Kalibreringskurver; på jakt etter statistisk signifikante datapar

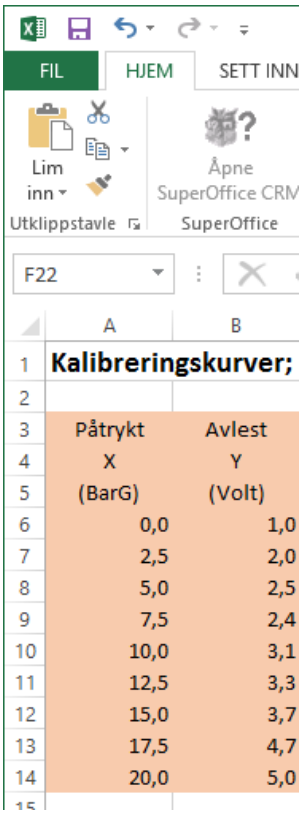
v/Rune Øverland, Trainor Elsikkerhet A/S

Den siste artikkelen om kalibrering og statistikk tar for seg praktisk bruk av Microsoft Excel for å:

- Legge inn rå-verdier fra kalibreringen
- Estimere en »Best Fit»-linje basert på rå-verdiene ved hjelp av lineær regresjon
- Estimere et usikkerhetsintervall for kalibreringslinjen (Best Fit-linjen»), også omtalt som «Confidence Interval» - CI
- Estimere et usikkerhetsintervall for kalibreringspunktene (dataparene), også omtalt som «Prediction Interval» - PI

Denne artikkel må sees i sammenheng med Microsoft Excel ® boka som jeg har laget: *Kalibreringskurve og usikkerhetsintervall CI og PI.xlsx*

Jeg kommer ikke til å vurdere kalibreringsresultatet opp mot Oljedirektoratets usikkerhetsgrenser (Måleforskriftens § 8). Det får jeg heller gjøre ved en senere anledning.



Figur 1: Datapar fra kalibrering av trykksensor

1. Legge inn kalibreringsverdier

Jeg tenker meg kalibrering av en trykksensor. Den skal kalibreres i området 0 til 20 barg. Og, sensoren gir ut et spenningssignal.

Trykket kaller jeg for x-verdi, og den avleste sensorspenningen for Y.

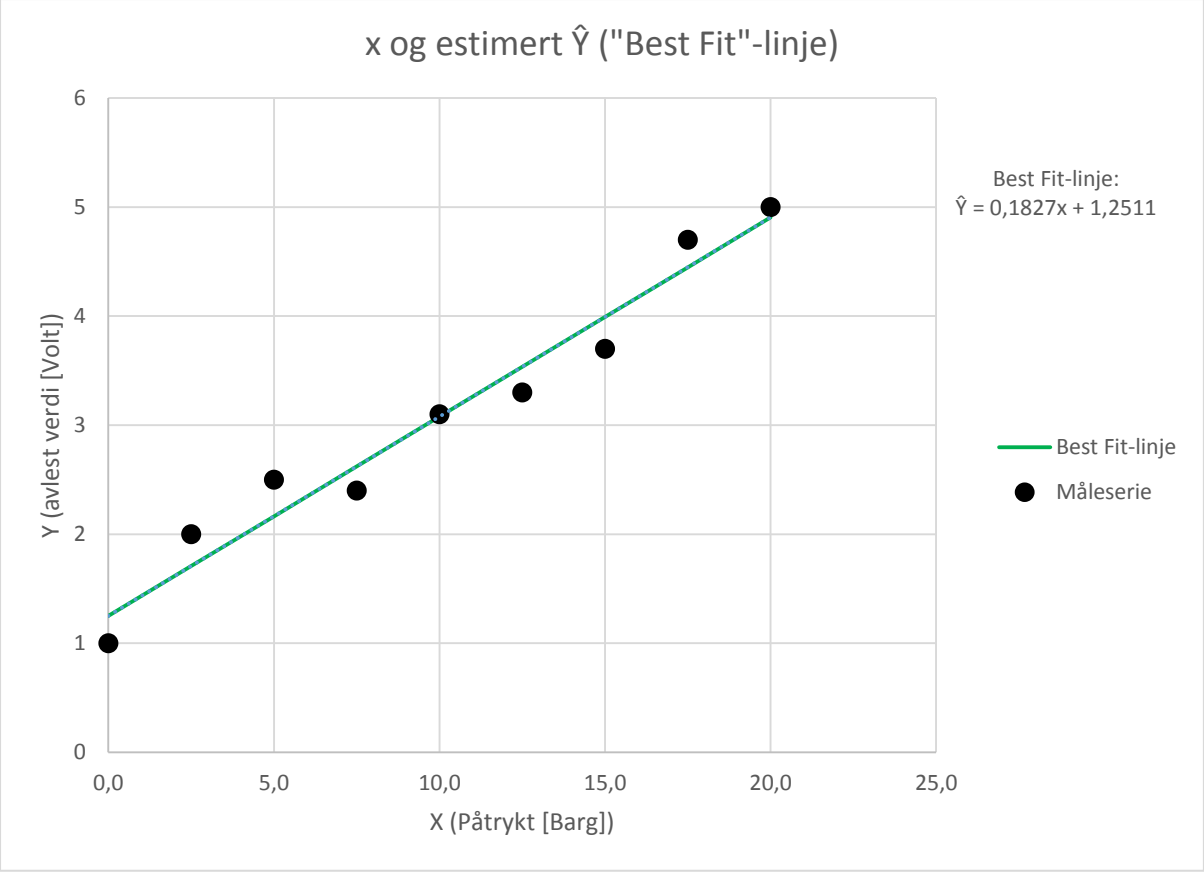
Dataparene fra kalibreringen har jeg lagt inn i radene 6 til og med 14 i kolonnene A og B i regnearket (Figur 1).

