

M24 Statistik 1: Sommersemester 2024

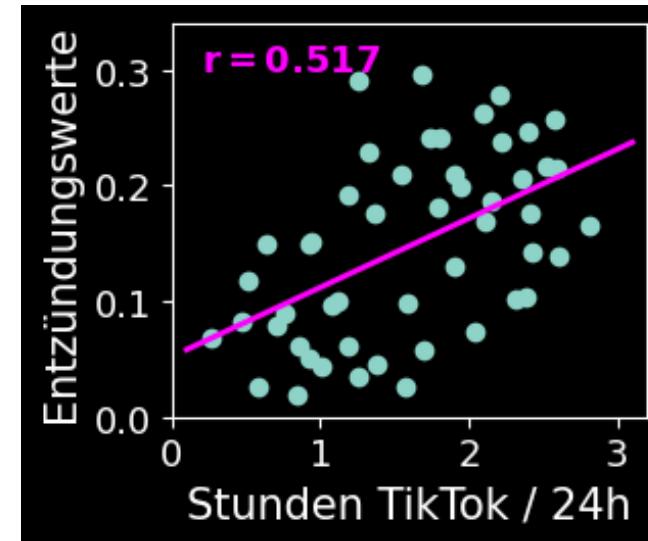
Vorlesung 06: Regression

Prof. Matthias Guggenmos

Health and Medical University Potsdam

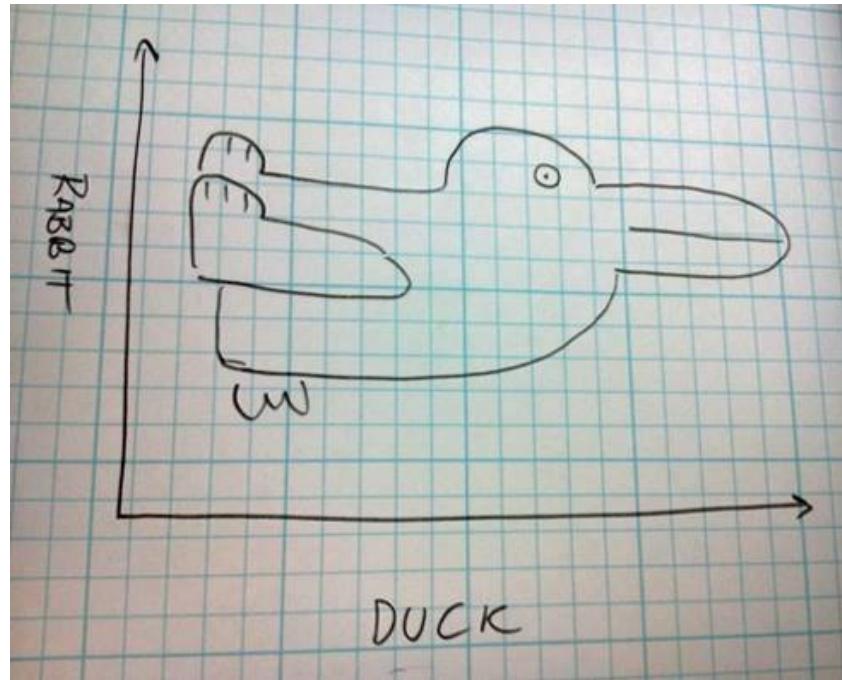


Kurze Erinnerung: beim letzten Mal fanden wir einen Zusammenhang von TikTok-Online-Zeit und Entzündungsparametern:



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



Bildnachweis¹

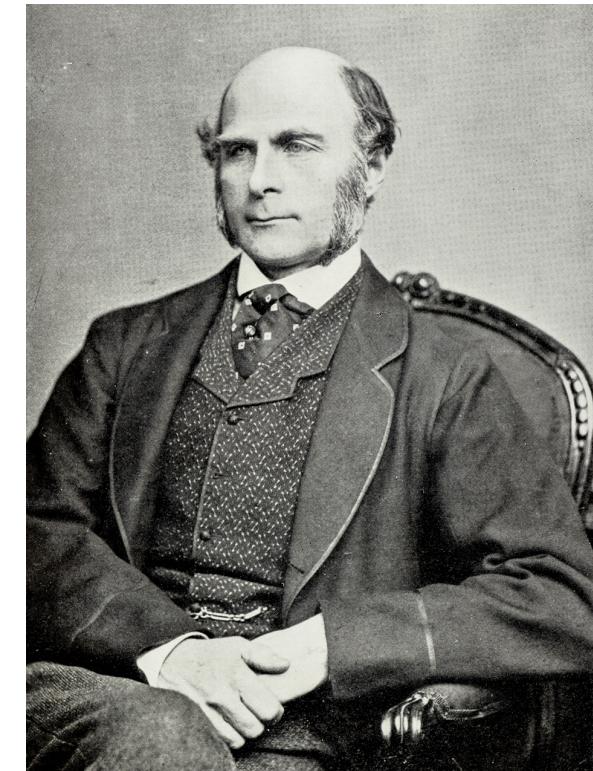
Woher kommt der Ausdruck “Regression”?

