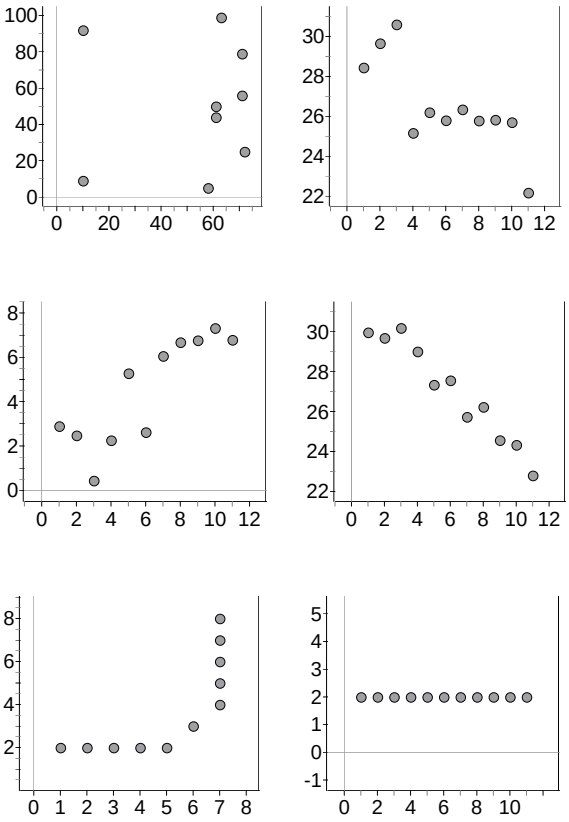
3 Analyse statistischer Daten zu zwei Merkmalen - Lösungshinweise

Aufgabe 3.1: Schätzen Sie den Wert eines Korrelationskoeffizienten für folgende Punktwolken und begründen Sie Ihre Schätzung:



Lösungsskizze Aufgabe 3.1

- Punktwolke oben links: Korrelationskoeffizient nahe 0, da kein Zusammenhang hinsichtlich größer (kleiner) werdenden Abszissenwerten und größer (kleiner) werdenden Ordinatenwerten erkennbar ist

- Punktwolke oben rechts: relativ großer negativer Korrelationskoeffizient, da in der Tendenz mit größer (kleiner) werdenden Abszissenwerten kleiner (größer) werdende Ordinatenwerte einhergehen.
- Punktwolke mittig links: relativ großer positiver Korrelationskoeffizient, da in der Tendenz mit größer (kleiner) werdenden Abszissenwerten größer (kleiner) werdende Ordinatenwerte einhergehen.
- Punktwolke mittig rechts: sehr großer negativer Korrelationskoeffizient, die Punkte kommen in guter Näherung auf einer Geraden mit negativer Steigung zu liegen.
- Punktwolke unten rechts: positiver Korrelationskoeffizient r, mit größer werdenden Abszissenwerten gehen durchschnittlich größer werdende Ordinatenwerte einher: Zwar haben die Abszissenwerte bis zum fünften Wert immer den gleichen Ordinatenwert zugeordnet (hier also keine einhergehende Erhöhung der Ordinatenwerte), allerdings werden dem siebten Abszissenwert fünf zunehmend größere Ordinatenwerte zugeordnet, so dass sich im Gesamten betrachtet ein positiver Korrelationszusammenhang ergibt. Im Fall des ausgezählten und resistenten Korrelationskoeffizienten ergeben sich entgegen der Anschauung Werte nahe 1, da der mittlere Punkt im Zentrum des Mediankreuzes liegt, alle anderen Punkte im 1. und 3. Quadranten des Mediankreuzes.
- Punktwolke unten rechts: Hier ergibt sich keine sinnvolle Schätzung für einen Korrelationskoeffizienten, da bei unterschiedlichen Abszissenwerten die Ordinatenwerte immer denselben Wert haben. Der Korrelationskoeffizient nach Bravais und Pearson ließe sich nicht berechnen, da beim Dividieren durch die Standardabweichung der Ordinatenwerte durch Null geteilt werden müsste, der augezählte wie auch der resistente Korrelationskoeffizient haben den Wert 0, da alle Punkte auf dem Mediankreuz liegen.

