QQ-plot og test for normalfordelte residualer

Her er sakset lidt fra "Lærebog i matematik A3 stx" og "MAT A2 stx" 😊

Fordelingsfunktionen F for en normalfordelt stokastisk variabel er kontinuert og voksende på $\mathbb{R}$ med værdimængde $]0;1[$. Funktionen har dermed en omvendt funktion F^{-1}, defineret på $]0;1[$, som vi kalder *fraktilfunktionen* hørende til F.

Fraktilfunktionen hørende til Φ betegner vi med Φ^{-1}.

Eksempel 4.4.1

For den normalfordelte stokastiske variabel $X \sim N(25,4)$ vil vi bestemme tallet q, så

$$P(X \leq q) = 0{,}80$$

Når F er fordelingsfunktionen hørende til X, skal vi løse ligningen $F(q) = 0{,}80$.

Vi får med CAS

$$F(q) = 0{,}80 \quad \Leftrightarrow \quad q = F^{-1}(0{,}80) \approx 28{,}37$$

QQ-plot/fraktilplot

Antag, at $X \sim N(\mu, \sigma)$ med fordelingsfunktion F. Da er

$$P(X \leq x) = F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Og ved at anvende den inverse fordelingsfunktion Φ^{-1} til standardnormalfordelingen fås

$$\Phi^{-1}(F(x)) = \Phi^{-1}(\Phi(\frac{x-\mu}{\sigma})) = \frac{x-\mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$$

som viser, at punkterne $(x, \Phi^{-1}(F(x)))$ ligger på en ret linje med hældningskoefficient $\frac{1}{\sigma}$ og konstantled $-\frac{\mu}{\sigma}$.

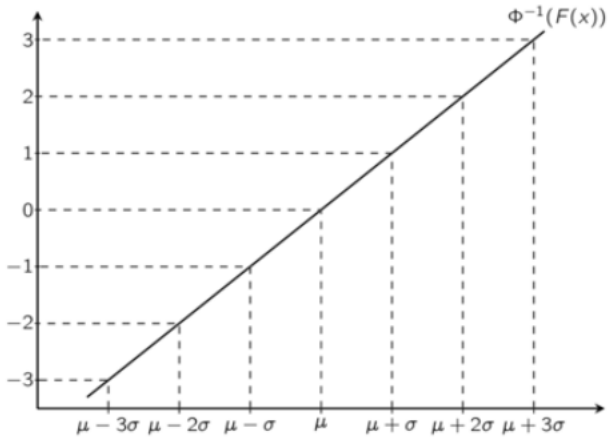
Nogle udvalgte funktionsværdier fremgår af Tabel 4.4.1 nedenfor.

x	$\mu-3\sigma$	$\mu-2\sigma$	$\mu-\sigma$	μ	$\mu+\sigma$	$\mu+2\sigma$	$\mu+3\sigma$
$\Phi^{-1}(F(x))$	-3	-2	-1	0	1	2	3

Tabel 4.4.1

Den rette linje, som er graf for $\Phi^{-1}(F(x))$, tegner vi typisk i et *fraktilplot*.

I et fraktilplot er begge koordinatakser ækvidistante, og figur 4.4.1 viser den generelle opbygning af et fraktilplot.



Figur 4.4.1

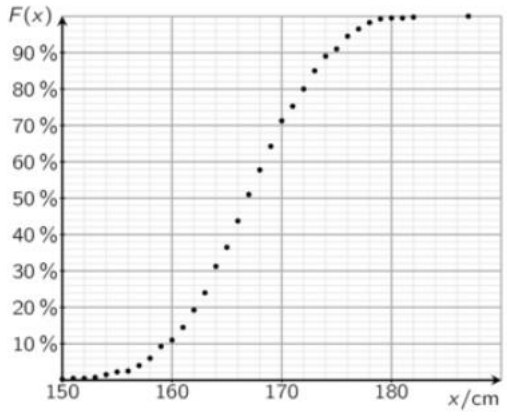
Eksempel 4.4.6

En opmåling af højderne på 400 gymnasiepiger har resulteret i de kumulerede frekvenser i Tabel 4.4.2.

