Opdracht 12a
---

enkelvoudige lineaire regressie

Kan de leeftijd waarop een kind begint te spreken voorspellen hoe zijn score zal zijn bij een latere test op verstandelijke vermogens? Een studie over de cognitieve ontwikkeling van jonge kinderen registreerde voor ieder van 21 kinderen de leeftijd (in maanden) waarop het eerste woord werd gesproken en de score van de Gesell Aanpassingstoets die veel later werd afgenomen. Hieronder staan de resultaten. (Gegevens uit N. R. Draper en J. A. John, 'Influential observations and outliers in regression', Technometrics, 23 (1981), blz. 21-26.)

geval	leeftijd	score
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

a. Teken een spreidingsdiagram leeftijd vs. score. Zorg dat ook de kleinste-kwadratenlijn wordt getekend. Lijkt de samenhang lineair te zijn? Mag de kleinste-kwadratenlijn bepaald worden? Onderzoek vervolgens de residuen (de verschillen tussen de waargenomen waarden en de waarden die voorspeld worden door de kleinste-kwadratenlijn). Teken twee spreidingsdiagrammen: geval vs. residu en leeftijd vs. residu. Het gemiddelde van de residuen is altijd gelijk aan 0. Teken nu in elk van de twee spreidingsdiagrammen ook de lijn residu=0. Zijn er opvallend verdachte patronen of abnormale waarnemingen?

b. Bepaal b1 en b0 en geef de vergelijking van de kleinste-kwadratenlijn.

c. We doen verder onderzoek naar de residuen. Teken een normaal-kwantiel-plot van de residuen. Vormen de punten ongeveer een rechte lijn? Zijn de residuen normaal verdeeld? Bepaal ook s, de standaardfout van de residuen.

d. Bepaal s_b1, de standaardfout voor b1, en bepaal s_b0, de standaardfout voor b0.

e. We verwachten dat een kind dat op jonge leeftijd al het eerste woord spreekt later een hoge score zal hebben. Geef voor beta1 een 95%-betrouwbaarheidsinterval. Formuleer H_0 en H_a en bewijs dat de score ne-

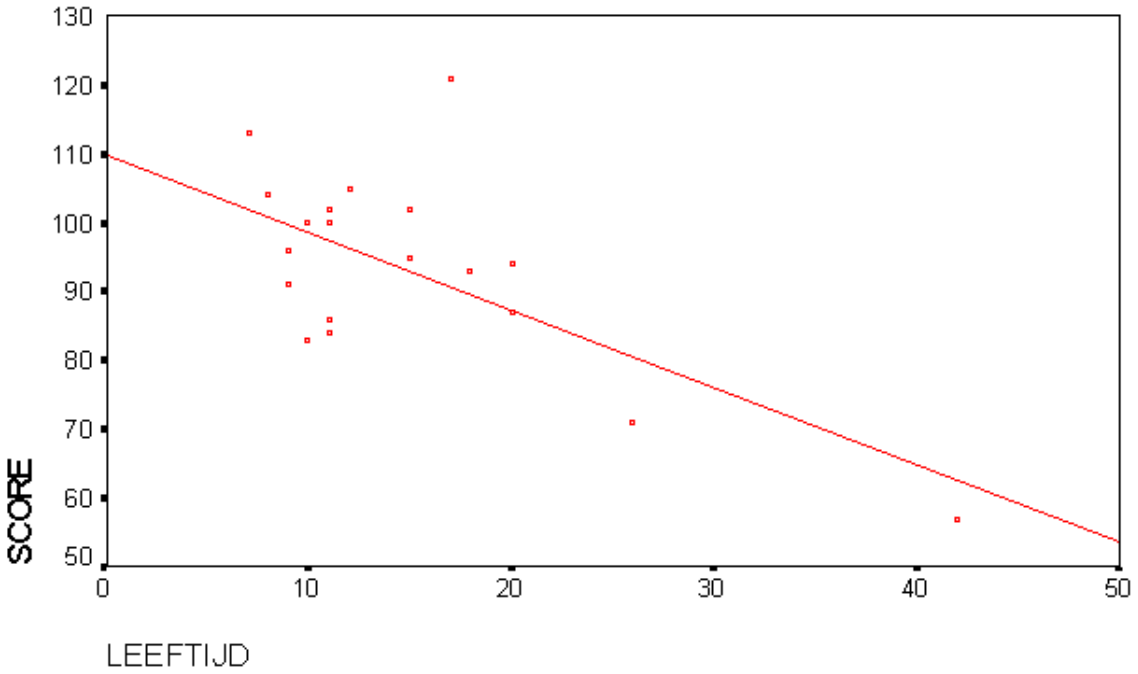
gatief samenhangt met de leeftijd.

