«Best Fit»-linje med usikkerhetsintervall (CI)

v/Rune Øverland, Trainor Elsikkerhet AS

1. Innledning

Denne artikkelen utleder formel for usikkerhetsintervallet CI (Confidence Interval) som omslutter en «Best Fit»-linje.

CI = Estimert Y ± [usikkerhetsintervall] [1]

Vi skal vise at den komplette formelen vil se slik ut:

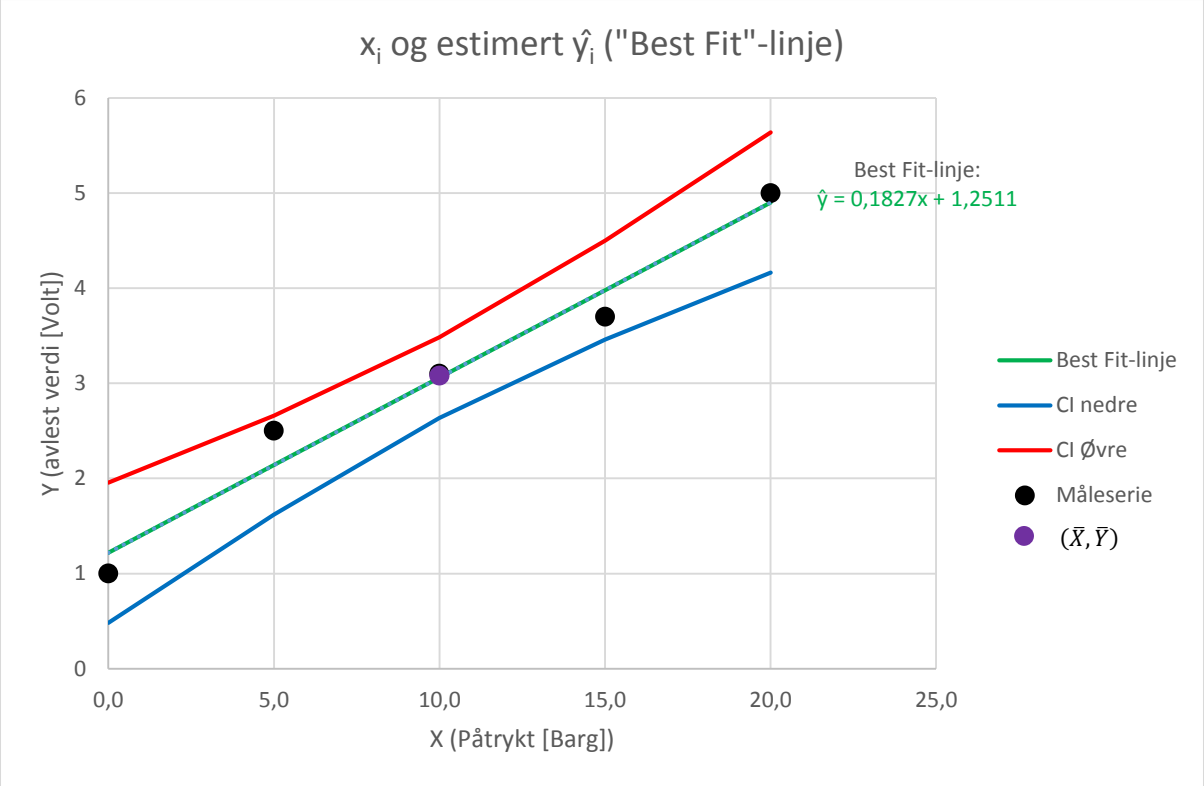
$$CI = [bx + a] \pm \left[T_{(\alpha/2,v)} \cdot \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-2)}} \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right] \quad [2]$$