x	$h(x)$	$F(x)$	x	$h(x)$	$F(x)$
150	1	0,25 %	167	29	51,00 %
151	1	0,50 %	168	27	57,75 %
152	0	0,50 %	169	26	64,25 %
153	1	0,75 %	170	28	71,25 %
154	3	1,50 %	171	16	75,25 %
155	3	2,25 %	172	19	80,00 %
156	1	2,50 %	173	20	85,00 %
157	6	4,00 %	174	16	89,00 %
158	8	6,00 %	175	8	91,00 %
159	13	9,25 %	176	14	94,50 %
160	7	11,00 %	177	8	96,50 %
161	14	14,50 %	178	7	98,25 %
162	19	19,25 %	179	4	99,25 %
163	19	24,00 %	180	1	99,50 %
164	29	31,25 %	181	0	99,50 %
165	21	36,50 %	182	1	99,75 %
166	29	43,75 %	187	1	100,00 %

Tabel 4.4.2

De kumulerede frekvenser giver os punkterne på sumkurven vist i figur 4.4.3.



Figur 4.4.3

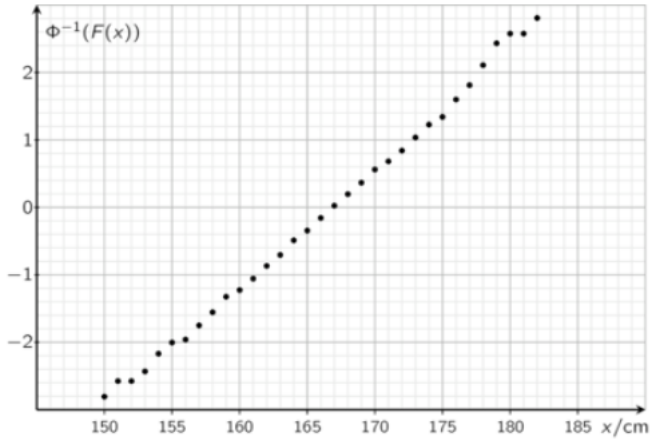
Punkterne ligger ret symmetrisk, og formen på en tegnet kurve gennem punkterne minder meget om grafen for fordelingsfunktionen for en normalfordeling. Vi vil derfor modellere data ved hjælp af en normalfordeling.

Vi tegner et fraktilplot hørende til de kumulerede data. Med tallene fra Tabel 4.4.2 , får vi Tabel 4.4.3.

x	$\Phi^{-1}(F(x))$	x	$\Phi^{-1}(F(x))$	x	$\Phi^{-1}(F(x))$
150	-2,80703	161	-1,05812	172	0,841621
151	-2,57583	162	-0,86872	173	1,036433
152	-2,57583	163	-0,7063	174	1,226528
153	-2,43238	164	-0,48878	175	1,340755
154	-2,17009	165	-0,34513	176	1,598193
155	-2,00465	166	-0,15731	177	1,811911
156	-1,95996	167	0,025069	178	2,108358
157	-1,75069	168	0,195502	179	2,432379
158	-1,55477	169	0,365149	180	2,575829
159	-1,32552	170	0,560703	181	2,575829
160	-1,22653	171	0,682378	182	2,807034

Tabel 4.4.3

Vi indsætter tabellens tal i et fraktilplot som vist i figur 4.4.4.



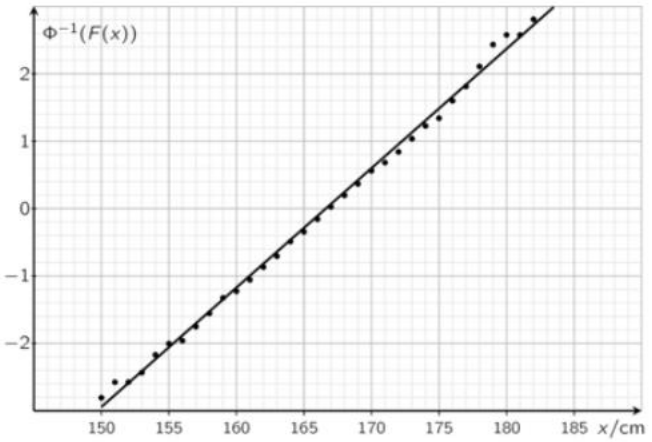
Figur 4.4.4

Punkterne beskriver med god tilnærmelse en ret linje, og lineær regression giver os sammenhængen

$$y = 0{,}1771x - 29{,}50$$

Heraf får vi

$$\sigma = \frac{1}{0{,}1771} = 5{,}65 \quad \text{og} \quad \mu = \frac{29{,}50}{0{,}1771} = 166{,}6$$

