Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

Fragen:

- Wie hängen `x` und `y` zusammen?

Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

Fragen:

- Wie hängen `x` und `y` zusammen?

  Korrelation: `cor(x,y)` $= 0.97$

  Positiver Zusammenhang! 🧐

Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

Fragen:

- Wie hängen `x` und `y` zusammen?

  Korrelation: `cor(x,y)` $= 0.97$
  Positiver Zusammenhang!

- Wie können wir einen bestimmten Wert von `y` vorhersagen?

Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

Fragen:

- Wie hängen `x` und `y` zusammen?

  Korrelation: `cor(x,y)` $= 0.97$
  Positiver Zusammenhang! 🧐

- Wie können wir einen bestimmten Wert von `y` vorhersagen? 🤔

  Wir brauchen ein Modell!

Ausgangssituation

Wir haben Daten erhoben:

x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

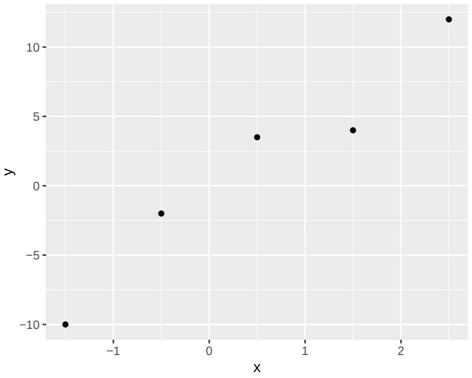
Fragen:

- Wie hängen `x` und `y` zusammen?

  Korrelation: `cor(x,y)` $= 0.97$
  Positiver Zusammenhang! 🧐

- Wie können wir einen bestimmten Wert von `y` vorhersagen? 🤔

  Wir brauchen ein Modell!



Ein lineares Modell scheint passend...

Lineares Modell

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n$$

Lineares Modell

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n$$

- i ist der Index für eine Beobachtung
- n ist die Anzahl Beobachtungen
- y_i ist die abhängige (zu erklärende) Variable für die Beobachtung i
- x_i ist der Regressor (die erklärende Variable) für die Beobachtung i
- u_i ist der Fehlerterm (der Messfehler) für die Beobachtung i
- β_0 und β_1 sind unbekannte Parameter, die geschätzt werden
  - β_0 ist der Achsenabschnitt (auch Intercept genannt)
  - β_1 ist der Steigungsparameter

Modellannahmen

Modellannahmen

SLR.1 **Modell** In der Population hängt die abhängige Variable y von dem Regressor x und dem Fehler u in folgender Form ab:

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0,\ \beta_1 \in \mathbb{R}.$$

Modellannahmen

SLR.1 **Modell** In der Population hängt die abhängige Variable y von dem Regressor x und dem Fehler u in folgender Form ab:

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0,\ \beta_1 \in \mathbb{R}.$$

SLR.2 **Stichprobe** Die Daten $\{(y_i, x_i), i = 1, \ldots, n\}$ sind die Realisation einer Zufallsstichprobe von Zufallsvariablen, die gemäß dem Modell in Annahme SLR.1 generiert wurden.

Modellannahmen

SLR.1 **Modell** In der Population hängt die abhängige Variable y von dem Regressor x und dem Fehler u in folgender Form ab:

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0,\ \beta_1 \in \mathbb{R}.$$

SLR.2 **Stichprobe** Die Daten $\{(y_i, x_i), i = 1, \dots, n\}$ sind die Realisation einer Zufallsstichprobe von Zufallsvariablen, die gemäß dem Modell in Annahme SLR.1 generiert wurden.

SLR.3 **Information im Regressor** Die Daten $\{x_i, i = 1, \dots, n\}$ sind nicht alle gleich.

Modellannahmen

SLR.1 **Modell** In der Population hängt die abhängige Variable y von dem Regressor x und dem Fehler u in folgender Form ab:

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0,\ \beta_1 \in \mathbb{R}.$$

SLR.2 **Stichprobe** Die Daten $\{(y_i, x_i), i = 1, \dots, n\}$ sind die Realisation einer Zufallsstichprobe von Zufallsvariablen, die gemäß dem Modell in Annahme SLR.1 generiert wurden.

SLR.3 **Information im Regressor** Die Daten $\{x_i, i = 1, \dots, n\}$ sind nicht alle gleich.

