Statistisk behandling av kalibreringsresultatene – Del 4.

v/ Rune Øverland, Trainor Elsikkerhet AS

Denne artikkelserien handler om statistisk behandling av kalibreringsresultatene.

Dennne artikkelen tar for seg hvor godt en rett linje ($\hat{Y} = \hat{a}_1 \cdot x + \hat{a}_0$) er tilpasset tallgrunnlaget (dataparene vi oppnådde ved kalibrering). Det skal bestemmes en R^2-verdi. Denne verdien ligger i intervallet 0 til 1. Desto bedre «Best Fit»-linjen er tilpasset dataparene, desto nærmere er R^2-verdien 1. Og, dersom linjen går perfekt igjennom alle datapar, vil R^2-verdien være lik 1.

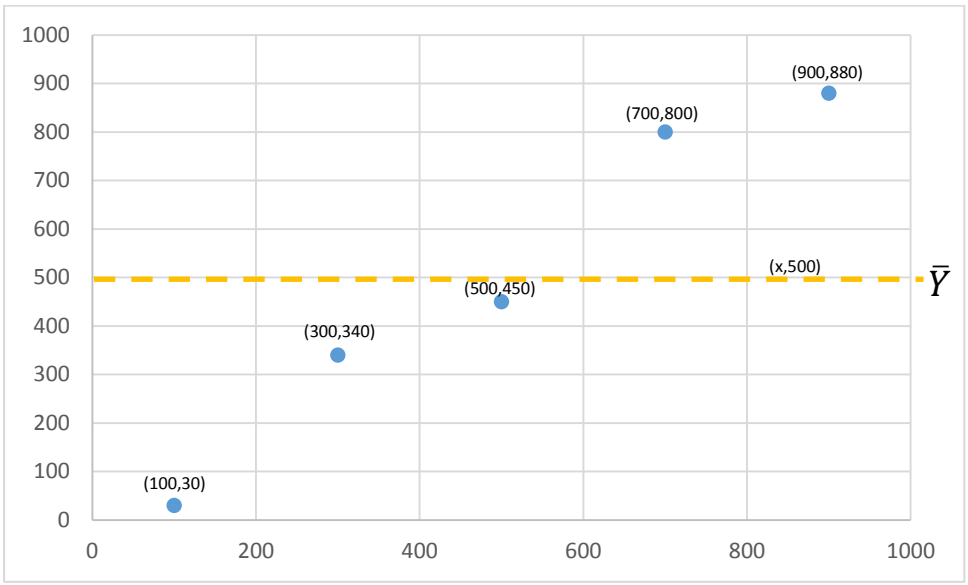
Sentralverdi av dataverdi i et utvalg

Vi tenker oss et utvalg med datapar. I vårt eksempel har vi 5 datapar.

Datapar	X	y
1	100	30
2	300	340
3	500	450
4	700	800
5	900	880

Tabell 1: Et datasett bestående av 5 datapar.

Grafisk vil blir dataparene bli illustrert slik av Microsoft Excel:



Figur 1: Grafisk presentasjon av datasettet, og linje som representerer gjennomsnittsverdien av observasjonene (oransje stiplet linje). X-verdiene fins langs den horisontale aksen, mens y-verdiene fins langs den vertikale.

Sentralt i statistikken er at i et utvalg er gjennomsnittsverdien den verdien som best representerer observasjonene i utvalget. La oss derfor kalkulere sentralverdien for dette utvalget:

Den aritmetiske gjennomsnittsverdien er summen av enkeltverdiene delt på antall elementer.

