum x_i}{\sum x_i^2}\right) = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$b_1 \left(\frac{\sum x_i^2}{\sum x_i^2} - \frac{\bar{x} \sum x_i}{\sum x_i^2}\right) = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$b_1 \frac{\sum x_i^2 - \bar{x} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$\longrightarrow b_1 = \frac{\sum x_i^2}{\sum x_i^2 - \bar{x} \sum x_i} \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i} \overset{(\sum x_i = n\bar{x})}{=} \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \overset{:n}{=} \frac{\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2}$$

KEIN KLAUSUR STOFF

Herleitung der Regressionskoeffizienten

- Zwischenergebnis:

$$b_1 = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2}$$

- Um zu erkennen, dass der Zähler der Kovarianz und der Nenner der Varianz entspricht, betrachten wir nochmal die Formeln der (Ko)Varianz:

$$Cov(X,Y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum\left(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}\right) = \frac{1}{n}\sum x_i y_i - \bar{y}\frac{1}{n}\sum x_i - \bar{x}\frac{1}{n}\sum y_i + \bar{x}\bar{y} =$$

$$\overset{\left(\frac{1}{n}\sum x_i = \bar{x}\right)}{\underset{\left(\frac{1}{n}\sum y_i = \bar{y}\right)}{=}} \frac{1}{n}\sum x_i y_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

$$Var(X) = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \frac{1}{n}\sum x_i^2 - 2\bar{x}\frac{1}{n}\sum x_i + \bar{x}^2 =$$

$$\overset{\left(\frac{1}{n}\sum x_i = \bar{x}\right)}{=} \frac{1}{n}\sum x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

- Es gilt also tatsächlich:

$$b_1 = \frac{Cov(X,Y)}{Var(X)}$$

KEIN KLAUSUR STOFF

Intuition hinter der Regressionssteigung

- Die Formel für die Steigung bei der einfachen Regression

$$b_1 = \frac{Cov(X,Y)}{Var(X)}$$

.. erinnert an die Formel der Korrelation, bei der die Kovarianz ebenfalls standardisiert wird (mit $\frac{1}{s_X s_Y}$)

- Der entscheidende Unterschied ist, dass bei der Korrelation eine Standardisierung bezüglich *beider* Variablen vorgenommen wird, bei der Regression aber nur bezüglich der *unabhängigen* Variable.
- In der Folge wird bei der Regression folgende Frage beantwortet:

> Was ist die Auswirkung einer Änderung der unabhängigen Variable X um 1 (**einheitslos!**) auf die abhängige Variable Y (**in deren Rohwerteinheiten!**).

- Auch hier wird wieder deutlich, dass bei der Regression eine feste Rollenverteilung vorgenommen wird: nur die unabhängige Variable wird standardisiert.
- Da die Steigung also von der Varianz der als unabhängig deklarierten Variable abhängt, ist es nicht zulässig anzunehmen, dass $\frac{1}{b_1}$ einfach die Steigung wäre, wenn Y die unabhängige und X die abhängige Variable ist. Für die umgekehrte Steigung müssten wir schließlich auch die Varianz von Y berücksichtigen!

Beweis, dass $R^2 = r^2$ bei einfacher Regression

- Ausgestattet mit der Formel für den Regressionskoeffizienten, lässt sich nun auch beweisen, dass bei der einfachen Regression gilt: $R^2 = r^2$

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{\frac{1}{n}\sum(\hat{y}_i - \bar{y})^2}{Var(Y)} \overset{(\hat{y}_i = b_0 + b_1 x_i)}{\underset{(\bar{y} = b_0 + b_1 \bar{x})}{=}} \frac{\frac{1}{n}\sum\left[(b_0 + b_1 x_i) - (b_0 + b_1 \bar{x})\right]^2}{Var(Y)} = \frac{\frac{1}{n}\sum(b_1 x_i - b_1 \bar{x})^2}{Var(Y)} =$$

$$= b_1^2 \frac{\frac{1}{n}\sum(x_i - \bar{x})^2}{Var(Y)} = b_1^2 \frac{Var(X)}{Var(Y)} \qquad \left(\text{NB sieht man, dass gilt:} \quad Var(\hat{Y}) = b_1^2 Var(X)\right)$$

- Nun $b_1 = \frac{Cov(X,Y)}{Var(X)}$ einsetzen:

$$R^2 = \frac{Cov^2(X,Y)}{Var^2(X)} \frac{Var(X)}{Var(Y)} = \frac{Cov^2(X,Y)}{Var(X)Var(Y)}$$

