Lösungen zu den Übungen zu Korrelation und linearer Regression Serie 11

Aufgabe 1 Wir nehmen nochmals Ihre Daten zur Hand. Auf Moodle sind die Daten im File `Stats_Height_Weight.xlsx` zu finden. Wir möchten überprüfen, ob es einen linearen Zusammenhang zwischen der Grösse und dem Gewicht eines Menschen gibt, und eine lineare Regression durchführen. Im Theorieteil ganz am Schluss ist die gesamte Analyse aufgeschrieben, und auf Moodle finden Sie das zugehörige Jupyter Notebook `Stats_Height_Weight.ipynb`.

(a) Lesen Sie die Theorie auf den Seiten 7 bis 10 gut durch und versuchen Sie die Befehle im Jupyter Notebook zu verstehen und zu reproduzieren.

(b) Was ist nach Modell die Vorhersage für Ihr Gewicht? Stimmt dieser Wert einigermassen mit Ihrem tatsächlichen Gewicht überein?

(c) Wie gross ist die Abweichung zwischen Ihrem tatsächlichen und Ihrem geschätzen Gewicht?

Lösungen:

Die Werte sind persönlich und können nicht allgemein angegeben werden.

Aufgabe 2 Die folgende Tabelle enthält die Länge und das Gewicht von 9 Schlangen. Auf Moodle sind diese Daten im File `Snakes.xlsx` gespeichert.

Nummer	Länge X [cm]	Gewicht Y [g]
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145

(a) Stellen Sie die Daten in einem Streudiagramm dar.

(b) Ist es aufgrund des Plots sinnvoll, von einem linearen Zusammenhang zwischen X und Y auszugehen?

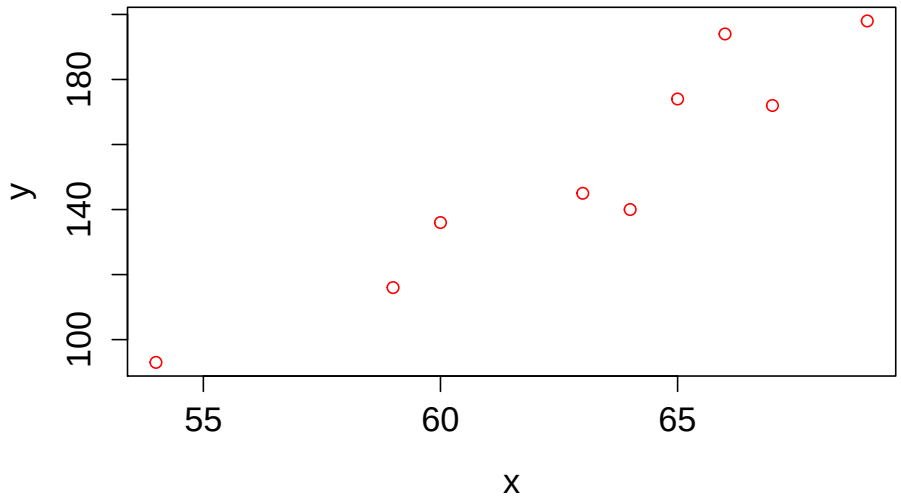
(c) Berechnen Sie von Hand die Regressionsgerade mit Y als abhängige und X als erklärende Variable. Berechnen Sie ebenfalls die Residuen von Hand.

(d) Die Residuen müssen normalverteilt sein. Diese Annahme kann man mit einem Q-Q-Plot überprüfen. Sind Sie nun in der Lage den zugehörige Q-Q-Plot von Hand zu zeichnen? Falls es nicht klappt, machen Sie sich keine Sorgen, wir kommen darauf züruck.

(e) Führen Sie nun eine lineare Regression mit Python durch. Erhalten Sie die gleichen Werte für die Koeffizienten wie bei der Berechnung von Hand?

Lösung:

(a) Das Scatterplot kann z.B. folgendermassen aussehen:



(b) Ja, wir sehen einen linearen Zusammenhang in den Daten.