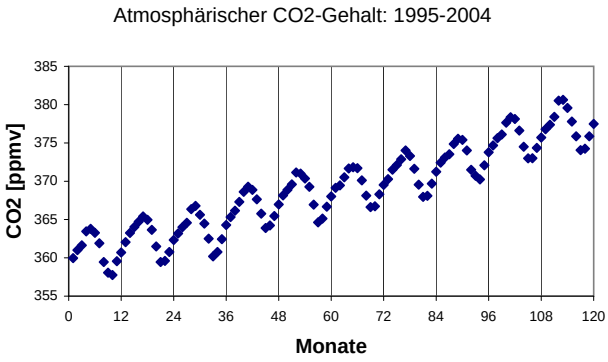
Aufgabe 3.2: Untersuchen Sie mit Rechnerunterstützung Datensätze, die wir in den Formaten xls (Excel), ftm (fathom) und txt (Textdatei zum Einlesen in potentielle Statistiksoftware) im Zusatzmaterial für dieses Buch zur Verfügung stellen[a], auf einen funktionalen Zusammenhang. Die Datensätze betreffen:

- die Eigenschaften von Studierenden,
- die deutsche Fußball-Bundesliga und
- den weltweiten CO_2-Anstieg.

[a] Die Bezugsquelle ist im Vorwort angegeben.

Lösungsskizze Aufgabe 3.2 Exemplarische Überlegungen zur Analyse des Datensatzes zum weltweiten CO_2-Anstieg:

- Hintergrundinformationen: Der vorliegenden Datensatz stellt einen Ausschnitt aus den Originaldaten (Datenquelle: http://cdiac.ornl.gov/) einer CO_2-Messstation auf Mauna Loa (ein Vulkan der hawaiianischen Inselgruppe) dar, der die Messwerte aus dem Zehnjahreszeitraum von 1995-2004 umfasst. Der Blick auf das Streudiagramm zeigt deutlich, dass



sich zwei Trends überlagern: Einerseits ein Anstieg in dem gewählten Zehnjahreszeitraum, der durchschnittlich betrachtet in augenscheinlich guter Näherung als linear charakterisiert werden könnte, und ein zyklischer Trend, dessen Zyklus sich über einen zwölfmonatigen Zeitraum erstreckt.

Eine Erklärung für die periodische Veränderung des atmosphärischen CO_2-Gehalts liefert die jahreszeitlich bedingte Veränderung der Vegetation vor Ort. Von Frühjahr bis Herbst sind Bäume und sonstige Grünpflanzen belaubt. Durch die Photosynthese wird CO_2 umgewandelt und der CO_2-Gehalt der Atmosphäre sinkt in diesem Zeitraum. Im Zeitraum von Spätherbst bis Frühjahresende trägt dann verrottendes organisches Material (z. B. abgeworfenes Laub) zur Steigerung des CO_2-Gehalts in der Atmosphäre bei. Für die zyklische Schwankung ist die Tatsache von Bedeutung, dass die Landmasse mit der Vegetation ungleich auf Nord- und Südhalbkugel der Erde verteilt ist. Eine Erklärung für den aufsteigenden, im gewählten Zehnjahreszeitraum augenscheinlich nahezu linearen Trend, kann

im steigenden weltweiten Verbrauch fossiler Brennstoffe vermutet werden: Das Verbrennen von Fossilbrennstoffen wie Kohle und Erdöl geht nicht in die natürliche CO_2-Bilanz ein, da die dadurch freigesetzten CO_2-Mengen nicht aus einer vorangehenden photosynthetischen Bindung hervorgehen. So werden zusätzliche Mengen an Kohlendioxid in der Atmosphäre frei.