2. Estimere «Best Fit»-linje $\hat{Y}$ = bx + a

Neste oppgave er å estimere koeffisientene a og b som inngår i en rett linje:

$\hat{Y} = bx + a$

Linjen skal være av typen «Best Fit». Dette er en linje som best passer til dataparene. Jeg bruker Excel-funksjonen RETTLINJE. Funksjonen vil returnere verdiene på koeffisientene a og b.



Figur 2: Grafisk presentasjon av Excel-resultatene. Her presenteres dataparene (x,y) som sorte prikker, og den estimerte rette linjen fra regrasjonsanalysen (Ŷ = 0,1827x + 1,2511).

De sorte prikkene er dataparene (fra området A6:B14).

Den grønne linjen er «Best-Fit»-linjen. Dette er det beste estimatet for en rett linje som passer til dataparene. Excel har beregnet at «Best Fit»-linjen har koeffisientene b= 0,1827 og a = 1,2511. Den elektriske sensorspenningen endres med 0,1827 volt per bar jeg endrer trykket.

Naturlig nok er kvalitetsparameteren R^2 = 1 for «Best Fit»-linjen.

Hvordan Excel kom frem til verdiene på koeffisientene a og b vises i punktene nedenfor.

3. Usikkerhetsintervall for koeffisienten b

Verdien på b, som jeg estimerer basert på dataparene, er omgitt av et usikkerhetsintervall.

Forenklet kan jeg si at b uttrykker gjennomsnitt av stigningstallet på Ŷ. Men, ut i fra tankegangen om at dataparene er hentet fra en populasjon, må jeg ta hensyn til at en x-verdi og y-verdi er beheftet med naturlige variasjoner.

Dette gjør at stigningskoeffisienten b kan innta både lavere og høyere verdi enn b. Jeg skal bestemme et intervall hvor det er 100*(1-α) % = 95 % sannsynlig å finne b-verdien.

4. Lineær regresjon

I celle B16 bruker jeg Excel-funksjonen RETTLINJE for å bestemme verdier for «Best Fit»-linjen.

=RETTLINJE(B6:B14;A6:A14;SANN;SANN)

Området B6:B14 peker på y-verdiene, og området A6:A14 peker på x-verdiene.

Den første SANN tilhører konfigurasjonen 'konst'. SANN innebærer at jeg ønsker at funksjonen RETTLINJE skal beregnes på normal måte. Hadde jeg valgt USANN, hadde jeg tvunget «Best Fit»-linjen til å gå gjennom dataparet (0:0).

Den andre SANN tilfører konfigurasjonen 'statistikk'. Jeg ønsker at Excel skal beregne tilleggsverdier. Resultatet fra Excel vises i celle B16. Imidlertid, verdiene for tilleggsverdier vises ikke. For å få frem disse, gjøres følgende:

- Her; avmerk området B16:C20.
- Klikk på funksjonsknappen F2 på tastaturet.
- Trykk deretter ned disse tre knappene samtidig: CTRL + SHIFT + ENTER.