Hvor:

x	Vi ønsker å bestemme usikkerhetsgrenser for denne x-verdi. Dette kan for eksempel være kalibreringstrykket eller kalibreringstemperaturen. x er den uavhengige variabelen.
b	b forteller om stigningstallet til den rette regresjonslinjen.
a	a forteller om skjæringspunktet til den rette regresjonslinjen i Y-aksen (når x = 0)
$T_{\alpha/2,v}$	Student t-verdien avhenger antall datapar som inngår i regresjonsanalysen.
α	α forteller om usikkerhetsintervallet: 100(1- α) prosent. Normalt er $\alpha = 0{,}05$ det vil si at konfidensintervallet er 95 prosent.
/2	I Student t-funksjonen skal vi ha 2-sidig usikkerhetsgrense; det vil si 2,5 % i hver ende.
v	Dette er «Degrees of Freedom. I vårt tilfelle har vi to variabler (x og y) og derav er v = n-2
n	n er antall datapar som inngår i regresjonsanalysen.
$\sum_{i=1}^{n}$	Summasjonstegn. Vi repeterer summasjonen n antall ganger.
y_i	Dette er den avleste responsen (y-verdien) i dataparet. Dette kan for eksempel være sensorspenning eller instrumentsignal. y er den avhengige variabelen (avhenger av x-variabelen).
$\hat{y}_i$	Dette er den beregnede, forventede responsverdien basert på regresjonslikningen og x-verdien vi har valgt.
$\bar{X}$	Dette er aritmetisk sentralverdi (gjennomsnittsverdi) av alle x-verdier ifra dataparene våre.
x_i	Dette er en individuell x-verdi i dataparet.

Tabell 1: Symboldefinisjoner

«Best Fit»-linjen $\hat{y}$ =bx+a er basert på et sett med datapar. Linjen har et rotasjonspunkt i $(\bar{X}, \bar{Y})$. Linjen har koeffisientene b (stigningstall) og a (a er y-verdi når x = 0). Koeffisienten a har en slik verdi at regresjonslinjen går gjennom regresjonslinjens rotasjonspunktet $(\bar{X}, \bar{Y})$.



Figur 1: Grafisk presentasjon av usikkerhetsintervallene (øvre og nedre CI) fra Microsoft Excel. «Best Fit»-linjen har et rotasjonspunkt i $(\bar{X}, \bar{Y})$.

I vårt eksempel er en trykksensor kalibrert i fem punkter i området 0 til 20 Barg. Sensoren er kalibrert for x_1 = 0 Barg, x_2 = 5 Barg, x_3 = 10 Barg, x_4 = 15 Barg, og x_5 = 20 Barg. x kalles for den uavhengige variabelen. For hvert kalibreringspunkt har vi gjort en avlesning (y) av sensorspenningen. y kalles den avhengige variabelen. Observasjonene fra kalibreringen er lagt inn i tabell 2.

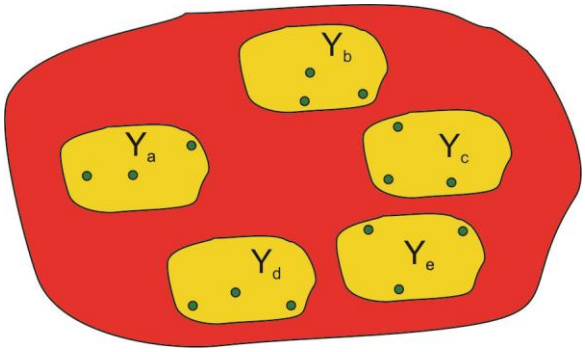
X (Barg)	Y (Volt)	Datapar
0,0	1,0	(x_1,y_1)
5,0	2,5	(x_2,y_2)
10,0	3,1	(x_3,y_3)
15,0	3,7	(x_4,y_4)
20,0	5,0	(x_5,y_5)

Tabell 2: Responsverdier (Y) fra 5 verdier (x).

2. Populasjon av datasett (Sub-populasjoner)

For trykksensoren har vi etablert en database for tidligere kalibreringer. Det er etablert 5 sub-populasjoner Y_a, Y_b, Y_c, Y_d, og Y_e med datasett.

