Opdracht 4a

Spreidingsdiagram, kleinste-kwadraten regressielijn, correlatiecoefficient

In 1738 werd in de haven van Stockholm voor een aantal landen voor elk land geregistreerd hoeveel schepen het gebied als herkomst hadden, en hoeveel schepen het gebied als bestemming hadden. Dit leverde de onderstaande gegevens op (Bron:J. Th. Lindblad, Sweden’s trade with the Dutch Republic, 1738-1795 (Assen 1982) Hfst. III.1.)).

gebied	aangekomen schepen	vertrokken schepen
Denemarken	41	22
Duitsland	7	9
Engeland	67	107
Frankrijk	29	33
Portugal	24	19
Nederland	23	30
Rusland	103	90
Spanje	3	1

a. Maak een spreidingsdiagram van de gegevens. Zorg dat in de grafiek ook de kleinste-kwadraten regressielijn wordt getekend.

b. Bepaal de vergelijking van de kleinste-kwadraten regressielijn.

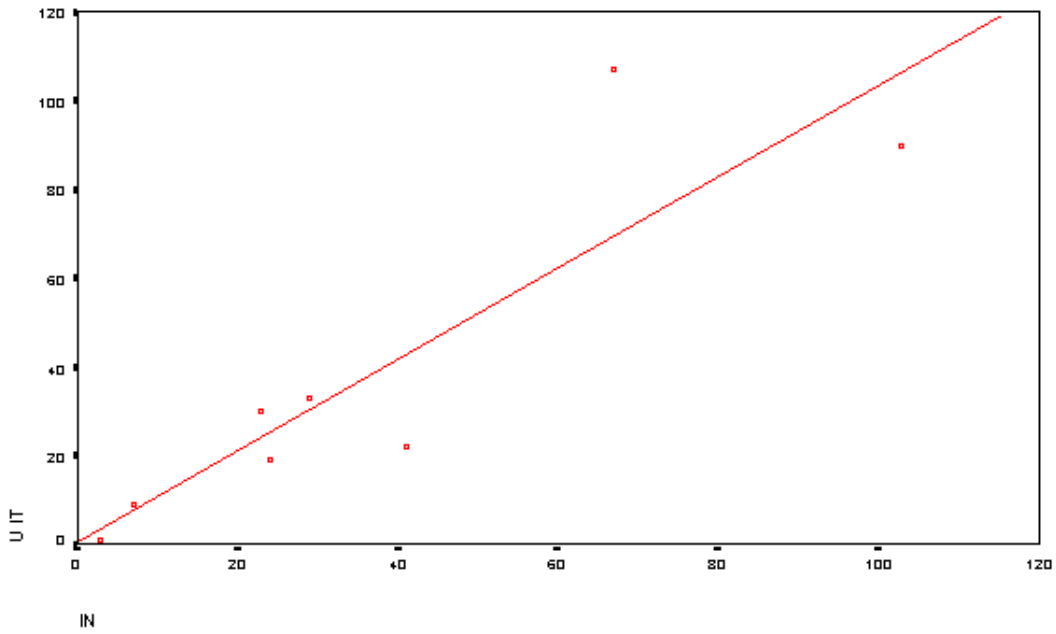
c. Stel het aantal aangekomen schepen is gelijk aan 23. Wat is dan volgens de tabel het aantal vertrokken schepen? En wat is volgens de kleinste-kwadraten regressielijn het aantal vertrokken schepen?

d. Bereken de correlatiecoefficient r tussen de aangekomen schepen en de vertrokken schepen.

e. Hoeveel procent van de variantie van het aantal vertrokken schepen wordt verklaard door het aantal aangekomen schepen?

Opdracht 4a - berekening

a. Het aantal aangekomen schepen is uiteindelijk afhankelijk van het aantal vertrokken schepen, en het aantal vertrokken schepen is uiteindelijk afhankelijk van het aantal aangekomen schepen. De x-as kan het aantal aangekomen schepen aangeven en de y-as het aantal vertrokken schepen, maar net andersom mag ook.



