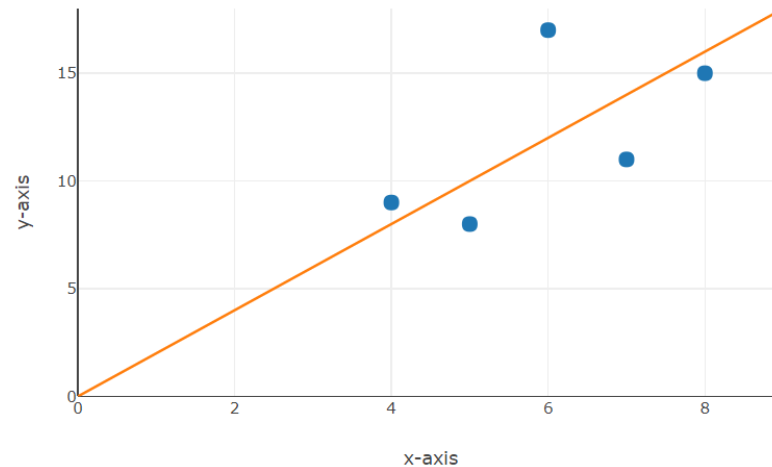


Teil 6: Regressionen

Linear Regression

Lineare Regression

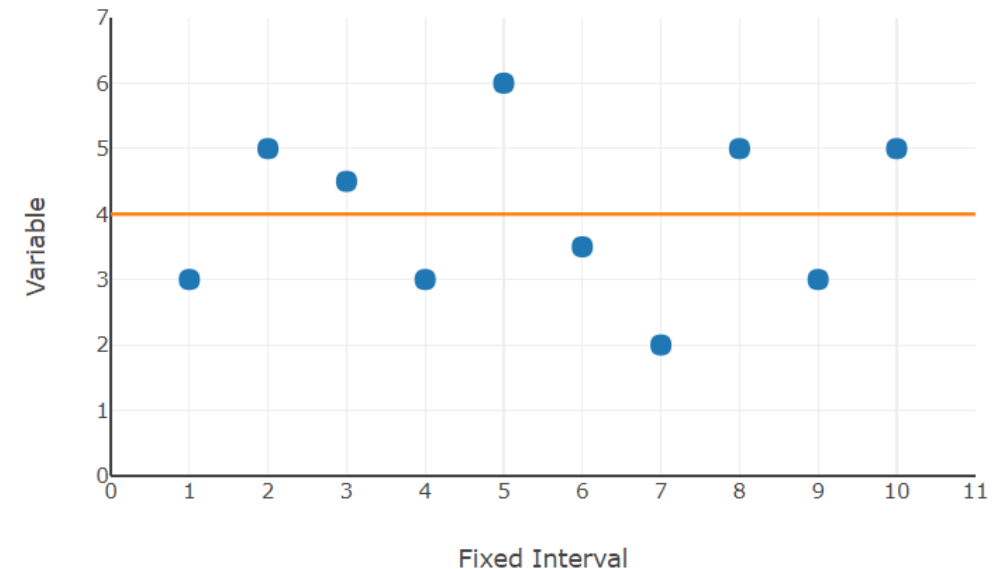
- Das Ziel der **Regression** ist das Modellieren einer Gleichung oder Formel, die die Beziehung zwischen Variablen **am besten beschreibt**.



$$y = 2x$$

Lineare Regression

- Wie finden wir eine Regressionsgerade (**Best-Fit-Linie**)?
- Betrachten wir ein Dataset mit nur einer Variablen
- Die am besten passende Linie ist hier nur der Mittelwert der Datenpunkte



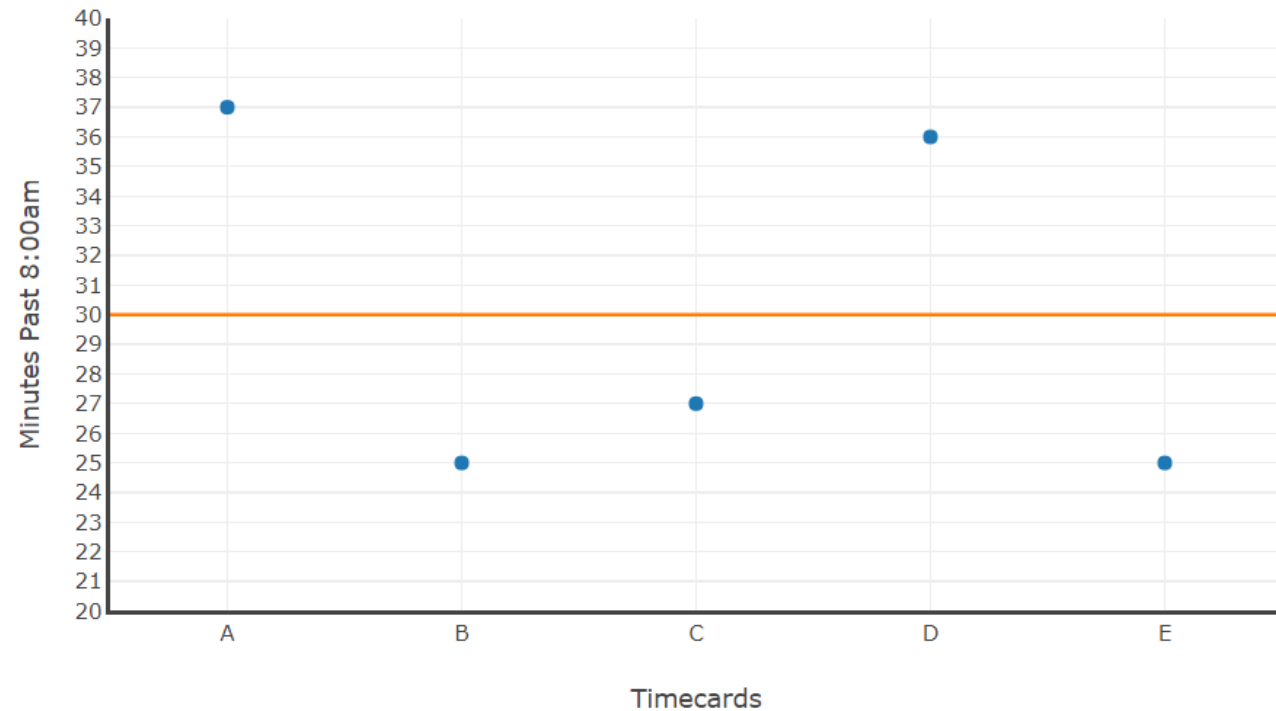
“Best fit” Verstehen

- Eine Betriebsleiterin möchte wissen, wann ihre Mitarbeiter zur Arbeit erscheinen
- Die Schicht beginnt um 8:30 Uhr
- Sie nimmt fünf zufällige Zeitkarten und zeichnet die Ankunftsminuten in ein Diagramm ein



