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:

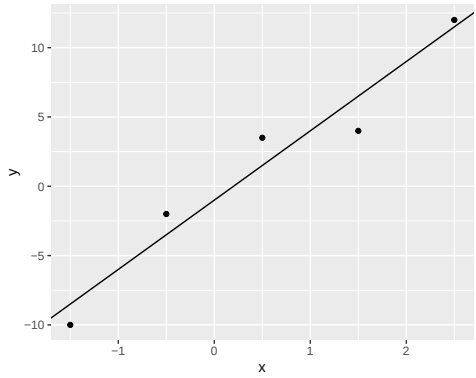
x	-0.5	-1.5	0.5	1.5	2.5
y	-2.0	-10.0	3.5	4.0	12.0

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 (**S**imple **L**inear **R**egression Model)

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

Lineares Modell (**S**imple **L**inear **R**egression Model)

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

- 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

5 Modellannahmen

	Stichwort	Annahme	wäre z.B. verletzt, wenn...
SLR.1	**Modell**	$y = \beta_0 + \beta_1 x + u$ mit den Parametern $\beta_0,\ \beta_1 \in \mathbb{R}$	der Zusammenhang ist exponentiell ($y = \beta_0 e^{\beta_1 x} + u$)
SLR.2	**Stichprobe**	$\{(y_i, x_i), i = 1, \dots, n\}$ zufällig gemäß SLR.1 generiert	aus der Grundgesamtheit der Wahlberechtigten wurden nur Studierende befragt
SLR.3	**Information im Regressor**	$\text{Var}(x) > 0$	Experiment immer mit den exakt gleichen Parametern durchgeführt
SLR.4	**Bedingte Erwartung**	$\mathbb{E}(u \mid x) = 0$	es gibt einen systematischen Messfehler
SLR.5	**Homoskedastizität**	$\text{Var}(u \mid x) = \sigma^2$	die Körpergröße von Kleinkindern hat eine geringere Varianz als die von Erwachsenen

Schätzwerte bestimmen

"Kleinste Quadrate" Methode:

$$\arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$$

Schätzwerte bestimmen

"Kleinste Quadrate" Methode:

$$\arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$$

Schätzer:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = \frac{\mathsf{Cov}(Y, X)}{\mathsf{Var}(X)}$$

Mehrere Regressoren (**M**ultiple **L**inear **R**egression Model)

Wir sind nicht auf nur einen Regressor beschränkt:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_K x_{i,K} + u_i, \quad i = 1, ..., n$$

Mehrere Regressoren (**M**ultiple **L**inear **R**egression Model)

Wir sind nicht auf nur einen Regressor beschränkt:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_K x_{i,K} + u_i, \quad i = 1, ..., n$$

Wir werden uns anschauen:

- Wie verändern sich die Annahmen?
- Wie werden hier die Parameter geschätzt? Und was bedeutet *teris paribus*?
- Was ist die Normalgleichung? Und was bedeutet Multikollinearität?
- Wie führt man eine multiple Regression in `R` durch? Wie interpretiert man den `R` Output?
- Und ganz wichtig: Was besagt das Gauss-Markov-Theorem?

Mehr dazu nächste Woche! 👍