(c) Um die Koeffizienten zu bestimmen, verwenden wir die folgenden Formeln:

$$\begin{aligned}\widehat{\beta_1} &= \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y}}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2} \\ \widehat{\beta_0} &= \overline{y} - \widehat{\beta_1}\overline{x}\end{aligned}$$

Für die Mittelwerte erhalten wir $\overline{x} = 63$ und $\overline{y} = 152$. Weiter ist:

$$\begin{aligned}\frac{1}{9}\left(\sum_{i=1}^{9} x_i^2\right) - \overline{x}^2 &= \frac{1}{9}(60^2 + 69^2 + \cdots + 63^2) - 63^2 \\ &= 3988.111111 - 3969 = 19.\overline{1}\end{aligned}$$

$$\begin{aligned}\frac{1}{9}\left(\sum_{i=1}^{9} x_i y_i\right) - \overline{x}\cdot\overline{y} &= \frac{1}{9}(60\cdot 136 + \cdots + 63\cdot 145) - 63\cdot 152 \\ &= 9713.444444 - 9576 = 137.\overline{4}\end{aligned}$$

Die gesuchten Koeffizienten lauten:

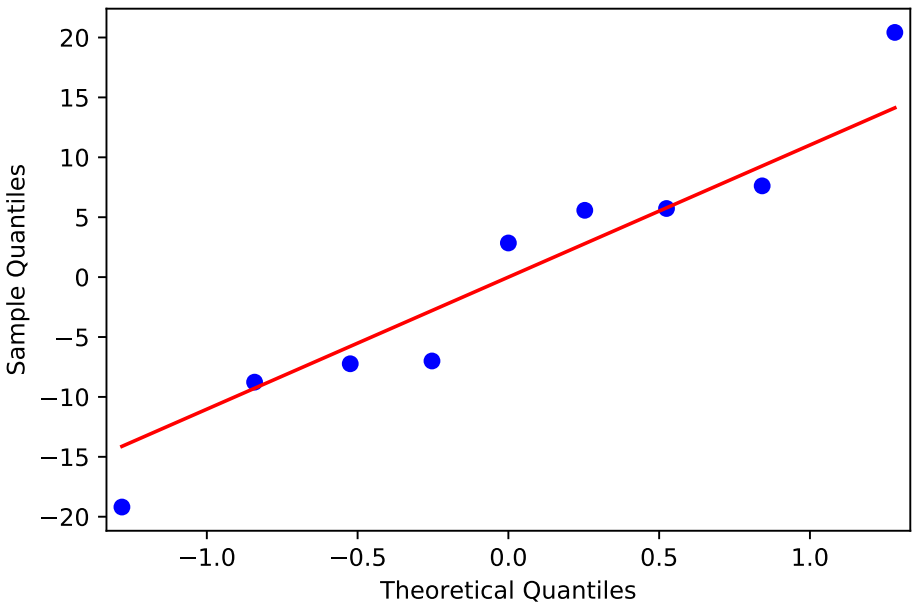
$$\hat{\beta}_1 = \frac{137.\overline{4}}{19.\overline{1}} \approx 7.192 \quad \text{und} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x} \approx -301.087$$

Um die Residuen zu berechnen, bilden wir die Differenzen zwischen den gemessennen und den modellierten y-Werte. Die Residuen sind in folgender Tabelle aufgelistet:

Nummer	Länge X [cm]	Gewicht Y [g]	Geschätze Werte $\hat{y}$	Residuen ϵ
1	60	136	130.433	5.567
2	69	198	195.161	2.839
3	66	194	173.585	20.415
4	64	140	159.201	−19.201
5	54	93	87.281	5.719
6	67	172	180.777	−8.777
7	59	116	123.241	−7.241
8	65	174	166.393	7.607
9	63	145	152.009	−7.009