Slik ser det ut etter at det ovennevnte er gjort:

	A	B	C	D	E	F	G
1	**Kalibreringskurver; på jakt etter statistisk signifikante datapar**						
2							
3	Påtrykt	Avlest	Estimert				
4	X	Y	Ŷ	Confidence (CI)		Prediction (PI)	
5	(BarG)	(Volt)	(Ŷ = bx +a)	Nedre	Øvre	Nedre	Øvre
6	0,0	1,0	1,25111111	0,85388661	1,64833561	0,49252059	2,00970164
7	2,5	2,0	1,70777778	1,37753715	2,03801841	0,98201532	2,43354024
8	5,0	2,5	2,16444444	1,89195068	2,43693821	1,4630707	2,86581819
9	7,5	2,4	2,62111111	2,39009329	2,85212894	1,93478647	3,30743575
10	10,0	3,1	3,07777778	2,86235254	3,29320301	2,39654338	3,75901218
11	12,5	3,3	3,53444444	3,30342662	3,76546227	2,8481198	4,22076909
12	15,0	3,7	3,99111111	3,71861735	4,26360487	3,28973737	4,69248486
13	17,5	4,7	4,44777778	4,11753715	4,77801841	3,72201532	5,17354024
14	20,0	5,0	4,90444444	4,50721994	5,30166895	4,14585392	5,66303497
15							
16	b	0,18266667	1,25111111	a			
17	SE_b	0,01411367	0,16798631	SE_a			
18	R^2	0,95988749	0,2733101	SE_Y			
19	F	167,509137	7	v			
20	ssreg	12,5126667	0,52288889	ssresid			

Figur 3: Resultater for beregning og presentasjon av tilleggsverdier fra celle B16.

Excel viser alle tilleggsverdier fra regresjonsberegningen av RETTLINJE. Utenfor cellene B16:B20 og C16:C20 har jeg skrevet en liten forklaringer på tilleggsinformasjon. Ellers henviser jeg til Excel for nærmere forklaring.

5. Estimert verdi av $\hat{Y}$ («Best Fit»)

I området C6:C14 har jeg beregnet estimatet av $\hat{Y}$ = bx+a for ulike verdier på x. For celle C4 har jeg:

=$B$16*A6+$C$16

Tips: Jeg skiller mellom relative cellereferanser (her A6) og absolutte cellereferanser ved bruk av $-tegn (her $B$16 og $C$16). Ønsker jeg en absolutt referanse, sett skrivemerket på verdien (B16) og trykk på funksjonsknappen F4 for å få satt inn $-tegn for kolonne og rad.

Deretter er det bare å velge celle C6 og trekke denne nedover for å kopiere innholdet.

6. Tilleggsparametre

Jeg trenger å bestemme verdier på tre tilleggsparametre;

- Gjennomsnittsverdi av x. Denne har jeg lagt inn i celle B22 ved bruk av funksjonen =GJENNOMSNITT(A6:A14)
- Student T-verdi for 95 % (1-α) usikkerhetsintervall og v = frihetsgrader (her v=7). Alfa (α) er 0,05; det vil si 5 %. T-verdien har jeg lagt inn i celle B23 med funksjonen =TINV(0,05;C19).
- Summen av kvadrert avvik $\sum_i (x_i - \bar{x})^2$, her SSxx. Denne har jeg lagt inn i celle B24 med funksjonen =AVVIK.KVADRERT(A6:A14)

Excel-funksjonen TINV er ulik fra funksjonen T.INV. Funksjonen TINV brukes for 2-sidige usikkerheter, mens T.INV brukes for 1-sidig.

22	xm	10,0
23	t.95	2,36462425
24	SSxx	375

7. Diagnostikk CI (Confidence Interval) Beregning av usikkerhetsintervall for den rette linjen $\hat{Y}$ = bx + a

I område D6:D14 har jeg beregnet minimumsverdi på estimert $\hat{Y}$-verdi (95 % usikkerhet), og i område E6:E14 har jeg beregnet maksimumsverdi på estimert $\hat{Y}$.

