CHARLOTTE FRESENIUS
HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Statistik II

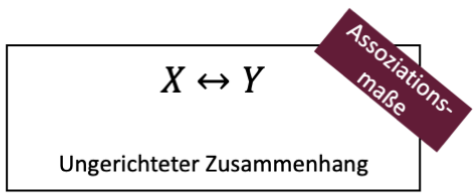
Einheit 3: Einfache lineare Regression (1)

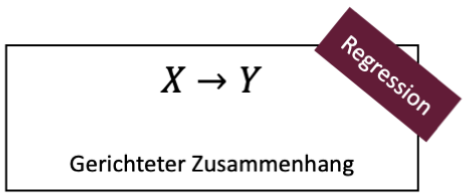
Wintersemester 2025 | Prof. Dr. Stephan Goerigk

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Zusammenhänge - Korrelation vs. Regression





$$X \rightarrow Y$$

Unabhängige Variable (UV) Abhängige Variable (AV)

X „erklärt" Y
Y durch X „vorhergesagt"

Woher kommt die Richtung?

Welche Variable (Y) durch welche Variable (X) erklärt werden wird anhand inhaltlichen Kontexts (Theorie) entschieden.

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Zusammenhänge - Korrelation vs. Regression

Ungerichtete Zusammenhänge

- Wir haben uns bereits mit ungerichteten Zusammenhängen zwischen 2 Variablen beschäftigt
- ungerichtet: es könnte X auf Y wirken, Y aber auch auf X, oder die beiden könnten einfach parallel auftreten
- Um solche Zusammenhänge zu messen, gibt es sogenannte Assoziationsmaße z.B. Varianz oder Korrelation.

Gerichtete Zusammenhänge

- Jetzt wollen wir einen Schritt weiter gehen und unterstellen, dass der Zusammenhang eine Wirkrichtung hat
- D.h. X wirkt auf Y und eben nicht Y wirkt auf X
- Wenn wir diese Richtung unterstellen, kann man die lineare Regression anwenden

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Zusammenhänge - Korrelation vs. Regression

- Die Schreibweise $X \rightarrow Y$ soll also andeuten X wirkt auf Y (bzw. sagt Y vorher)
- Dabei ist X die unabhängige Variable (UV) und Y ist die abhängige Variable (AV)

Beispiel:

- Die Intensität des Sports wirkt sich auf den Puls aus, aber nicht umgekehrt
- Wenn wir so einen gerichteten Zusammenhang untersuchen, dann stellt sich die Frage, wo die Richtung herkommt

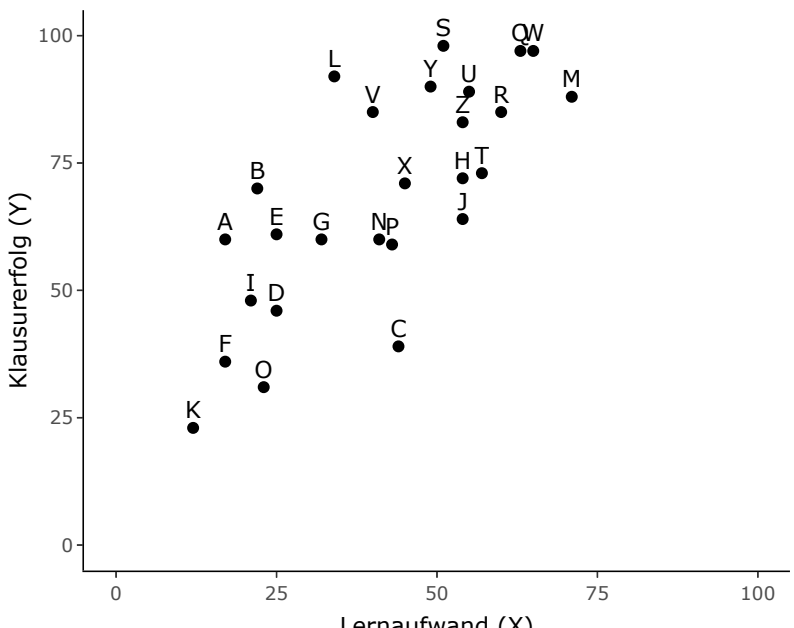
$\rightarrow$ Typischerweise müssen wir die Richtung aus der Theorie oder aus Plausibilitätsüberlegungen herleiten

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Einfache lineare Regression, Beispiel

Gegeben: Lernaufwand X (Stunden) und Klausurerfolg Y (Punkte: 0-100) von n = 26 Studierenden (hier Zeilen 1-13).