Voor b. en d. gebruiken en berekenen we eerst de volgende gegevens:

x	sqr(x)	y	sqr(y)	xy
41	1681	22	484	902
7	49	9	81	63
67	4489	107	11450	7169
29	841	33	1089	957
24	576	19	361	456
23	529	30	900	690
103	10610	90	8100	9270
3	9	1	1	3
297	18784	311	22465	19510

```
sum_x      :    297
sum_sqr(x) :  18783
sum_y      :    311
sum_sqr(y) :  22465
sum_xy     :  19510

x_mean = 297 / 8 = 37.125
y_mean = 311 / 8 = 38.875

b.        sum_xy - ((sum_x * sumy) / n)
   b = --------------------------------
       sum_sqr(x) - (sqr(sum_x) / n)

            19510 - ((297 * 311) / 8)
     = --------------------------------
            18783 - (sqr(297) / 8)

     = 1.0267182338

   a = y_mean - (b             * x_mean)
```

= 38.875 - (1.0267182338 * 37.125)

= 0.75808556925

y_dakje = 0.75808556925 + (1.0267182338 * x)

c. Als het aantal aangekomen schepen gelijk is aan 23, is het aantal vertrokken schepen volgens de tabel gelijk aan 30, en volgens de kleinste-kwadraten regressie lijn gelijk aan 0.75808556925 + (1.0267182338 * 23) = 24.372604947

d.
$$r = \frac{sum\_xy - (n * x\_mean * y\_mean)}{sqrt((sum\_sqr(x)-(n*sqr(x\_mean)))*(sum\_sqr(y)-(n*sqr(y\_mean))))}$$

$$= \frac{19510 - (8 * 37.125 * 38.875)}{sqrt((18783-(8 * sqr(37.125))) * (22465 - (8 * sqr(38.875))))}$$

= 0.88777566211

e. Het percentage van de variantie van het aantal vertrokken schepen dat verklaard wordt door het aantal aangekomen schepen is gelijk aan het kwadraat van de correlatie tussen het aantal aangekomen schepen en het aantal vertrokken schepen.

sqr(r) = sqr(0.88777566211) = 0.78814562623 = 78.814562623%

Opdracht 4a - S-PLUS

Plaats de gegevens in de tabel. Noem de kolom van aangekomen schepen ’binnen’ en de kolom van de vertrokken schepen ’buiten’.

a. Kies >Graph >2D Plot. Kies onder Axes Type voor Linear, en onder Plot Type voor Fit - Linear Least Squares (x, y1, y2, ...). Klik op >OK. Selecteer onder Data Columns en achter x Columns de variabele ’binnen’, en achter y Columns de variabele ’buiten’. Klik op >OK.

   In de grafiek wordt de lineaire regressielijn weergegeven door de lineaire lijn.

b. Kies >Statistics >Regression >Linear. Selecteer achter Dependent de variabele ’buiten’, en achter Independent de variabele ’binnen’. Klik op >OK.

   In het Report-venster vinden we onder Coefficients een tabel. In deze tabel vinden we in rij ’binnen’ kolom Value de waarde voor b, en in rij (Intercept) kolom Value vinden we de waarde voor a. Deze waarden moeten ingevuld worden in de vergelijking y = a + bx.

c.

d. Kies >Statistics >Data Summaries >Correlations. Selecteer achter Variables < ALL >. Klik op >OK.

   In het Report-venster vinden we een tabel met de correlatie tussen ’binnen’ en ’buiten’ in rij ’binnen’ en kolom ’buiten’, of rij ’buiten’ en kolom ’binnen’.

e.

Opdracht 4a - SPSS

Plaats de gegevens in de tabel. Noem de kolom van aangekomen schepen ’binnen’ en de kolom van de vertrokken schepen ’buiten’.

a. Kies nu >Graphs >Scatter. Kies in het window ’Scatterplot’ voor ’Simple’. Klik vervolgens op ’Define’. Nu kom je in het window ’Simple Scatterplot’. Verplaats ’binnen’ naar XAxis en ’buiten’ naar YAxis. Klik nu op >OK. In het output-window verschijnt nu het spreidingsdiagram. Selecteer het diagram door er met de linkermuisknop op te klikken. Kies nu >Edit >SPSS Chart Object >Open. Je komt nu in de SPSS Chart Editor. Kies daar >Chart >Options. Zet onder Fit Line ’Total’ aan. Klik op >Fit Options. Klik op >Linear Regression en daarna op >Continue. Klik tenslotte op >OK. Sluit de SPSS Chart Editor af.

b. Kies >Statistics >Regression >Linear. Je komt nu in het window ’Linear Regression’. Breng ’binnen’ naar independent en ’buiten’ naar dependent. Klik op >OK. Kijk nu in het output-window in de tabel ’Coefficient’. In rij BINNEN kolom B vinden we b. In rij (constant) kolom B vinden we a. Deze waarden moeten ingevuld worden in de vergelijking y = a + bx.

c.