- Vergleiche mit r^2:

$$r^2 = \left[\frac{Cov(X,Y)}{s_X s_Y}\right]^2 = \frac{Cov^2(X,Y)}{s_X^2 s_Y^2} = \frac{Cov^2(X,Y)}{Var(X)Var(Y)}$$

KEIN KLAUSUR STOFF

Ausblick: Multiple Regression

- Gibt es mehr als eine **unabhängige Variable** (auch **Prädiktoren** genannt), handelt es sich nicht mehr um eine einfache Regression, sondern um eine **multiple Regression**:

1. Prädiktor 2. Prädiktor usw.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ \ldots$$

Konstante

Koeffizient für den ersten Prädiktor

Koeffizient für den zweiten Prädiktor

- Jeder Prädiktor $X_1, X_2, \ldots X_n$ hat einen eigenen Regressionskoeffizienten $b_1, b_2, \ldots b_n$
- Multiple Regression wird ausführlich in Statistik 2 behandelt.

"Formula Notation": Formalisierung von Regressionsmodellen

- Da Regressionen heute ausschließlich mit dem Computer berechnet werden, hat sich eine eigene Sprache etabliert, um Regressionsmodelle zu definieren (bekannt als *Formula Notation*):

Unabhängige Variablen (Independent Variables)

y-Achsenabschnitt

$$DV \sim 1 + IV1 + IV2 + \ldots \quad \longleftrightarrow \quad \hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots$$

Trennt abhängige und unabhängige Variablen

Abhängige Variable (Dependent Variable)

- Der Ausdruck "`DV ~ 1 + IV1 + IV2`" kann der Statistiksoftware als *String* übergeben werden; so wird definiert, welches Regressionsmodell gerechnet werden soll.
- `DV`, `IV1`, `IV2` sind dabei die gewählten Variablennamen — beliebige Ausdrücke sind möglich

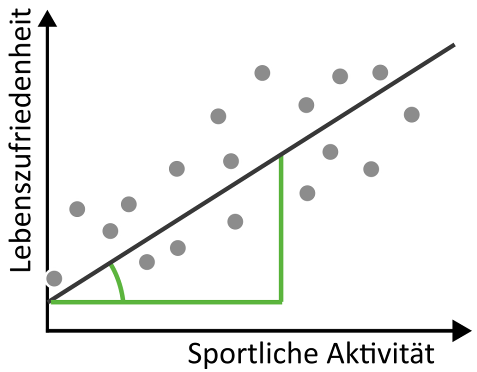
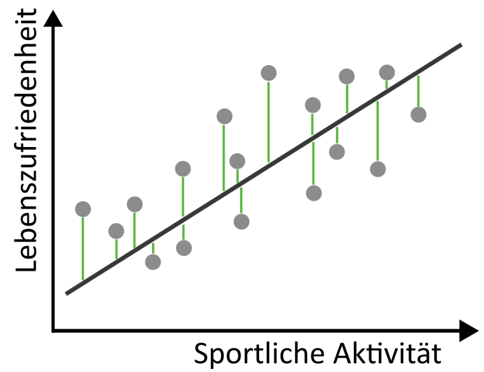
---

Beispiel

"`satisfaction ~ 1 + physical_activity`"

Dies wäre eine mögliche Definition unserer einfachen Regression mit sportlicher Aktivität als unabhängiger und Lebenszufriedenheit als abhängiger Variable.