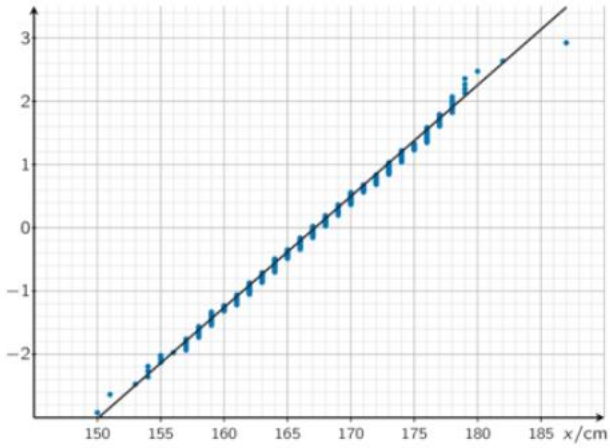


Figur 4.4.5

Ser vi på hyppighederne i Tabel 4.4.2, kan vi få en mistanke om, at de målte pigehøjder indeholder outliers. Vi benytter derfor et program til at tegne et boksplot ud fra de 400 rådata. Boksplottet, som er vist i figur 4.4.6, afslører fem outliers.

Figur 4.4.6

Ud fra rådata kan vi ligeledes benytte et program til at tegne et fraktilplot.



Figur 4.4.7

Test for normalfordelte residualer

Når vi ud fra et punktplot grafisk vurderer, om modellens forudsætninger er opfyldt, benytter vi som kriterium, at:

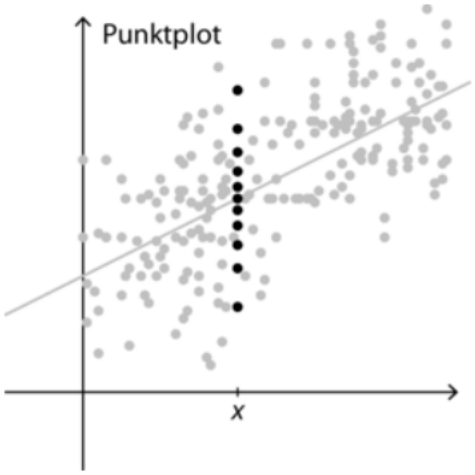
> En lineær tendens i datasættet viser sig ved punkternes små og tilfældige variationer omkring regressionslinjen.

I den videregående statistik udbygges dette kriterium med følgende krav til den statistiske fordeling af residualerne:

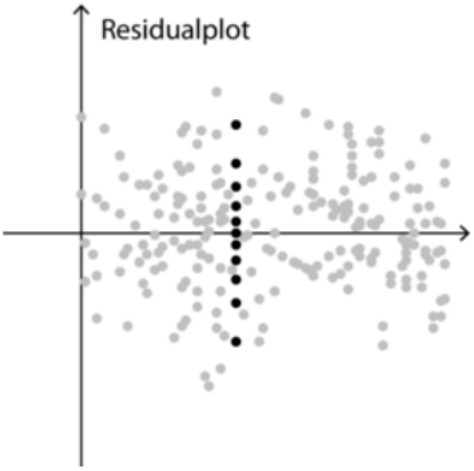
> **Forudsætning for modellering med lineær regression**
>
> I en lineær regressionsmodel skal residualerne være normalfordelte med middelværdien 0 og med samme spredning σ.

I et datasæt med sammenhørende observationer x og y kan der til den samme x-værdi være observeret forskellige y-værdier. F.eks. kan vi forestille os en måling, der gentages, og hvor der hver gang aflæses lidt forskellige resultater.

Usikkerheden ved måling og stikprøveudtagning giver altså, for fastholdt x, en variation i y. Da residualerne beregnes ud fra y, gælder på samme måde om residualerne, at de kan variere for fastholdt x. Det er denne variation, der forudsættes normalfordelt med middelværdi 0 og samme spredning σ. Se figur 6 og 7.