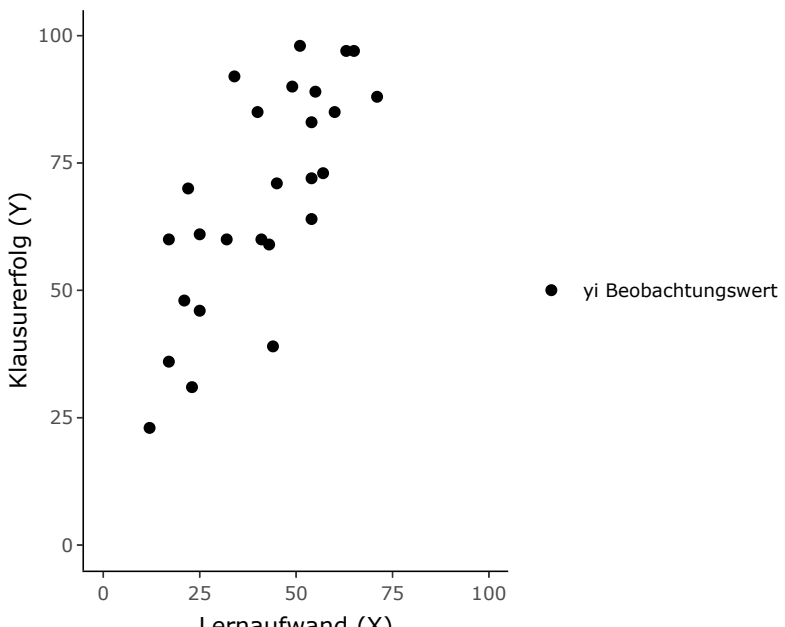
ID	X	Y
A	17	60
B	22	70
C	44	39
D	25	46
E	25	61
F	17	36
G	32	60
H	54	72
I	21	48
J	54	64
K	12	23
L	34	92
M	71	88

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



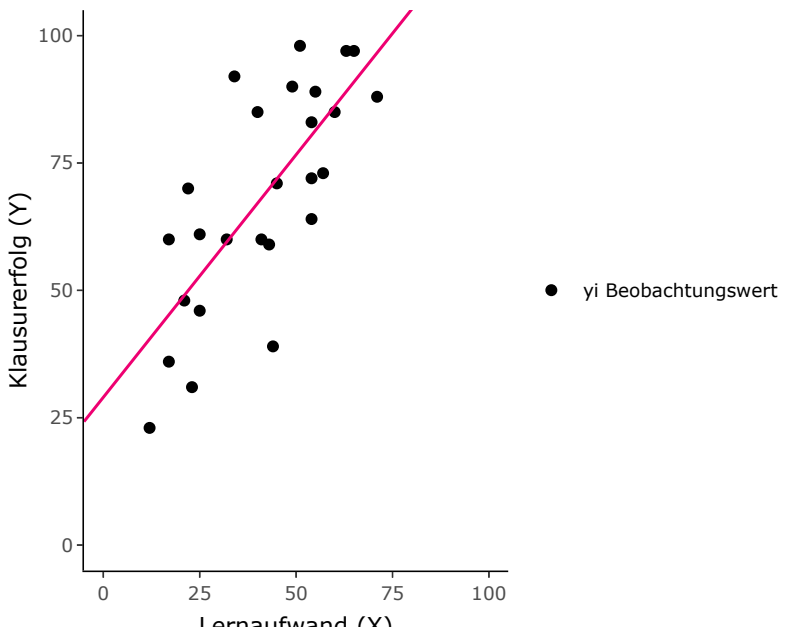
- Jeder Punkt repräsentiert eine Kombination aus X und Y Werten
- Wir könnten also sagen, jeder Punkt ist eine Person aus unserem Beispiel
- Es gilt, in der Regression eine Funktion zu finden, die diese Daten möglichst genau widerspiegelt

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



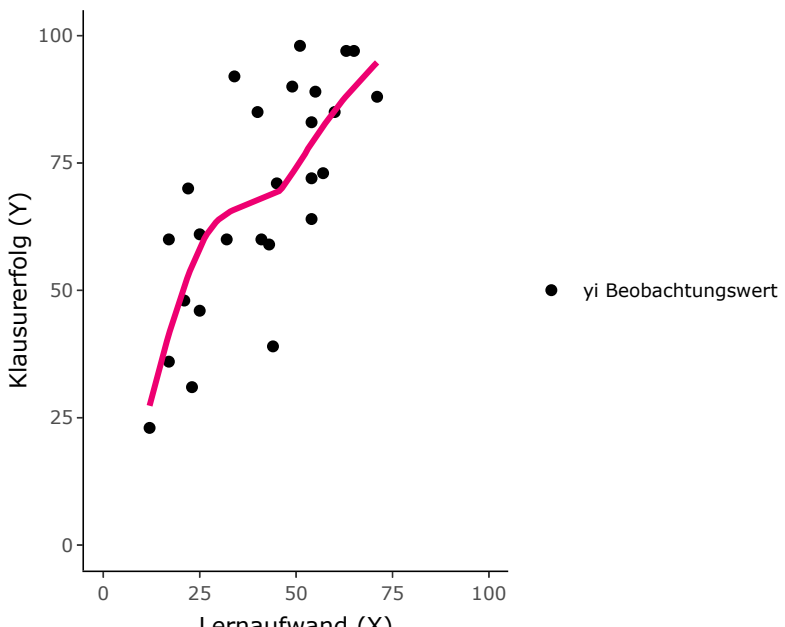
- Im Falle der *linearen* Regression wird unterstellt, dass diese Funktion linear, also eine Gerade ist