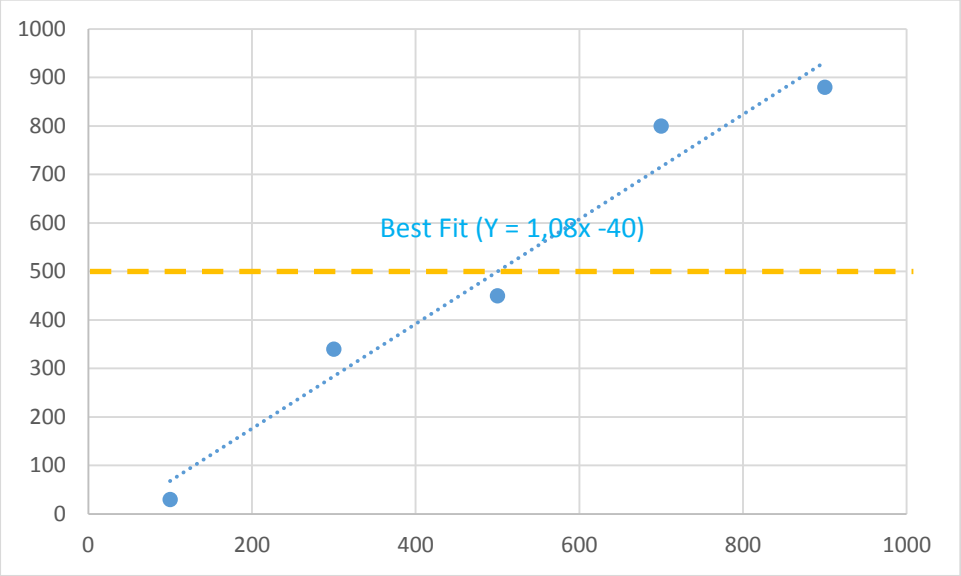
$$\bar{Y} = \frac{y_{i,1} + y_{i,2} + \ \ldots + y_{i,N}}{n} = \frac{1}{n}\sum_{j=1}^{N} y_{i,j} \qquad [1]$$

I vårt eksempel er sentralverdien

$$\bar{Y} = \frac{30+340+450+800+880}{5} = 500 \qquad [2]$$

Denne sentralverdien er presentert i figur 1 som stiplet oransje linje.

Vi ber Excel tegne inn en «Best Fit»-regresjonslinje (kalibreringskurve). Denne går «gjennom» datasettet på en slik måte at summen av avvikene fra enkeltpunktene og inn mot regresjonslinjen blir minst mulig. Her er regresjonslinjen tegnet inn som en prikket blå linje.



Figur 2: Innsatt «Best Fit»-linje (kalibreringskurve for vårt datasett.

Kalibreringskurven («Best Fit»-linjen) kan beskrives slik:

$\hat{Y}$ = (1,08 · x) - 40 [3]

Spørsmålet er: hvorfor er det enkeltverdier som avviker mye fra sentralverdien? Hvorfor er noen verdier høyere enn sentralverdien, og hvorfor er noen verdien lavere enn sentralverdien?

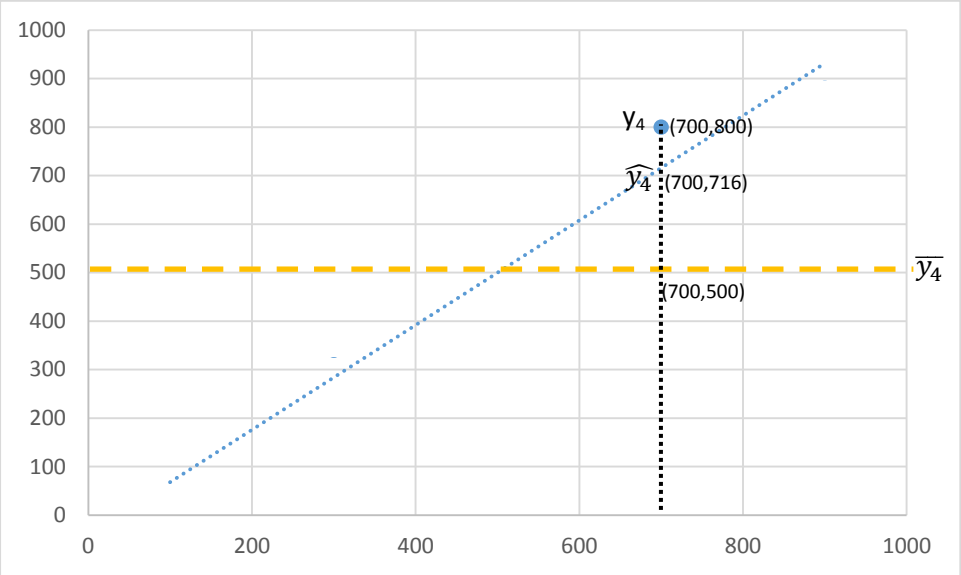
Vi ser i vårt eksempel at ved strømningsraten X_4 = 700 m3/h fikk vi en høyere y-verdi enn forventet.