“Best fit” Verstehen

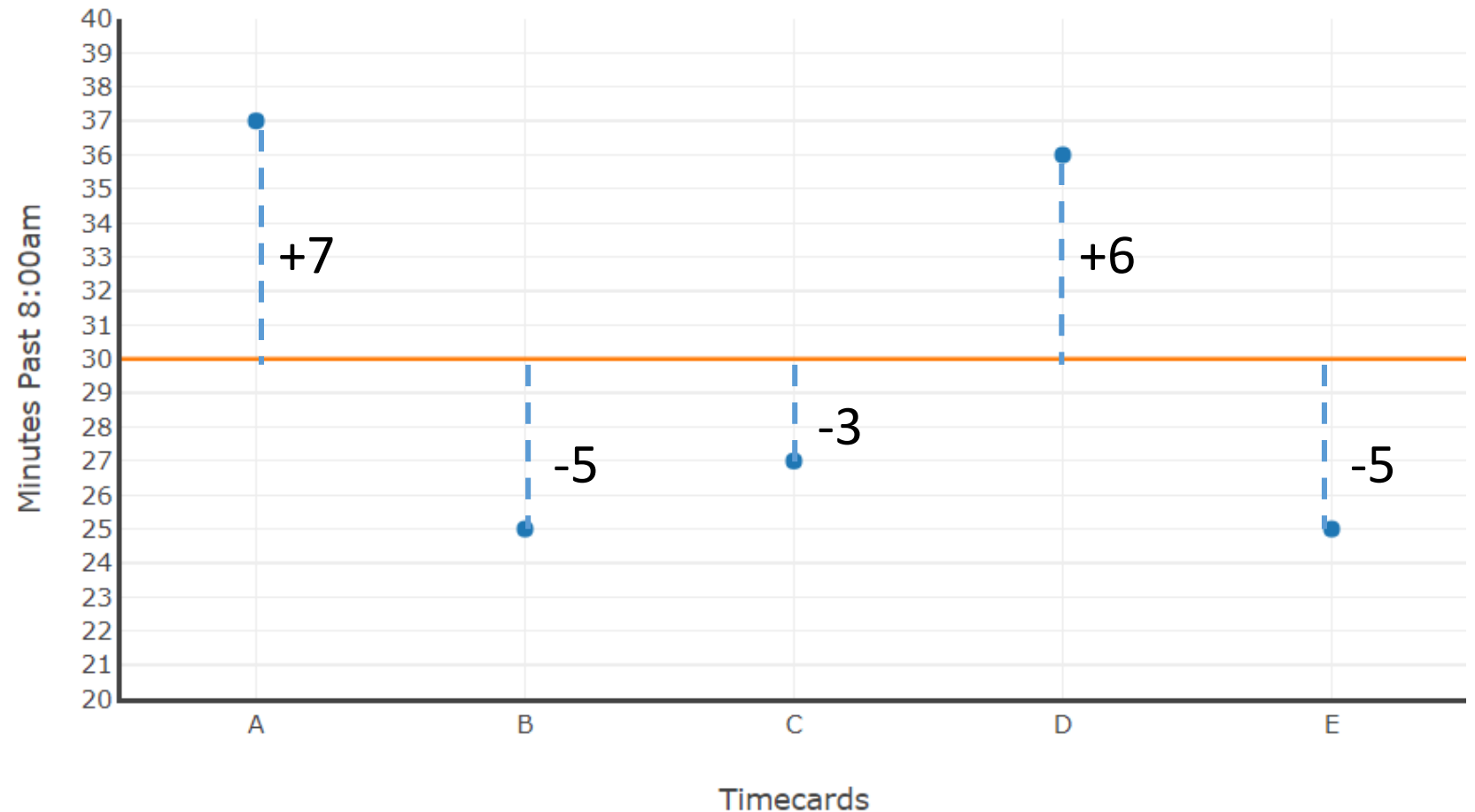
Stempel karte	Minuten nach 8:00am
A	37
B	25
C	27
D	36
E	25
Total:	150
Mittel	30



“Best fit” Verstehen

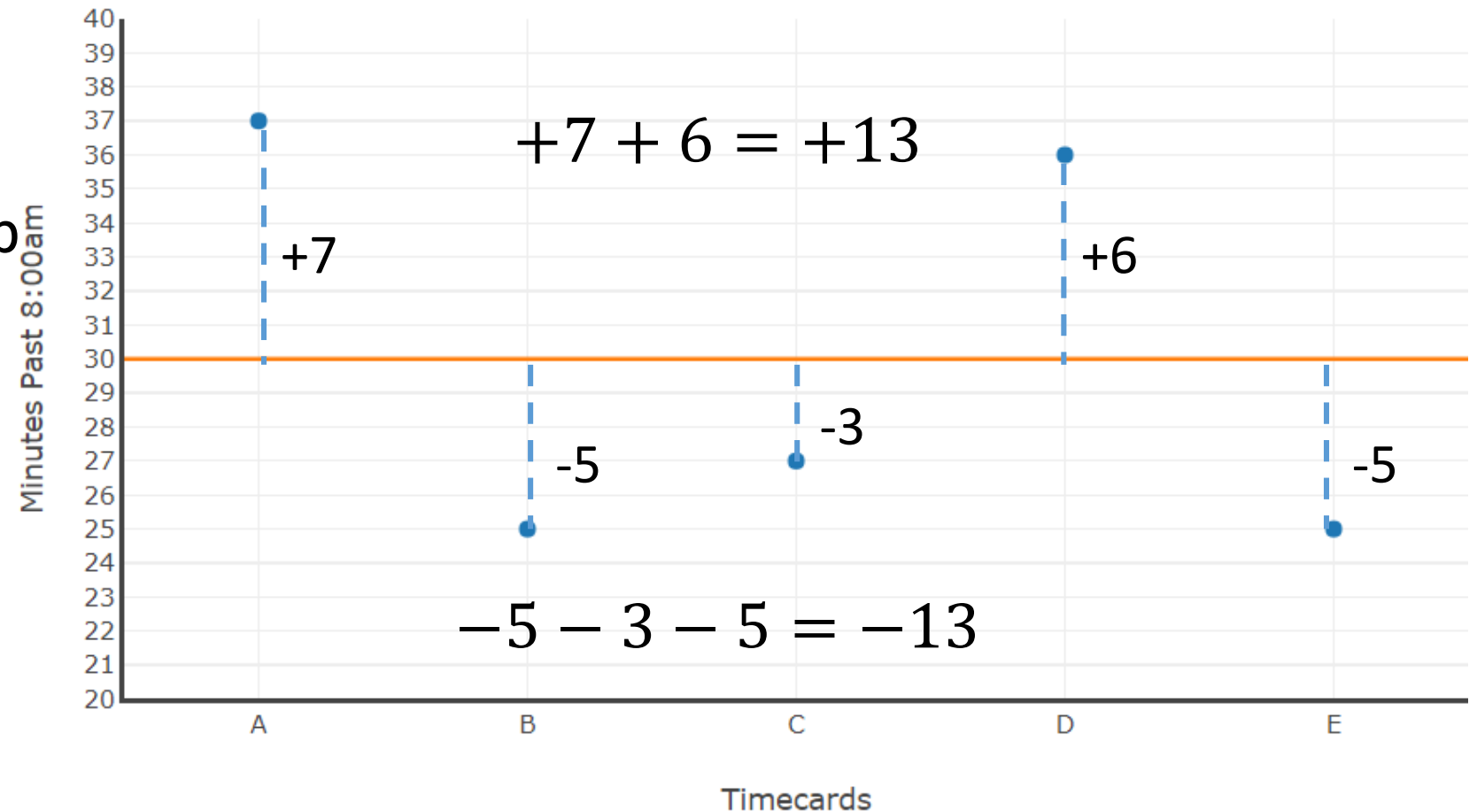
Was macht
 $y = 30$ zur
best-fit line?