Y_a (0 Barg, Y_1)	Dette er datasett hvor det et påtrykt 0 Barg, med avleste sensorverdier Y_1.
Y_b (5 Barg, Y_2)	Dette er datasett hvor det et påtrykt 5 Barg, med avleste sensorverdier Y_2.
Y_c (10 Barg, Y_3)	Dette er datasett hvor det et påtrykt 10 Barg, med avleste sensorverdier Y_3.
Y_d (15 Barg, Y_4)	Dette er datasett hvor det et påtrykt 15 Barg, med avleste sensorverdier Y_4.
Y_e (20 Barg, Y_5)	Dette er datasett hvor det et påtrykt 20 Barg, med avleste sensorverdier Y_5.



Figur 2: Vi tenker oss en populasjon (rødt område som vi for eksempel kan navngi Y). I dette eksemplet har vi fordelt dataparenes y-verdier (elementene) på 5 sub-populasjoner/utvalg (oransje områder. Sub-populasjonene er navngitt Y_a, Y_b, Y_c, Y_d, og Y_e.

For hver sub-populasjon kan vi beregne den aritmetiske sentralverdien for responsen. For eksempel for subpopulasjon Y_a.

$$\overline{U_a} = \frac{y_{a(1)} + y_{a(2)} + \ldots + y_{a(N)}}{N} = \frac{\sum_{i=1}^{N} y_{a(i)}}{N} \quad [3]$$

Vi summerer alle responsverdier ($y_{a(i)}$) i en sub-populasjon og dividerer med antall elementer (N) i sub-populasjonen for å bestemme den aritmetiske sentralverdien i sub-populasjonen ($\overline{U_a}$).

Den aritmetiske sentralverdien i sub-populasjonen ($\overline{U_a}$) er vurdert til å være *den* enkeltverdi som best representerer en *gruppe* av verdier.

Vi er i et dilemma. For en sub-populasjon skal vi ha uendelig mange (N ≈ ∞) med datapar. I praksis er dette umulig. Sub-populasjonen består kanskje av et par hundre eller tusen datapar. I praksis kan vi derfor ikke med 100 prosent sikkerhet si at den kalkulerte sentralverdien vi har funnet, er sentralverdien for sub-populasjonen, men heller er sentralverdien for utvalget.

Estimert aritmetisk sentralverdi for sub-populasjon Y_a:

$$\mathrm{E}(\overline{U_a}) = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} \quad [4]$$

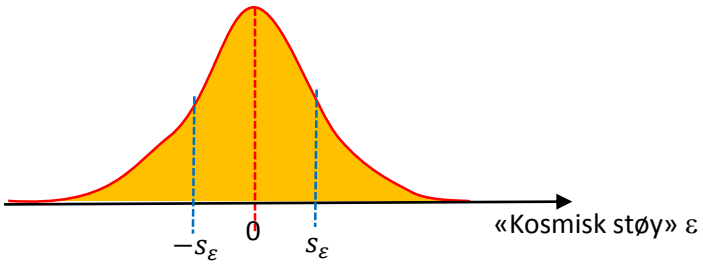
Uansett, vi plotter våre fem sentralverdier inn i et diagram (figur 3). Vi trekker en rett linje gjennom sentralverdiene.

I figur 3 kan vi teoretisk omtale dette som en Populasjonskarakteristikk. Men, i praksis burde vi snakke om en Utvalgskarakteristikk.

3. «Kosmisk støy» ε

Vår trykksensor har korrelasjon mellom trykk (kalt x-verdi) og signalrespons (kalt y-verdi). Sekundært har instrumentet flere mindre korrelasjoner som vi ikke har kontroll på. Dette kan være i hvilken grad sensorresponsen er temperaturavhengig, i hvilken grad sensorresponsen er avhengig av stabiliteten på krafttilførsel, i hvilken grad instrumentet reagerer på elektromagnetisk interferens (EMC/EMI), i hvilken grad instrumentet reagerer på vibrasjoner og så videre.

Så selv om vi holder kalibreringstrykket stabilt, for eksempel 10,0 Barg, så vil vi likevel over tid observere en signalrespons som varierer noe. I den videre diskusjon kaller jeg denne variasjonen for «kosmisk støy».



