

M24 Statistik 1: Wintersemester 2023/24

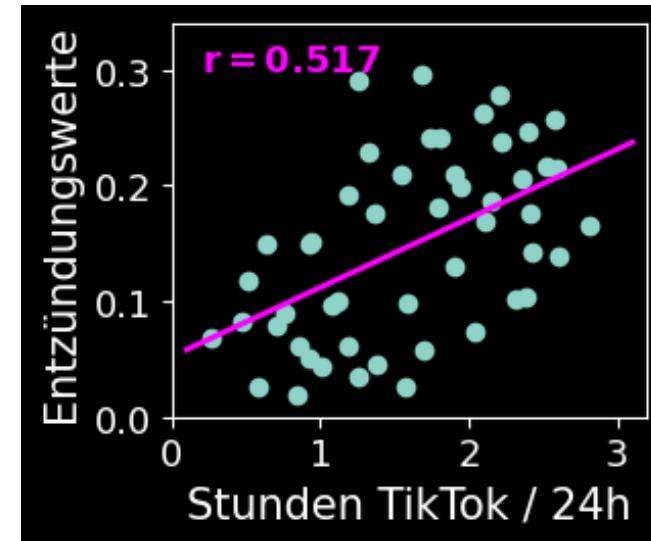
# Vorlesung 06: Regression

Prof. Matthias Guggenmos

Health and Medical University Potsdam

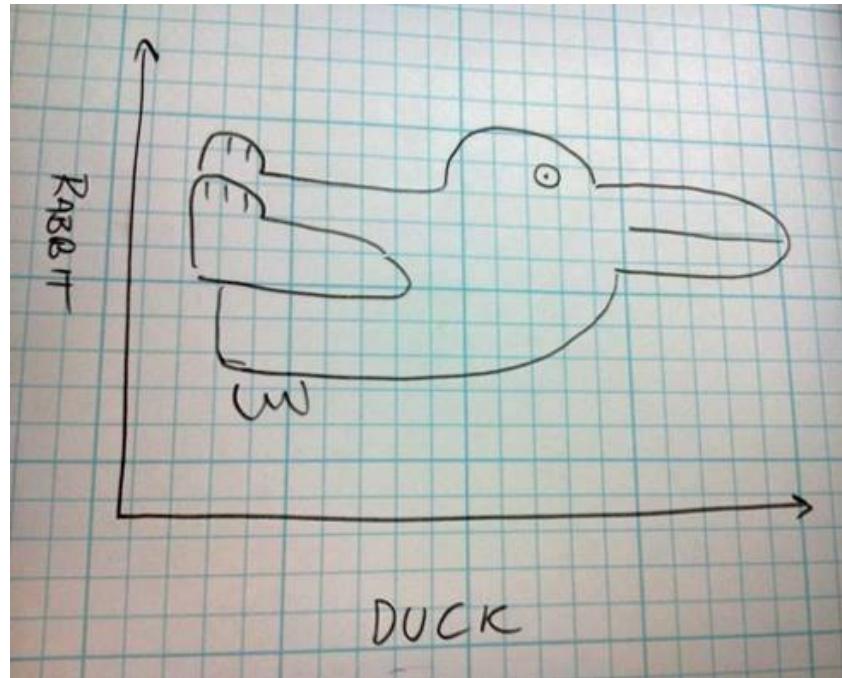


Kurze Erinnerung: beim letzten Mal fanden wir einen Zusammenhang von TikTok-Online-Zeit und Entzündungsparametern:



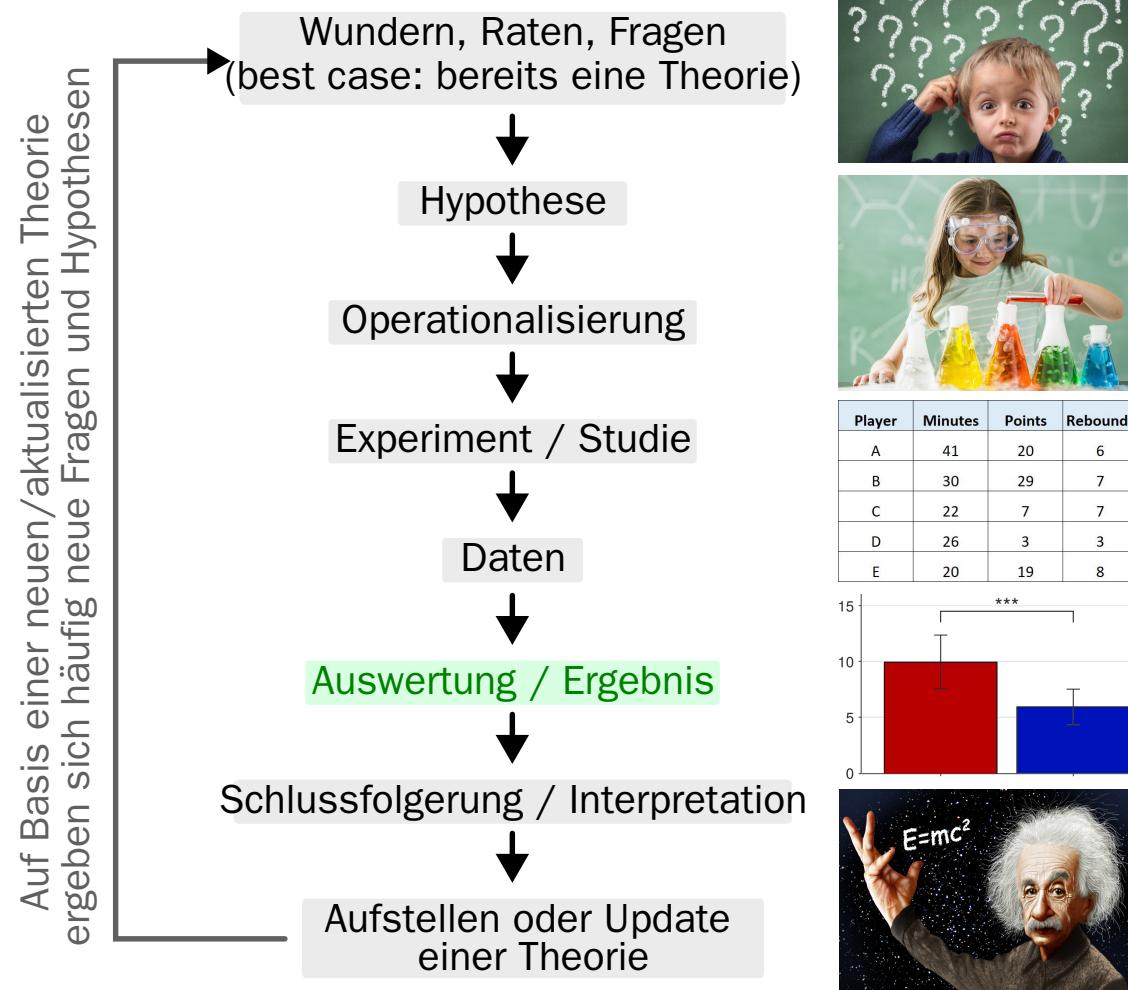
Bei der Interpretation stellt sich einerseits die Kausalitätsfrage, andererseits, wie stark der Zusammenhang tatsächlich ist. Da die Pearson-Korrelation lediglich den **Grad der Linearität** beurteilt, fragen Sie sich: um wie viel erhöhen sich die Entzündungsparameter pro Stunde zusätzliche Zeit auf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

# Regression



Bildnachweis<sup>1</sup>

# Der Forschungsprozess



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



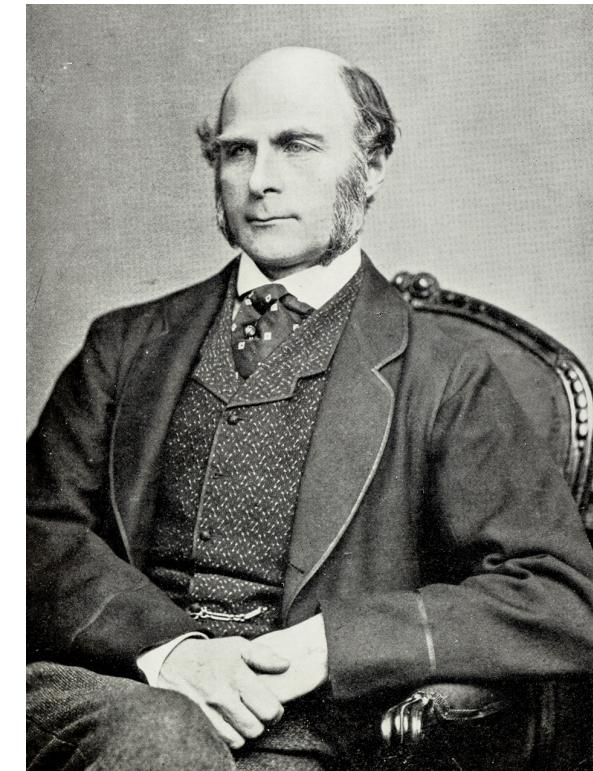
# Woher kommt der Ausdruck “Regression”?

- 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.

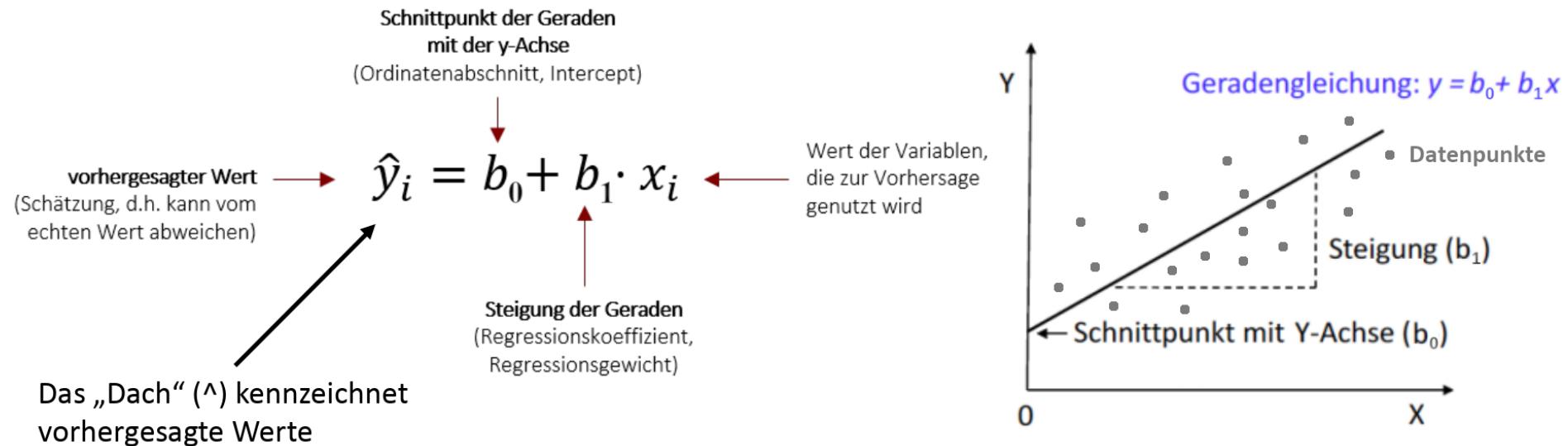
Sigmund Freud (1900). “Traumdeutung”.