Betrachten wir die
Abweichungen
(Residuen)



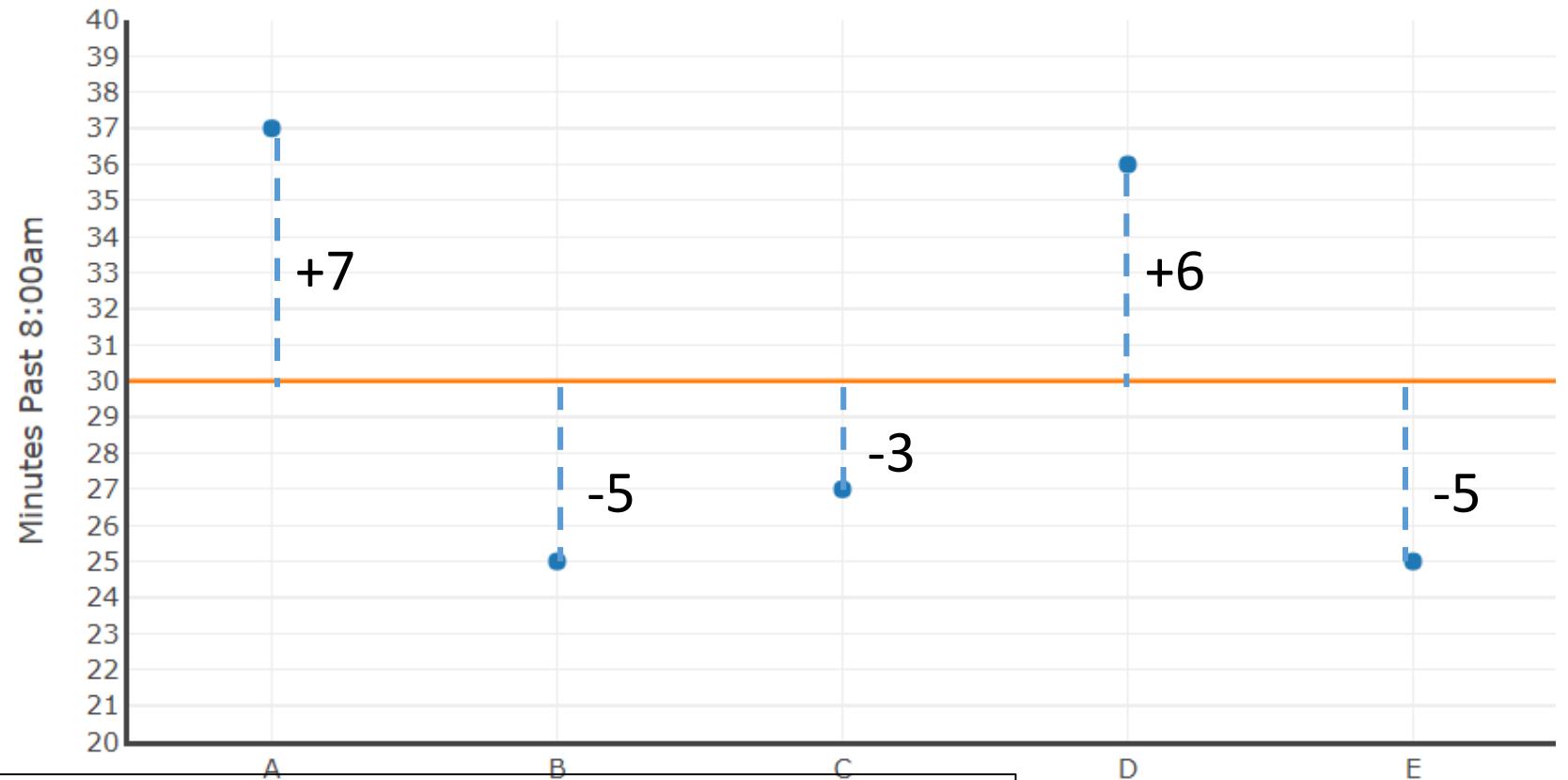
“Best fit” Verstehen

Wir sehen, dass die Summe der „Abstände“ oberhalb der Geraden, die Summe derjenigen unter der Geraden ausgleicht



“Best fit” Verstehen

Abweichung (E)	Abweichung im Quadrat(SE)
+7	49
-5	25
-3	9
+6	36
-5	25
Quadrat-summe der Abweichungen (SSE)	144