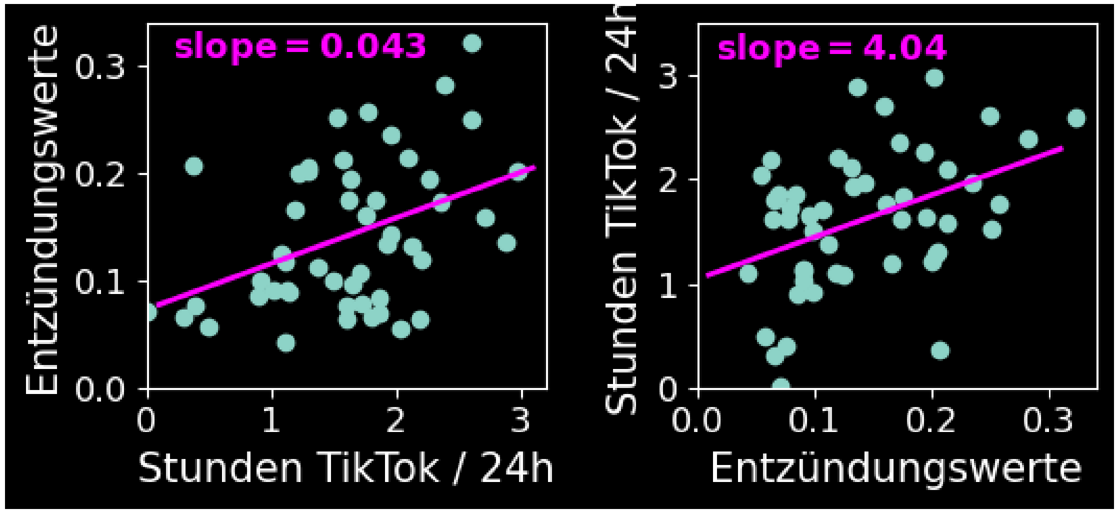
Regression: Erklärung versus Vorhersage

	Erklärung	Vorhersage
Ziel	Zusammenhänge zwischen Variablen untersuchen: - Hängen die Variablen X und Y zusammen? - Ist der Zusammenhang positiv oder negativ? - Wie stark ist der Zusammenhang?	Wie gut kann Variable Y durch Variable X vorhergesagt werden?
Interessante Größe	Steigung b_1	Bestimmtheitsmaß R^2
Visuelle Hervorhebung der interessanten Größe		
Beispiel	Regression von Lebenszufriedenheit auf sportliche Aktivität. Der Regressionskoeffizient sei $b_1 = 0{,}5$. - Der Zusammenhang ist positiv. - Eine Erhöhung von sportlicher Aktivität um den Wert 1 führt im Schnitt zu einer Erhöhung der Lebenszufriedenheit um den Wert $0{,}5$.	Regression von Lebenszufriedenheit auf sportliche Aktivität. Das Bestimmtheitsmaß sei $R^2 = 0{,}4$. - Sportliche Aktivität hat eine gute Vorhersagekraft für Lebenszufriedenheit. - Sportliche Aktivität erklärt 40% der Varianz von interindividueller Lebenszufriedenheit.

[[Zusammenfassung]]

- Die lineare Regression erweitert die Korrelation zu einer **Vorhersageanalyse**: wenn Variablen korrelieren, lässt sich eine Variable aus der anderen vorhersagen.
- Die Vorhersage basiert auf einer **Regressionsgerade**, die alle Datenpunkte so gut wie möglich repräsentiert.
- Die Regressionsgerade wird durch den **Achsenabschnitt** b_0 und die **Steigung** b_1 beschrieben.
- Die standardisierte Form des Steigungs-Koeffizienten wird **Beta** oder **Beta-Gewicht** genannt und ist identisch dem Korrelationskoeffizienzen (bei einfacher Regression).
- Das **Bestimmtheitsmaß** R^2 bemisst die Vorhersagegenauigkeit der Regression.
- Die Regression kann sowohl der **Vorhersage** einer Variable Y auf Basis einer Variable X dienen, als auch der **Erklärung** bzw. Beschreibung eines Zusammehangs von X und Y.

Sie führen nun eine Regressionsanalyse bezüglich des Zusammenhangs von TikTok-Online-Zeit und Entzündungswerten durch. Einmal mit TikTok-Online-Zeit und einmal mit Entzündungswerten als unabhängiger Variable:



Es zeigt sich, dass 1 Stunde zusätzlicher TikTok-Konsum mit einer Erhöhung des Entzündungsparameters um 0,043 verbunden ist. Umgekehrt ist eine Erhöhung des Entzündungswertes um 1 mit 4,04 Stunden — bzw. etwas praktikabeler, eine Erhöhung des Entzündungswertes um 0,1 mit 0,404 Stunden (24 Minuten) — verbunden.

Diese “rohen” Effektstärken zeigen: es handelt sich um ein substantiellen Zusammenhang!

Fußnoten

1. https://flowingdata.com/2014/06/25/duck-vs-rabbit-plot/
2. Krashniak A, Lamm E (2021) Francis Galton's regression towards mediocrity and the stability of types. Studies in History and Philosophy of Science Part A 86:6–19.
3. https://de.wikipedia.org/wiki/Datei:R2values.svg