- 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<sup>2</sup>.



# Regression

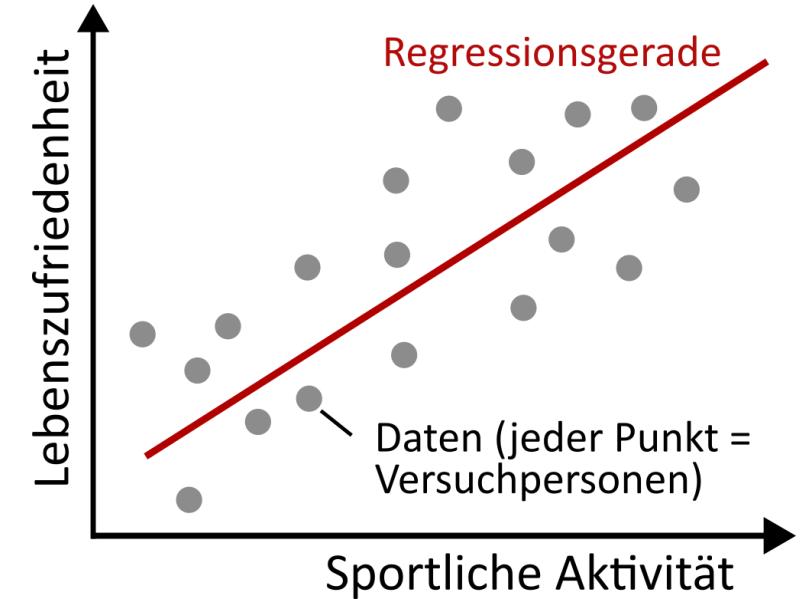
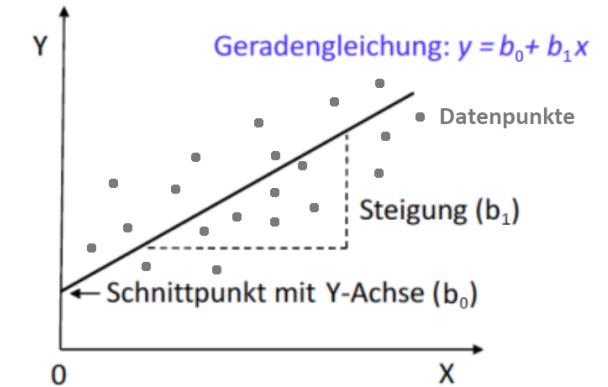
- Dem Wortsinn nach ist Ziel der **Regression** eine abhängige Variable auf eine oder mehrere unabhängige Variablen zurückzuführen (auf diese zu *regredieren*)
- Eingängiger ist aber die umgekehrte Formulierung: Ziel der Regression ist es, auf Basis der unabhängigen Variablen die eine abhängige Variable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



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

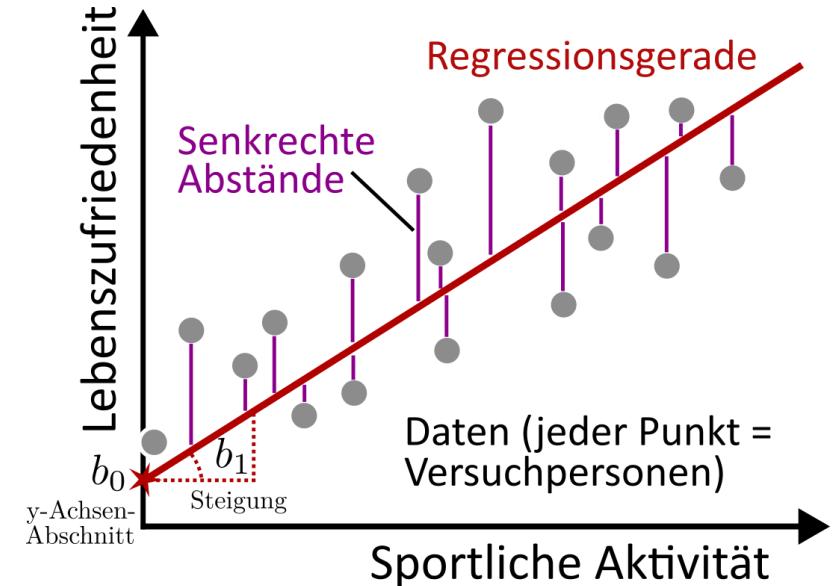
# Regression

- 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 ( $\hat{x}_i = \frac{1}{b_1}y_i - \frac{b_0}{b_1}$ ) – 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