Figur 3: «Kosmisk støy» som gjør at instrumentresponsen Y varierer til tross for at den påtrykte verdien (x) er konstant. Vi antar at denne variasjonen er normaltfordelt (Gauss-kurve). Den har variansen Var(ε). Vi kan beskrive et usikkerhetsintervall s_ε for den «kosmiske» støyen. Variasjonen er symmetrisk rundt den lokale 0-verdien.

«Kosmisk støy» er normalfordelt, med sentralverdi = 0

$$\bar{\epsilon} = 0 \qquad [5]$$

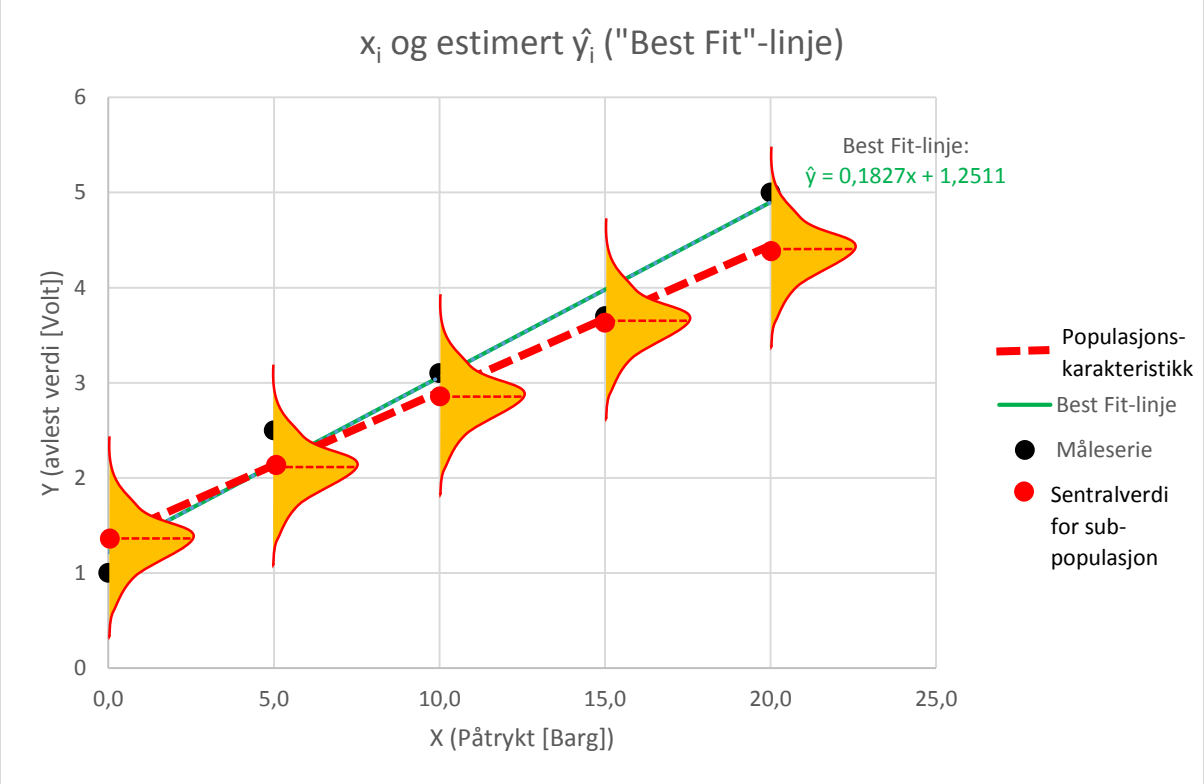
Variansen for den «kosmiske» støyen er:

$$\text{Var}(\varepsilon) = \frac{(\varepsilon_1 - \bar{\epsilon})^2 + (\varepsilon_2 - \bar{\epsilon})^2 + \ldots + (\varepsilon_n - \bar{\epsilon})^2}{n-2} \qquad [6]$$

Standardavviket for den «kosmiske støyen» er kvadratroten av variansuttrykket:

$$s_\varepsilon = \sqrt{\text{Var}(\varepsilon)} = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}{(n-2)}} \qquad [7]$$

NB: Det er viktig å legge merke til at for en gitt x_i-verdi, er variasjonen i responsen y_i (Var (y_i)) lik variasjonen for den «kosmiske støyen» Var (ε).



