Dr. Quapp: Statistik für Mathematiker mit SPSS

Weitere Hinweise zur 3. Übung – Lineare Regression

2.) Die folgende Tabelle enthält die Größe (H in cm) und das Gewicht (G in kg) von 30 elfjährigen Mädchen (datei *hoehe30.sav* in D:).

H	G	H	G	H	G	H	G	H	G
135	26	133	31	143	36	140	33	149	44
146	33	149	36	146	35	143	42	147	36
153	55	141	32	141	28	148	32	155	36
154	50	164	47	136	28	149	32	135	30
139	32	146	37	154	36	141	29	137	31
131	25	149	46	151	48	137	34	152	47

Untersuchen Sie zunächst die Merkmale G und H separat.
a) Bestimmen Sie für G und H die Häufigkeitstabellen und Histogramme.
b) Bestimmen Sie für beide Merkmale Modalwert, Median, arithmetisches Mittel und die Quartile der Ordnung 1/4 bzw. 3/4.
c) Bestimmen Sie die Standardabweichungen, Schiefen, Wölbungen beider Merkmale und die mittleren absoluten Abweichungen.
Untersuchen Sie die Abhängigkeit der Merkmale G und H.
d) Zeichnen Sie das Streuungsdiagramm.
e) Bestimmen Sie durch lineare Regression die beste Funktion $G = a + bH$.
f) Klassifizieren Sie G und H in je 3 Klassen und Stellen Sie die Kontingenztafel auf.

Die Werte können von SPSS-Nutzern geladen werden; danach gehe man wieder auf sein home-directory, um beim Abspeichern nicht die vorhandene Datei zu beschädigen.
zu a) und b) und c) Wie gehabt gehe man zu Analysieren, deskriptive Statistik, und klicke die entsprechenden Felder an.
Bei c) ist direkt zu berechnen die mittleren absoluten Abweichung:

$$\frac{1}{n} \sum_{i=1}^{n} Abs(h_i - median(H))$$

im Fenster Transformieren, Berechnen, und durch Bilden der kumulativen Summe. Es ergibt sich für H: 6.27 und für G: 5.83. Beide Werte sind geringfügig kleiner als die entsprechenden Werte der Standardabweichung.
zu d) Das Streuungsdiagramm gibt die Möglichkeit, die numerisch skalierten Werte für H und G untereinander grafisch zu vergleichen. Es zeigt, daß es sinnvoll sein kann, eine lineare Abhängigkeit anzunehmen. Außerdem kann man erwägen, das Wertepaar $(164, 47)$ als Ausreißer zu eliminieren? (Das muß dann aber in einer Zusatznote berichtet werden.)
zu f) Im Fenster Analysieren, deskriptive Statistik, Kreuztabellen kann eine solche aufgerufen werden. Wenn die Daten nicht grössere Häufigkeiten aufweisen, ist dies aber ziemlich witzlos. Deshalb sollte man vorher eine Klassifizierung vornehmen. Im Fenster $-->$ Transformieren, $-->$ Bereichseinteilung, kann mit neuem Namen, Klick bei gleichen Percentilen, und 2 Trennwerten eine wie gefordert 3-Klasseneinteilung erreicht werden. Es ergibt sich mit G in den Zeilen, H in den Spalten:

	1	2	3
1	10	2	0
2	2	5	4
3	0	2	5

Auch diese kleine Tabelle zeigt deutlich, daß eine (zu erwartende) Abhängigkeit zwische G und H besteht.

3.) Erzeugen Sie je 100 normalverteilte Zufallszahlen X_i∼N(0,1) und U_i∼N(0, 0.5)
(verwenden Sie den Befehl RV.Normal(...) im Fenster Berechnen).
a) Stellen Sie die Dichten von X und U grafisch dar.
Erzeugen Sie je 100 weitere Zufallszahlen mittels
$Y_i \sim$ **10**$* X_i$ **+ 3 + 0.1**$*U_{(i+1)}$**,**
$Z_i \sim$ **10**$* X_i$ **+ 3 + 0.5**$*U_i$ **sowie**
$W_i \sim$ **10**$* X_i$ **+ 3 + 2.0**$*U_{(i-1)}$ **.**
(Hinweis: schräg zu den Zeilen kann man Variable mit den Befehlen Leads und Lag verwenden.)
b) Bestimmen Sie die Korrelationen zwischen den neuen Variablen und X_i.
c) Bestimmen Sie die beste lineare und die beste quadratische Anpassung von Y_i, Z_i und W_i zu X_i . Wie sinnvoll ist letzteres?

Zuerst muß dem System von SPSS mitgeteilt werden, daß 100 "Fälle" zu behandeln sind: Durch Anklicken eines Feldes in der 100sten Zeile des Datenfeldes und Belegen mit einem Wert. Im Fenster Transformieren, Berechnen kann der Befehl RV.Random geladen werden.
zu a) Das Resultat sind 100 zufällig verteilte Werte, die sich ungeordnet auf einer Zahlen-Achse befinden. Man kann diese sofort als Dichte veranschaulichen durch Aufruf des Histogramms für X oder U, welches automatisch eine Klasseneinteilung, und die Sortierung in die Klassen vornimmt. (Es zeigt sich, daß bei 100 noch keine allzu perfekte "Zufälligkeit" erreicht ist!)
Bei der Berechnung von Y und W ist zu beachten, daß Werte schräg zu den Zeilen verknüpft werden sollen: Zeile i mit i+1 oder i-1. Bei W soll zurückgegriffen werden, dazu dient LAG, welches im Fenster Transformieren, Berechnen vorhanden ist: W=10∗X+3+2∗LAG(U) . Bei Y soll vorausgegriffen werden. Dies leistet LEAD, das allerdings extra aufgerufen werden muß: Es ist im Fenster Transformieren, Zeitreihen versteckt; (analog zur kumulativen Summe).
zu c) Man rufe auf Regression, Kurvenanpassung, und klicke dort noch den quadratischen Fall an. Es zeigt sich , daß die quadratische Regression "vernünftig" arbeitet: Da in diesem Fall eine ziemlich klare lineare Abhängigkeit vorliegt, sind die berechneten Koeffizienten des quadratischen Gliedes faktisch Null. Bei dem vorgeschlagenem Fenster wird auch eine Grafik der Lösung mit ausgegeben.

4.) Gegeben sei folgende Tabelle:

x	2	2	2	7	7	7	7	7	7	7
y	1	3	3	3	6	6	6	9	9	9

a) Bestimmen Sie den Pearsonschen und den Spearmanschen Korrelationskoeffizienten.
b) Bestimmen Sie durch lineare Regression die beste Funktion $y = a + b\,x$.
c) Geben Sie eine Kreuztabelle an.

Die kleine Tabelle ist so konstruiert, daß die Werte der Variablen auch gleichzeitig ihre Platz-Ziffern=Ränge sind. In Statistik, Korrelation, Bivariat klicke man entsprechende Felder an. Es resultiert dann

$$\rho_{\text{Pearson}} = \rho_{\text{Spearman}} = 0.75$$

d.h. die beiden Werte sind gleich.