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



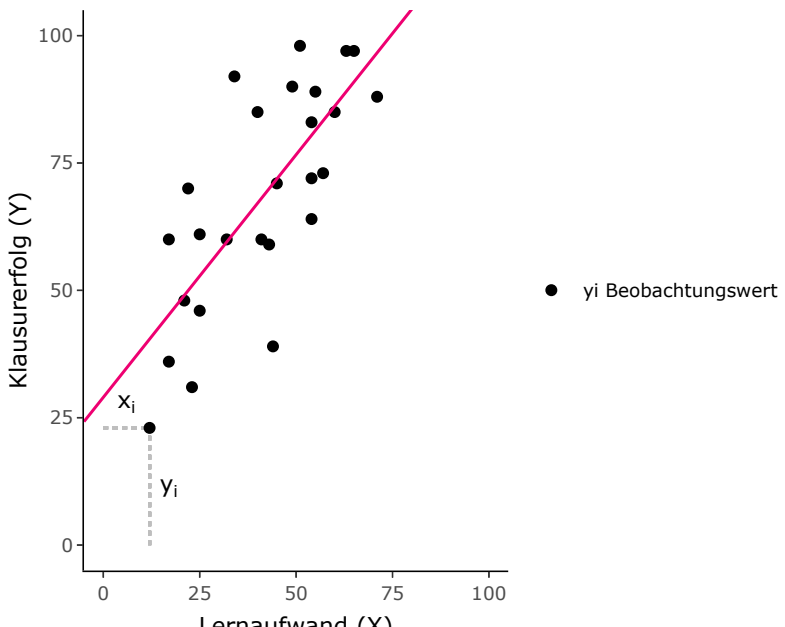
- Theoretisch wären allerdings auch andere Funktionen denkbar.
- Diese beschreiben die vorliegenden Daten ggf. besser, sind aber nicht so leicht interpretierbar/generalisierbar.

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



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

ABER:

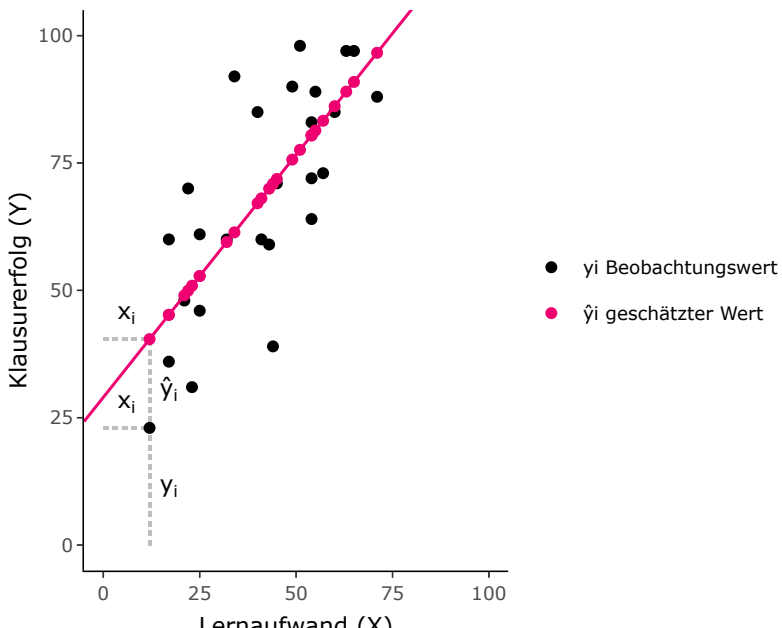
- Für jeden gegebenen X Wert lässt sich ein Punkt auf der Geraden finden, der einen anderen Y Wert hat

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



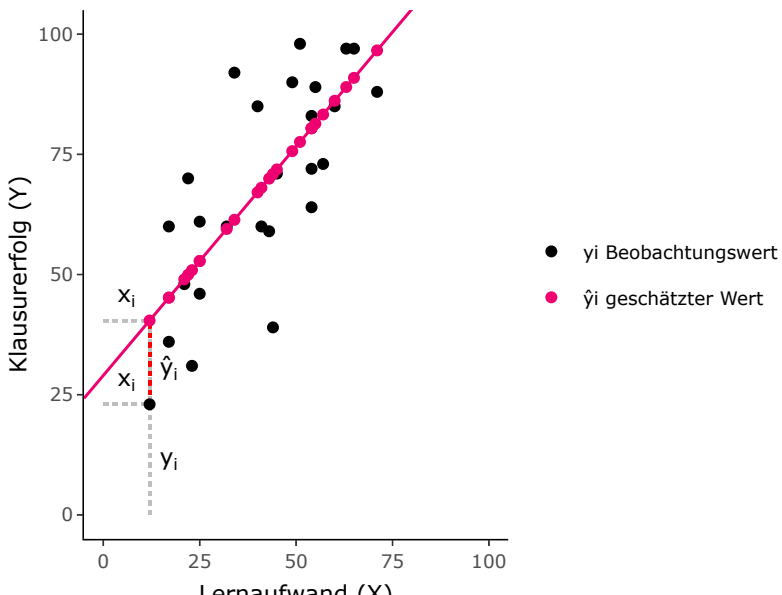
- Der pinke Punkt ist der gemäß der linearen Funktion geschätzte Y Wert für den Punkt X
- Es ist also der Wert, den man unter Annahme eines linearen Zusammenhangs **erwarten** würde
- Diese Punkte haben den X Wert gemeinsam aber sind unterschiedlich im Y Wert.