- Vorgehensweise bei der Modellierung: Aufgrund der Überlagerung zweier Trends bietet sich ein schrittweises Vorgehen an, indem die Trends getrennt funktional angepasst werden. Erweitert und passt man die im Buch auf S. 60 genannte Modellstrukturgleichung $Daten = Fit + Residuen$ entsprechend an, so ergibt sich die Modellstrukturgleichung

$$Daten = Fit_{Anstieg} + Fit_{zyklischer\ Trend} + Residuen$$

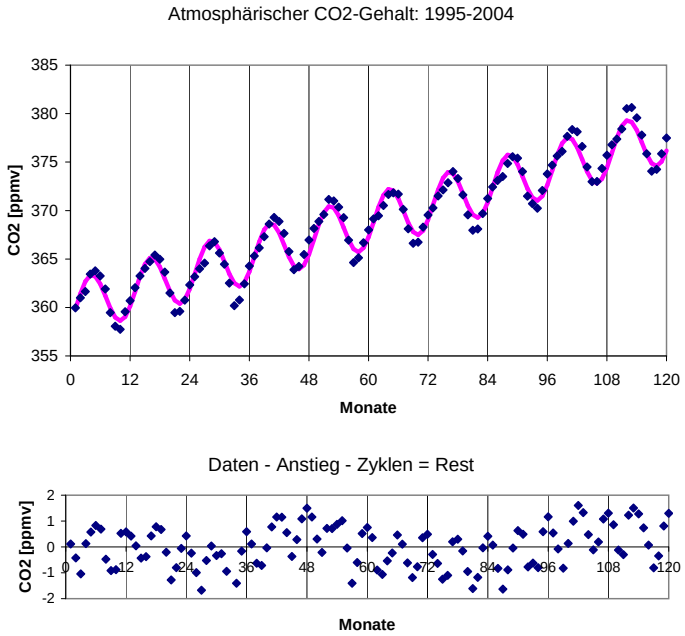
  Die Modellierung wird in zwei aufeinander folgenden Schritten vollzogen.

- Modellierung des Anstiegs: Möchte man den durchschnittlichen Anstieg der CO_2-Daten im gewählten Zehnjahreszeitraum modellieren, kann eine Gerade herangezogen werden, deren Steigung den gleich bleibenden, durchschnittlichen Zuwachs pro Zeiteinheit repräsentiert und der Ordinatenabschnitt einen Modellierungsstartwert. Wählt man hierfür die Regressionsgerade aus, so ergeben sich für die Steigung $m = 0,148$ und den Ordinatenabschnitt $b = 359,94$ (siehe Excel-Datei in den Online-Zusatzmaterialien). Die Regressionswerte werden von den entsprechenden Datenwerten subtrahiert, so dass sich zyklische Residuen ergeben, welche die jahreszeitlich bedingten Veränderungen repräsentieren.

- Modellierung des zyklischen Trends: Die Tatsache, dass die CO_2-Daten einem periodisch wiederkehrenden Zyklus unterliegen, legen die Überlegung nahe, die Residuen mit z. B. einer Sinusfunktion $f(x) = u \cdot sin(v \cdot x + w)$ zu modellieren. Dabei bietet sich die Verwendung von Software an, die die Anpassung der Funktionsparameter durch Schieberegler ermöglicht (siehe Excel-Datei in den Online-Zusatzmaterialien). Bei einer guten Näherung, welche durch die Minimierung der Summe der absoluten Abweichungen kontrolliert werden kann, ergeben sich für den Amplitudenparameter $A = 2,8$, die Winkelgeschwindigkeit $v = 0,524$ und den Nullphasenwinkel $w = -0,611$. Insgesamt ergibt sich mit den beiden Modellierungsschritten die funktionale Beschreibung der Daten, die in der folgenden Abbildung zu sehen ist:

  Zur Bedeutung der Funktionsparameter:

  - Der verwendete Amplitudenparameter A ist ein Maß für die Weite der zyklischen jahreszeitlichen Schwankungen. Dadurch wird die Ober- und Untergrenze des Korridors festgelegt, in welchem die modellierten CO_2-Werte jahreszeitlich bedingt periodisch schwanken. Ausgehend davon, dass die Funktionswerte von $sin\ x$ in einem Korridor zwischen -1 und $+1$ schwanken, wird mit Blick auf die Daten deutlich, dass A einen Wert in der Größenordnung von ungefähr 3 annehmen muss.
  - Mit dem Parameter v wird die Periodendauer bestimmt. Die Periodendauer ist in diesem Beispiel die Zeit, in welcher der jahreszeitlich bedingte Zyklus der atmosphärischen CO_2-Gehaltsänderung einmal durchlaufen wird. Somit ist v ein Parameter, mit dem die Monats- bzw. Jahreslänge beschrieben wird. Seine Größenordnung



kann durch folgende Überlegung bestimmt werden: Der Parameter v bezeichnet die Winkelgeschwindigkeit der Kreisbewegung, deren Projektion die modellierte harmonische Schwingung ergibt. Für diese gilt: $v = \frac{2\pi}{T}$, wobei T die Dauer einer vollen Schwingung bezeichnet. Damit ergibt sich $T = 12$ bezogen auf die Zeiteinheit Monat. Daraus folgt: $v = \frac{2\pi}{12} = \frac{\pi}{6} \approx 0,52$.

– Mit dem Nullphasenwinkel w wird der Phasenwinkel zum Anfangszeitpunkt der Schwingung festgelegt. Hier entspricht dies der zeitlichen Abweichung (Vielfaches der Zeiteinheit Monat) zwischen einer angenommenen Schwingung, deren Zyklus im Januar beginnt, und der auf der Datenbasis angepassten Schwingung.

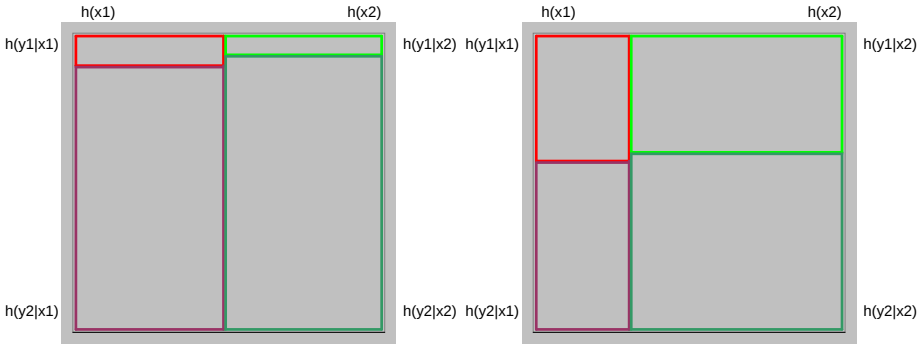
Wiederum können die Residuen hinsichtlich dieser Modellanpassung betrachtet werden. Augenscheinlich streuen die Reste hier schon „recht zufällig“. Man erkennt vage einen Resttrend, der zu tiefer gehenden Überlegungen anstiften kann (noch ein Zyklus mit einer Periode, die länger als ein Jahr ist?). Das Kriterium für eine weitergehende funktionale Anpassung sollte jedoch sein, dass die verwendete Funktionenklasse hinsichtlich der verwendeten Funktionsparameter im Sachkontext interpretierbar sein sollte (vgl. Buch, S. 76). Eine alternative Art der Datenmodellierung ergäbe sich durch das Anwenden von Glättungsmethoden (z.B. Eichler & Vogel, 2009). Diese haben wir jedoch in diesem Buch nicht besprochen, so sei diese Möglichkeit an dieser Stelle lediglich der Vollständigkeit halber erwähnt.

Aufgabe 3.3: Gegeben sind die Datensätze zu Studierenden (Münster und Freiburg) zu der Parteipräferenz, dem bevorzugten Beförderungsmittel sowie zum Erhalt von BAföG.
Sind Grünen-Wähler umweltbewusster? Sind BAföG-Bezieher SPD-Wähler?