Figur 4: Populasjonskarakteristikk (rød stiplet strek) og «Best Fit»-linje (basert på et tilfeldig uttrekk av datapar som i vist i tabell 2 (sorte prikker)) fra hver av de fem sub-populasjonene (med sin lokale sentralverdi røde prikker).

4. Populasjonens hypotetiske likning

Vi forutsetter i den videre diskusjon av at populasjonen har en lineær karakteristikk som kan beskrives slik:

$U_{Y|x}$=βx+α [8]

Populasjonskarakteristikken er en rett linje med stigningstallet β, hvor linjen skjærer y-aksen i α. Denne har jeg tegnet inn i figur 3 som en stiplet rød strek. Det bemerkes at dette er en hypotetisk tenkt linje. Vi kan aldri vite med sikkerhet hvor denne ligger, men vi vet at denne linjen fins.

Den «kosmiske» støyen (ε) er uavhengig av x-verdien (se figur 3), og vil inngå i respons-karakteristikken.

$U_{Y|x}$=βx+α + ε [9]

ε-parameteren er/har ikke én fast verdi. Den varierer i størrelse for hver gang vi avleser et datapar [x_i,(y_i+ε)]. Variasjonen i figur 3 og 4, her det oransje området, kan tilsvarende i figur 2 assosieres som variasjon i områdene Y_a, Y_b, Y_c, Y_d, og Y_e.

5. Lineær regresjon

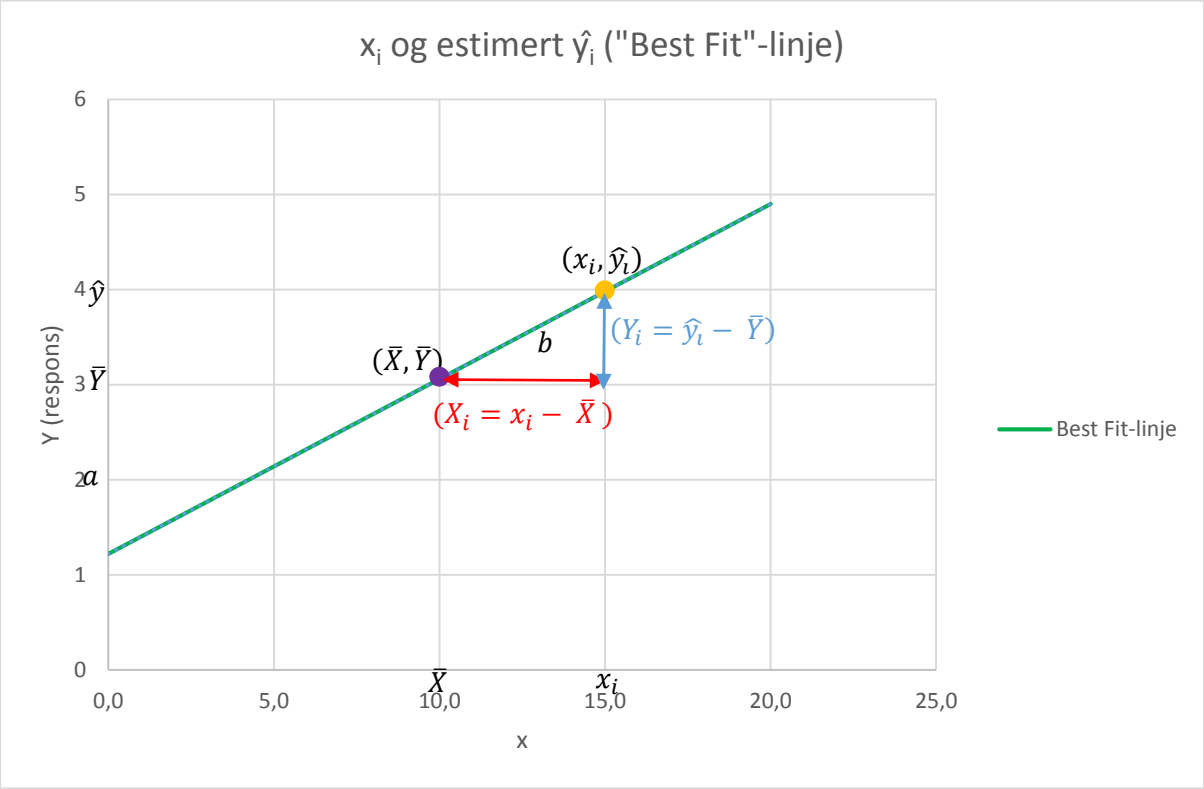
I figur 3 er det tegnet inn én lineær regresjonslinje (grønn heltrukken strek). Denne er beregnet ut i fra et tilfeldig tallsett fra de fem sub-populasjonene. I dette tilfellet har vi fem følgende datapar (se tabell 2).