- 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.

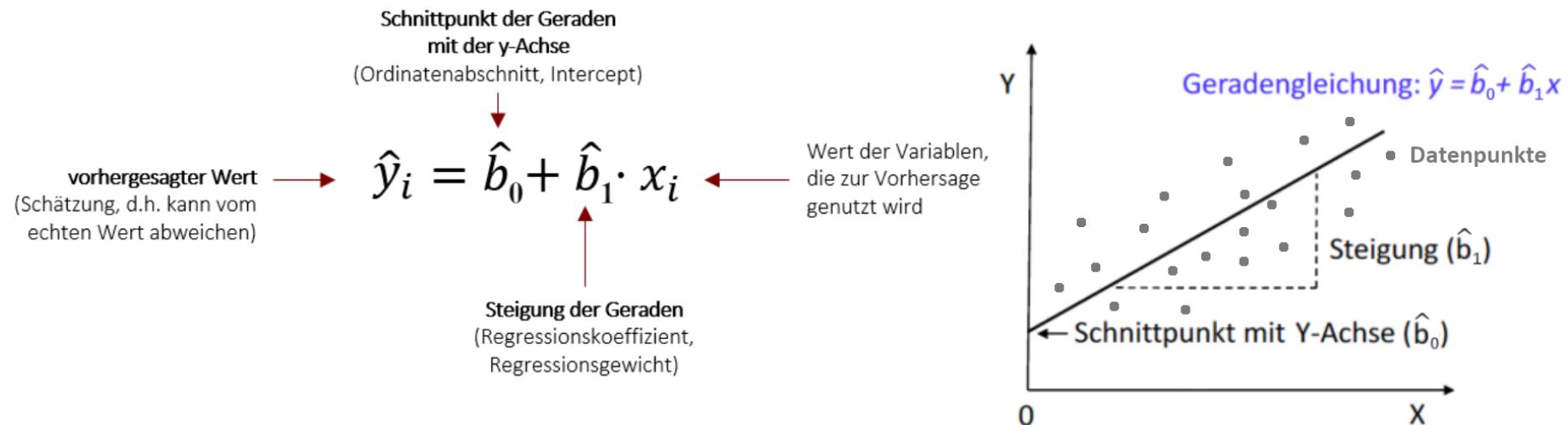
Sigmund Freud (1900). “Traumdeutung”.

- 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 favorisierte².



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)** = vorhersagende oder erklärende Variable(n) ("Ursache").
 - Abhängige Variable** = vorhergesagte oder erklärte Variable ("Auswirkung").



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

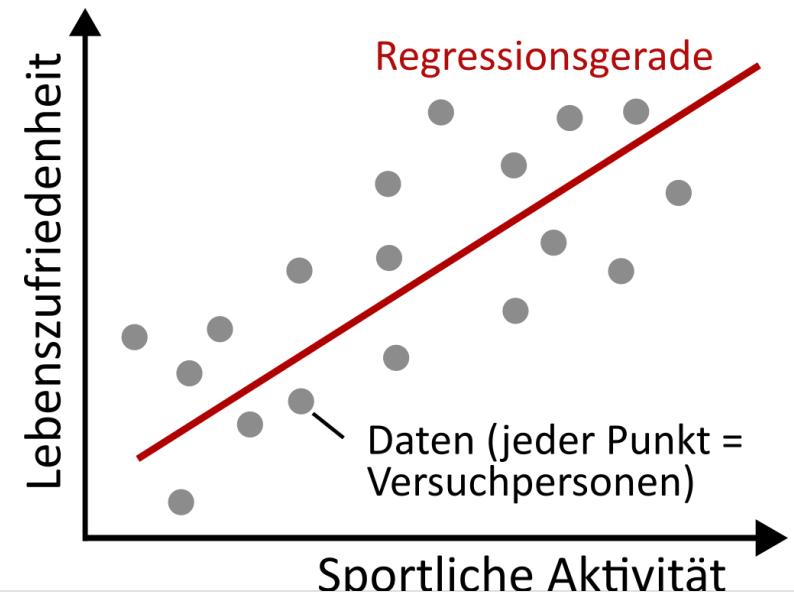
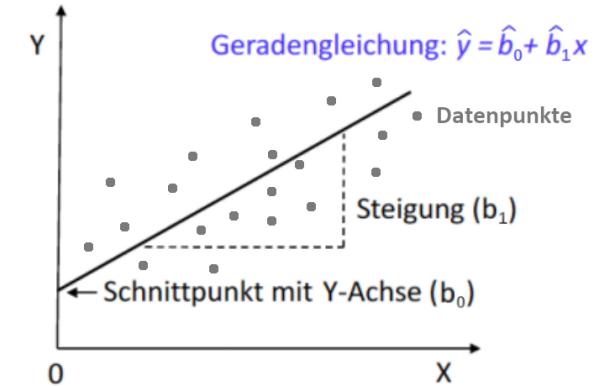
Regression