	Grüne	CDU/FDP	Summe
Auto	13	9	22
Fahrrad	116	132	248
Summe	129	141	270

	BAföG	kein BAföG	Summe
SPD	49	107	156
CDU/FDP	66	163	229
Summe	115	270	385

Lösungsskizze Aufgabe 3.3 In dieser Aufgabe geht es bei beiden Vierfeldertafeln um die Frage nach der Stärke eines Zusammenhangs zwischen nominalskalierten Merkmalen. Zur Aufgabenbearbeitung bietet sich das Applet „assoziationsmass.xls“ an, das in den online-Zusatzmaterialien zum Download bereit steht. Damit ergeben sich die folgenden Vierfeldertafeln, wobei im ersten Fall x1 für Grüne und x2 für CDU\FDP steht und im zweiten Fall x1 für BAföG bzw. x2 für kein BAföG: Beide Einheitsquadrate sind sich darin ähnlich, dass jeweils der Abstand der waag-

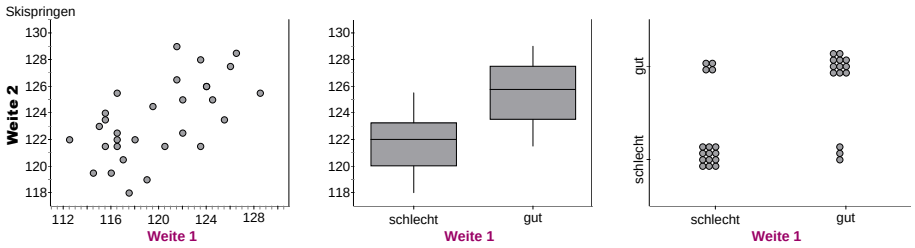


rechten Unterteilungen gering ist. Diese Eigenschaft weist grafisch jeweils auf einen geringen Zusammenhang der Merkmale hin. Die Werte der im Buch besprochenen Assoziationsmaße unterlegen das numerisch: $A = 0,04$ und $\rho = 1,64$ bzw. $1/\rho = 0,61$ im ersten Fall und ein noch geringerer Zusammenhang $A = 0,03$ und $\rho = 1,13$ bzw. $1/\rho = 0,88$ im zweiten Fall. Der Einfluss des Umfangs der Stichproben bzw. Teilstichproben auf die Interpretation der errechneten Werte soll an dieser Stelle noch nicht diskutiert werden.

Aufgabe 3.4: Untersuchen Sie durch Clusterung und möglichst variantenreich mit einem linearen Modell ohne Rechnerunterstützung den Zusammenhang zwischen erster und zweiter Sprungweite beim Skispringen. Versuchen Sie von dem hier gegebenen Springen in Innsbruck 2009 das Modell auf die im Netz verfügbaren Springen in Innsbruck anderer Jahre sowie anderer Orte zu übertragen.

Platz	Name	Weite1	Weite2	Platz	Name	Weite1	Weite2
1	Schmitt	128,5	125,5	16	Evensen	119	119
2	Loitzl	126,5	128,5	17	Watase	118	122
3	Schlierenzauer	126	127,5	18	Eggenhofer	117,5	118
4	Amman	125,5	123,5	19	Larinto	117	120,5
5	Morgenstern	124,5	125	20	Uhrmann	116,5	125,5
6	Kasai	124	126	21	Hilde	116,5	122,5
7	Neumayer	124	126	22	Ito	116,5	121,5
8	Hautamaeki	123,5	128	23	Schoft	116,5	122
9	Rosliakow	123,5	121,5	24	Stoch	116	119,5
10	Olli	122	125	25	Hocke	115,5	124
11	Koch	122	122,5	26	Koudelka	115,5	123,5
12	Vassilev	121,5	129	27	Lackner	115,5	121,5
13	Jacobsen	121,5	126,5	28	Yumoto	115	123
14	Malysz	120,5	121,5	29	Kofler	114,5	119,5
15	Kuettel	119,5	124,5	30	Tochimoto	112,5	122