Dataparene er tegnet inn som sorte prikker i figur 1 og 4.

Men, hva hvis vi hadde trukket ut/avlest andre Y-verdier når vi gjorde kalibreringen av instrumentet? Jo, da ville vi selvfølgelig kalkulert en regresjonslinje med andre koeffisienter b og a.

I og med «kosmisk støy» (ε), må vi tilsvarende konstruere et usikkerhetsintervall (CI – Confidence Interval). Dette intervallet skal fortelle oss at vi med 95 prosent sikkerhet forventer å finne «Best Fit»-linjen.

Med andre ord, koeffisientene b og a har et usikkerhetsintervall knyttet til seg. I denne artikkelen skal jeg nå vise uttrykket for disse.



Figur 5: Grafisk fremstilling av rotasjonspunktet $(\bar{X}, \bar{Y})$, og koeffisientene b og a for den rette regresjonslinjen $\hat{y}$ =bx+a. [10]

Vi skal senere i denne artikkelen vise at regresjonslinjen har et rotasjonspunkt i $(\bar{X}, \bar{Y})$. Vi har tidligere sakt at den enkeltverdi for best representerer en gruppe er den aritmetiske sentralverdien. I vårt tilfelle kan vi naturlig forklare hvorfor responsverdien for (x_i) var $\hat{y}_i$ og ikke $\bar{Y}$. Den enkle årsaken er sensorkarakteristikken mellom den uavhengige variabelen x og den avhengige variabelen y. Vi snakker om det forklarlige residual. Avviket mellom $\hat{y}_i$ og $\bar{Y}$ har en naturlig forklaring grunnet instrumentkarakteristikken, og kan beregnes.

For en tilfeldig x-verdi (x_i) kan vi beregne den forventede responsverdien $(\hat{y})$. Et utgangpunkt for å beregne den forventede $\hat{y}_i$ er å bruke regresjonslikningen. Ulempen er at ennå ikke kjenner a-verdien.