- 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

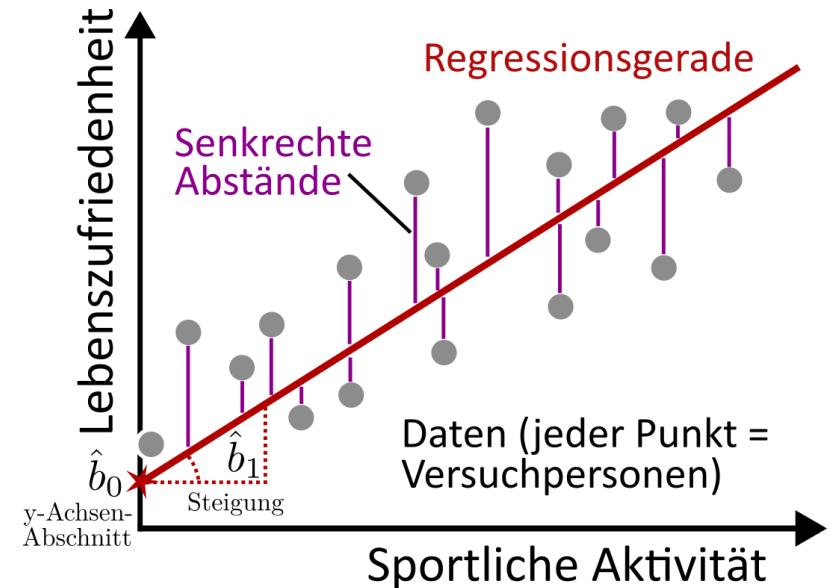
Regressionsgleichung zu invertieren ($x_i = \frac{1}{\hat{b}_1} \hat{y}_i - \frac{\hat{b}_0}{\hat{b}_1}$) – im Allgemeinen ist $\frac{1}{\hat{b}_1}$ nicht die Steigung, wenn die Rollen von X und Y vertauscht werden.

- 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

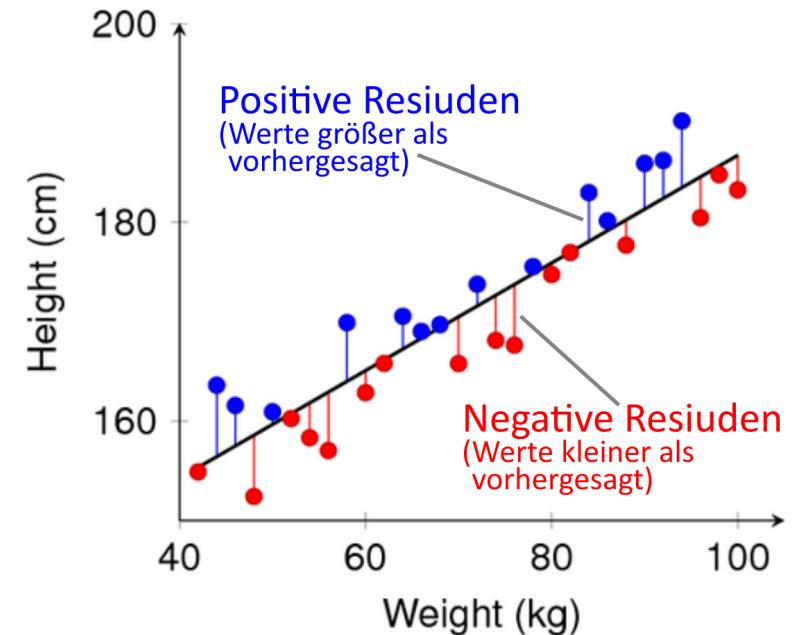
- Ziel der Regression ist es, die Gerade zu finden, die die Datenpunkte möglichst gut abbildet – es gibt jedoch verschiedene Definitionen dessen, was “möglichst gut” heißt.
- Die häufigste Variante ist die **Methode der kleinsten Quadrate**, bei der die Gerade so gewählt wird, dass die Summe der quadrierten **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 \hat{b}_0 (engl. *intercept*)
 - Steigung \hat{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 hängen gar nicht zusammen
- Zusammenhang ist nichtlinear
- Einfluss von Störvariablen
- Messungenauigkeit