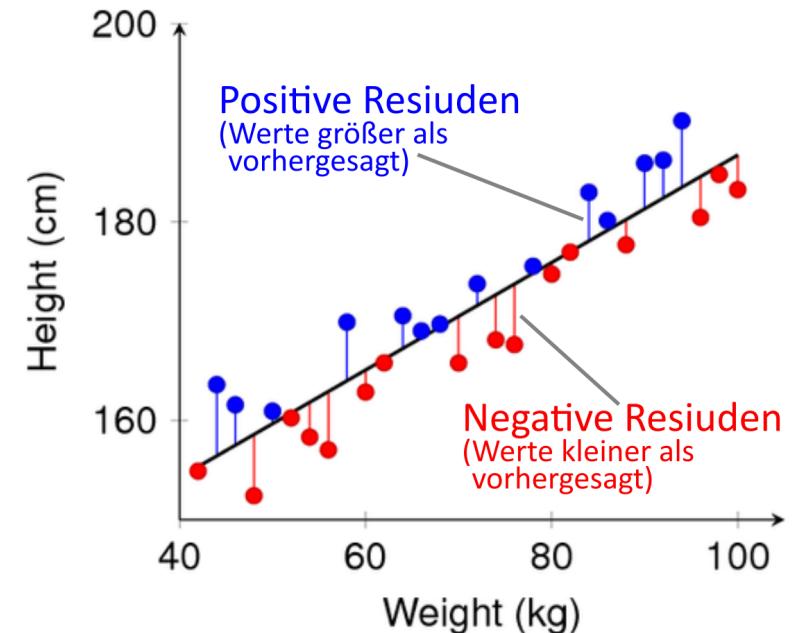
- Ziel der Regression ist es, die Gerade zu finden, die die Datenpunkte möglichst gut abbildet – es gibt jedoch verschiedene Definitionen dessen, was “möglichst gut” heißt
- Die häufigste Variante ist die **Methode der kleinsten Quadrate**, bei der die Gerade so gewählt wird, dass die **Summe der quadrierten 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*)
- Exakt 0 wären die senkrechten Abstände nur, wenn alle Punkte auf einer perfekten Gerade liegen. Dies ist eigentlich nie der Fall. Die verbleibenden senkrechten Abstände der Datenpunkte von der gefitteten Geraden werden **Residuen** genannt.



# 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**  $\Delta\hat{y}_i$ :

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



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

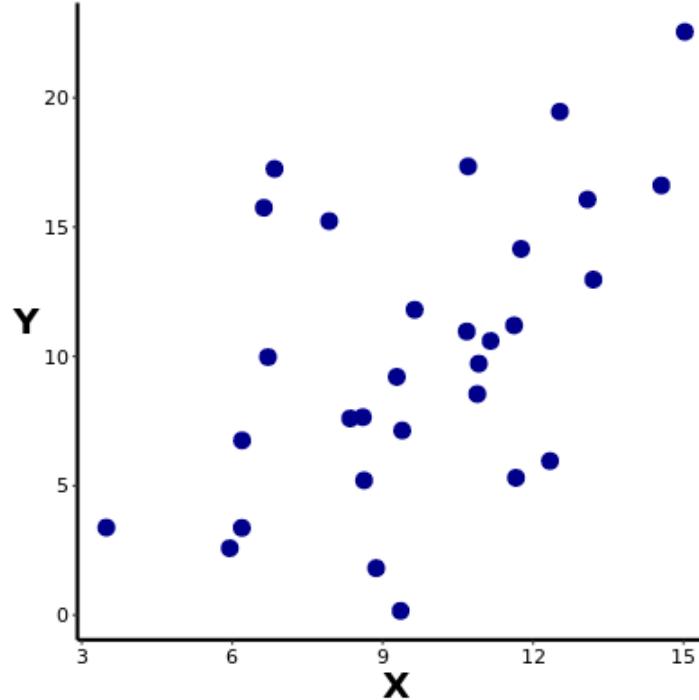
[https://shiny.zoology.ubc.ca/whitlock/Guessing\\_correlation/](https://shiny.zoology.ubc.ca/whitlock/Guessing_correlation/)

## Guessing correlation coefficients

What is the correlation coefficient ( $r$ ) for these data?