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



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

Formel für das Residuum:

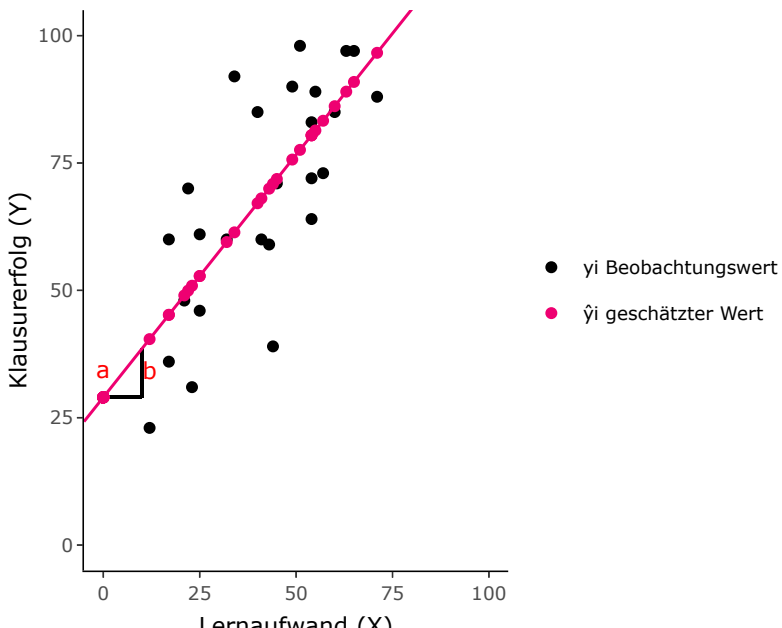
$$\varepsilon_i = \hat{y}_i - y_i$$

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion

$X \rightarrow Y$ Regressionsfunktion und Beobachtungswerte



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

a : Y-Achsenabschnitt
b : Steigungsparameter

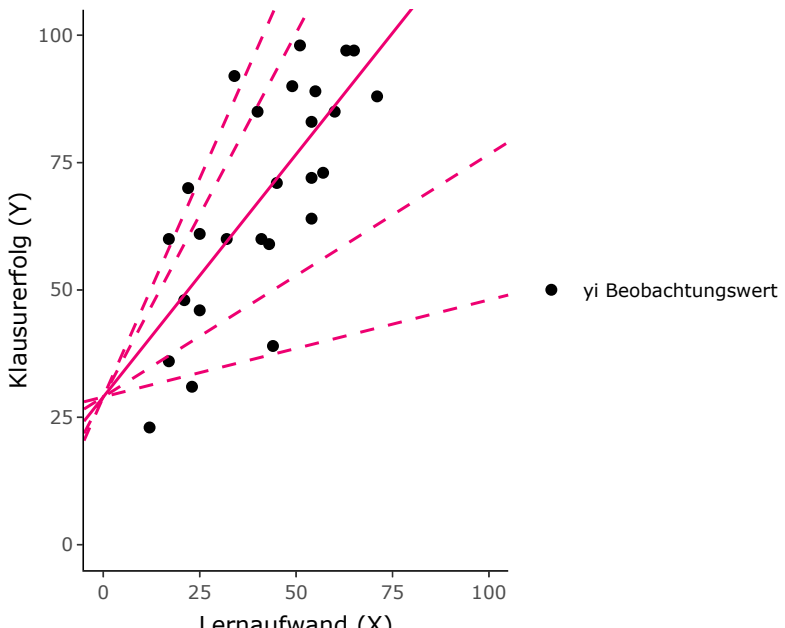
Interpretation:

a : Wert, den Y hat, wenn $X = 0$ ist
b : Veränderung von Y bei Zunahme von X um 1 Einheit

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Residuen und Zielfunktion



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

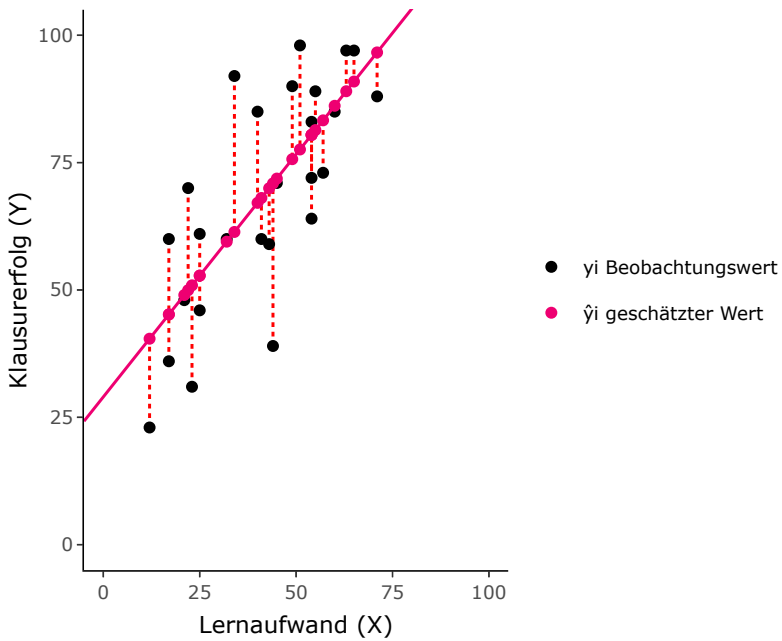
a : Y-Achsenabschnitt
b : Steigungsparameter

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

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Residuen und Zielfunktion



Ziel:

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

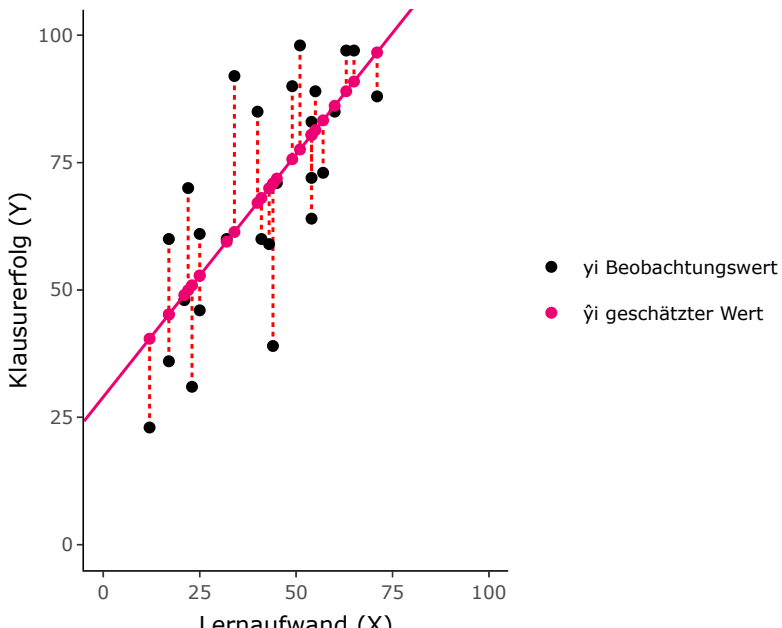
Bildliche Vorstellung:

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

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Residuen und Zielfunktion



Es liegt ein Optimierungsproblem vor:

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

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

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Bestimmung der zu schätzenden Parameter

- Schätzung von a und b $\rightarrow$ **Methode der kleinsten Quadrate**
- Ziel: Summe der quadrierten Residuen minimieren

Analytische Lösung des Optimierungsproblems:

1. Y-Achsenabschnitt (a)

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

1. Y-Steigungsparameter (b)

$$b = \frac{\sigma_{XY}^2}{\sigma_X^2} = r_{XY} \cdot \frac{\sigma_y}{\sigma_x}$$

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Bestimmung der zu schätzenden Parameter

- Schätzung von a und b $\rightarrow$ **Methode der kleinsten Quadrate**
- Ziel: Summe der quadrierten Residuen minimieren

Analytische Lösung des Optimierungsproblems:

1. Y-Achsenabschnitt (a)

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

1. Y-Steigungsparameter (b)

$$b = \frac{\sigma_{XY}^2}{\sigma_X^2} = r_{XY} \cdot \frac{\sigma_y}{\sigma_x}$$

To Do - wir benötigen:

- Mittelwert von X : $\bar{x}$
- Mittelwert von Y : $\bar{y}$
- Kovarianz von XY : σ_{XY}^2
- Varianz von X : σ_X^2

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Bestimmung der zu schätzenden Parameter

X	17	22	44	25	25	17	32	54	21	54	12	34	71	41	23	43	63	60	51	57	55	40	65	45	49	54
Y	60	70	39	46	61	36	60	72	48	64	23	92	88	60	31	59	97	85	98	73	89	85	97	71	90	83

Mittelwert von $X = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Mittelwert von $Y = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Kovarianz von $XY = \sigma_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$

Varianz von $X = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Bestimmung der zu schätzenden Parameter

X	17	22	44	25	25	17	32	54	21	54	12	34	71	41	23	43	63	60	51	57	55	40	65	45	49	54
Y	60	70	39	46	61	36	60	72	48	64	23	92	88	60	31	59	97	85	98	73	89	85	97	71	90	83

Mittelwert von $X = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ = 41.31

Mittelwert von $Y = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ = 68.35

Kovarianz von $XY = \sigma_{xy}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$ = 274.01

Varianz von $X = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ = 287.66

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Bestimmung der zu schätzenden Parameter

X	17	22	44	25	25	17	32	54	21	54	12	34	71	41	23	43	63	60	51	57	55	40	65	45	49	54
Y	60	70	39	46	61	36	60	72	48	64	23	92	88	60	31	59	97	85	98	73	89	85	97	71	90	83