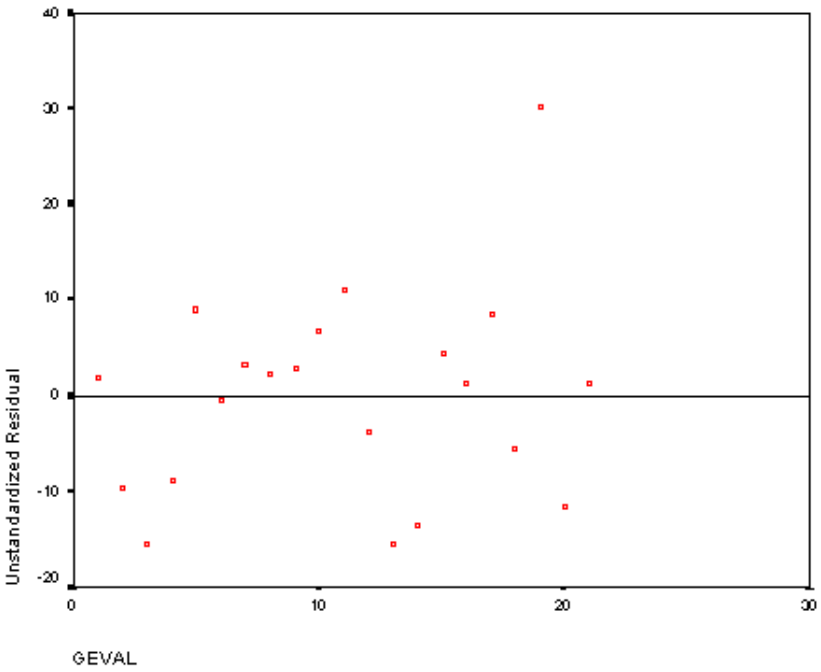
f. De constante beta0 representeert de gemiddelde score van kinderen die op de leeftijd van 0 maanden al het eerste woord spraken. Geef voor beta0 een 95%-betrouwbaarheidsinterval. Formuleer H_0 en H_a en bewijs dat beta0 positief is.

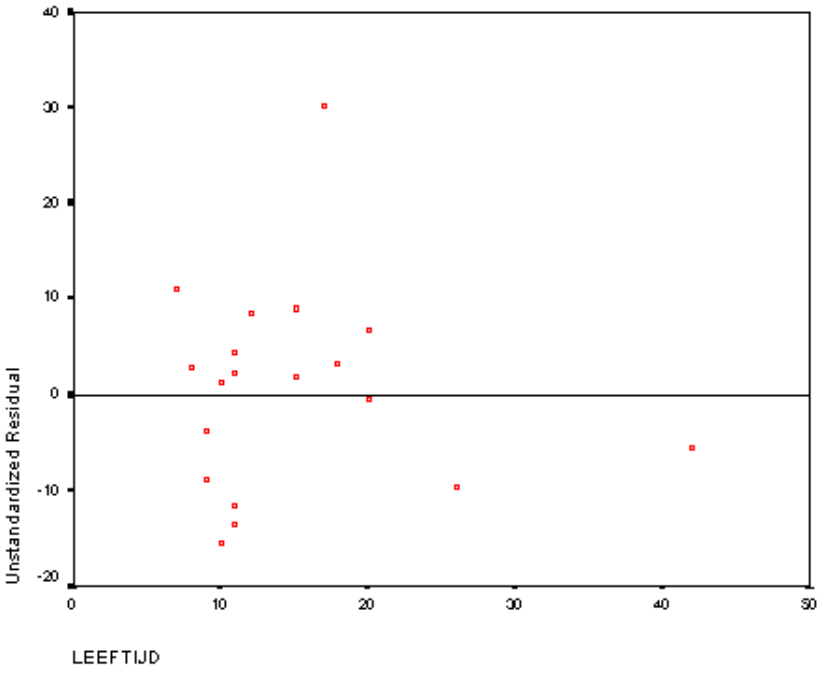
Opdracht 12a - berekening

Beschouw voor e. het kader op bladzijde 533.