- 0.21
- 0.51
- 0.82

Simulate new data



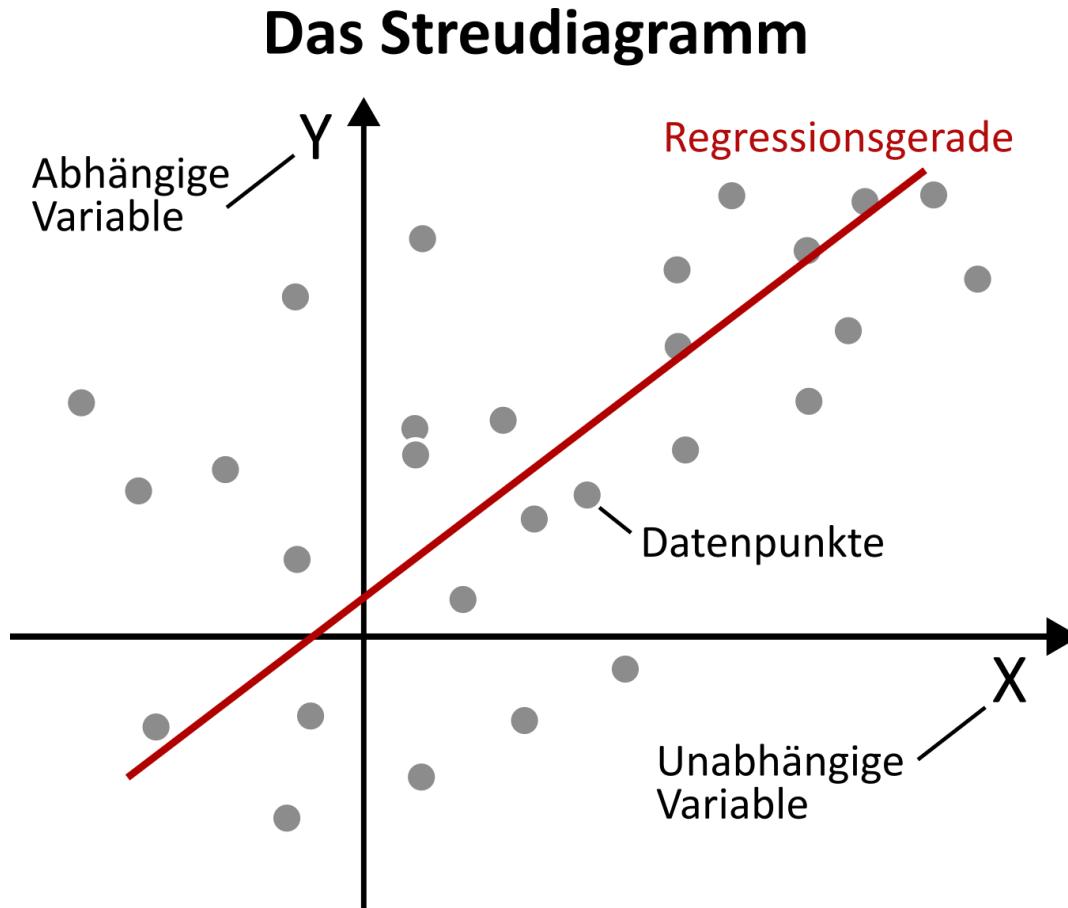
The correlation coefficient ( $r$ ) describes the strength and direction of the association between two numerical variables. The goal of this page is to allow you to develop an intuition about how strong an association is implied by a given correlation coefficient. Each time you press the button marked "Simulate new data", the app will randomly choose a new correlation coefficient and randomly draw 30 data points with that  $r$ . Choose which of the three values best matches the correlation coefficient of the new data.

Developed by Michael Whitlock (<http://www.zoology.ubc.ca/~whitlock/>) at the University of British Columbia. Look here (<http://www.zoology.ubc.ca/~whitlock/kingfisher/KFhomepage.htm>) for more apps.

This web-page and its code is released on a Creative Commons Zero (<https://wiki.creativecommons.org/wiki/CC0>) agreement, meaning that it is freely available for use, re-use, and modification. We request that you give credit, when possible. This work is in the public domain.

# Streudiagramm bei Regression

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



[https://phet.colorado.edu/sims/html/graphing-slope-intercept/latest/graphing-slope-intercept\\_all.html?locale=de](https://phet.colorado.edu/sims/html/graphing-slope-intercept/latest/graphing-slope-intercept_all.html?locale=de)

[https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression\\_all.html?locale=de](https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_all.html?locale=de)