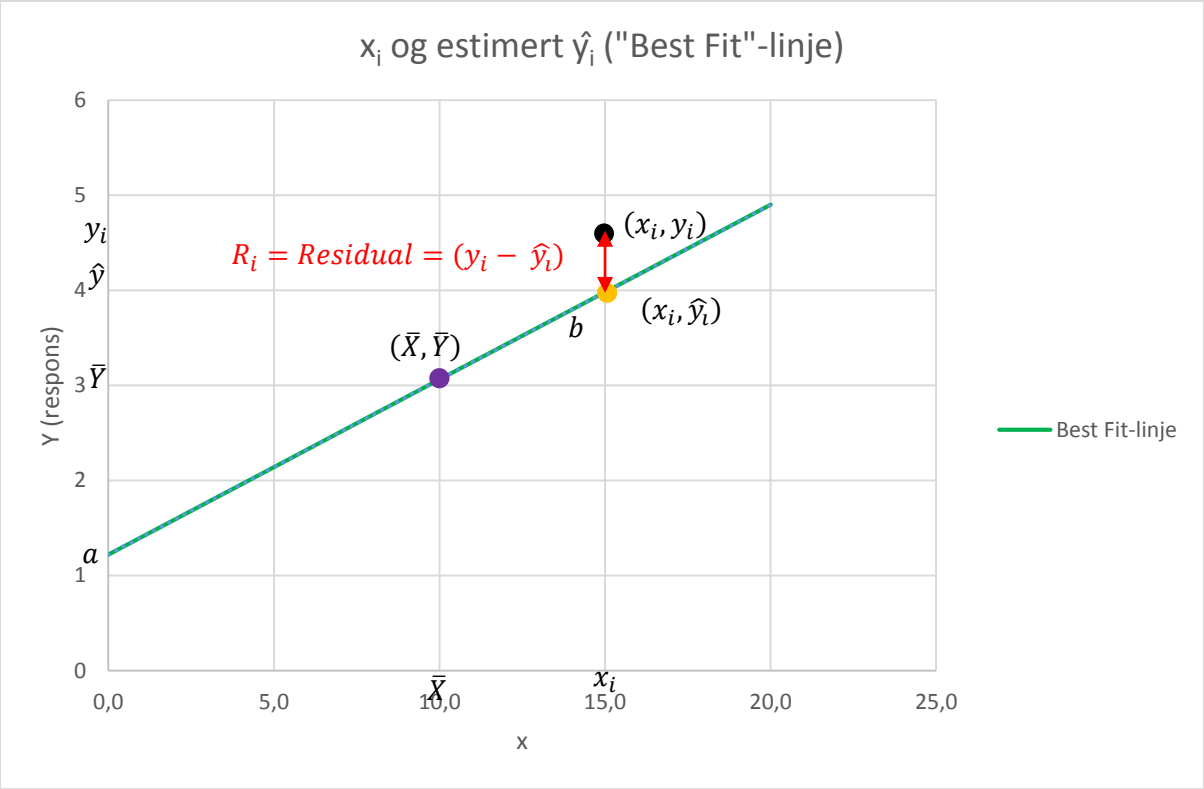
Men, denne gangen tar vi utgangspunkt i rotasjonspunktet for linjen. Vi har nå alternativ måte å beregne den forventede responsverdien (for det oransje punktet):

$\hat{y}_i = \bar{Y} + b(x_i - \bar{X})$ [11]

Fordelen med denne likningen [11] i stedet for $\hat{y}$=*bx+a*, er at nå har vi «kvittet oss med» a-koeffisienten i likningen.

Vi har nå kun én variabel utfordring; b. Vi forutsetter at våre påtrykte trykkverdier (x-verdier) er nøyaktige (x-verdien er ikke beheftet med varians).

5. Variasjon i avlest responsverdi (Y_i) og forvent responsverdi ($\hat{y}_i$)



Figur 6: For på påtrykte x-verdien (x_i) har vi her et avvik mellom den forventede responsverdien ($\hat{y}_i$) og den faktiske avleste responsverdien (y_i).

Men, vår observasjon (y_i) lå ikke på den forventede responsen ($\hat{y}_i$). Dette skyldes «kosmisk støy». Ja, vi har en naturlig forklaring på avviket mellom (y_i) og ($\hat{y}_i$), men vi har ingen muligheter for å beregne denne på forhånd. Vi vet at den avleste verdien (y_i) har en fordeling som vist i figur 3.

Vi beregner det uforklarlige Residual for våre fem datapar fra tabell 2 (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), og (x_5, y_5).

Matematisk uttrykker vi variansen for den uforklarlige Residual slik for våre 5 datapar:

$$S = \sum_{i=1}^{5}(Residual_i)^2 = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 \quad [12]$$

Vi erstatter $\hat{y}_i$ med $[\bar{Y} + b(x_i - \bar{X})]$, og får

$$S = \sum_{i=1}^{5}(y_i - [\bar{Y} + b(x_i - \bar{X}])^2 \quad [13]$$

Vi rydder litt i likningen, og får

$$S = \sum_{i=1}^{5}([y_i - \bar{Y}] + b[x_i - \bar{X}])^2 \quad [14]$$

Vi setter

$y_i - \bar{Y} = Y_i$	Y_i er total residual for et punkt. Avvik mellom observasjon (y_i) og sentralverdien ($\bar{Y}$).	[15]
$x_i - \bar{X} = X_i$	X_i er endringen som gjøres fra påvirkningspunktet (x_i) og sentralverdien ($\bar{X}$).	[16]