a.







b. geval	x_i	y_i	sqr(x_i)	x_i*y_i
1	15	95	225	1425
2	26	71	676	1846
3	10	83	100	830

4	9	91	81	819
5	15	102	225	1530
6	20	87	400	1740
7	18	93	324	1674
8	11	100	121	1100
9	8	104	64	832
10	20	94	400	1880
11	7	113	49	791
12	9	96	81	864
13	10	83	100	830
14	11	84	121	924
15	11	102	121	1122
16	10	100	100	1000
17	12	105	144	1260
18	42	57	1764	2394
19	17	121	289	2057
20	11	86	121	946
21	10	100	100	1000
	302	1967	5606	26864

sum_x = 302
sum_y = 1967
sum_sqr(x) = 5606
sum_xy = 26864
n = 21

x_gemiddeld = sum_x/n = 302/21 = 14.38
y_gemiddeld = sum_y/n = 1967/21 = 93.67

$$b1 = \frac{sum\_xy - (1/n)*sum\_x*sum\_y}{sum\_sqr(x) - (1/n)*sqr(sum\_x)}$$

$$= \frac{26864 - (1/21)*302*1967}{5606 - (1/21)*sqr(302)}$$

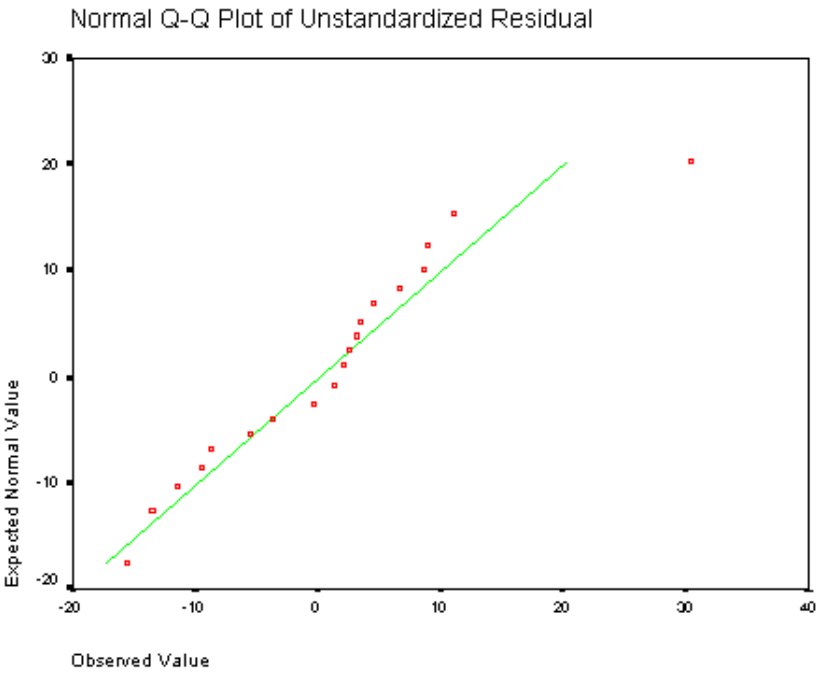
$$= -1.13$$

b0 = y_gemiddeld - (b1*x_gemiddeld)
= 93.67 - (-1.13*14.38)
= 109.92

De vergelijking van de kleinste-kwadratenlijn is derhalve:

y_dakje = 109.92 - 1.13 * x

c.