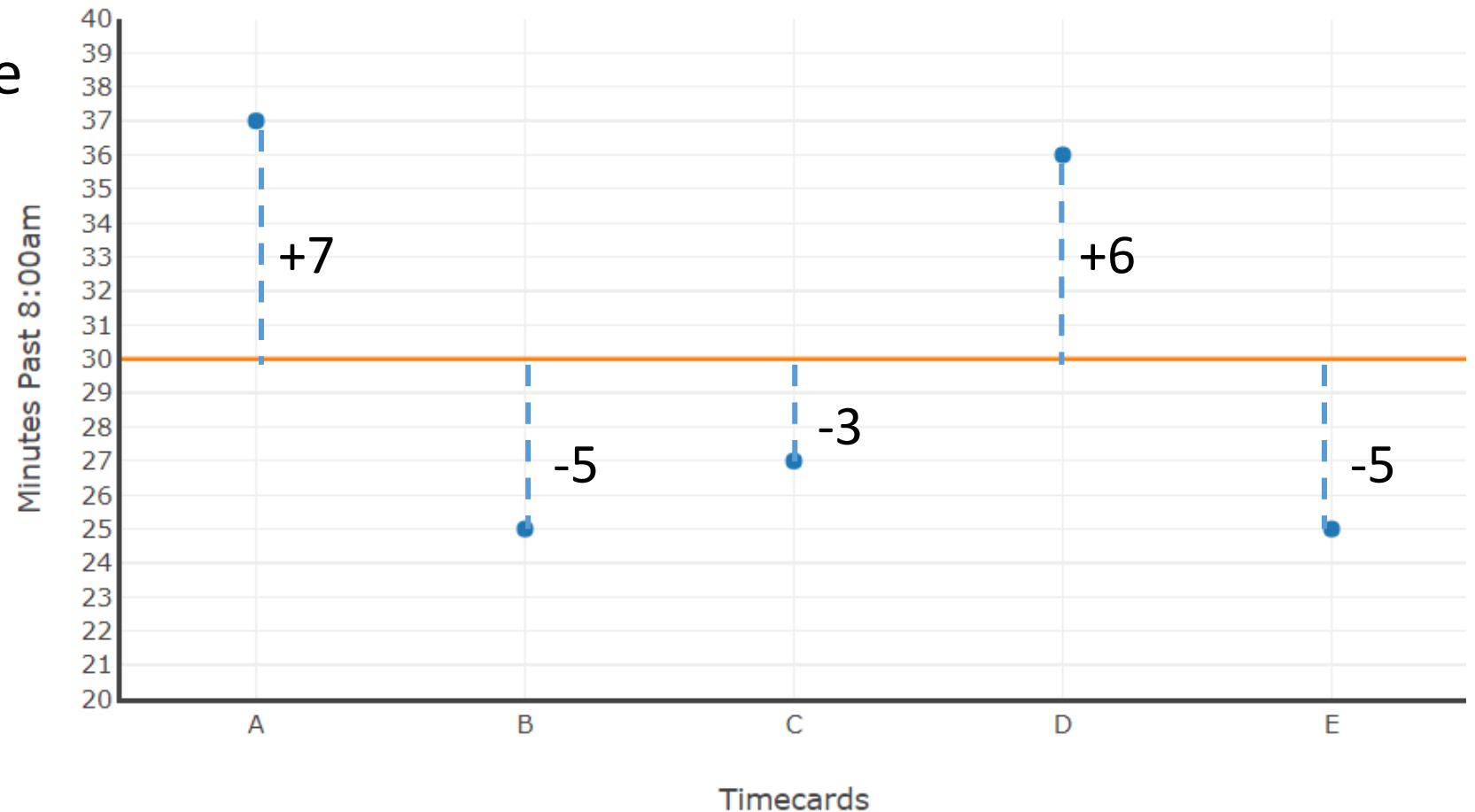


Wir wollen, dass SSE so klein wie möglich ist

“Best fit” Verstehen

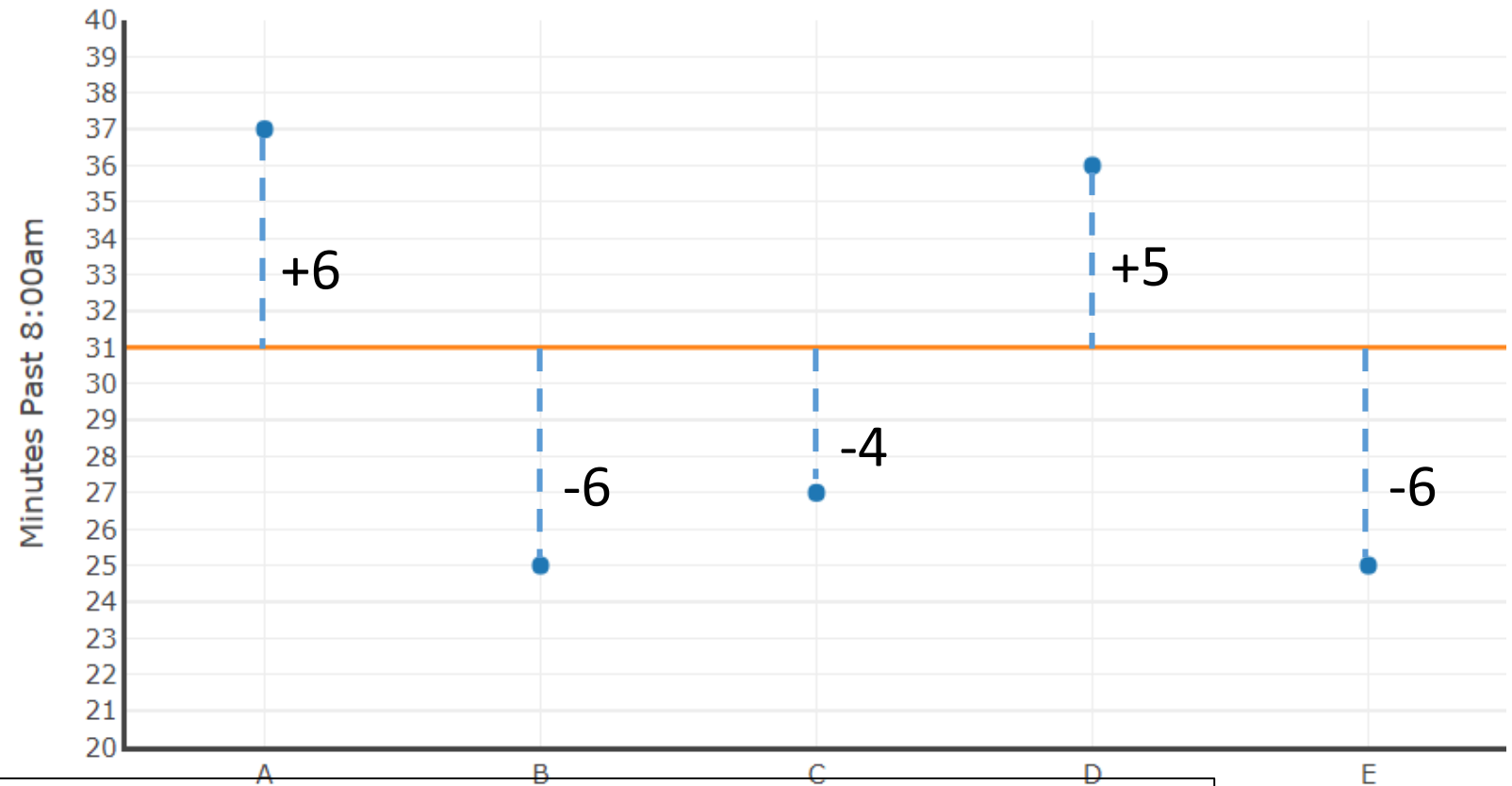
Was ist, wenn wir die
Linie verschieben?
Setzen wir sie
stattdessen auf $y =$
31

Wie wirkt sich das
auf SSE aus?



“Best fit” Verstehen

Abweichung (E)		Abweichung im Quadrat(SE)	
+7	+6	49	36
-5	-6	25	36
-3	-4	9	16
+6	+5	36	25
-5	-6	25	36
Quadrat-summe der Abweichungen (SSE)		144	149

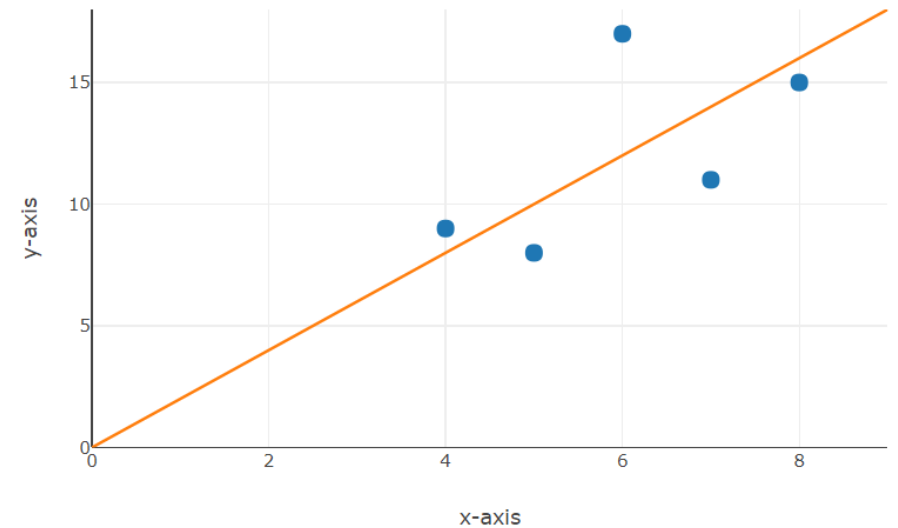


SEE wird beim Verschieben der Linie **größer**

Timecards

Lineare Regression

- Das ist es! Ziel der Regression ist es, die Gerade zu finden, die unsere Daten am besten beschreibt.
- Glücklicherweise müssen wir uns nicht aufs Ausprobieren verlassen.
- Wir haben Algebra!