Figur 6



Figur 7

Forudsætningen betyder, at

- for fastholdt x skal datapunkterne være fordelt ligeligt til begge sider af regressionslinjen (middelværdien er 0).
- for fastholdt x skal datapunkternes tæthed om regressionslinjen aftage på naturlig måde, når afstanden til linjen øges (residualerne er normalfordelt).
- uanset x skal datapunkternes variation om regressionslinjen være den samme (residualernes spredning er konstant σ).

Tal, der er fordelte med de identiske fordelinger:

$$N(0,\sigma), N(0,\sigma), \ldots, N(0,\sigma)$$

er – betragtet under et – normalfordelt

$$N(0,\sigma).$$

I praksis kan vi derfor undersøge, om forudsætningen er opfyldt, ved at afbilde residualerne på enten normalfordelingspapir eller i CAS.

Eksempel 3

Sammenhængen mellem den tid x (timer), der bruges til at læse op til en matematikprøve, og resultatet af prøven y (point mellem 0 og 100), påstås at være lineær.

Vi vil undersøge påstanden og udvælger derfor, efter en matematikprøve, tilfældigt 10 elever og noterer tid brugt på at læse op til prøven og prøveresultat:

Elev nr.	1	2	3	4	5	6	7	8	9	10
Forberedelse (x)	10	15	8	7	13	15	20	10	5	5
Testresultat (y)	78	83	75	77	80	85	95	83	65	68

Tabel 2

Lineær regression med CAS giver regressionsligningen

$$f(x) = 1{,}636x + 61{,}227.$$

Vi skal undersøge, om modelforudsætningen er opfyldt, og udregner derfor residualerne. Residualet hørende til elev nr. 4 er

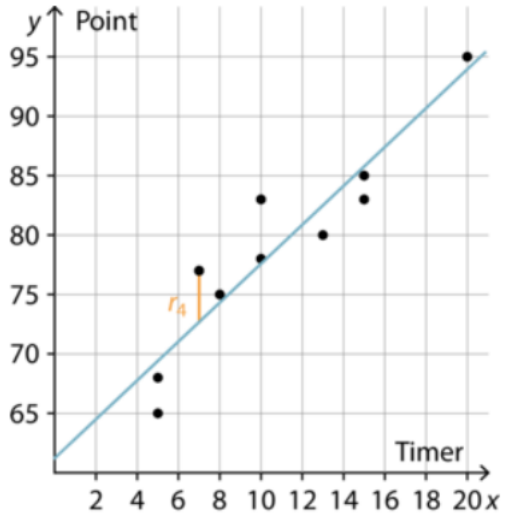
$$r_4 = y_{\text{data}} - y_{\text{model}} = 77 - (1{,}636 \cdot 7 + 61{,}227) \approx 4{,}3.$$

På samme måde beregner vi de andre 9 residualer. Resultatet er vist i tabellen nedenfor.

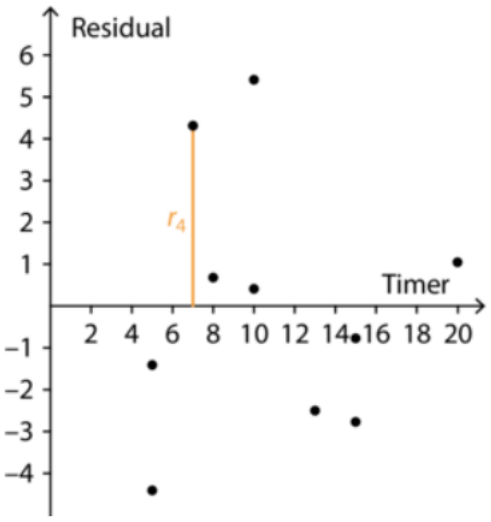
Elev nr.	1	2	3	4	5	6	7	8	9	10
Residual	0,4	−2,8	0,7	4,3	−2,5	−0,8	1,0	5,4	−4,4	−1,4

Tabel 3

Figur 9 og 10 viser hhv. et punktplot og et residualplot. Begge plot viser en lineær tendens i datasættet.