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

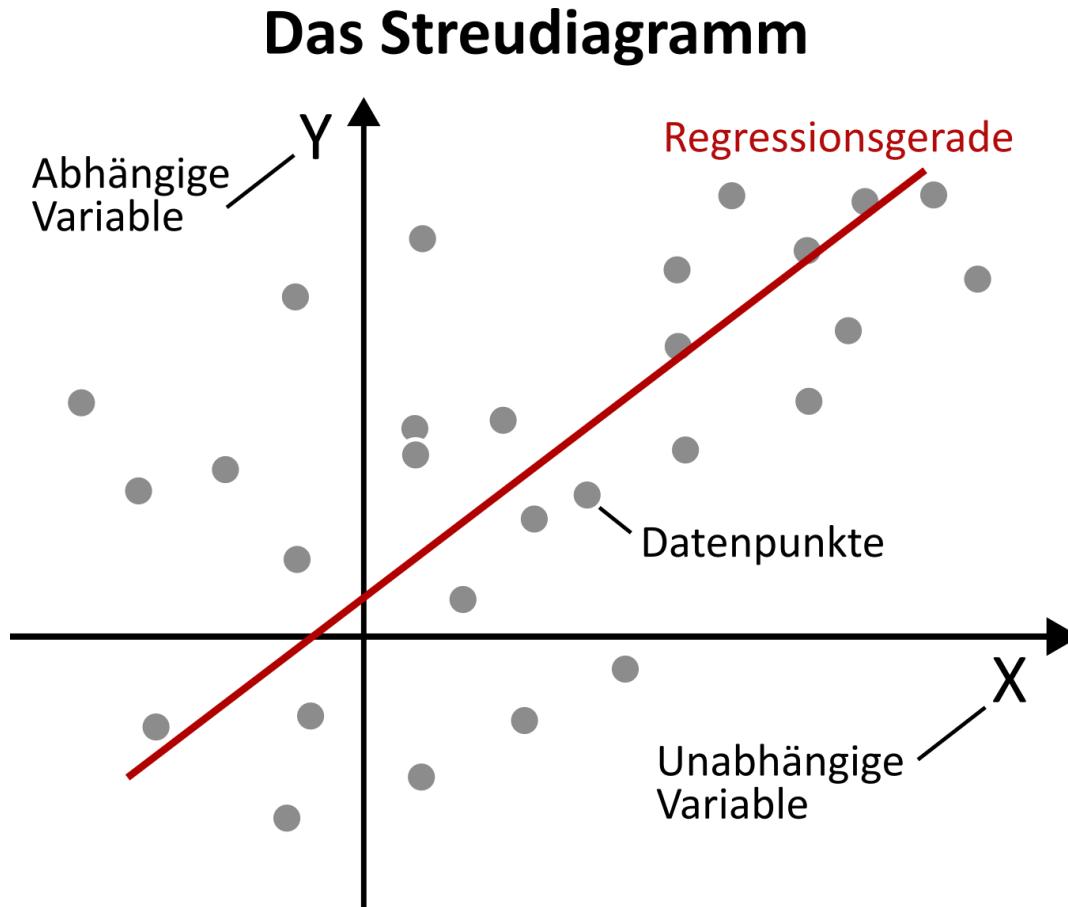
$$\text{Residuum: } \Delta\hat{y}_i = \hat{\epsilon}_i = \hat{y}_i - y_i$$



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

Streudiagramm bei Regression

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



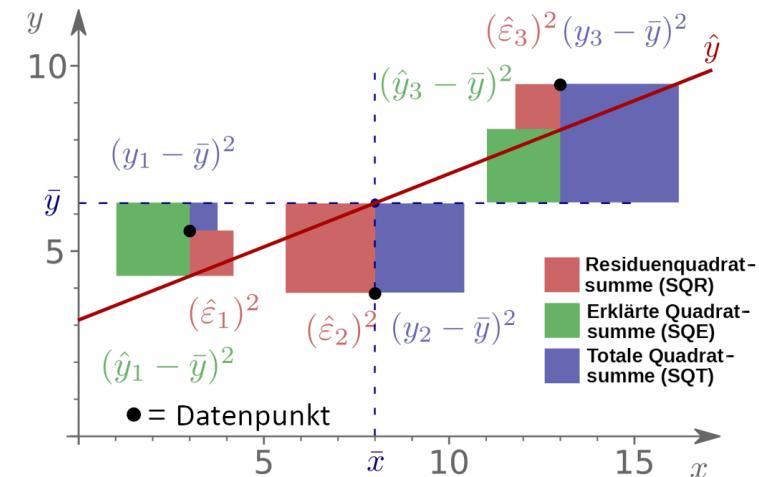
https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_all.html?locale=de

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 lässt sich in Bezug setzen zur **totalen Quadratsumme (SQT)** und zur **erklärten Quadratsumme (SQE)**.