SLR.4 **Bedingte Erwartung** Der Erwartungswert des Fehlers u bedingt auf x ist null:

$$\mathbb{E}(u \mid x) = 0.$$

Modellannahmen

SLR.1 **Modell** In der Population hängt die abhängige Variable y von dem Regressor x und dem Fehler u in folgender Form ab:

$$y = \beta_0 + \beta_1 x + u, \quad \beta_0,\ \beta_1 \in \mathbb{R}.$$

SLR.2 **Stichprobe** Die Daten $\{(y_i, x_i), i = 1, \dots, n\}$ sind die Realisation einer Zufallsstichprobe von Zufallsvariablen, die gemäß dem Modell in Annahme SLR.1 generiert wurden.

SLR.3 **Information im Regressor** Die Daten $\{x_i, i = 1, \dots, n\}$ sind nicht alle gleich.

SLR.4 **Bedingte Erwartung** Der Erwartungswert des Fehlers u bedingt auf x ist null:

$$\mathbb{E}(u \mid x) = 0.$$

SLR.5 **Homoskedastizität** Der Fehler u hat für jeden Wert des Regressors x die gleiche Varianz:

$$\mathrm{Var}(u \mid x) = \sigma^2.$$

Beispiele, bei denen die Annahmen verletzt sind

Beispiele, bei denen die Annahmen verletzt sind

SLR.1 **Modell** Der Fehler ist multiplikativ: $y = (\beta_0 + \beta_1 x)u$.

Beispiele, bei denen die Annahmen verletzt sind

SLR.1 **Modell** Der Fehler ist multiplikativ: $y = (\beta_0 + \beta_1 x)u$.

SLR.2 **Stichprobe** Die Stichprobe ist nicht zufällig aus der Grundgesamtheit der Wahlberechtigten gezogen, weil nur Studierende befragt wurden.

Beispiele, bei denen die Annahmen verletzt sind

SLR.1 **Modell** Der Fehler ist multiplikativ: $y = (\beta_0 + \beta_1 x)u$.

SLR.2 **Stichprobe** Die Stichprobe ist nicht zufällig aus der Grundgesamtheit der Wahlberechtigten gezogen, weil nur Studierende befragt wurden.

SLR.3 **Information im Regressor** Das Gehalt soll anhand des Bildungsgrades untersucht werden, die Umfrage wurde aber ausschließlich unter Bachelorabsolventen gemacht.

Beispiele, bei denen die Annahmen verletzt sind

SLR.1 **Modell** Der Fehler ist multiplikativ: $y = (\beta_0 + \beta_1 x)u$.

SLR.2 **Stichprobe** Die Stichprobe ist nicht zufällig aus der Grundgesamtheit der Wahlberechtigten gezogen, weil nur Studierende befragt wurden.

SLR.3 **Information im Regressor** Das Gehalt soll anhand des Bildungsgrades untersucht werden, die Umfrage wurde aber ausschließlich unter Bachelorabsolventen gemacht.

SLR.4 **Bedingte Erwartung** Es besteht ein quadratischer Zusammenhang zwischen y und x:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + (\beta_2 x^3 + u) \\ &= \beta_0 + \beta_1 x + \tilde{u} \end{aligned}$$

Dann gilt: $\mathbb{E}(\tilde{u} \mid x) = \mathbb{E}(\beta_2 x^3 + u \mid x) = \mathbb{E}(\beta_2 x^3 \mid x) + \mathbb{E}(u \mid x) = \beta_2 x^3$

Beispiele, bei denen die Annahmen verletzt sind

SLR.1 **Modell** Der Fehler ist multiplikativ: $y = (\beta_0 + \beta_1 x)u$.

SLR.2 **Stichprobe** Die Stichprobe ist nicht zufällig aus der Grundgesamtheit der Wahlberechtigten gezogen, weil nur Studierende befragt wurden.

SLR.3 **Information im Regressor** Das Gehalt soll anhand des Bildungsgrades untersucht werden, die Umfrage wurde aber ausschließlich unter Bachelorabsolventen gemacht.

SLR.4 **Bedingte Erwartung** Es besteht ein quadratischer Zusammenhang zwischen y und x:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + (\beta_2 x^3 + u) \\ &= \beta_0 + \beta_1 x + \tilde{u} \end{aligned}$$

Dann gilt: $\mathbb{E}(\tilde{u} \mid x) = \mathbb{E}(\beta_2 x^3 + u \mid x) = \mathbb{E}(\beta_2 x^3 \mid x) + \mathbb{E}(u \mid x) = \beta_2 x^3$

SLR.5 **Homoskedastizität** Modellierung der Körpergröße in Abhängigkeit des Alters: Die Körpergröße von Neugeborenen (24cm - 57,5cm) hat eine geringere Varianz als die von Erwachsenen (54,6cm - 251cm).