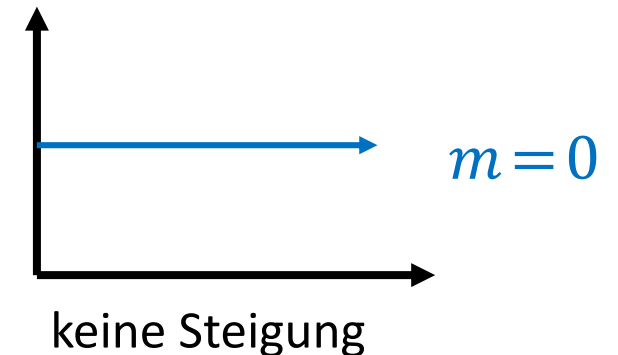
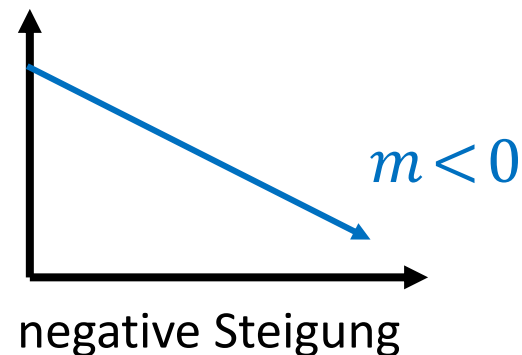
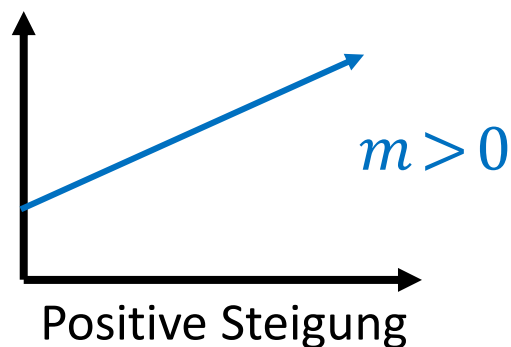


Lineare Regression

- Erinnern wir uns an die Gleichung einer Geraden

$y = mx + b$ bei der

- m ist die **Steigung** darstellt und
- b der Punkt ist, an dem die Gerade die y-Achse schneidet
wenn $x = 0$ (b ist der y-Achsenabschnitt)



Lineare Regression

- In einer linearen Regression, in der wir versuchen, die Beziehung zwischen Variablen zu formulieren, wird $y = mx + b$

$$\hat{y} = b_0 + b_1x$$

- Unser Ziel ist es, den Wert einer **abhängigen Variablen (y)** auf der Basis einer **unabhängigen Variablen (x)** vorherzusagen.

Lineare Regression

$$\hat{y} = b_0 + b_1 x$$

- Wie man b_1 and b_0 ableitet:

$$b_1 = \rho_{x,y} \frac{\sigma_y}{\sigma_x}$$

$\rho_{x,y}$ = Pearson Korrelationskoeffizienz
 σ_x, σ_y = Standardabweichung

$$= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}} \cdot \frac{\sqrt{\frac{\sum (y - \bar{y})^2}{n}}}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Lineare Regression

$$\hat{y} = b_0 + b_1 x$$

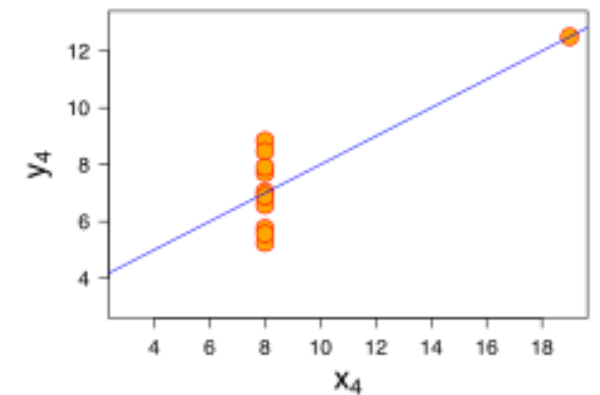
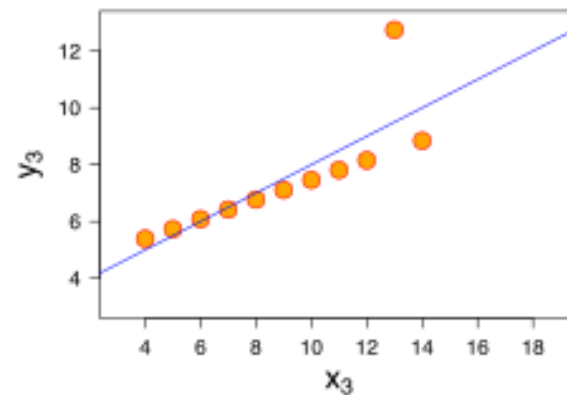
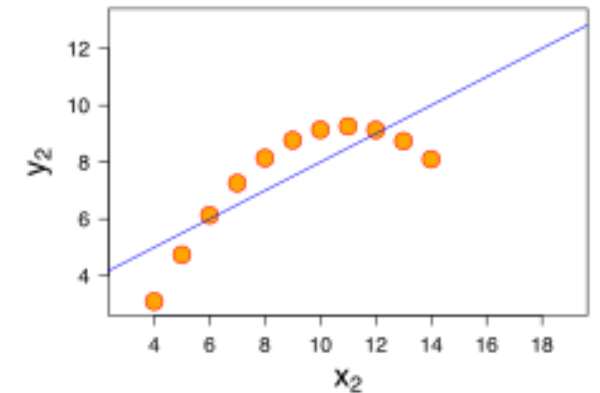
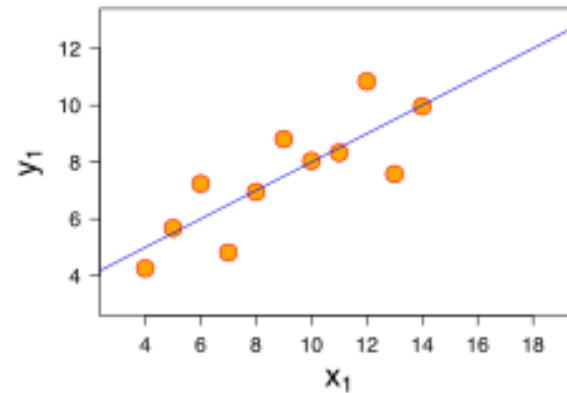
- Wie man b_1 and b_0 ableitet:

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

Einschränkungen der linearen Regression

- Das Anscombe-Quartett zeigt die Fallen auf, wenn man sich auf reine Berechnungen verlässt.
- Jedes Diagramm führt zur selben, berechnete Regressionsgeraden.



Lineare Regression

Beispiel

Regression Übung #1

- Ein Manager möchte die Beziehung zwischen der Anzahl der Stunden, die eine Anlage pro Woche in Betrieb ist, und der wöchentlichen Produktion herausfinden.



Regression Übung #1

- Hier ist die **unabhängige Variable** x die Betriebsstunden und die **abhängige Variable** y ist das Produktionsvolumen.



Regression Übung #1

- Der Manager entwickelt die folgende Tabelle:

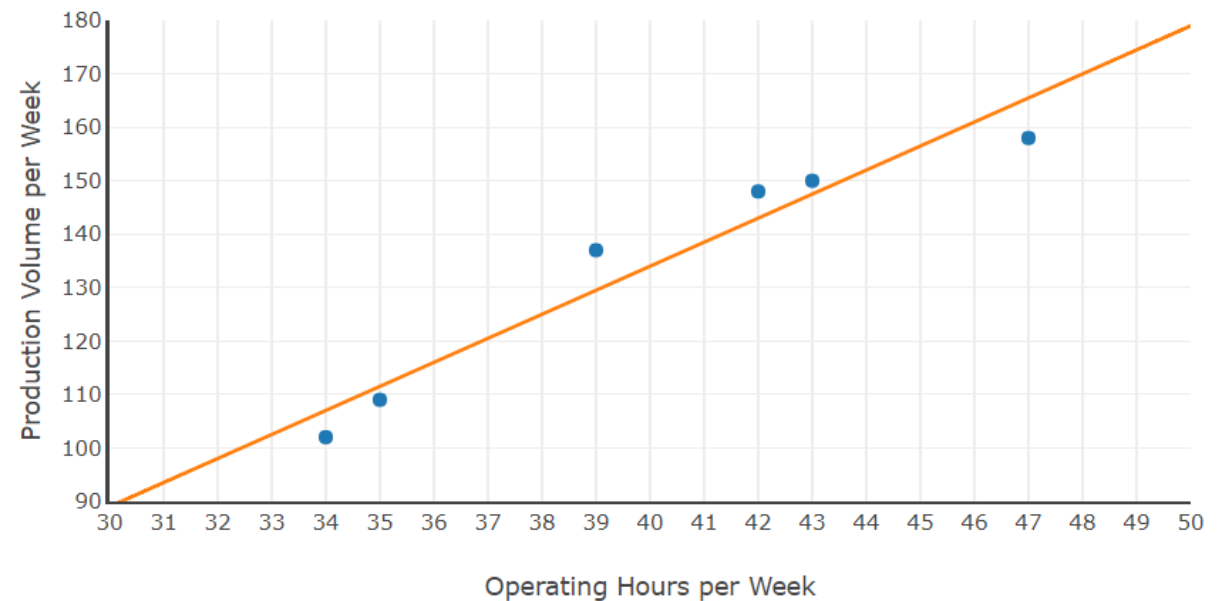
Produktions- stunden (x)	Produktions- volumen (y)
34	102
35	109
39	137
42	148
43	150
47	158