Totale Quadratsumme	$SQT = \sum (y_i - \bar{y})^2$	Abstandsquadrate der Datenpunkte y_i vom Mittelwert \bar{y}	Gesamte Variabilität der Daten y_i .
Erklärte Quadratsumme	$SQE = \sum (\hat{y}_i - \bar{y})^2$	Abstandsquadrate der Vorhersagen \hat{y}_i vom Mittelwert \bar{y}	Variabilität, die durch das Regressionsmodell (\hat{y}_i) erklärt wird.
Residuenquadratsumme	$SQR = \sum (\hat{y}_i - y_i)^2$	Abstandsquadrate der Vorhersagen \hat{y}_i von den Datenpunkten y_i	Variabilität, die durch das Regressionsmodell (\hat{y}_i) <u>nicht</u> erklärt wird.

Es gilt

$$SQT = SQE + SQR$$

Die totale Quadratsumme (SQT) ist die Summe aus der erklärten Quadratsumme (SQE) und der Residuenquadratsumme (SQR).

$$SQR = SQT - SQE$$

Der Teil der Datenvarianz (SQT), der nicht durch das Modell erklärt wird (SQE), entspricht der Residuenquadratsumme (SOR).

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: $R^2 = \rho^2$

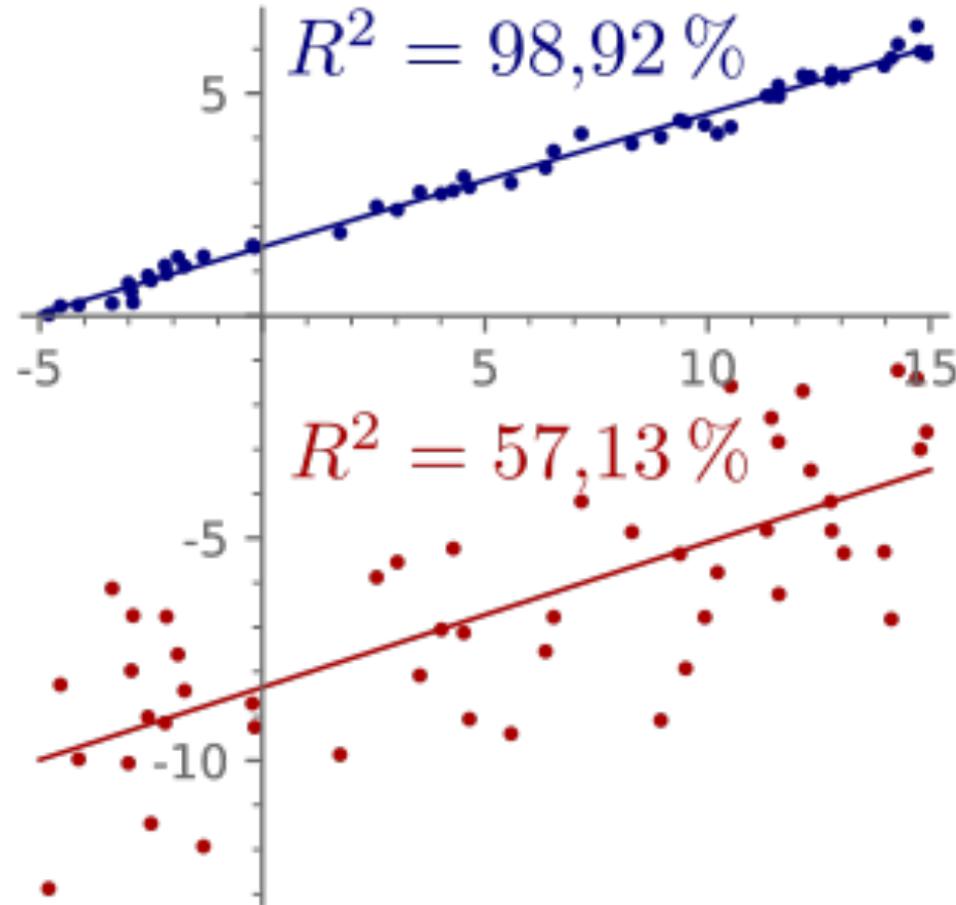
Das Bestimmtheitsmaß ist bei einer einfachen Regression also identisch dem quadrierten Korrelationskoeffizienten zwischen den Variablen X und Y !



Lebenszufriedenheit und sportliche Aktivität haben eine Korrelation von $\rho = 0.8$.

Beispiel ⇒ Sportliche Aktivität erklärt $\rho^2 = 0.64 \hat{=} 64\%$ 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.³

- 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 im Bild links). Der Maximalwert von R^2 ist dann 100%.