# y-Achsenabschnitt $b_0$

[ToDo] Einige Beispielspielplots für b\_0

# Steigung $b_1$

[ToDo] Einige Beispiele für  $b_1$

# 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)**:

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

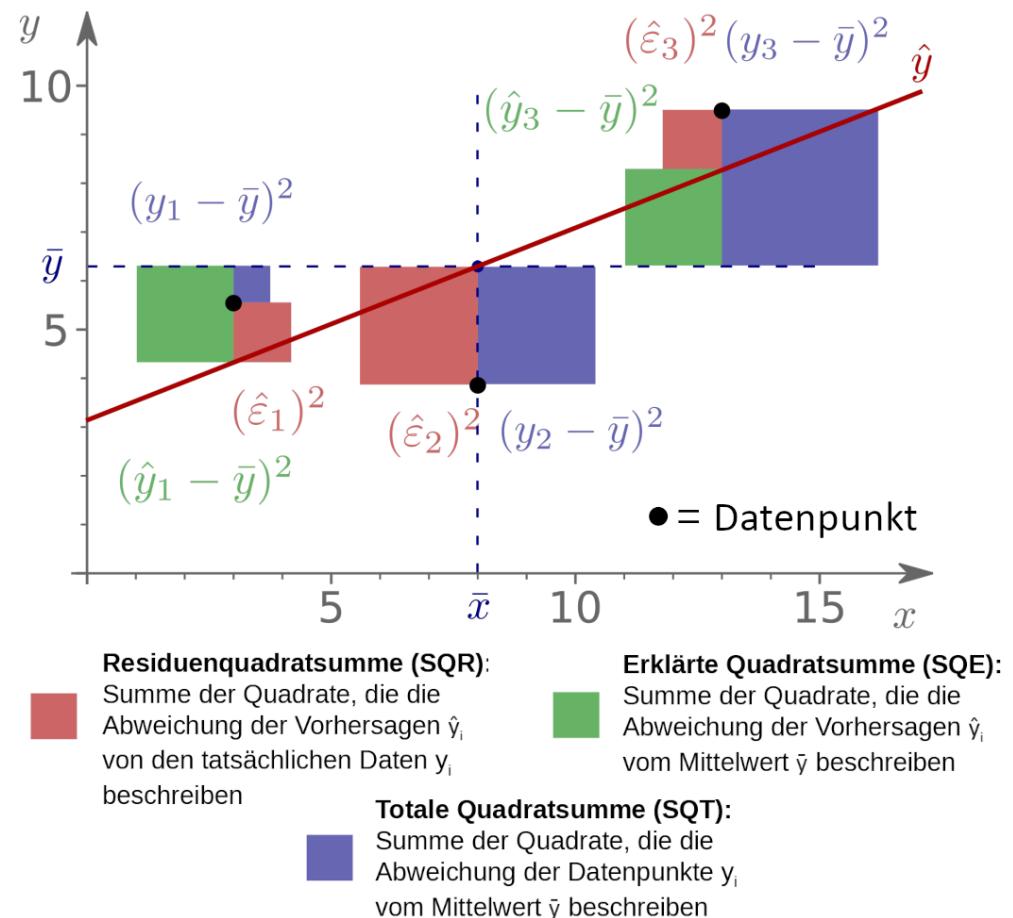
$$SQT = SQT + SQR$$

- 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:  $R^2 = r^2$

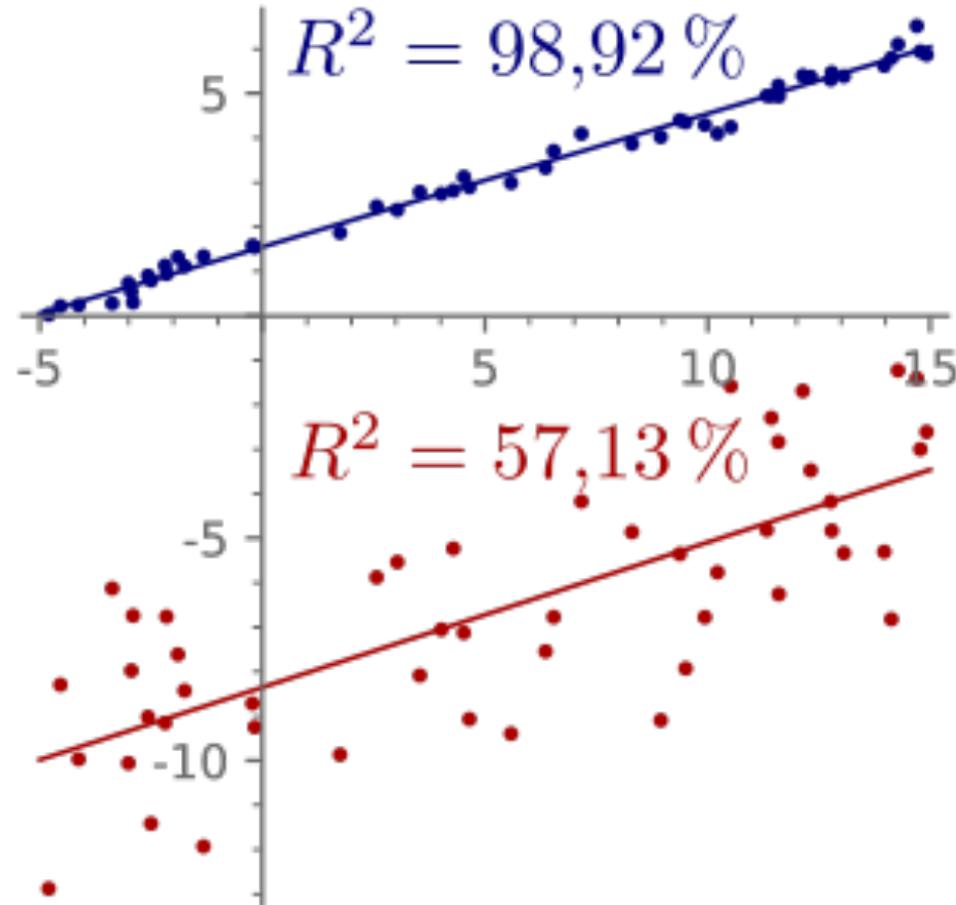
Das Bestimmtheitsmaß ist bei einer einfachen Regression also identisch dem quadrierten Korrelationskoeffizienten zwischen den Variablen X und Y!