	A	B	C	D	E
1	Kalibreringskurver; på jakt etter statistisk signifikar				
2					
3	Påtrykt	Avlest	Estimert		
4	X	Y	Ŷ	Confidence (CI)	
5	(BarG)	(Volt)	(Ŷ = bx +a)	Nedre	Øvre
6	0,0	1,0	1,25111111	0,85388661	1,64833561
7	2,5	2,0	1,70777778	1,37753715	2,03801841
8	5,0	2,5	2,16444444	1,89195068	2,43693821
9	7,5	2,4	2,62111111	2,39009329	2,85212894
10	10,0	3,1	3,07777778	2,86235254	3,29320301
11	12,5	3,3	3,53444444	3,30342662	3,76546227
12	15,0	3,7	3,99111111	3,71861735	4,26360487
13	17,5	4,7	4,44777778	4,11753715	4,77801841
14	20,0	5,0	4,90444444	4,50721994	5,30166895

Figur 4: Resultater for beregning av nedre og øvre grenseverdier for den estimerte «Best Fit»-linjen.

«Best Fit»-linjen ($\hat{Y}$ =bx+a) har et rotasjonspunkt i $(\bar{X}, \bar{Y})$. Dette finner du igjen i figuren nedenfor som et lilla punkt.

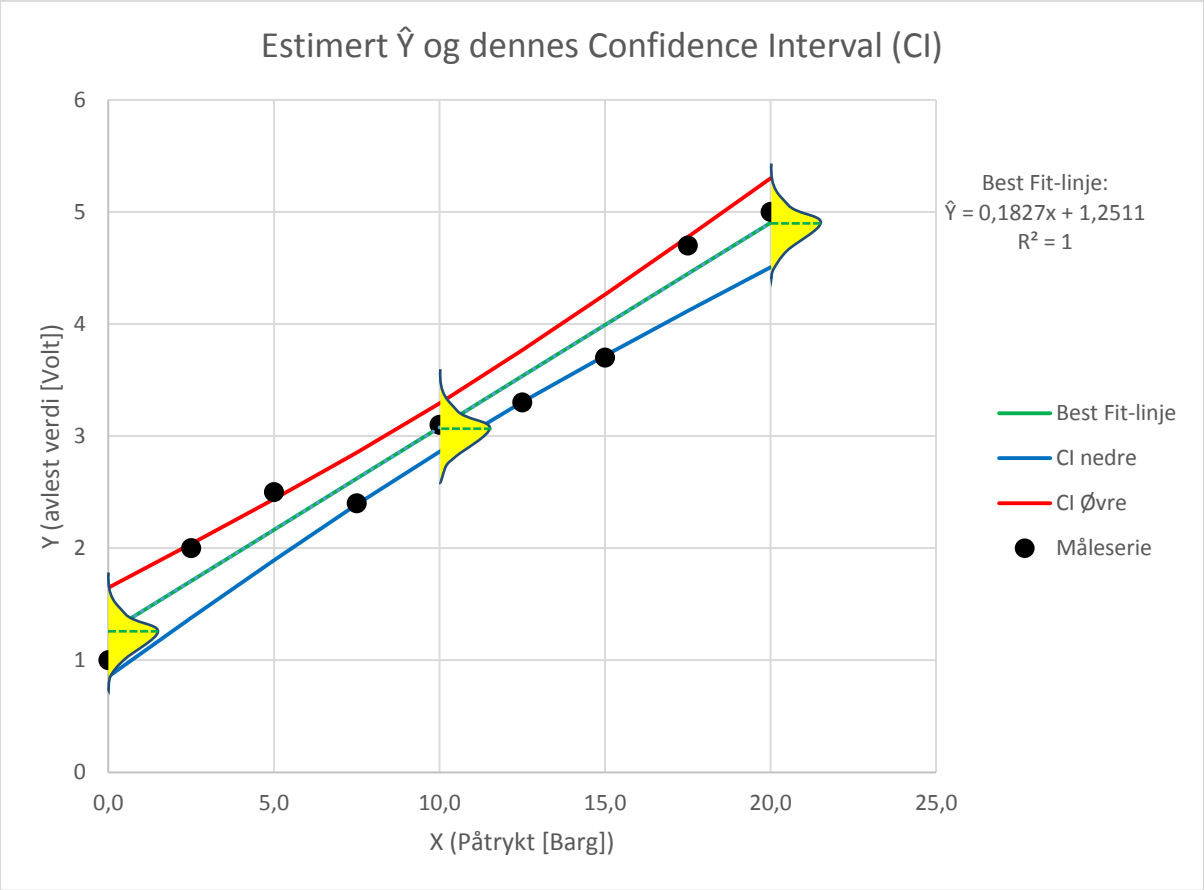
Usikkerhetsintervallet CI for koeffisienten b beskrives med følgende likning:

$$= \hat{Y} \pm T_{(\alpha,v)} * SE_y * \sqrt{\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum_i (x_i - \bar{x})^2}}$$

For celle D6 ser det slik ut:

=C6-(tVerdi*SEy*ROT(1/(df+2)+(A6-xm)^2/SSxx))

N kjenner jeg som antall frihetsgrader v (df «Degrees of Freedom») + antall variabler (her x og y); det vil si N = 7 + 2.