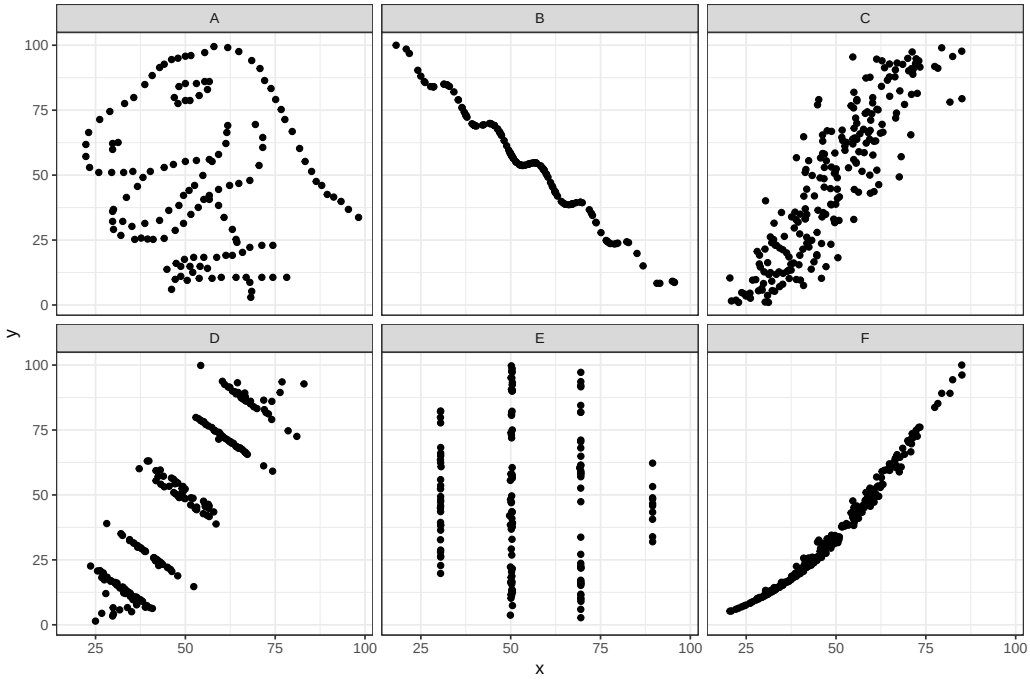
(d) Die Residuen sind in diesem Fall die Messwerte (sample quantiles)! Als ersten müssen wir die Residuen der Grosse nach sortieren, und dann die kumulierten Wahrscheinlichkeiten berechnen. Für den Q-Q-Plot brauchen wir ebenfalls die geglätteten Wahrscheinlichkeiten und die zugehörigen theoretische Quantile der Standardnormalverteilung (mit Hilfe der Tabelle ablesen).

Neue Nr.	Sortierte Residuen	kumulierte Wahrscheinlichkeiten	geglättete Wahrscheinlichkeiten	theoretische Quantile
1	−19.201	$0.\overline{1}$	$0.0\overline{5}$	−1.645
2	−8.777	$0.\overline{2}$	$0.1\overline{6}$	−0.955
3	−7.241	$0.\overline{3}$	$0.2\overline{7}$	−0.585
4	−7.009	$0.\overline{4}$	$0.3\overline{8}$	−0.255
5	2.839	$0.\overline{5}$	0.5	0.0
6	5.567	$0.\overline{6}$	$0.6\overline{1}$	0.285
7	5.719	$0.\overline{7}$	$0.7\overline{2}$	0.59
8	7.607	$0.\overline{8}$	$0.8\overline{3}$	0.965
9	20.415	1	$0.9\overline{4}$	1.56



(e) siehe Jupyter Notebook auf Moodle für die Lösung

Aufgabe 3 Betrachten Sie die folgenden 6 Streudiagramme.



(a) Ist ein linearer und/oder monotoner Zusammenhang zwischen den Variablen y und x in den Datensätzen A-F ersichtlich? Achtung: Jeder Datensatz muss separat betrachtet werden.

	A		B		C		D		E		F	
	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein
linear	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
monoton	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

(b) Schätzen Sie den zugehörigen Pearsons Korrelationskoeffizienten anhand der Streudiagramme.

(c) Auf Moodle stehen Ihnen die 6 Datensätzen A-F in einen File Names `Dino.xlsx` zur Verfügung. Achtung: Sie müssen jeweils die richtigen Zeilen selektieren. Der Datensatz befindet sich in dem sogenannten „long“ Format. Berechnen Sie für jeden Datensatz den zugehörigen Pearsons Korrelationskoeffizienten. Sie können dafür die folgenden Befehle in Python verwenden:

```
> from scipy import stats
# use the names that you choose to replace data.x and data.y
> Pearsons_r, p_value = stats.pearsonr(data.x, data.y)
> print(Pearsons_r)
```

Die erste Ausgabe ist Pearsons Korrelationskoeffizient und die zweite Ausgabe ist der zugehörige p-Wert. In diesem Fall lautet die Nullhypothese, dass es keinen linearen Zusammenhang zwischen den Daten gibt.

(d) Vergleichen Sie nun die Ergebnisse von Python mit Ihren Schätzungen.