Lebenszufriedenheit und sportliche Aktivität haben eine Korrelation von  $r = 0,8$ .

Beispiel   ⇒ Sportliche Aktivität erklärt  $r^2 = 0,64 \hat{=} 64\%$  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.<sup>3</sup>

- Das Bestimmtheitsmaß  $R^2$  gibt an, wie gut sich die Variable  $Y$  mit einer linearen Gleichung basierend auf  $X$  vorhersagen lässt.
- Der Maximalwert 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 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:

$$\rightarrow b_0 = \bar{Y} - b_1 \bar{X}$$

Achsen-abschnitt

Regressionskoeffizient  
Mittelwert des Prädiktors (x)  
Mittelwert des Kriteriums (y)

Regressionskoeffizient  
 $\rightarrow b_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$   
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  $\text{Var}(X)\text{Var}(Y)$  im Nenner steht, ist es beim Regressionskoeffizienten lediglich die Varianz der unabhängigen Variable  $\text{Var}(X)$
- Wäre stattdessen  $Y$  die unabhängige Variable, stünde  $\text{Var}(Y)$  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:

$$\hat{x}_i = \frac{1}{b_1} y_i - \frac{b_0}{b_1}$$

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

# Zusammenhang Regression $\leftrightarrow$ Korrelation

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

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

- Es gilt also

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


---



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 Ausdruck  $\frac{s_X}{s_Y} b_1$  wird auch **standardisierter Regressionskoeffizient**  $\beta_1$  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**  $\beta$  durch die Transformation  $\beta_1 = \frac{s_X}{s_Y} b_1$
- Im Gegensatz zu  $b_1$  ist  $\beta_1$  unabhängig von der Skalierung von  $X$  und  $Y$  (also z.B. ob die Einheit als  $cm$  oder  $m$  gewählt wurde)  $\Rightarrow \beta$ -Koeffizienten lassen sich besser zwischen verschiedenen Regressionen vergleichen

**Defin** Standardisierung einer Variable  $X$ := Variable  $X$  durch ihre Standardabweichung  $s_X$  teilen  
**ition** 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?

$\beta_1$  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:

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

- 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):

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

$$\rightarrow b_0 = \frac{1}{n} \sum (y_i - b_1 x_i) = \frac{1}{n} \sum y_i - \frac{b_1}{n} \sum x_i = \bar{y} - b_1 \bar{x}$$

- ... jetzt benötigen wir noch  $b_1$



# Herleitung der Regressionskoeffizienten

- SQR nach  $b_1$  ableiten und gleich Null setzen:

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

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

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

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

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

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

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

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



# 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:

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

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

- Es gilt also tatsächlich:

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



# 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:  $R^2 = r^2$

$$\begin{aligned}
 R^2 &= \frac{Var(\hat{Y})}{Var(Y)} = \frac{\frac{1}{n} \sum (\hat{y}_i - \bar{y})^2}{Var(Y)} \stackrel{(\hat{y}_i = b_0 + b_1 x_i)}{=} \frac{\frac{1}{n} \sum [(b_0 + b_1 x_i) - (b_0 + b_1 \bar{x})]^2}{Var(Y)} = \frac{\frac{1}{n} \sum (b_1 x_i - b_1 \bar{x})^2}{Var(Y)} = \\
 &= b_1^2 \frac{\frac{1}{n} \sum (x_i - \bar{x})^2}{Var(Y)} = b_1^2 \frac{Var(X)}{Var(Y)} \quad \left( \text{NB sieht man, dass gilt: } Var(\hat{Y}) = b_1^2 Var(X) \right)
 \end{aligned}$$

- 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  $r^2$ :

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



# 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**:

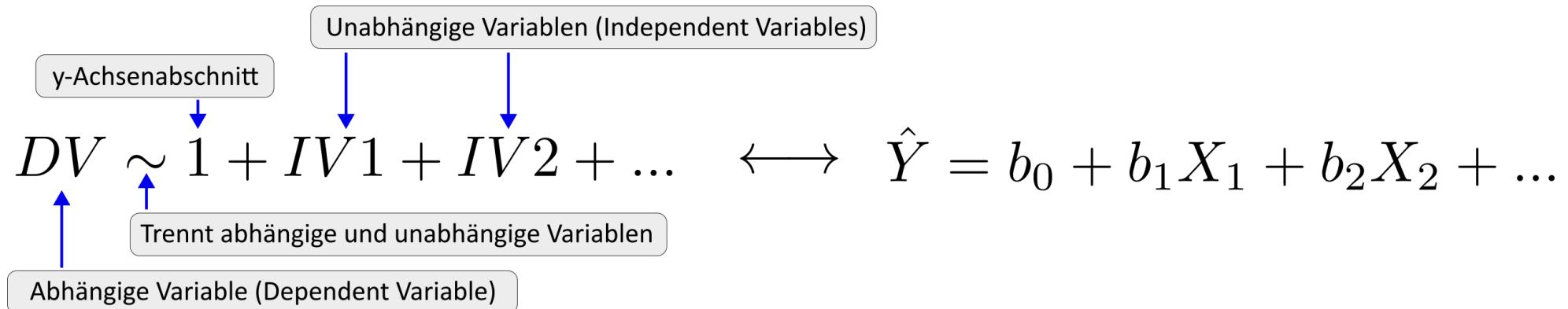
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots$$