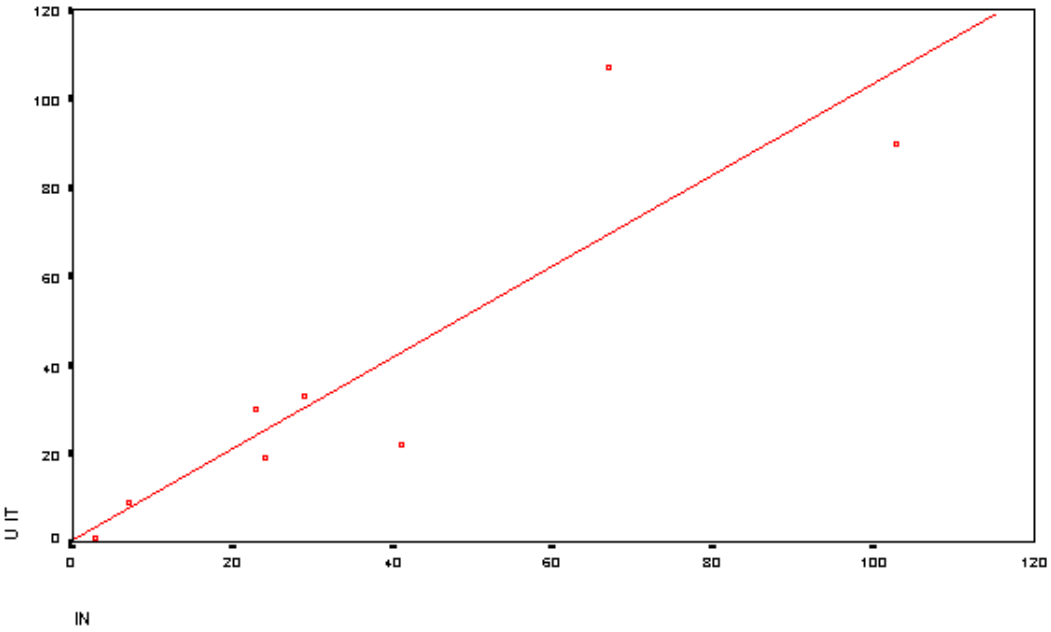
d. Kies >Statistics >Correlate >Bivariate. Je komt dan in het window ’Bivariate Correlations’. Breng ’binnen’ en ’buiten’ naar rechts onder ’Variables’. Onder Correlation Coefficients moet gekozen zijn voor Pearson. Klik op >OK. In het output-window vind je in de tabel ’Correlations’ de correlatie tussen ’binnen’ en ’buiten’ in rij BINNEN kolom BUITEN, of rij BUITEN kolom BINNEN.

e.

Opdracht 4a - verslag

In 1738 werd in de haven van Stockholm voor een aantal landen voor elk land geregistreerd hoeveel schepen het gebied als herkomst hadden, en hoeveel schepen het gebied als bestemming hadden.

a. Maak een spreidingsdiagram van de gegevens. Zorg dat in de grafiek ook de kleinste-kwadraten regressielijn wordt getekend.



b. Bepaal de vergelijking van de kleinste-kwadraten regressielijn.

   y_dakje = 0.75808556925 + (1.0267182338 * x)

c. Stel het aantal aangekomen schepen is gelijk aan 23. Wat is dan volgens de tabel het aantal vertrokken schepen? En wat is volgens de kleinste-kwadraten regressielijn het aantal vertrokken schepen?

   Als het aantal aangekomen schepen gelijk is aan 23, is het aantal vertrokken schepen volgens de tabel gelijk aan 30, en volgens de kleinste-kwadraten regressie lijn gelijk aan 0.75808556925 + (1.0267182338 * 23) = 24.372604947

d. Bereken de correlatiecoefficient r tussen de aangekomen schepen en de vertrokken schepen.

   r = 0.88777566211

e. Hoeveel procent van de variantie van het aantal vertrokken schepen wordt verklaard door het aantal aangekomen schepen?

   Het percentage van de variantie van het aantal vertrokken schepen dat verklaard wordt door het aantal aangekomen schepen is gelijk aan het kwadraat van de correlatie tussen het aantal aangekomen schepen en het aantal vertrokken schepen.

   sqr(r) = sqr(0.88777566211) = 0.78814562623 = 78.814562623%