Analytische Form der Regressionskoeffizienten (einfache Regression)

Die **Regressionskoeffizienten** \hat{b}_0 (Achsenabschnitt) und \hat{b}_1 (Steigung) lassen sich analytisch herleiten:

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

- 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 $\hat{\rho}$ der symmetrische Ausdruck $\sigma_X \sigma_Y$ im Nenner steht, ist es beim Steigungskoeffizienten \hat{b}_1 lediglich die Varianz der unabhängigen Variable $\text{Var}(X)$.
- Wäre stattdessen Y die unabhängige Variable, stünde $\text{Var}(Y)$ im Nenner, und \hat{b}_1 hätte i.d.R. einen anderen Wert.
 - Dies ist auch der Grund, weshalb die Regressionsgleichung nicht einfach invertiert werden darf:

$$x_i = \frac{1}{\hat{b}_1} \hat{y}_i - \frac{\hat{b}_0}{\hat{b}_1}$$

$\left(\dots \text{ und } \frac{1}{\hat{b}_1} \text{ im Allgemeinen } \mathbf{nicht} \text{ der Steigungskoeffizient für } Y \text{ als unabhängige Variable ist.} \right)$

Zusammenhang Regression \leftrightarrow Korrelation

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

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \underbrace{\frac{\sigma_Y}{\sigma_Y}}_1 \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_X} = \frac{\sigma_Y}{\sigma_X} \underbrace{\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}}_{\hat{\rho}} = \frac{\sigma_Y}{\sigma_X} \hat{\rho}$$

- Es gilt also

$$\hat{b}_1 = \frac{\sigma_Y}{\sigma_X} \hat{\rho} \quad \text{bzw.} \quad \hat{\rho} = \frac{\sigma_X}{\sigma_Y} \hat{b}_1$$



Sind die Standardabweichungen σ_X und σ_Y bekannt, kann aus der Steigung \hat{b}_1 der Regression immer auch der Korrelationskoeffizient $\hat{\rho}$ bestimmt werden (und umgekehrt).

- Der Ausdruck $\frac{\sigma_X}{\sigma_Y} \hat{b}_1$ wird auch **standardisierter Regressionskoeffizient** $\hat{\beta}_1$ genannt:

$$\hat{\beta}_1 = \frac{\sigma_X}{\sigma_Y} \hat{b}_1$$

- Bei der einfachen Regression ist der standardisierte Regressionskoeffizient identisch mit dem Korrelationskoeffizienten: $\hat{\beta}_1 = \hat{\rho}$

Standardisierter Regressionskoeffizient

- Wie gesehen erhält man den **standardisierten Regressionskoeffizienten** $\hat{\beta}$ durch die Transformation $\hat{\beta}_1 = \frac{\sigma_X}{\sigma_Y} \hat{b}_1$.
- Im Gegensatz zu \hat{b}_1 ist $\hat{\beta}_1$ unabhängig von der Skalierung von X und Y (also z.B. ob die Einheit als cm oder m gewählt wurde) $\Rightarrow \hat{\beta}$ -Koeffizienten lassen sich besser zwischen verschiedenen Regressionen vergleichen.

Interpretation im Kontext der Regressionsgleichungen $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ bzw. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

\hat{b}_1 Um welchen Wert ändert sich Y bei einer Änderung von X um den Wert 1?

$\hat{\beta}_1$ Um welchen Wert ändert sich Y bei einer Änderung von X um eine Standardabweichung σ_x ?

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

$$\hat{\beta}_1 = \frac{\sigma_X}{\sigma_Y} \hat{b}_1 = \frac{1}{1} \hat{b}_1 = \hat{b}_1$$

Definition **Standardisierung einer Variable** X = Variable X durch Stichprobenstandardabweichung σ_x teilen. Für die Variable X gilt nach der Standardisierung $\sigma_x = 1$.

Intuition hinter der Regressionssteigung

- Die Formel für die Steigung bei der einfachen Regression

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

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

- 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 unabhängigen Variable abhängt, ist es nicht zulässig anzunehmen, dass $\frac{1}{\hat{b}_1}$ einfach die Steigung wäre, wenn Y die unabhängige und X die abhängige Variable ist. Für die umgekehrte Steigung müssten wir schließlich die Varianz von Y berücksichtigen!

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**:

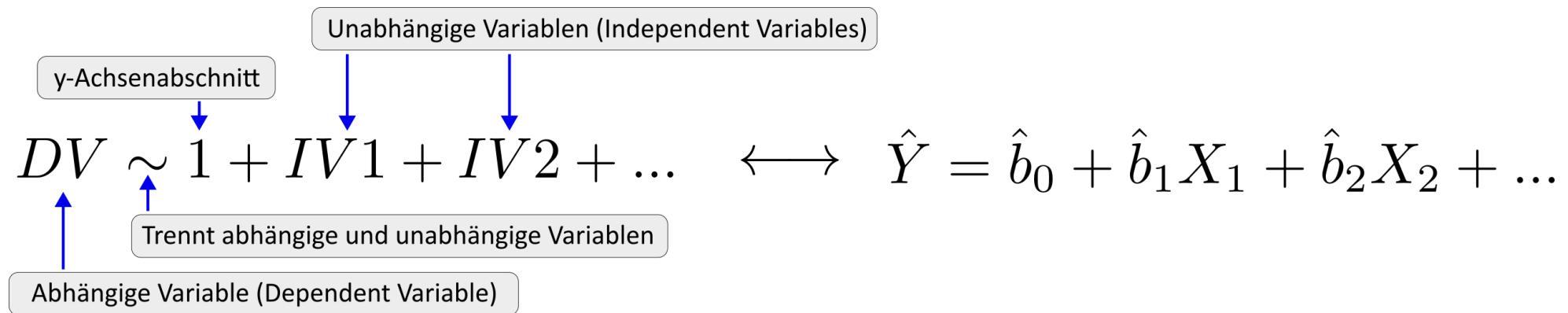
$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots$$