Ellers observerer vi at y_1, y_2 y_3 er lavere enn gjennomsnittet, mens y_4 og y_5 har verdier høyere enn sentralverdien.

La oss analysere det fjerde dataparet mer i detalj.

Den estimerte verdien ved x_4 = 700 m^3/h er:

$\hat{Y}_4$= 1,08 · 700 – 40 = 716 m^3/h [4]



Figur 3: Vurderinger av det fjerde dataparet (700,800) i forhold til estimert verdi (700,716) og utvalgets sentralverdi (700,500).

Vi kan trekke følgende slutninger for det fjerde dataparet:

- Basert på et gjennomsnitt av alle y-verdier i utvalget, forventet vi at y_4-verdien ville være $\bar{y}$ = $\bar{y}_4$ = 500 m^3/h.
- Basert på kalibreringskurven, og dennes «Best Fit»-linje, estimerte vi en forventet verdi på $\hat{y}_4$ = 716 m^3/h.
- Basert på observasjon avleste vi en instrumentrespons på y_4 = 800 m^3/h.

Differansen mellom gjennomsnittsverdien $\bar{y}_4$ og den estimerte forventede verdien $\hat{y}_4$ kan forklares ved bruk av regresjons-modellen. Denne differansen kalles «Residual». Og, som vanlig når vi skal kvantifisere variasjonen av residual for utvalget, kan dette skrives slik:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 \quad [5]$$

Variasjonen mellom forventede verdier og gjennomsnittsverdien kalles for SSR («Sum Squared Residuals»).

Differansen mellom den observerte verdien y_4 og den estimerte forventede veriden $\hat{y}_4$ kan ikke forklares ved bruk av regresjons-modellen. Denne differansen kalles «Error». Og, som vanlig vi også kvantifisere denne variasjonen for utvalget, og dette kan skrives slik:

[6]

$$SSE = \sum_{1=1}^{n} (y_i - \hat{y}_i)^2$$

Variasjonen mellom den obserteverdien og den estimerte forventede verdien kalles for SSE («Sum Squared Error»).

Vi har følgende sammenheng:

SST = SSR + SSE [7]

Den totale variasjonen er summen av variasjonen av «Residual» og variasjonen av «Error».

Ideelt sett ønsker vi at SSE skal være så liten som mulig, og aller helst null. Når SSE er null innebærer det at alle estimerte verdier ($\hat{y}$) er lik verdien på observasjonen (y).

Forenklet, når SSE ~ 0, kan vi skrive:

SST $\sim$ SSR [8]

R-kvadrat

Et kvalitetsmål for hvor godt den «Best Fit»-linjen er i forhold til dataparene kan nå beskrives slik:

$R^2 = \frac{SSR}{SST} = \frac{(SST-SSE)}{SST} = 1 - \frac{SSE}{SST}$ [9]

Vi ser at når SSE når seg null, vil R^2 nærmere seg 1.