1. Prädiktor                    2. Prädiktor                    usw.

- Jeder Prädiktor  $X_1, X_2, \dots, X_n$  hat einen eigenen Regressionskoeffizienten  $b_1, b_2, \dots, b_n$
- 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*):



- Der Ausdruck “ $DV \sim 1 + IV1 + IV2$ ” 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

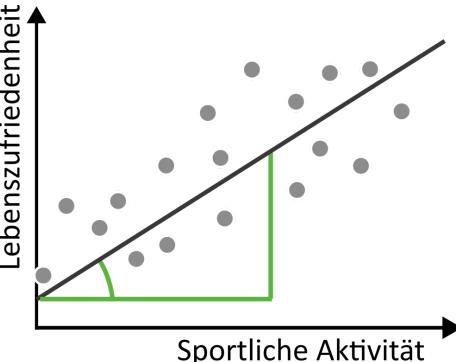
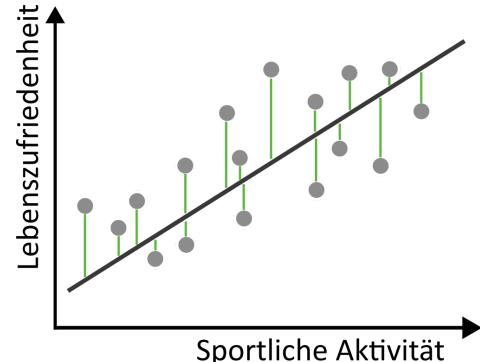


“satisfaction ~ 1 + physical\_activity”

**Beispiel**

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

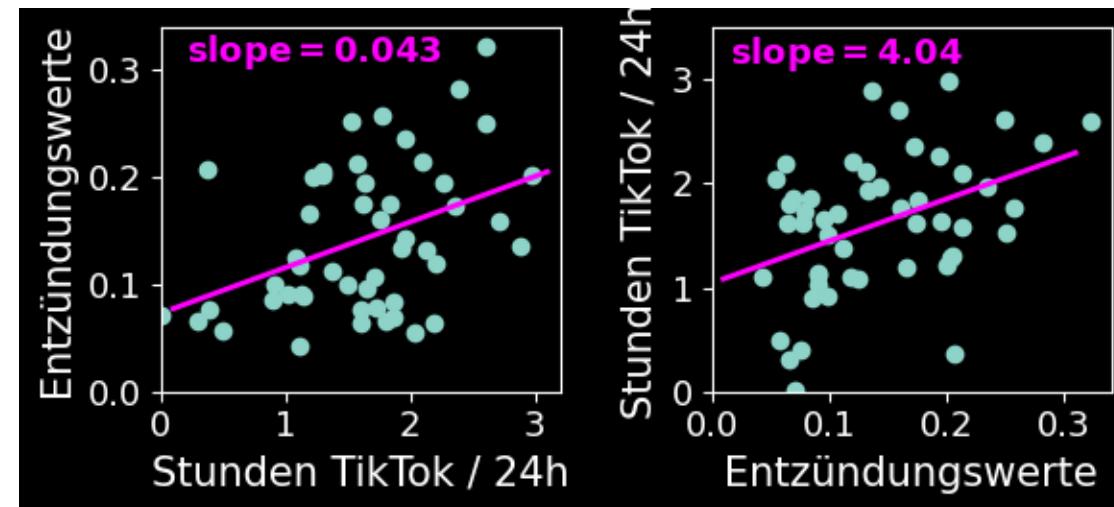
# Regression: Erklärung versus Vorhersage

	Erklärung	Vorhersage
Ziel	<p>Zusammenhänge zwischen Variablen untersuchen:</p> <ul style="list-style-type: none"> <li>- Hängen die Variablen <math>X</math> und <math>Y</math> zusammen?</li> <li>- Ist der Zusammenhang positiv oder negativ?</li> <li>- Wie stark ist der Zusammenhang?</li> </ul>	Wie gut kann Variable $Y$ durch Variable $X$ vorhergesagt werden?
Interessante Größe	Steigung $b_1$	Bestimmungsmaß $R^2$
Visuelle Hervorhebung der interessanten Größe		
Beispiel	<p>Regression von Lebenszufriedenheit auf sportliche Aktivität. Der Regressionskoeffizient sei <math>b_1 = 0,5</math>.</p> <ul style="list-style-type: none"> <li>- Der Zusammenhang ist positiv.</li> <li>- Eine Erhöhung von sportlicher Aktivität um den Wert 1 führt im Schnitt zu einer Erhöhung der Lebenszufriedenheit um den Wert 0,5.</li> </ul>	<p>Regression von Lebenszufriedenheit auf sportliche Aktivität. Das Bestimmtheitsmaß sei <math>R^2 = 0,4</math>.</p> <ul style="list-style-type: none"> <li>- Sportliche Aktivität hat eine gute Vorhersagekraft für Lebenszufriedenheit.</li> <li>- Sportliche Aktivität erklärt 40% der Varianz von interindividueller Lebenszufriedenheit.</li> </ul>

# [ 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 Korrelationskoeffizienten (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 praktikabler, 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>