1. Prädiktor 2. Prädiktor usw.

- Jeder Prädiktor X_1, X_2, \dots, X_n hat einen eigenen Regressionskoeffizienten $\hat{b}_1, \hat{b}_2, \dots, \hat{b}_n$
- 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*):



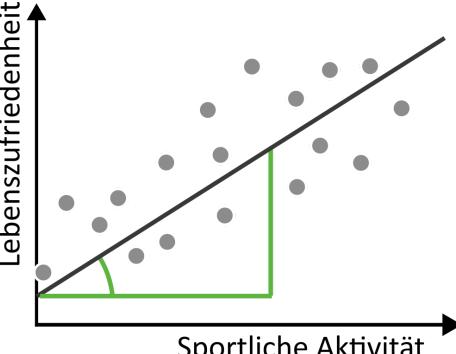
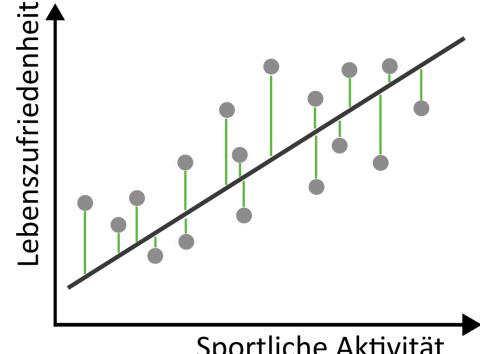
- Der Ausdruck “ $DV \sim 1 + IV1 + IV2$ ” kann der Statistiksoftware als *String* übergeben werden; so wird definiert, welches Regressionsmodell gerechnet werden soll.
- DV, IV1, IV2 sind dabei die gewählten Variablennamen – beliebige Ausdrücke sind möglich

 Beispiel

“satisfaction ~ 1 + physical_activity”

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

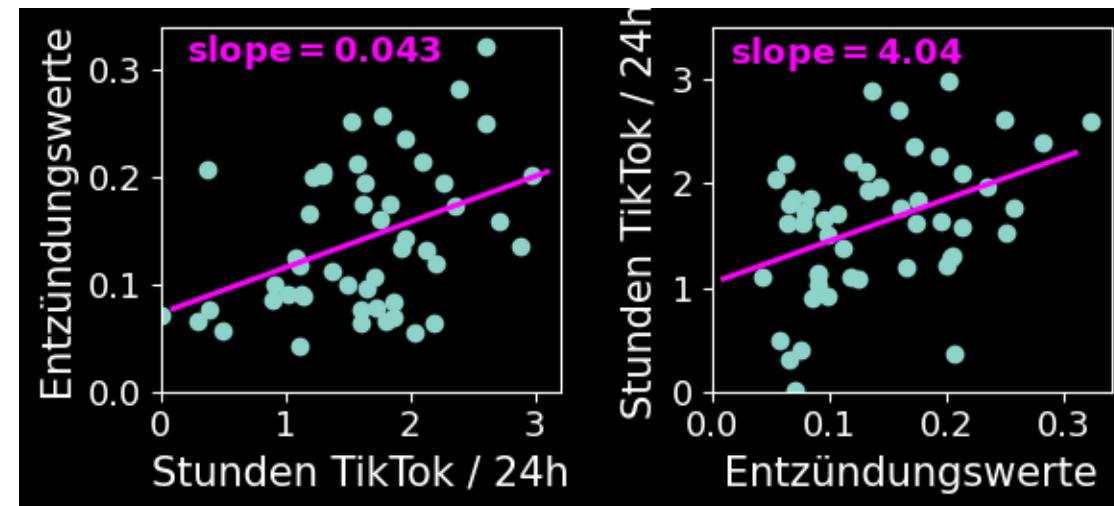
Regression: Erklärung versus Vorhersage

	Erklärung	Vorhersage
Ziel	<p>Zusammenhänge zwischen Variablen untersuchen:</p> <ul style="list-style-type: none"> - 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 \hat{b}_1	Bestimmtheitsmaß R^2
Visuelle Hervorhebung der interessanten Größe		
Beispiel	<p>Regression von Lebenszufriedenheit auf sportliche Aktivität. Der Regressionskoeffizient sei $\hat{b}_1 = 0.5$.</p> <ul style="list-style-type: none"> - 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. 	<p>Regression von Lebenszufriedenheit auf sportliche Aktivität. Das Bestimmtheitsmaß sei $R^2 = 0.4$.</p> <ul style="list-style-type: none"> - 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** \hat{b}_0 und die **Steigung** \hat{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!