geval	x_i	y_i	y_dakje_i	e_i	sqr(e_i)
1	15	95	92.97	2.03	4.12
2	26	71	80.54	- 9.54	91.01
3	10	83	98.62	-15.62	243.98
4	9	91	99.75	- 8.75	76.56
5	15	102	92.97	9.03	81.54
6	20	87	87.32	- 0.32	0.10
7	18	93	89.58	3.42	11.70
8	11	100	97.49	2.51	6.30
9	8	104	100.88	3.12	9.73
10	20	94	87.32	6.68	44.62
11	7	113	102.01	10.99	120.78
12	9	96	99.75	- 3.75	14.06
13	10	83	98.62	-15.62	243.98
14	11	84	97.49	-13.49	181.98
15	11	102	97.49	4.51	20.34
16	10	100	98.62	1.38	1.90
17	12	105	96.36	8.64	74.65
18	42	57	62.46	- 5.46	29.81
19	17	121	90.71	30.29	917.48
20	11	86	97.49	-11.49	132.02
21	10	100	98.62	1.38	1.90
					2308.60

sum_sqr(e_i) = 2308.60
s = sqrt(sum_sqr(e_i)/(n-2)) = sqrt(2308.60/(21-2)) = 11.02

d. De standaardfout voor de helling b1 van de kleinste-kwadratenregressie-lijn is:

$$s_{b1} = \frac{s}{\ldots}$$

sqrt(sum_sqr(x_i-x_gemiddeld))

$$= \frac{11.02}{\text{sqrt}(1262.86)}$$

$$= 0.31$$

De standaardfout voor de constante b0 is:

$$s\_b0 = s \cdot \text{sqrt}\left(\frac{1}{n} + \frac{\text{sqr}(x\_gemiddeld)}{\text{sum\_sqr}(x\_i\text{-}x\_gemiddeld)}\right)$$

$$= 11.02 \cdot \text{sqrt}\left(\frac{1}{21} + \frac{\text{sqr}(14.38)}{1262.86}\right)$$

$$= 5.07$$

e. Voor een betrouwbaarheid van 95% geeft tabel E bij 19 vrijheidsgraden
t_ster = 2.093

Het betrouwbaarheidsinterval voor beta1 is:

b1 +/- t_ster * s_b1 =

-1.13 +/- 2.093 * 0.31 =

-1.13 +/- 0.65 =

(-1.78,-0.48)

Voor toetsing van

H_0: beta1=0
H_a: beta1<0

berekenen we de t-grootheid:

$$t = \frac{b1}{s\_b1} = \frac{-1.13}{0.31} = -3.65$$

Het aantal vrijheidsgraden is n-2 = 21-2 = 19.

Als H_a: beta1 < 0, dan de P-waarde is P(T<=t). P(T<=t) heeft dezelfde waarde als P(T>=-t). Dus P(T<=-3.65) heeft dezelfde waarde als P(T>=3.65). In tabel E is gegeven P(T>=t). Voor P(T>=3.65) en 19 vrijheidsgraden zien we dat t ligt tussen 3.579 en 3.883, en p ligt tussen 0.001 en 0.0005. Er geldt P(T>=3.65) = 0.0005 < p < 0.001.

f. Voor een betrouwbaarheid van 95% geeft tabel E bij 19 vrijheidsgraden
t_ster = 2.093

Het betrouwbaarheidsinterval voor beta1 is:

b0 +/- t_ster * s_b0 =

109.92 +/- 2.093 * 5.07 =

109.92 +/- 10.61   =

(99.31,120.53)

Voor toetsing van

H_0: beta0=0
H_a: beta0>0

berekenen we de t-grootheid:

$$t = \frac{b0}{s\_b0} = \frac{109.92}{5.07} = 21.68$$

Het aantal vrijheidsgraden is n-2 = 21-2 = 19.

In tabel E is gegeven P(T>=t). Voor P(T>=21.68) en 19 vrijheidsgraden zien we dat t groter is dan 3.883. Er geld P(T>=21.68) = p < 0.0005.

Opdracht 12a - SPSS

Voer de gegevens in in een tabel. Noem de kolommen 'geval', 'leeftijd' en 'score'.

Kies >Statistics >Regression >Linear. Plaats 'leeftijd' in Independent(s) en 'score' in Dependent. Klik op >Statistics. Zorg dat >Model Fit aan staat en dat onder Regression Coefficients >Estimates en >Confidence intervals aanstaan. Klik op >Continue. Klik vervolgens op >Save. Zorg dat onder Residuals 'Unstandardized' aan staat. Klik op >Continue. Klik op >OK.

In het output-window vinden we vijf tabellen, en in de data editor vinden we de residuen in de kolom 'res_1'.

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 'leeftijd' naar XAxis en 'score' 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.