Eksempel: Beregning av R-kvadrat

La oss gå tilbake til vår strømningsmåler som har vært på kalibrering, og ser hvor godt vår kalibreringskurve er tilpasset våre datapar.

Vi tenker oss et scenario hvor vi har gjort en kalibrering i fem punkter for vår mengdemåler. I det nederste punktet påtrykte vi en strømningsrate på 100,00 m3/h, og avleste instrumentresponsen 99,90 m3/h. Deretter kalibrerte vi i punkt 2, og så videre. Verdiene er satt inn i tabell 1.

Sub-populasjon 1 (X_1 , $y_{1.1}$)	Sub-populasjon 2 (X_2 , $y_{2.1}$)	Sub-populasjon 3 (X_3 , $y_{3.1}$)	Sub-populasjon 4 (X_4 , $y_{4.1}$)	Sub-populasjon 5 (X_5 , $y_{5.1}$)
(100 , 99,90)	(300 , 300,45)	(500 , 500,83)	(700 , 700,25)	(900 , 899,50)

Tabell 2: Datapar fra kalibrering.

Her viser vi deg hvorledes vi steg-for-steg manuelt går frem for å komme frem til R-kvadrat.

Datapar		Best Fit	Residual	Variasjon SSE
(x_i)	(y_i)	$\hat{y}_i$ = 0,9995x+0,436	$\hat{u}_i = y_i - \hat{y}_i$	$(\hat{u}_i)^2$
100	99,90	100,39	-0,49	0,236
300	300,45	300,29	+0,16	0,027
500	500,83	500,19	+0,64	0,415
700	700,25	700,09	+0,16	0,027
900	899,50	899,97	-0,49	0,236
			$\sum = 0$	**SSE** =$\sum$ =0,941

Tabell 3: I første og andre kolonne har vi data fra kalibreringen av instrumentet. I tredje kolonne har vi «interpolerte», estimerte verdier basert på «Best Fit»-formel. I fjerde kolonne har vi kalkulert Residual mellom avlest y-verdi og verdi basert på «Best Fit». I den femte kolonnen har vi kvadrert denne verdien.

I den nederste raden i tabellen har vi i kolonne fire summert «error»-verdiene. I vårt eksempel ble summen av de individuelle verdiene null. Summen av de kvadrerte bidragene ble SSE = 0,941.

Datapar		Spredning	Variasjon SST
(X_i)	(y_i)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
100	99,90	-400,28	160228,9
300	300,45	-199,74	39894,5
500	500,83	0,64	0,4
700	700,25	200,06	40025,6
900	899,50	399,31	159451,7
	$\bar{Y}$ = 510,186		**SST** =$\sum(y_i - \bar{Y})^2$ = 399601,0

Tabell 4: Beregning av sentralverdi for observasjonene i datasettet, samt den totale variasjonen (SST)

Vi kalkulerer sentralverdien (gjennomsnittet) av observasjonene $\bar{Y}$ i utvalget.

$\bar{Y}$ = 510,186

Vi kalkulerer det totale variasjonstallet:

SST = 399601,0

Vi setter verdiene for SSE og SST inn i formelen for R-kvadrat:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0,941}{399601,0} = 0,99999 \quad [10]$$

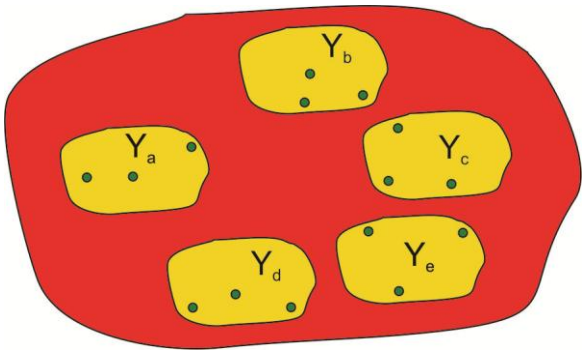
$R^2 \sim 1$

I tilfellet har vi svært sterk korrelasjon mellom kalibreringskurven («Best Fit»-linjen) og dataparene fra observasjonen.