Herleitung der Regressionskoeffizienten

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

$$SQR = \sum (\hat{y}_i - y_i)^2 = \sum (\hat{b}_0 + \hat{b}_1 x_i - y_i)^2 \stackrel{!}{=} \min$$

- Um das Minimum von SQR in Abhängigkeit von \hat{b}_0 und \hat{b}_1 zu finden, setzen wir die Ableitungen von SQR nach den Parametern gleich Null (Infinitesimalrechnung@Schule 😊)
- Zunächst leiten wir SQR nach \hat{b}_0 ab (Kettenregel):

$$\frac{dSQR}{d\hat{b}_0} = \sum 2(\hat{b}_0 + \hat{b}_1 x_i - y_i) = 2n\hat{b}_0 + 2 \sum (\hat{b}_1 x_i - y_i) = 0$$

$$\rightarrow \hat{b}_0 = \frac{1}{n} \sum (y_i - \hat{b}_1 x_i) = \frac{1}{n} \sum y_i - \frac{\hat{b}_1}{n} \sum x_i = \bar{y} - \hat{b}_1 \bar{x}$$

- ... jetzt benötigen wir noch \hat{b}_1



Herleitung der Regressionskoeffizienten

- SQR nach \hat{b}_1 ableiten und gleich Null setzen:

$$\frac{dSQR}{d\hat{b}_1} = \sum 2(\hat{b}_0 + \hat{b}_1 x_i - y_i)x_i = 2\hat{b}_0 \sum x_i + 2\hat{b}_1 \sum x_i^2 - 2 \sum x_i y_i = 0$$

$$\rightarrow \hat{b}_1 = \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\hat{b}_0 \sum x_i}{\sum x_i^2} \stackrel{(\hat{b}_0 \text{ einsetzen})}{=} \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\bar{y} \sum x_i}{\sum x_i^2} + \hat{b}_1 \frac{\bar{x} \sum x_i}{\sum x_i^2}$$

- Alle \hat{b}_1 -Terme auf die linke Seite bringen und einige Umformungen vornehmen:

$$\hat{b}_1 - \hat{b}_1 \frac{\bar{x} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\bar{y} \sum x_i}{\sum x_i^2}$$

$$\hat{b}_1 \left(1 - \frac{\bar{x} \sum x_i}{\sum x_i^2} \right) = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$\hat{b}_1 \left(\frac{\sum x_i^2}{\sum x_i^2} - \frac{\bar{x} \sum x_i}{\sum x_i^2} \right) = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$\hat{b}_1 \frac{\sum x_i^2 - \bar{x} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2}$$

$$\rightarrow \hat{b}_1 = \frac{\sum x_i^2}{\sum x_i^2 - \bar{x} \sum x_i} \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2} = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i} \stackrel{(\sum x_i = n\bar{x})}{=} \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \stackrel{(\text{vgl.})}{=} \frac{\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2}$$



Herleitung der Regressionskoeffizienten

- Zwischenergebnis:

$$\hat{b}_1 = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2}$$

- 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 $R^2 = \hat{\rho}^2$ bei einfacher Regression

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

$$\begin{aligned}
 R^2 &= \frac{Var(\hat{Y})}{Var(Y)} = \frac{\frac{1}{n} \sum (\hat{y}_i - \bar{y})^2}{Var(Y)} \stackrel{\substack{(\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i) \\ (\bar{y} = \hat{b}_0 + \hat{b}_1 \bar{x})}}{=} \frac{\frac{1}{n} \sum [(\hat{b}_0 + \hat{b}_1 x_i) - (\hat{b}_0 + \hat{b}_1 \bar{x})]^2}{Var(Y)} = \frac{\frac{1}{n} \sum (\hat{b}_1 x_i - \hat{b}_1 \bar{x})^2}{Var(Y)} \\
 &= \hat{b}_1^2 \frac{\frac{1}{n} \sum (x_i - \bar{x})^2}{Var(Y)} = \hat{b}_1^2 \frac{Var(X)}{Var(Y)} \quad \left(\text{NB sieht man, dass gilt: } Var(\hat{Y}) = \hat{b}_1^2 Var(X) \right)
 \end{aligned}$$

- Nun $\hat{b}_1 = \frac{Cov(X, Y)}{Var(X)}$ 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 $\hat{\rho}^2$:

$$\hat{\rho}^2 = \left[\frac{Cov(X, Y)}{\sigma_X \sigma_Y} \right]^2 = \frac{Cov^2(X, Y)}{\sigma_X^2 \sigma_Y^2} = \frac{Cov^2(X, Y)}{Var(X)Var(Y)}$$



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>