Lösungsskizze Aufgabe 3.4 Bei der Clusterung werden die erzielten Werte des ersten Sprungdurchgangs in zwei Gruppen eingeteilt: „schlecht“ für unter dem arithmetischen Mittel (der Median wäre ebenso möglich) des ersten Durchgangs und „gut“ für darüber. Auf dieser Basis lassen sich folgende Abbildungen erstellen, je nachdem, ob die zweite Sprungweite mit den metrisch skalierten Merkmalsausprägungen beibehalten oder in gleicher Weise wie der erste Sprung geclustert wurde: Die beiden Abbildungen rechts zeigen deutlich den Trend, dass gu-



te Springer im zweiten Sprung überwiegend gut abschneiden und schlechte Springer des ersten Durchgangs wiederum schlecht. Aber es gibt auch Ausnahmen, die sich mit dem rechten Punktgruppendiagramm quantifizieren lassen: drei gute Springer des ersten Durchgangs sind nur unterdurchschnittlich weit im zweiten Durchgang gesprungen und vier schlechte Springer im ersten Durchgang haben sich im zweiten Sprung so gesteigert, dass sie über dem arithmetischen Mittel des zweiten Sprungs lagen. Außer der groben Clusterung in „gut“ und „schlecht“ können natürlich auch feinere Unterteilungen vorgenommen werden, wie z.B. die Drittelung der Datenreihen durch das 33%-Perzentil (was auch für eine Analyse über die Median-Median-Gerade benötigt wird) oder die entsprechende Viertelung.

Da beide Korrelationskoeffizienten, der ausgezählte wie der resistente, nicht von der Werten der einzelnen Datenpunkte abhängen, sondern lediglich von deren Anzahl hinsichtlich der Einteilung durch ein Mittelkreuz, eignet sich das rechte Diagramm zum händischen Berechnen dieser Korrelationskoeffizienten:

$$r_z = \frac{23-7}{30} \approx 0,53 \qquad r_{rst} = sin\left[\left(\frac{23-7}{30}\right) \cdot \frac{1}{2}\pi\right] \approx 0,74$$

Mit Rechnerunterstützung lässt sich der Korrelationskoeffizient nach Bravais und Pearson ermitteln zu $r = 0,61$.

Die Einpassung einer Geraden nach Augenmaß könnte für eine erste Näherung durchaus zielführend sein, jedoch besteht ohne Computerunterstützung (Betrachtung der Residuen, Minimierung der Summe der Residuenbeträge) keine Möglichkeit zur objektiven Kontrolle (eine zwar geringe, aber doch vorhandene Fehlleitung ergäbe sich im vorliegenden Datensatz dadurch, dass das Datum $(124;126)$ zweimal auftritt und somit den gleichen Datenpunkt im Diagramm erzeugt. Die Regressionsgerade ist in der praktischen Durchführung nur mit Computerunterstützung zu ermitteln.

Die Median-Median-Gerade lässt sich ohne Computerunterstützung in folgenden Schritten ermitteln:

1. Clusterung des ersten Merkmals (Weite 1) in drei gleich große Teile.
2. In allen Clustern wird der jeweilige Medianpunkt berechnet, dessen Koordinaten sich aus den jeweiligen Medianen der Weite 1 und der Weite 2 in dem betreffenden Cluster ergeben. Dabei ergibt sich:
   Cluster 1 (im ersten Sprung schwache Springer): Medianpunkt $M_1(115,5|122)$; Cluster 2 (im ersten Sprung mittlere Springer): Medianpunkt $M_2(119,25|122,25)$; Cluster 3 (im ersten Sprung starke Springer): Medianpunkt $M_3(124,25|125,75)$.
3. Bestimmung der Steigung a der Geraden durch die beiden äußeren Medianpunkte: $a = \frac{125,75-122}{124,25-115,5} = \frac{3}{7} \approx 0,429$. Diese Steigung entspricht der Steigung der Median-Median-Geraden.
4. Bestimmung der Schnittpunkte der drei Geraden mit der Steigung a durch die drei Medianpunkte mit der Ordinatenachse (Weite 2-Achse): $S_1(0|72,50)$, $S_2(0|71,14)$ und $S_3 = S_1$ (mit Rundungen).
5. Bestimmung des arithmetischen Mittels der Koordinaten der drei Schnittpunkte: $S(0|\frac{2 \cdot 75,5+71,14}{3})$, d.h. ungefähr $S(0|72,05)$. Der Schnittpunkt S legt den Ordinatenabschnitt der Median-Median-Geraden fest.
6. Es ergibt sich als Median-Median-Gerade $Weite\ 2 = 0,429 \cdot Weite\ 1 + 72,05$.