   Kies nu >Graphs >Scatter. Kies in het window 'Scatterplot' voor 'Simple'. Klik vervolgens op 'Define'. Nu kom je in het window 'Simple Scatterplot'. Klik op >Reset. Verplaats 'geval' naar XAxis en 'Unstandardized Residual [res_1]' 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 >Reference Line. Kies >Y scale en daarna op >OK. Achter 'Position of Line(s)' moet de waarde 0 zijn ingevuld. Klik op >Add. Klik tenslotte op OK. Sluit de SPSS Chart Editor af.

   Kies nu >Graphs >Scatter. Kies in het window 'Scatterplot' voor

’Simple’. Klik vervolgens op ’Define’. Nu kom je in het window ’Simple Scatterplot’. Klik op >Reset. Verplaats ’leeftijd’ naar XAxis en ’Unstandardized Residual [res_1]’ 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 >Reference Line. Kies >Y scale en daarna op >OK. Achter ’Position of Line(s)’ moet de waarde 0 zijn ingevuld. Klik op >Add. Klik tenslotte op OK. Sluit de SPSS Chart Editor af.

b. De vierde tabel in het output-window heet ’Coefficients’. Onder de kolom ’Unstandardized Coefficient’ en de subkolom ’B’ vinden we in rij ’LEEFTIJD’ b1 en in rij ’(Constant)’ b0. We zien dat b1=-1.127 en b0= 109.874. De vergelijking van de kleinste-kwadratenlijn is derhalve:

   y_dakje = 109.874 - 1.127 * x

c. Kies >Graphs >Q-Q. Plaats ’Unstandardized Residual [res_1]’ onder Variables. Klik op >OK. In het output-window vind je achtereenvolgens twee grafieken. De grafiek ’Normal Q-Q Plot of Unstandardize Residual’ is de eerste grafiek. Deze grafiek is de grafiek die gevraagd wordt

   De tweede tabel in het output-window heet ’Model Summary’. Onder ’Std. Error of the Estimate’ vinden we s. We zien dat s=11.0229.

d. De vierde tabel in het output-window heet ’Coefficients’. Onder de kolom ’Unstandardized Coefficient’ en de subkolom ’Std. Error’ vinden we in rij ’LEEFTIJD’ de standaardfout voor b1 en in rij ’(Constant)’ de standaardfout voor b0. We zien dat s_b1=0.310 en b0=5.068.

e. De vierde tabel in het output-window heet ’Coefficients’. Onder de kolom ’95% Confidence Interval for B’ vinden we in rij ’LEEFTIJD’ de onder- en bovengrens van het betrouwbaarheidsinterval voor beta1 in respectievelijk de subkolom ’Lower Bound’ en de subkolom ’Upper Bound’. Deze is (-1.776,-0.478).

   Onder de kolom ’t’ vinden we in de rij ’LEEFTIJD’ onder de t-waarde (-3.633). Onder de kolom ’Sig.’ in de rij ’LEEFTIJD’ vinden we de tweezijdige P-waarde (0.002). Omdat we willen toetsen of de score negatief samenhangt met de leeftijd bepalen we de linkseenzijdige P-waarde. Daar de t-waarde negatief is, is de P-waarde gelijk aan 0.002/2 = 0.001.

f. De vierde tabel in het output-window heet ’Coefficients’. Onder de kolom ’95% Confidence Interval for B’ vinden we in rij ’(Constant)’ de onder- en bovengrens van het betrouwbaarheidsinterval voor beta0 in respectievelijk de subkolom ’Lower Bound’ en de subkolom ’Upper Bound’. Deze is (99.267,120.481).