Lösung:

	A		B		C		D		E		F	
	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein	ja	nein
linear	☐	☒	☐	☒	☒	☐	☐	☒	☐	☒	☐	☒
monoton	☐	☒	☐	☒	☒	☐	☐	☒	☐	☒	☒	☐

(b) Meine Schätzwerte für r sind

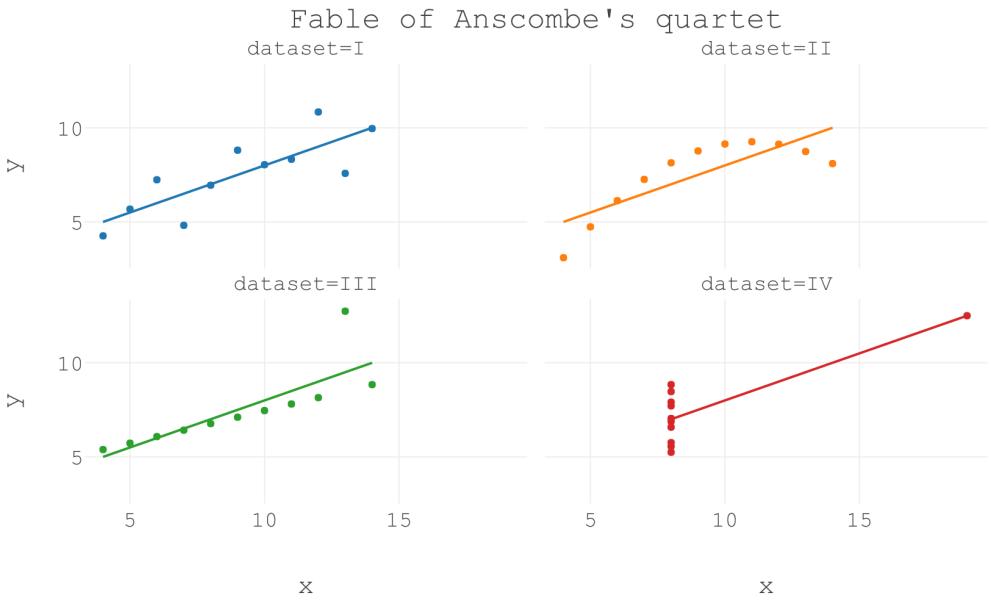
	A	B	C	D	E	F
$-1.0 \leq r < -0.8$		☒				
$-0.8 \leq r < -0.5$						
$-0.5 \leq r < 0.0$	☒				☒	
$0.0 \leq r < 0.5$						
$0.5 \leq r < 0.8$						
$0.8 \leq r \leq 1.0$			☒	☒		☒

(c) siehe Jupyter Notebook auf Moodle für die Lösungen

(d) Meine grobe Schätzungen stimmen mit den exakten Werten überein. Eigentlich bräuchte ich mehr Übung! Aber zum Glück gibt es Software und ich muss das fast nie nur schätzen.

Aufgabe 4 Die 4 berühmten Datensätze names *Anscombe's Quartet* vom Mathematiker Anscomb sind auf Moodle im File `anscombes_quartet.xlsx` gespeichert. Lösen Sie diese Aufgabe nur mit Python.

(a) Visualisieren Sie die 4 Datensätzen in einer Graphik. Ihr Ergebnis muss nicht genau gleich aussehen, aber als Beispiel ist der folgende Plot gemeint:



(b) Berechnen Sie für jeden Datensatz den zugehörigen Pearsons Korrelationskoeffizienten und vergleichen Sie die Werte, die Sie erhalten. Achtung: Sie müssen jeweils die richtigen Zeilen selektieren. Die Daten befindet sich in dem sogenannten „long“ Format.

(c) Führen Sie für jeden Datensatz eine lineare Regressionsanalyse mit Python durch. Vergleichen Sie die Koeffizienten der verschiedenen Modellierungen miteinander.

(d) Machen Sie Q-Q-Plots um die Normalität der Residuen zu überprüfen (muss für jeden Datensatz separat ausgeführt werden). Falls es nicht klappt, machen Sie sich keine Sorgen, wir kommen darauf züruck.

(e) Was sind Ihre Erkenntnisse?

Lösungen:

siehe Jupyter Notebook auf Moodle für die Lösungen