Regression Übung #1

- Stelle die Daten zunächst grafisch dar

Produktions- stunden (x)	Produktions- volumen (y)
34	102
35	109
39	137
42	148
43	150
47	158

gibt es einen linearen Zusammenhang?



Regression Übung #1

$$\hat{y} = b_0 + b_1 x$$

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Führe nun die Kalkulationen durch

	Produktions- stunden (x)	Produktions- volumen (y)	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	34	102	-6	-32	192	36
	35	109	-5	-25	125	25
	39	137	-1	3	-3	1
	42	148	2	14	28	4
	43	150	3	16	48	9
	47	158	7	24	168	49
	\bar{x}, \bar{y}	40	134	Summe:		558
					$\Sigma(x - \bar{x})(y - \bar{y})$	$\Sigma(x - \bar{x})^2$

Regression Übung #1

$$\begin{aligned}\hat{y} &= b_0 + b_1 x \\ b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1 \bar{x}\end{aligned}$$

- Führe nun die Kalkulationen durch

Produktions- stunden (x)	Produktions- volumen (y)
34	102
35	109
39	137
42	148
43	150
47	158
\bar{x}, \bar{y}	40
	134

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{558}{124} = \mathbf{4.5}$$

$$b_0 = \bar{y} - b_1 \bar{x} = 134 - (4.5 \times 40) = \mathbf{-46}$$

$$\hat{y} = \mathbf{-46 + 4.5x}$$

Summe:	558	124
	$\sum (x - \bar{x})(y - \bar{y})$	$\sum (x - \bar{x})^2$

Regression Übung #1

- Basierend auf dieser Formel, kann man nun berechnen, wie lange die Anlage laufen sollten, wenn der Manager 125 Einheiten pro Woche produzieren möchte:

Produktions- stunden (x)	Produktions- volumen (y)
34	102
35	109
39	137
42	148
43	150
47	158

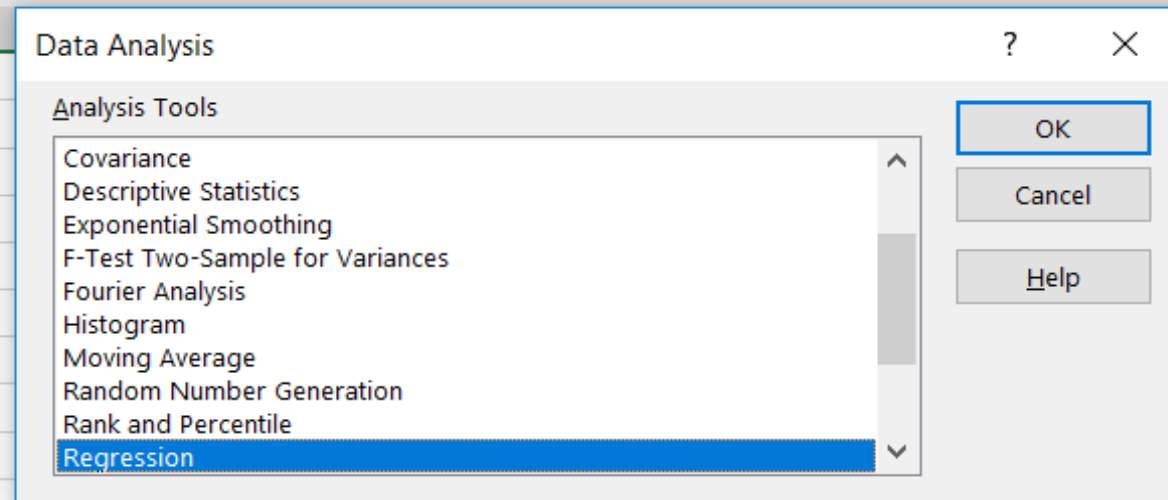
$$\hat{y} = b_0 + b_1 x$$

$$125 = -46 + 4.5x$$

$$x = \frac{171}{4.5} = \mathbf{38 \text{ Stunden pro Woche}}$$

Regressionen mit Excel Data Analysis

	A	B	C						
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.966875047							
5	R Square	0.934847357							
6	Adjusted R Square	0.918559196							
7	Standard Error	6.614378278							
8	Observations	6							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	2511	2511	57.39428571	0.00162772			
13	Residual	4	175	43.75					
14	Total	5	2686						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	-46	23.91250292	-1.923679849	0.126733563	-112.3917517	20.39175167	-112.3917517	20.39175167
18	X Variable 1	4.5	0.593988704	7.575901644	0.00162772	2.85082297	6.14917703	2.85082297	6.14917703
19									



Lineare Regressionen mit Python

```
>>> from scipy.stats import linregress
>>> x = [34, 35, 39, 42, 43, 47]
>>> y = [102, 109, 137, 148, 150, 158]
>>> slope = round(linregress(x,y).slope,1)
>>> intercept = round(linregress(x,y).intercept,1)
>>> print(f'y = {intercept} + {slope}x')
y = -46.0 + 4.5x
```

Multiple Regressionsanalyse

Lineare vs. multiple Regressionsanalyse

- In der linearen Regression haben wir eine unabhängige Variable, die sich auf eine abhängige Variable beziehen kann mit der Formel:

$$\hat{y} = b_0 + b_1x$$

Lineare vs. multiple Regressionsanalyse

- Mit der multiplen Regression können wir mehrere unabhängige Variablen gleichzeitig mit einer abhängigen Variablen vergleichen.
- Jede unabhängige Variable erhält einen Index: x_1 , x_2 , x_3 usw.