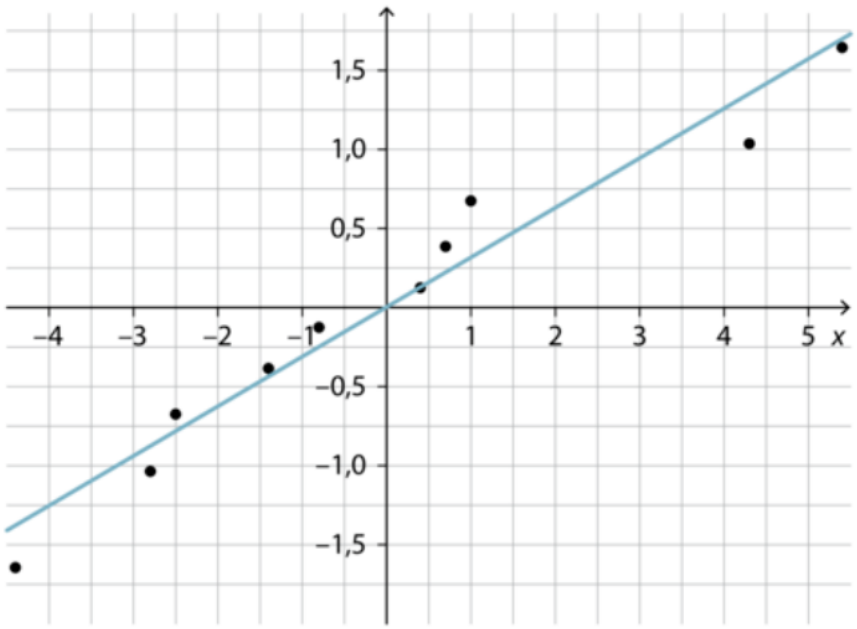


Figur 9



Figur 10

Vi undersøger også, om residualerne er normalfordelt, ved i CAS at afsætte punkterne i et QQ-plot. Resultatet er vist på figur 11, hvor vi ser, at punkterne tilnærmelsesvist er på en ret linje. Det betyder, at residualerne er tæt på at være normalfordelt.



Figur 11

Samlet har vi flere indikationer af, at der tilnærmelsesvist er en tale om en lineær sammenhæng mellem x og y.

Residualspredning

Datapunkternes variation om regressionsligningen beskrives vha. *residualspredningen*

$$s = \sqrt{\frac{r_1^2 + r_2^2 + \cdots + r_n^2}{n-2}}.$$

Tallet *s* er et estimat for residualernes underliggende spredning og fortolkes som datapunkternes gennemsnitlige afstand til regressionslinjen.
Residualspredningen viser derfor, hvor meget datapunkterne kan forventes at afvige fra regressionslinjen, når der udtages flere stikprøver.

En god lineær model er kendetegnet ved, at

- *s* er lille i forhold til *y*.
- hovedparten af residualerne har en numerisk værdi, der er under s.

I et normalfordelt talmateriale er ca. 68% af tallene inden for 1 spredning fra middelværdien og ca. 95% af tallene er inden for 2 spredninger fra middelværdien (jf. sætning 4 i afsnit 8.6). Da residualerne skal være normalfordelt med middelværdi 0, kan punkt nr. 2 nuanceres til, at

- hovedparten af residualerne er i intervallet [–s, s], og stort set alle residualer er i intervallet [–2s, 2s].

Et residual med en numerisk værdi, der adskiller sig markant fra *s*, er tegn på en outlier i datasættet.

Eksempel 4

Vi vender tilbage til eksempel 3, der handler om sammenhængen mellem forberedelse og testresultatet til en matematikprøve.

Residualspredningen kan beregnes ud fra listen af residualer. Da datasættet indeholder n = 10 observationer, giver formlen:

$$s = \sqrt{\frac{0{,}41^2 + (-2{,}77)^2 + 0{,}68^2 + \cdots + (-1{,}41)^2}{10 - 2}} = 3{,}27.$$

Tallet viser, at vi må forvente en variation i score på op til ca. 3 point, når vi sammenligner modellens forudsigelser med elevers faktiske resultater til matematikprøven.

Det mindste testresultat i undersøgelsen er 65 point og altså væsentlig større end residualspredningen. Vi ser også, at de fleste residualer er i intervallet [–s, s], og alle residualerne er i intervallet [–2s, 2s].

Da residualerne også er tilnærmelsesvist normalfordelt (eksempel 3), har vi tilsammen fundet flere klare indikationer på, at modellen er en god beskrivelse af sammenhængen mellem forberedelse og testresultat.