Figur 5: Grafisk presentasjon av usikkerhetsintervallet CI fra Excel, med påtegnet normalfordelingskurver for x-verdiene 0, 10 og 20 barg.

Den grønne «Best Fit»-linjen har en stigningskoeffisient b (celle B16 = 0,1827) med middelverdi som går gjennom punktet $(\bar{X}, \bar{Y})$. Den stiplede sorte linjen har en lavere stigningsverdi (celle B16 - B17 = 0,1827 - 0,0141 = 0,1686), mens den stiplede oransje linjen har en høyere stigningsverdi (celle B16 + B17 = 0,1827 + 0,0141 = 0,1968).

Våre beregnede CI-verdier sier at «Best Fit»-linjen med 95 % sikkerhet vil ligge innenfor intervallet beskrevet av $CI_{øvre}$ og CI_{nedre}.

8. Diagnostikk PI (Prediction Interval): Beregning av usikkerhets-intervall for dataparene. Ekstremverdi («uteligger») eller ikke?

I område F6:F14 har jeg beregnet minimumsverdi på dataparene (95 % usikkerhet), og i område G6:G14 har jeg beregnet maksimumsverdi på dataparene.

F6 =C6-(tVerdi*SEy*ROT(1+1/(df+2)+(A6-xm)^2/SSxx))

	A	B	C	D	E	F	G
1	**Kalibreringskurver; på jakt etter statistisk signifikante datapar**						
2							
3	Påtrykt	Avlest	Estimert				
4	X	Y	Ŷ	Confidence (CI)		Prediction (PI)	
5	(BarG)	(Volt)	(Ŷ = bx +a)	Nedre	Øvre	Nedre	Øvre
6	0,0	1,0	1,25111111	0,85388661	1,64833561	0,49252059	2,00970164
7	2,5	2,0	1,70777778	1,37753715	2,03801841	0,98201532	2,43354024
8	5,0	2,5	2,16444444	1,89195068	2,43693821	1,4630707	2,86581819
9	7,5	2,4	2,62111111	2,39009329	2,85212894	1,93478647	3,30743575
10	10,0	3,1	3,07777778	2,86235254	3,29320301	2,39654338	3,75901218
11	12,5	3,3	3,53444444	3,30342662	3,76546227	2,8481198	4,22076909
12	15,0	3,7	3,99111111	3,71861735	4,26360487	3,28973737	4,69248486
13	17,5	4,7	4,44777778	4,11753715	4,77801841	3,72201532	5,17354024
14	20,0	5,0	4,90444444	4,50721994	5,30166895	4,14585392	5,66303497

Figur 6: *Resultater for beregning av nedre og øvre grenseverdier for dataparene.*

Et resultat av en statistisk analyse betegnes som statistisk signifikant dersom det er lite sannsynlig at resultatet har oppstått tilfeldig.

Jeg ønsker å bestemme er intervall for dataparene uttrykt med 100*(1-α) % = 95 % usikkerhets-intervall. Sakt med andre ord, jeg ønsker å bestemme er intervall hvor det er 5 prosent sannsynlighet for at ett eller flere av dataparene våre kommer til ligge utenfor.

Så, selv om et datapar ligger utenfor dette intervallet, er dataparet ikke nødvendigvis en ekstremverdi. Ekstremverdier skal ellers gjøres ugyldige som datapar.