Lineare vs. multiple Regressionsanalyse

- Die Grundformel wird erweitert:

lineare Regression

$$\hat{y} = b_0 + b_1x$$

multiple Regression

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots$$

- b_1 ist der Koeffizient für x_1
- b_1 gibt die Veränderung von \hat{y} an, für eine gegebene Änderung von x_1 , wobei alles andere konstant bleibt

Lineare vs. multiple Regressionsanalyse

- Die Formeln für die Koeffizienten werden auch erweitert:

lineare Regression

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

Lineare vs. multiple Regressionsanalyse

- Die Formeln für die Koeffizienten werden auch erweitert:

Multiple Regression

$$b_1 = \frac{\sum(x_2 - \bar{x}_2)^2 \sum(x_1 - \bar{x}_1)(y - \bar{y}) - \sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \sum(x_2 - \bar{x}_2)(y - \bar{y})}{\sum(x_1 - \bar{x}_1)^2 \sum(x_2 - \bar{x}_2)^2 - (\sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2))^2}$$

$$b_2 = \frac{\sum(x_1 - \bar{x}_1)^2 \sum(x_2 - \bar{x}_2)(y - \bar{y}) - \sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \sum(x_1 - \bar{x}_1)(y - \bar{y})}{\sum(x_1 - \bar{x}_1)^2 \sum(x_2 - \bar{x}_2)^2 - (\sum(x_1 - \bar{x}_1)(x_2 - \bar{x}_2))^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

Multiple Regression

- Zum Beispiel könnte eine Gebrauchtwagenhändler wissen wollen, welche Variablen den Nettogewinn beeinflussen
- Er erstellt eine Liste von Faktoren auf, die mit dem Profit korrelieren könnten:

Preis Alter Marke
Farbe Stil



Multiple Regression

- Sie wollen die Korrelation jeder Variable zum Nettogewinn messen
- Allerdings könnten einige Faktoren untereinander miteinander korrelieren:



Multiple Regression

- Das Alter eines Autos könnte sich direkt auf den Verkaufspreis auswirken
- Man kann hier keinen Faktor verändern ohne dass es einen anderen beeinflusst
- Dies wird **Multikollinearität** genannt



Multiple Regressionsanalyse

Beispiel

Regression Übung #2



- Eine Apotheke liefert Medikamente an die umliegenden Gemeinden.
- Die Fahrer können pro Lieferung mehrere Stopps einlegen.
- Der Eigentümer möchte die „**Länge der Zeit**“, die eine Lieferung benötigt, basierend auf einer oder zwei zusammenhängenden Variablen vorhersagen.

Regression Übung #2



- Betrachten wir zuerst, welche Variablen sich auf die Lieferzeit auswirken könnten:
 - Anzahl der Stopps
 - die zu fahrende Distanz
 - Außentemperatur
 - Benzinpreise

Regression Übung #2



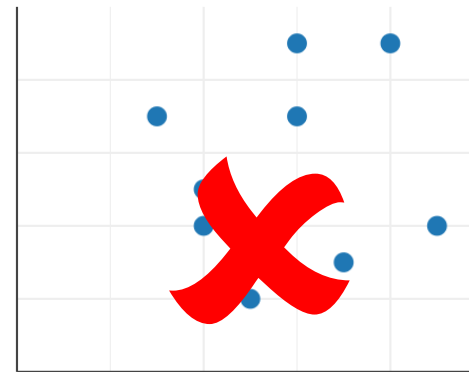
- Erstelle nun für jede Variable gegen die Lieferzeit ein Diagramm, um zu sehen, ob es eine Beziehung gibt



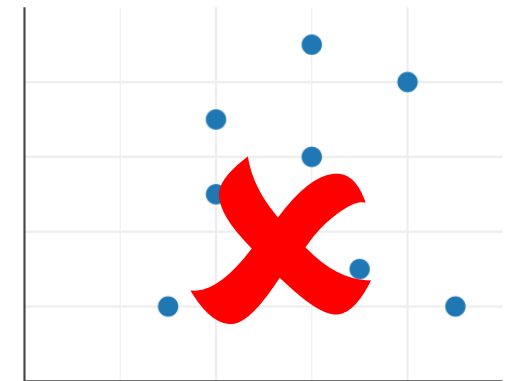
Zeit zu Entfernung



Zeit zu Stopps



Zeit zu Temperatur

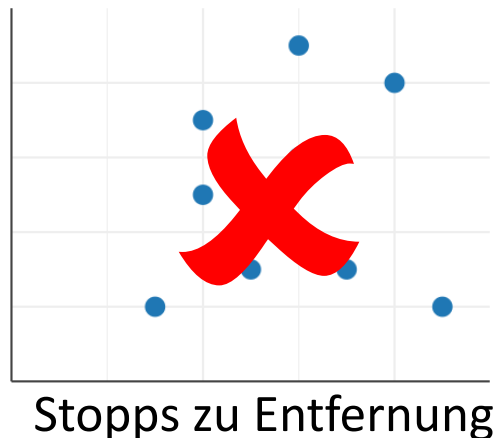


Zeit zu Benzinpreis

Regression Übung #2



- Sobald wir unsere Variablen x_1 und x_2 gewählt haben, testen wir normalerweise auf **Multikollinearität**
- Wir wollen wissen, ob unsere beiden unabhängigen Variablen eng miteinander verbunden sind
- Wenn ja, ist es sinnvoll, eine davon zu verwerfen!



Eine Lieferung könnte an einen weit entfernten Kunden oder an eine Gruppe von eng aneinander liegenden Haltestellen gehen.

Regression Übung #2



y = Lieferzeit (Minuten)

x_1 = Anzahl der Stopps

x_2 = Entfernung (km)

y	x_1	x_2	$(y - \bar{y})$	$(x_1 - \bar{x}_1)$	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$	$(x_1 - \bar{x}_1)(y - \bar{y})$	$(x_2 - \bar{x}_2)(y - \bar{y})$	$(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
29	1	8	-1	-1	1	2	4	1	-2	-2
31	3	4	1	1	1	-2	4	1	-2	-2
36	2	9	6	0	0	3	9	0	18	0
35	3	6	5	1	1	0	0	5	0	0
19	1	3	-11	-1	1	-3	9	11	33	3
\bar{y}	\bar{x}_1	\bar{x}_2				$\Sigma(x_1 - \bar{x}_1)^2$	$\Sigma(x_2 - \bar{x}_2)^2$	$\Sigma(x_1 - \bar{x}_1)(y - \bar{y})$	$\Sigma(x_2 - \bar{x}_2)(y - \bar{y})$	$\Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
30	2	6				4	26	18	47	-1

Regression Übung #2



y = Lieferzeit (Minuten)

x_1 = Anzahl der Stopps

x_2 = Entfernung (km)

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = 5$$

$$b_2 = \frac{(4)(47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = 2$$

\bar{y}	\bar{x}_1	\bar{x}_2
30	2	6

$\Sigma(x_1 - \bar{x}_1)^2$
4

$\Sigma(x_2 - \bar{x}_2)^2$	$\Sigma(x_1 - \bar{x}_1)(y - \bar{y})$	$\Sigma(x_2 - \bar{x}_2)(y - \bar{y})$	$\Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
26	18	47	-1

Regression Übung #2



y = Lieferzeit (Minuten)

x_1 = Anzahl der Stopps

x_2 = Entfernung (km)

$$\hat{y} = 8 + 5x_1 + 2x_2$$

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = 5$$

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \\ &= 30 - (5)(2) - (2)(6) \\ &= 30 - 10 - 12 = 8 \end{aligned}$$

$$b_2 = \frac{(4)(47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = 2$$

\bar{y}	\bar{x}_1	\bar{x}_2
30	2	6

$\Sigma(x_1 - \bar{x}_1)^2$
4

$\Sigma(x_2 - \bar{x}_2)^2$	$\Sigma(x_1 - \bar{x}_1)(y - \bar{y})$	$\Sigma(x_2 - \bar{x}_2)(y - \bar{y})$	$\Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
26	18	47	-1

Regression Übung #2



y = Lieferzeit (Minuten)

x_1 = Anzahl der Stopps

x_2 = Entfernung (km)