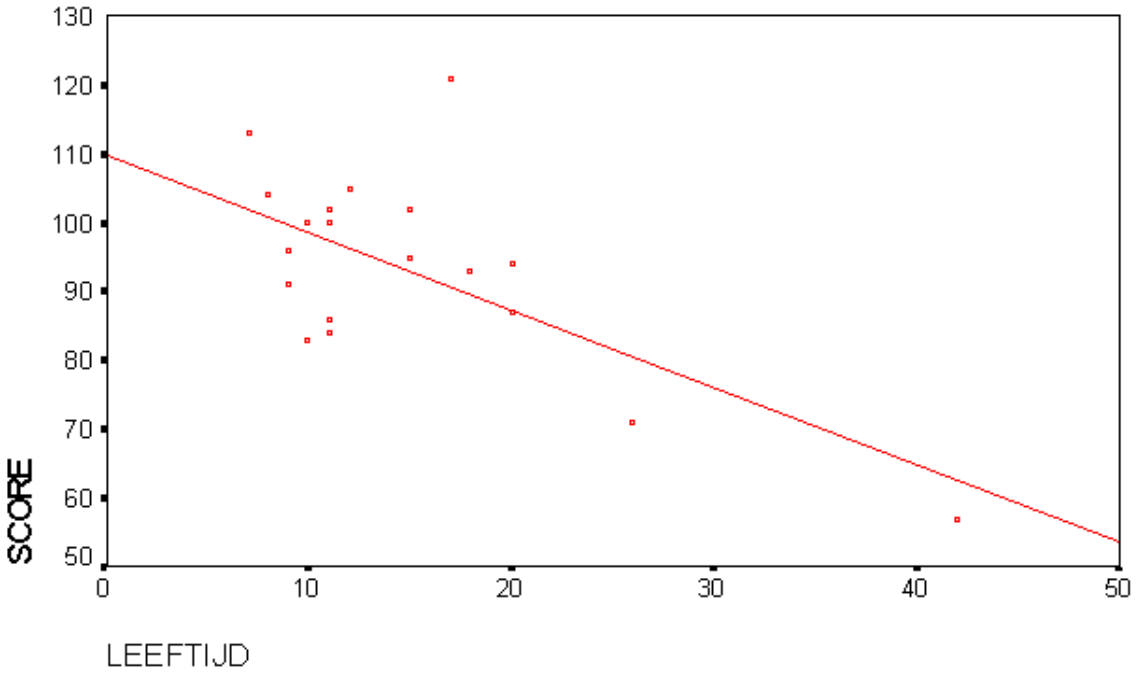
   Onder de kolom ’t’ vinden we in de rij ’LEEFTIJD’ onder de t-waarde (21.681). Onder de kolom ’Sig.’ in de rij ’LEEFTIJD’ vinden we de tweezijdige P-waarde (0.000). Omdat we willen toetsen of de gemiddelde score positief is, bepalen we de rechtseenzijdige P-waarde. Deze is gelijk aan 0.0000/2 = 0.0000.

Opdracht 12a - verslag

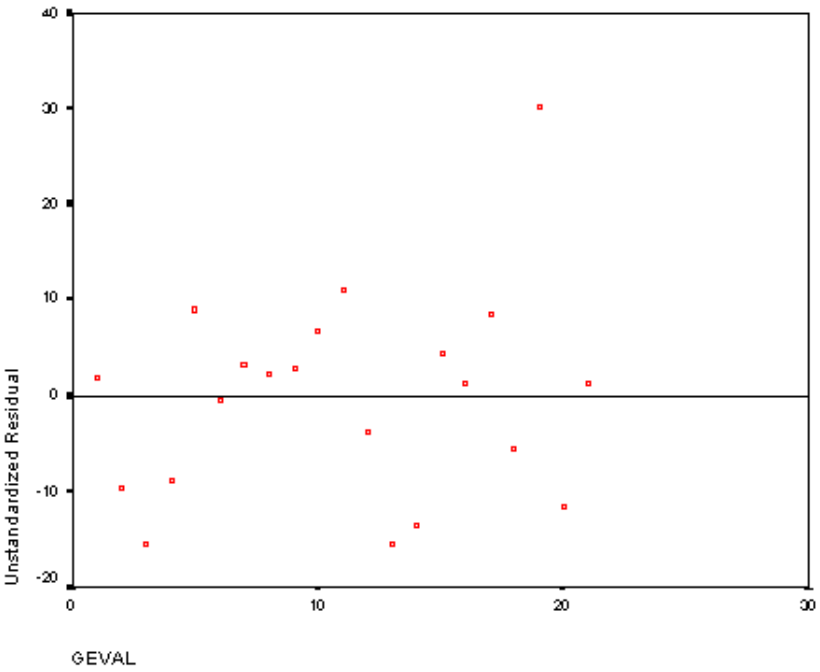
Kan de leeftijd waarop een kind begint te spreken voorspellen hoe zijn score zal zijn bij een latere test op verstandelijke vermogens? Een studie over de cognitieve ontwikkeling van jonge kinderen registreerde voor ieder