Usikkerhetsintervallet PI (Prediction Interval) beskrives med følgende likning:

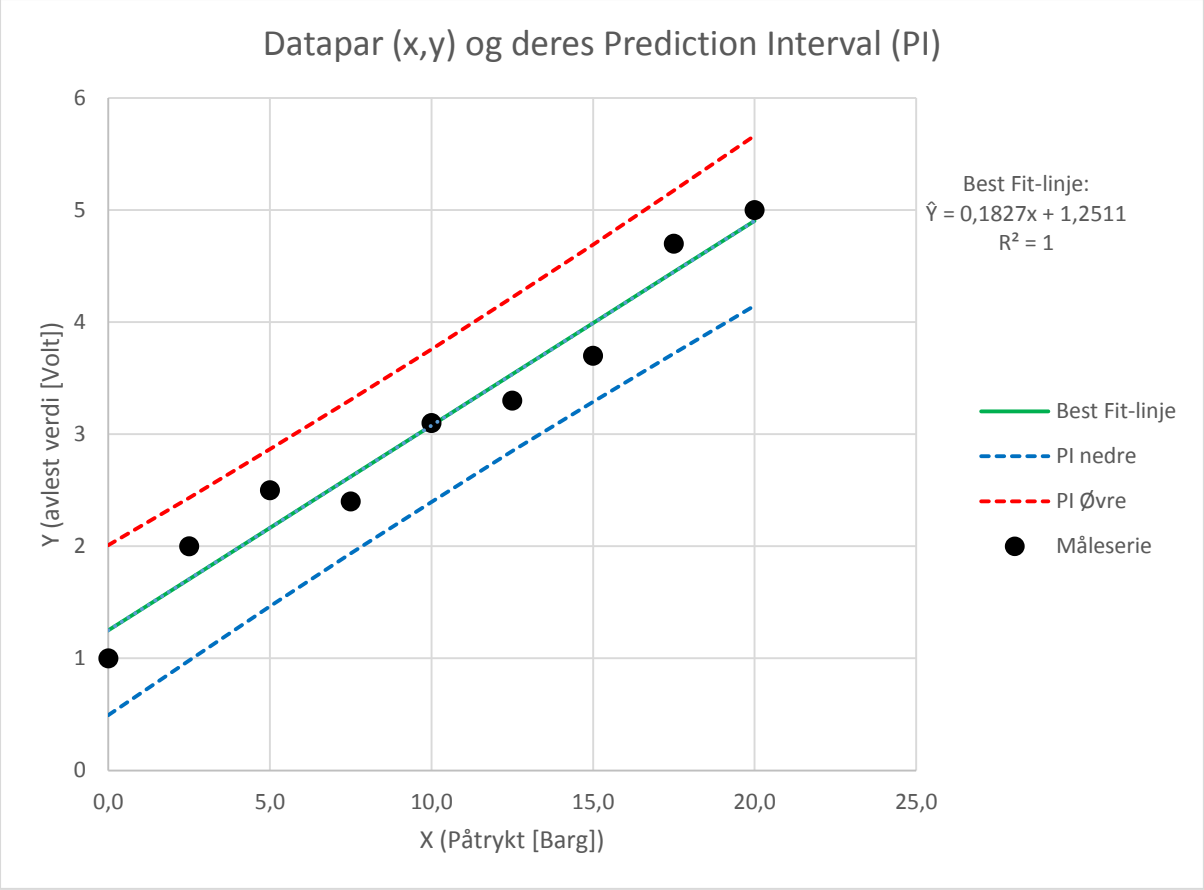
$$= \hat{Y} \pm T_{(\alpha,v)} * SE_y * \sqrt{1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum_i (x_i - \bar{x})^2}}$$

For celle F6 ser det slik ut:

=C6-(tVerdi*SEy*ROT(1+1/(df+2)+(A6-xm)^2/SSxx))

I mitt eksempel ligger alle datapar (A6:B14) innenfor «Prediction Interval» slik avgrenset av linjene $PI_{øvre}$ og PI_{nedre}. Sakt på en annen måte; jeg kan være rimelige sikker på at alle mine datapar ikke er ekstremverdier. Når de ikke er ekstremverdier, er avviket mellom verdi og estimert y begrunnet med en naturlig variasjon.

Det at dataparene ligger innenfor PI forteller i liten grad noen om nøyaktigheten (avvik til sann verdi).