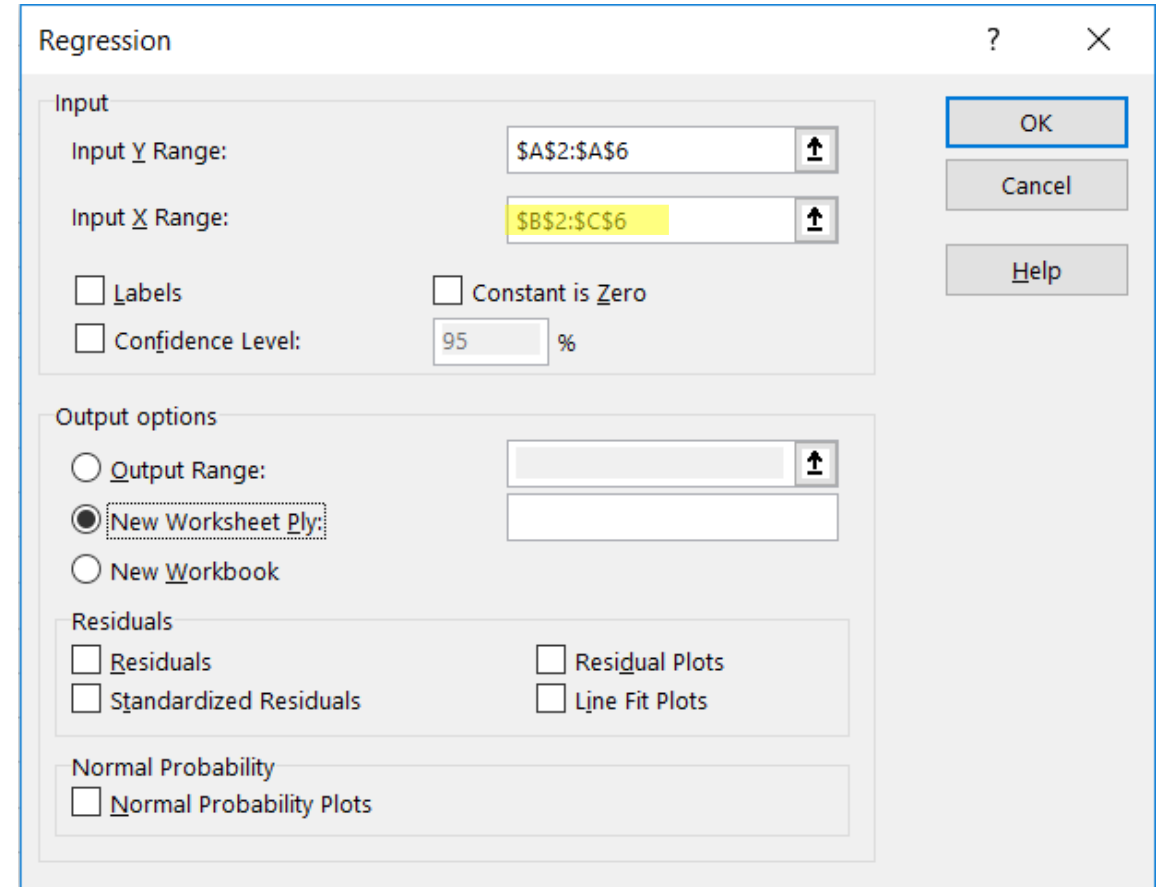
$$\hat{y} = 8 + 5x_1 + 2x_2$$

y	x_1	x_2
29	1	8
31	3	4
36	2	9
35	3	6
19	1	3

Basierend auf unserer Analyse haben Apothekenlieferungen eine feste Fahrtzeit von 8 Minuten, plus 5 Minuten für jeden Stopp, und 2 Minuten für jeden gefahrenen Kilometer

Multiple Regression in Excel

- Die Schritte sind die gleichen wie bei der linearen Regression,
- außer man wählt einen breiten x-Achsen Abschnitt aus



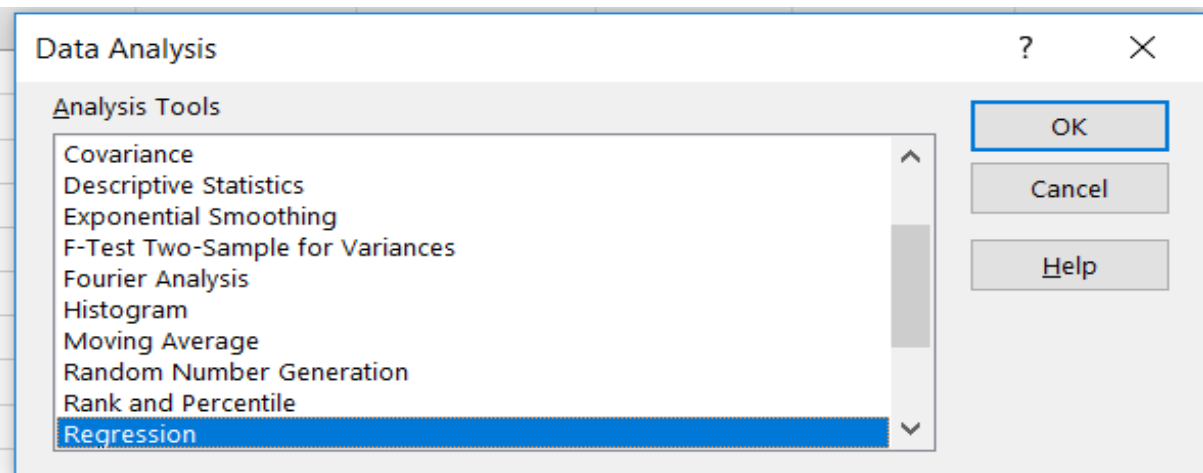
The image shows the 'Regression' dialog box in Microsoft Excel. The dialog is titled 'Regression' and has a standard Windows window with a question mark icon and a close button (X). It is divided into several sections:

- Input:**
 - Input Y Range:** A text box containing '\$A\$2:\$A\$6' with an upward arrow icon to its right.
 - Input X Range:** A text box containing '\$B\$2:\$C\$6' with an upward arrow icon to its right.
 - ☐ **Labels**
 - ☐ **Constant is Zero**
 - ☐ **Confidence Level:** A text box containing '95' followed by a '%' symbol.
- Output options:**
 - ☐ **Output Range:** A text box with an upward arrow icon to its right.
 - ☒ **New Worksheet Ply:** A text box.
 - ☐ **New Workbook**
- Residuals:**
 - ☐ **Residuals**
 - ☐ **Standardized Residuals**
 - ☐ **Residual Plots**
 - ☐ **Line Fit Plots**
- Normal Probability:**
 - ☐ **Normal Probability Plots**

On the right side of the dialog, there are three buttons: 'OK' (highlighted with a blue border), 'Cancel', and 'Help'.

Multiple Regression in Excel

	A	B	C						
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	1							
5	R Square	1							
6	Adjusted R Square	1							
7	Standard Error	1.25607E-15							
8	Observations	5							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	2	184	92	5.83119E+31	1.71492E-32			
13	Residual	2	3.15544E-30	1.57772E-30					
14	Total	4	184						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	8	2.11706E-15	3.77882E+15	7.00306E-32	8	8	8	8
18	X Variable 1	5	6.31078E-16	7.92295E+15	1.59304E-32	5	5	5	5
19	X Variable 2	2	2.47529E-16	8.07985E+15	1.53177E-32	2	2	2	2
20									



Multiple Regression in Python

```
>>> from sklearn.linear_model import LinearRegression
>>> x1,x2 = [1,3,2,3,1], [8,4,9,6,3]
>>> y = [29,31,36,35,19]
>>> reg = LinearRegression()
>>> reg.fit(list(zip(x1,x2)), y)
>>> b1,b2 = reg.coef_[0], reg.coef_[1]
>>> b0 = reg.intercept_
>>> print(f'y = {b0:.{3}} + {b1:.{3}}x1 + {b2:.{3}}x2')
y = 8.0 + 5.0x1 + 2.0x2
```


Als nächstes: Chi-Quadrat
Analyse