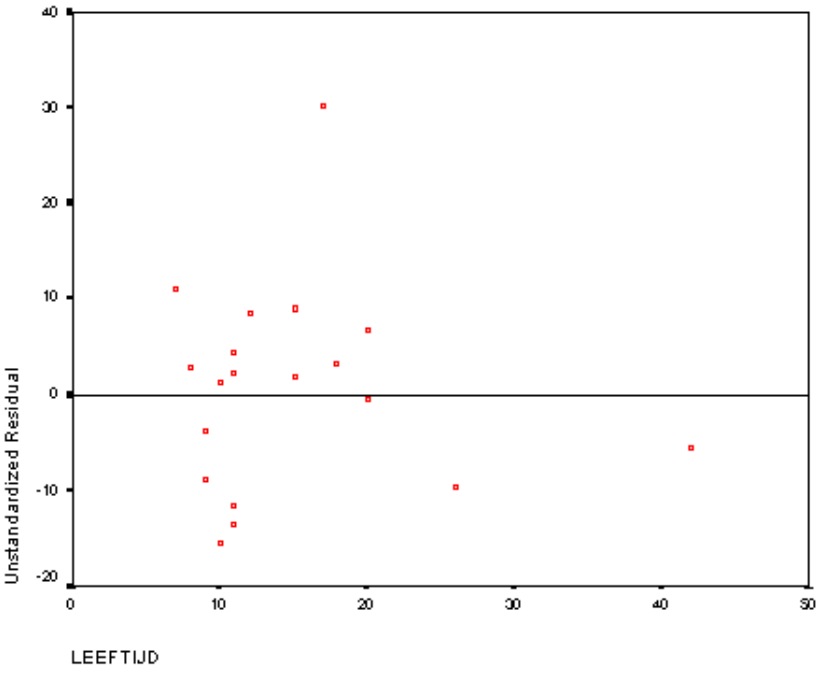
van 21 kinderen de leeftijd (in maanden) waarop het eerste woord werd gesproken en de score van de Gesell Aanpassingstoets die veel later werd afgenomen.

a. Teken een spreidingsdiagram leeftijd vs. score. Zorg dat ook de kleinste-kwadratenlijn wordt getekend. Lijkt de samenhang lineair te zijn? Mag de kleinste-kwadratenlijn bepaald worden? Onderzoek vervolgens de residuen (de verschillen tussen de waargenomen waarden en de waarden die voorspeld worden door de kleinste-kwadratenlijn). Teken twee spreidingsdiagrammen: geval vs. residu en leeftijd vs. residu. Het gemiddelde van de residuen is altijd gelijk aan 0. Teken nu in elk van de twee spreidingsdiagrammen ook de lijn residu=0. Zijn er opvallend verdachte patronen of abnormale waarnemingen?



De samenhang lijkt ruwweg lineair te zijn. We kunnen dus zonder problemen de vergelijking van de kleinste-kwadraten-lijn bepalen.