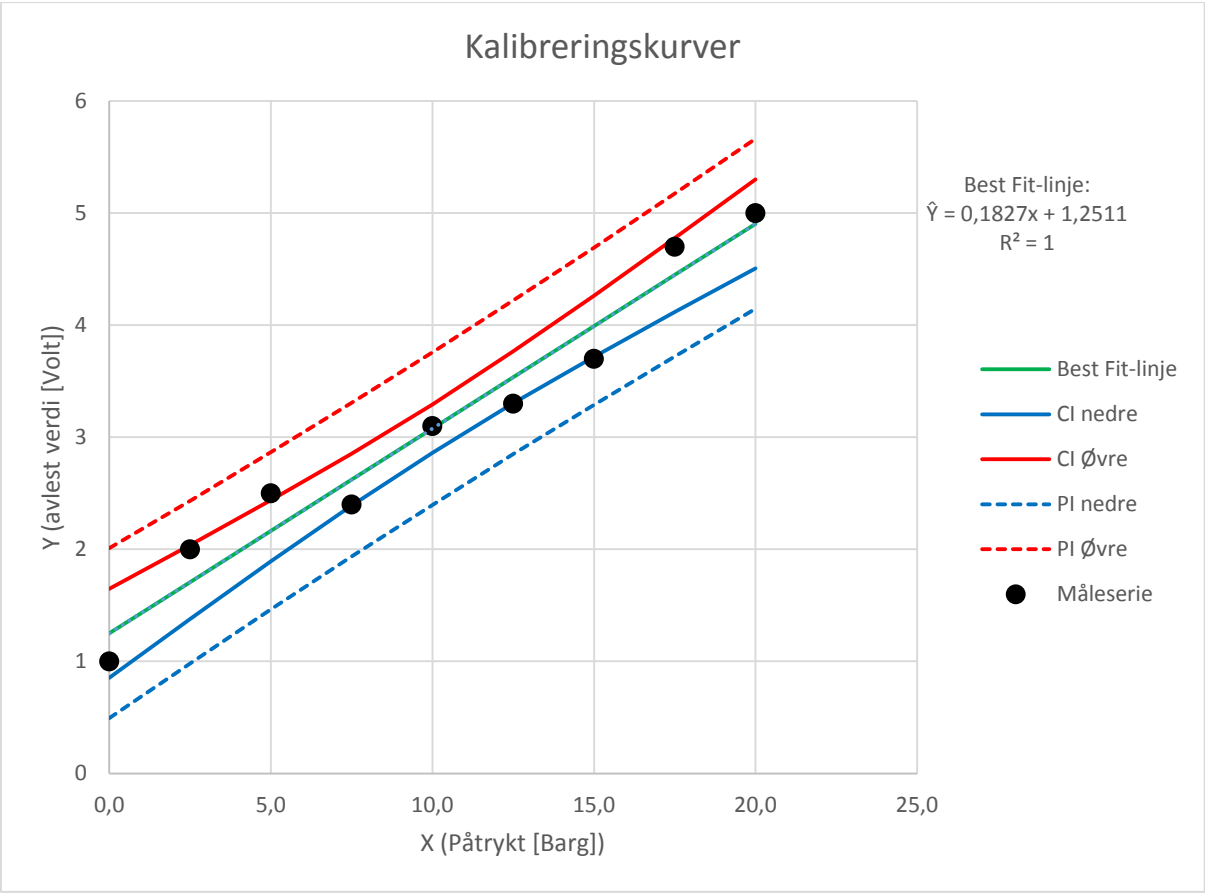
Figur 7: Grafisk presentasjon fra Excel av datapar og deres usikkerhetsintervall (PI)

9. Oppsummering

Jeg har i en artikkelserie sett på statistisk vurdering av kalibreringsdata (datapar).

Regresjon er når en stokastisk variabel (respons) antas å avhenge av andre variabler. Et resultat av en statistisk analyse betegnes som statistisk signifikant dersom det er lite sannsynlig at resultatet har oppstått tilfeldig.

Når jeg oppsummerer det hele, ser det slik ut:



Figur 8: Grafisk presentasjon fra Excel av datapar (måleserie med x- og y-verdier) og deres usikkerhetsintervall (PI), «Best Fit»-linjen Ŷ og dennes usikkerhetsintervall (CI).

Jeg har benyttet Microsoft Excel for statistisk behandling av dataparene, og i tillegg kalkulert usikkerhetsintervall (CI) for «Best Fit»-linjen og usikkerhetsintervall for dataparene (PI).

Ekstremverdier kan ha stor innflytelse på regresjonslinjer, men lar seg lett avdekke ved plotting av dataene. Dersom jeg bare har hadde sett på analyseresultatene og ikke plottet dataene, kunne jeg lett ha kommet til å dra feilaktige konklusjoner om sammenhengen mellom variablene.

Med vennlig hilsen

Trainor Elsikkerhet AS

Rune Øverland
Senioringeniør

Tønsberg februar 2015