$$b = \frac{\sigma_{xy}^2}{\sigma_x^2} = \frac{274.01}{287.66} = 0.95$$

$$a = \bar{y} - b \cdot \bar{x} = 29$$

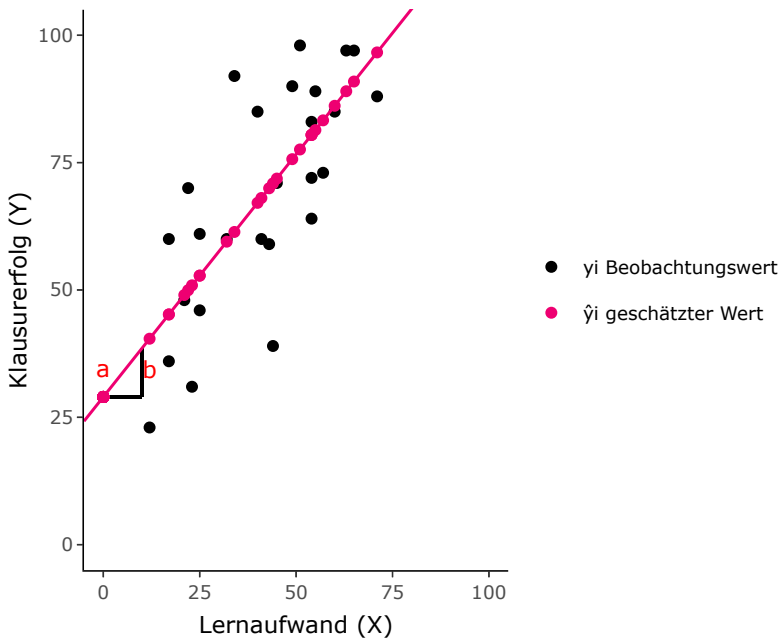
$$y = a + b \cdot x$$

$$y = 29 + 0.95 \cdot x$$

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion



$$\hat{y} = 29 + 0.95 \cdot x$$

- Es ergibt sich also der geschätzte Y Wert $(\hat{y})$ aus a plus b mal x

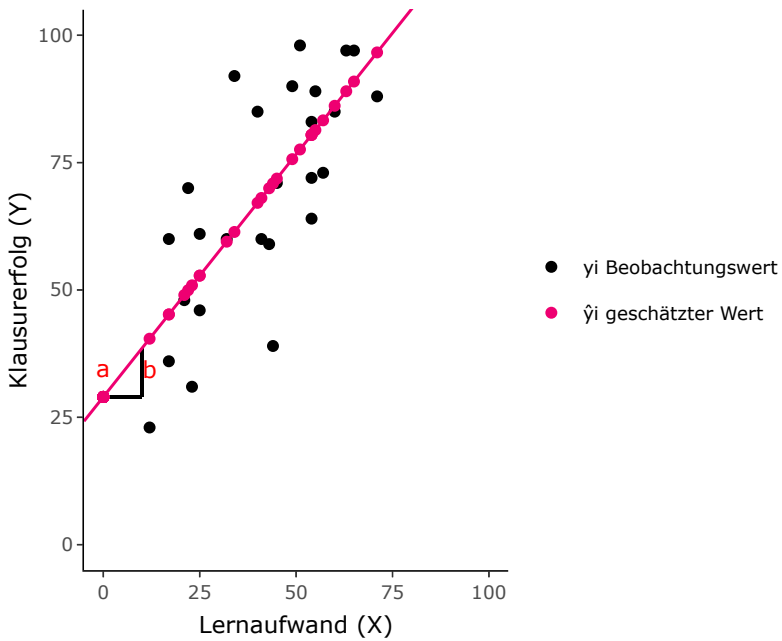
Nochmal zurück zu unserer Interpretation:

- a ist also der Wert wo $X = 0$ ist. Also hat jemand mit 0 auf der UV einen AV Wert von 29.
- Wenn wir nun um 1 Einheit X nach rechts gehen (in welcher Einheit die UV auch immmer gemessen wird), nimmt $\hat{y}$ um b zu.

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Lineare Regressionsfunktion



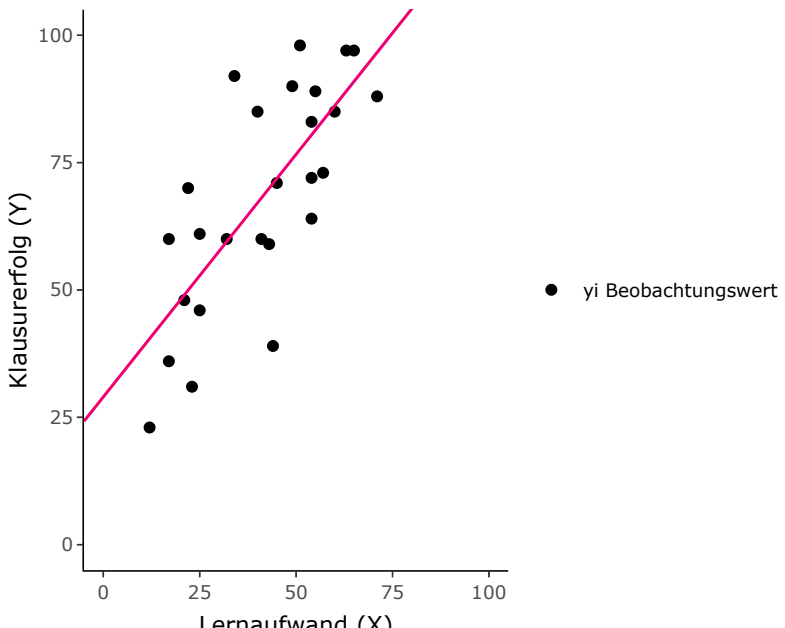
$$\hat{y} = 29 + 0.95 \cdot x$$

- Wir könnten nun ausrechnen, welchen Y Wert eine Person nach x Einheiten der UV hat.
- Welchen Wert erhalten wir z.B. für $X = 10$?