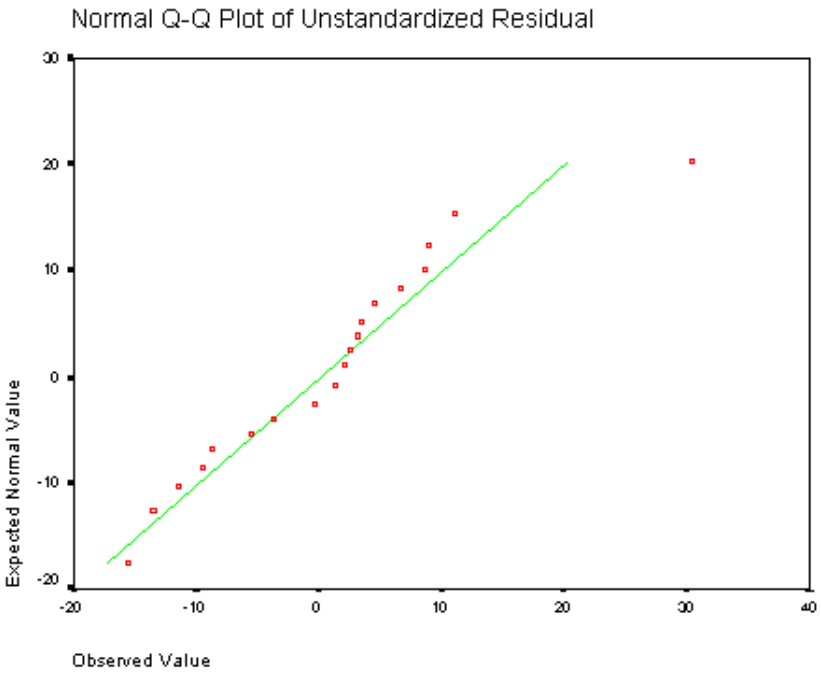


Het patroon in beide grafieken moet een ongestructureerde band zijn die gecentreerd is om de lijn residu=0 en symmetrisch om de lijn residu=0. Beide grafieken vertonen ruwweg dit patroon en tonen geen abnormale waarnemingen.

b. Bepaal b1 en b0 en geef de vergelijking van de kleinste-kwadratenlijn.

b1 = -1.127 en b0 = 109.874
y_dakje = 109.874 - 1.127 * x

c. We doen verder onderzoek naar de residuen. Teken een normaal-kwantiel-plot van de residuen. Vormen de punten ongeveer een rechte lijn? Zijn de residuen normaal verdeeld? Bepaal ook s, de standaardfout van de residuen.



De punten vormen ongeveer een rechte lijn. Aangenomen mag worden dat de residuen normaal verdeeld zijn. Gevolg is dat b1 en b0 normaal verdeeld zijn met gemiddelden beta1 en beta0. We mogen dus betrouwbaarheidsintervallen berekenen en significantietoetsen gebruiken voor beta1 en beta0.

s = 11.0229

d. Bepaal s_b1, de standaardfout voor b1, en bepaal s_b0, de standaardfout voor b0.

s_b1 = 0.310
s_b0 = 5.068

De kleinste-kwadratenlijn loopt door het punt met als x-waarde het gemiddelde van alle leeftijden, en als y-waarde het gemiddelde van alle scores. We zouden ook door elk punt afzonderlijk (waarbij de x-waarde een leeftijd is en de y-waarde de bijbehorende score) een lijn kunnen tekenen die dan zoveel mogelijk recht doet aan de overige punten. We krijgen dan evenveel lijnen als dat er punten zijn. Iedere lijn heeft z'n eigen b1 en b0. Nu is s_b1 de standaardfout van alle b1's en s_b0 is de standaardfout van alle b0's.

e. We verwachten dat een kind dat op jonge leeftijd al het eerste woord

spreekt later een hoge score zal hebben. Geef voor beta1 een 95%-betrouwbaarheidsinterval. Formuleer H_0 en H_a en bewijs dat de score negatief samenhangt met de leeftijd.

We kunnen met 95% zekerheid stellen dat beta1 ligt tussen -1.776 en -0.478.

H_0: beta1=0
H_a: beta1<0

De P-waarde is de kans, berekend onder de aanname dat H_0 waar is, dat t een waarde zou aannemen die even extreem of nog extremer is dan -3.633. Omdat we alleen geinteresseerd zijn in de negatieve samenhang en we dus linkseenzijdig toetsen, is de P-waarde gelijk aan 0.002/2 = 0.001. De P-waarde is kleiner dan alfa, want 0.001<0.05, dus wordt H_0 verworpen. Er is sprake van een negatieve samenhang.

f. De constante beta0 representeert de gemiddelde score van kinderen die op de leeftijd van 0 maanden al het eerste woord spraken. Geef voor beta0 een 95%-betrouwbaarheidsinterval. Formuleer H_0 en H_a en bewijs dat beta0 positief is.

We kunnen met 95% zekerheid stellen dat beta0 ligt tussen 99.267 en 120.481.

H_0: beta0=0
H_a: beta0>0

De P-waarde is de kans, berekend onder de aanname dat H_0 waar is, dat t een waarde zou aannemen die even extreem of nog extremer is dan 21.681. Omdat we willen toetsen of beta0 positief is en we dus rechtseenzijdig toetsen, is de P-waarde gelijk aan 0.000/2 = 0.000. De P-waarde is kleiner dan alfa, want 0.000<0.05, dus wordt H0 verworpen. Beta0 is positief.