Pålitelighet (standard usikkerhet) av SSE

Et par viktige spørsmål er hvordan vil gjennomsnittene i en serie utvalg fordele seg i forhold til gjennomsnittet i populasjonen de er trukket fra.

Jo flere antall medlemmer i utvalget (n) det er, jo større sjanse er det for at utvalgsgjennomsnittet er nær populasjonsgjennomsnittet. Uansett utvalgsstørrelse, er det størst sjanse for å treffe nært populasjonens gjennomsnitt og mindre sjanse for å treffe langt unna desto flere medlemmer vi har i utvalget.



Figur 4: Vi har en tenkt populasjon av SSE-verdier, som igjen er delt inn i fem utvalg; hver bestående av tre elementer.

Det er mest sannsynlig at utvalgsgjennomsnittet ligger nær populasjonsgjennomsnittet, og at denne sannsynligheten avtar lenger bort fra populasjonsgjennomsnittet. Denne sannsynlighetsfordelingen kalles normalfordelingen.

I dette tilfellet har standardavviket et spesielt navn; standard usikkerhet (eller standardfeil).

Standard usikkerhet (standardfeilen) er den gjennomsnittlige avstanden av utvalgsgjennomsnittene til populasjonsgjennomsnittet.

Hvordan regner vi så ut standardfeilen?

Standardfeilen er avhengig av standardavviket i populasjonsfordelingen og størrelsen på utvalget.

Problemet er at vi i virkeligheten ikke vet hva standardavviket i populasjonen er, og vi får dermed problemet med å regne ut standardfeilen. Da må vi ta sjansen på at standardavviket i utvalget er omtrent det samme som standardavviket i populasjonen. Vi får dermed en tilnærmingsformel for standardfeilen.

La oss gå tilbake til SSE, og tallet vi fant (=0,941).

Hadde vi hatt langt flere datapar, vil naturlignok summen av $(\text{Residual})^2$ øke. Vi må derfor «normalisere» denne summen i forhold til antall datapar som inngår i beregningsgrunnlaget.

For å beskrive en rett linje trengs obligatorisk 2 datapar. Vi vårt eksempel har vi 5 datapar. Innenfor statistikken fins begrepet «Degrees of Freedom». I vårt eksempel vil «Degrees of Freedom være:

Degrees of Freedom = Totalt antall datapar – Obligatoriske datapar [11]

I vårt tilfelle:

Degrees of Freedom = 5 – 2 = 3 [12]

Altså, desto flere enkeltpunkter «Degrees of Freedom», desto mer pålitelig vil den videre statistiske vurderingen være.

I vårt tilfelle har vi 3 stykk «Degrees of Freedom». Det er ikke spesielt mange, men i all bedre enn ett.

Estimert standard usikkerhet av regresjonen ($\hat{\sigma}$)

Estimert standard usikkerhet av regresjonen er en statistisk beskrivelse av variasjonene av enkeltverdiene.

$$\hat{\sigma} = \sqrt{\frac{SSE}{Degrees\ of\ Freedom}} = \sqrt{\frac{0{,}941}{3}} = 0{,}560 \quad [13]$$

Terminologi

Terminologi benyttet i denne artikkelen:

$\hat{a}_0$	Estimert koeffisient. Har verdien y når x = 0.
$\hat{a}_1$	Estimert koeffisient. Har verdien på stigningstallet til en rett linje
m^3/h	Kubikkmeter per time
n	Antall observasjoner
R^2	Verdi [-1,1] beskriver hvor godt en «Best Fit»-linje er i forhold til datasettet.
$\hat{\sigma}$	Estimert standard usikkerhet av regresjonen
SSE	Sum Squared Error
SSR	Sum Squared Residual
SST	Sum Squared Total (SST = SSE + SSR)
$\sum$	Summasjonstegn (viser summen av enkeltverdier)
$\widehat{u_i}$	Verdi som beskriver avviket mellom observert verdi og estimert forventet verdi. Brukes som beregningsgrunnlag for SSE.
x	Den uavhengige variablen, for eksempel en strømningsrate i et rør.
$X_1, \ldots X_5$	Den uavhengige variablen, for eksempel verdien på den påtrykte strømningsraten ved kalibrering (referanseverdien). Brukes også for å gruppere observasjonene i sub-populasjoner.
X_i	Den i-ende verdien på den uavhengige variablen.
y_i	Verdien på den i-ende observasjonen i en måleserie
$\widehat{y_i}$	Den estimerte forventede verdien på den i-ende observasjonen
$y_{i,1}$	Første observerte verdi i en sub-populasjon
$y_{i,2}$	Andre observerte verdi i en sub-populasjon
$y_{i,N}$	Siste observerte verdi i en sub-populasjon
$y_1, \ldots y_5$	Observasjon 1 til og med 5

y_4	Observasjon nummer fire
$\widehat{y_4}$	Estimert forventningsverdi for den fjerde observasjonen
$\hat{Y}$	Estimert verdi på observasjonen.
$\widehat{y_i}$	Estimert verdi på den i-ende observasjonen
$\bar{Y}$	Sentralverdien (gjennomsnittsverdi) av et utvalg

Neste artikkel . . . .

I den neste artikkelen skal vi fremdeles holde oss til kalibreringskurver.

Så langt har vi holdt oss til det første datasettet, kalibreringsserie 1. Her fikk etablert en kalibreringskurve. Vi estimerte to koeffisienter for den rette linjen $\hat{Y}_1 = \hat{a}_1 \cdot x + \hat{a}_0$.

Men, hva med kalibreringskurven $\hat{Y}_2$ og $\hat{Y}_3$ og så videre. Her vil vi å andre verdier på koeffisientene $\hat{a}_1$ og $\hat{a}_0$, samt R-kvadrat.

Det må derfor eksistere et bånd eller intervall som disse koeffisientene naturlig varierer innenfor. Da kan vi beskrive et konfidensintervall slikt uttrykt ved normalfordelingskurven og 95 prosent konfidensintervall (2 standard avvik). Så, når vi skal gjøre kalibreringsserie 16, da har vi et verktøy vi kan benytte for å vurdere godheten til våre estimerte koeffisienter i forhold til tidligere kalibreringsserier for å sjekke langstridstrender.

«Stay tuned for more fun!»

	$X_1 = 100{,}00$	$X_2 = 300{,}00$	$X_3 = 500{,}00$	$X_4 = 700{,}00$	$X_5 = 900{,}00$	$\hat{a}_1$	$\hat{a}_0$	R^2
1	99,90	300,45	500,83	700,25	899,50	0,9995	0,436	1
2	100,05	300,07	500,22	700,33	899,33			
3	100,17	300,00	500,11	698,80	900,45			
4	99,99	299,66	499,45	700,55	901,30			
5	99,86	299,80	499,30	700,22	899,56			
6	99,94	299,90	499,70	700,11	899,32			
7	100,01	299,65	499,90	699,85	900,32			
8	100,12	300,25	499,99	699,81	900,98			
9	100,05	300,28	500,50	700,05	899,78			
10	100,01	300,01	500,11	700,09	899,23			
11	99,88	299,95	501,00	699,50	900,23			
12	100,12	299,73	499,26	699,25	900,00			
13	100,00	300,22	499,83	700,68	900,25			
14	100,04	300,13	499,72	700,45	900,11			
15	99,86	299,90	500,08	700,06	899,64			
16	???	???	???	???	???	???	???	

Med vennlig hilsen

Trainor Elsikkerhet AS

Rune Øverland
Senioringeniør

Tønsberg oktober 2014