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung



Nach Aufstellen des Modells:

- Abstände zwischen Beobachtungswerten und Regressionsgerade unterschiedlich groß

Frage:

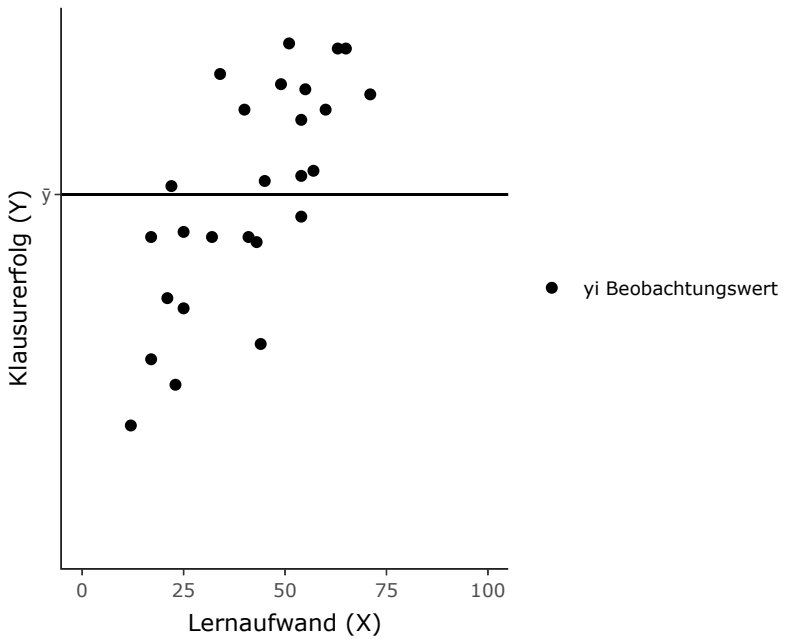
- Wie gut passt unser Modell auf die Beobachtungswerte?
- Maß zur Bestimmung der Passung:

$\rightarrow$ Das Bestimmheitsmaß (R^2)

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung



Frage:

- Wie gut passt unser Modell auf die Beobachtungswerte?
- horizontale Gerade = Mittelwert von Y (um welchen Werte streuen)

$\rightarrow$ **Gesamtvarianz**

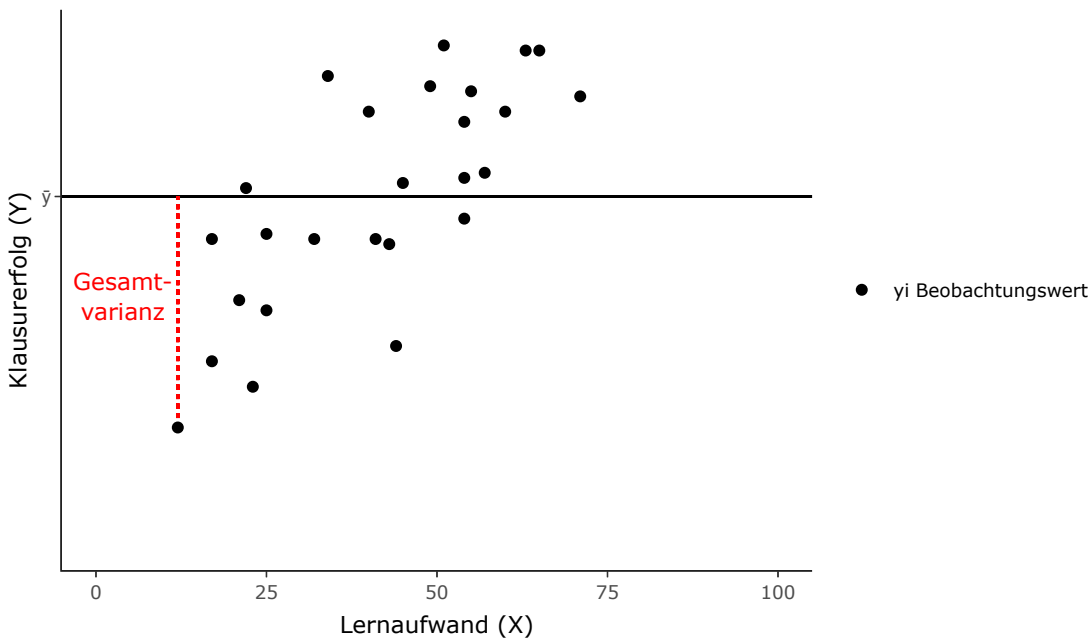
Regressionsgerade kann einen Anteil der Streuung um den Mittelwert erklären:

$\rightarrow$ **Aufgeklärte Varianz**

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

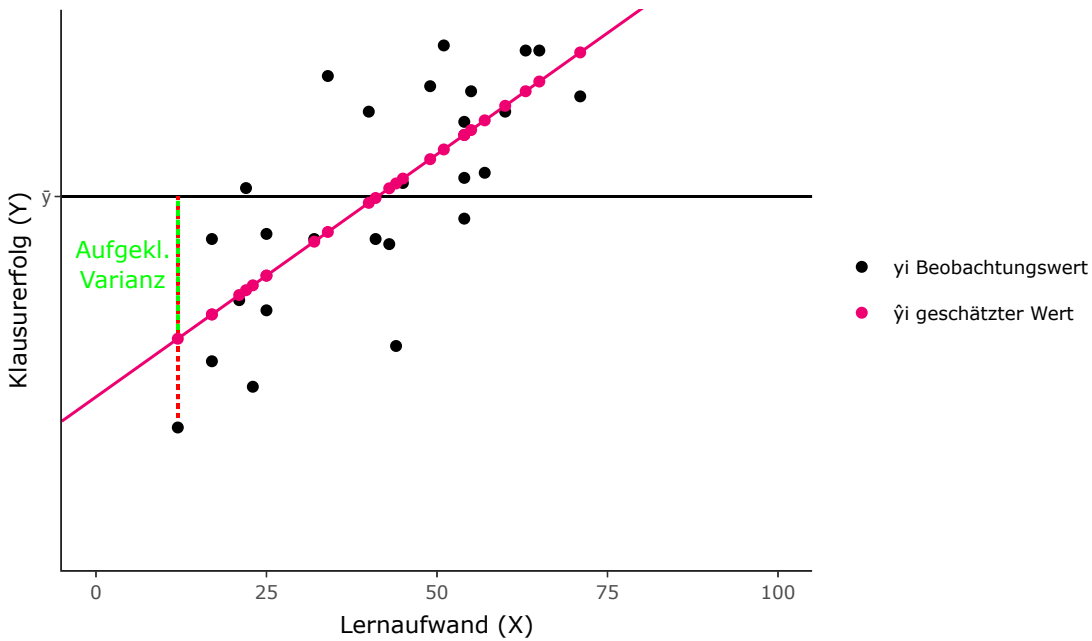
Modellpassung



Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung



Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung

Das Verhältnis aufgeklärter zu gesamter Streuung nennt sich Bestimmtheitsmaß (R^2)

$$R^2 = \frac{\text{erklärte Varianz}}{\text{gesamte Streuung}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = (\frac{s_{XY}}{s_X \cdot s_Y})^2$$

LINK zu interaktivem Regressionsbeispiel

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung

- $0 \leq R^2 \leq 1$
- Je näher R^2 an 1, desto besser passt sich Modell an Beobachtungspunkte an

X	17	22	44	25	25	17	32	54	21	54	12	34	71	41	23	43	63	60	51	57	55	40	65	45	49	54
Y	60	70	39	46	61	36	60	72	48	64	23	92	88	60	31	59	97	85	98	73	89	85	97	71	90	83

$$R^2 = \left(\frac{s_{xy}}{s_x \cdot s_y}\right)^2$$

$$R^2 = \left(\frac{274.01}{16.96 \cdot 21.74}\right)^2 = 0.74$$

Es können 74% der Streuung um den Mittelwert von Y durch die Gerade erklärt werden.

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Einsatz der Regression

Wozu können wir die Regression nutzen?

1. Als **Hypothesentest** für eine wissenschaftliche Hypothese (Inferenz):
    - Schritt 1: Mittels Regression Assoziation in der Stichprobe identifizieren
    - Schritt 2: Mittels Signifikanztest prüfen, ob Assoziation wahrscheinlich auch außerhalb Stichprobe vorliegt
2. Als **Vorhersagemodell** für neue Datenpunkte (Prädiktion):
    - Schritt 1: Mittels Stichprobendaten Regressionsmodell anpassen (X und Y bekannt)
    - Schritt 2: Mittels Modell neue Werte vorhersagen (X bekannt, Y unbekannt).

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Berechnen der Regression in R

```
model = lm(Y ~ X, data = sampledata) # Aufstellen des Modells

summary(model) # Anzeigen des Modelloutputs
```

```
##
## Call:
## lm(formula = Y ~ X, data = sampledata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.911  -9.051  -0.933   8.138  30.615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.9989     7.7979   3.719  0.00107 **
## X             0.9525     0.1751   5.439 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.85 on 24 degrees of freedom
## Multiple R-squared:  0.5521,    Adjusted R-squared:  0.5335
## F-statistic: 29.59 on 1 and 24 DF,  p-value: 1.372e-05
```

Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Berechnen der Regression in R

```
model = lm(Y ~ X, data = sampledata) # Aufstellen des Modells

summary(model) # Anzeigen des Modelloutputs
```

```
##
## Call:
## lm(formula = Y ~ X, data = sampledata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.911  -9.051  -0.933   8.138  30.615
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.9989     7.7979   3.719  0.00107 **
## X             0.9525     0.1751   5.439 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.85 on 24 degrees of freedom
## Multiple R-squared:  0.5521,    Adjusted R-squared:  0.5335
## F-statistic: 29.59 on 1 and 24 DF,  p-value: 1.372e-05
```

- Regressionskoeffizienten (a und b) stehen in der Spalte "Estimate"
- Für jeden Koeffizienten wird ein spezieller t-Test (Wald-Test) gerechnet
  - H_0 a und b = 0
  - H_1 a und b $\neq$ 0

Take-aways

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

- Zusammenhänge können neben Kovarianz/Korrelation auch mit der **Regression** quantifiziert werden.
- Regression ist sinnvoll, wenn aus den X-Werten auf die dazugehörigen Y-Werte **geschlossen** (diese vorhergesagt) werden soll.
- **Regressionsgerade** = graphische Veranschaulichung der Regressionsgleichung
- Regressionsgleichung ist definiert durch die **Regressionskoeffizienten** (Y-Achsenabschnitt und Steigung), welche aus Daten geschätzt werden müssen.
- **Y-Achsenabschnitt** ist der Startwert (wenn X = 0) und **Steigung** ist die Veränderung in der AV bei Zunahme der UV um 1 Einheit.
- Das **Bestimmtheitsmaß** (R^2) gibt an, wie viel Prozent (%) der Gesamtvarianz der AV durch die UV (also durch die das Regressionsmodell) aufgeklärt werden.