Y_i = Total Residual = Forklarlig Residual + Uforklarlig Residual = $(\hat{y}_i - \bar{Y}) + (y_i - \hat{y}_i) = y_i - \bar{Y}$ [17]

Vi skal nå beskrive variansutrykket slik:

$S = \sum_{i=1}^{5}([y_i - \bar{Y}] - b[x_i - \bar{X}])^2$ [18]

$S = \sum_{i=1}^{5}(Y_i - bX_i)^2$ [19]

Vi ønsker å bestemme en «Best Fit»-linje med den minste verdi på S, det vil si med minst variasjon. Dette finner vi å derivere funksjonen S og sette det deriverte uttrykket lik 0.

6. Bestemme utrykk for koeffisienten b

For å lettere komme frem til det deriverte uttrykket av S, bruker vi kjerneregelen.

$\frac{dS}{db} = \frac{dS}{dK} \cdot \frac{dK}{db}$ [20]

Vi definerer kjerneuttrykket:

$K = Y_i - bX_i$ [21]

Vi kan nå skrive variasjonsuttrykket [19] slik:

$S = \sum_{i=1}^{5}(K)^2$ [22]

Kjerneregelen sier at vi skal derivere funksjonen med hensyn til kjernen ($\frac{dS}{dK}$), og multipliserer dette med den deriverte av kjernen ($\frac{dK}{db}$).

Vi gjør altså derivasjonen av S i to etapper.

Vi deriverer funksjonen med hensyn til kjernen:

$\frac{dS}{dK} = 2\sum_{i=1}^{5}(K)^{2-1} = 2\sum_{i=1}^{5}(K)^1 = 2\sum_{i=1}^{5}K$ [23]

Vi deriverer kjernen med hensyn til b:

$\frac{dK}{db} = \frac{d(Y_i - bX_i)}{db} = 0 - 1X_i b^{1-1} = -X_i b^0 = -X_i 1 = -X_i$ [24]

Når vi gjør partiell derivasjon med hensyn til b, er X_i *og* Y_i å betrakte som konstanter. Og, deres deriverte er derfor 0.

Vi setter sammen uttrykkene, slik at vi kan bestemme den deriverte av S med hensyn på b, slik

$\frac{dS}{db} = \frac{dS}{dK} \cdot \frac{dK}{db} = [2\sum_{i=1}^{5}K] \cdot [-X_i] = 2\sum_{i=1}^{5}(Y_i - bX_i) \cdot [-X_i] = 2\sum_{i=1}^{5}(-X_iY_i + bX_i^2)$ [25]

Vi er interessert å bestemme den b-verdi som gjør det deriverte uttrykket lik 0.

$\frac{dS}{db} = 0$ [26]

$2\sum_{i=1}^{5}(-X_iY_i + bX_i^2) = 0$ [27]

Vi dividerer venstre og høyre side med 2.

$$\frac{2\sum_{i=1}^{5}(-X_iY_i+bX_i^2)}{2}=\frac{0}{2} \quad [28]$$

$$\sum_{i=1}^{5}(-X_iY_i+bX_i^2)=0 \quad [29]$$

Vi løser opp parentesen, og får:

$$\sum_{i=1}^{5}(-X_iY_i)+b\sum_{i=1}^{5}(X_i^2)=0 \quad [30]$$

Vi rydder litt til, og løser dette med hensyn til b, og får

$$b\sum_{i=1}^{5}(X_i^2)=\sum_{i=1}^{5}(X_iY_i) \quad [31]$$

Vi dividerer med $\sum_{i=1}^{5}(X_i^2)$ på venstre og høyre side, og får:

$$b=\frac{\sum_{i=1}^{5}(X_iY_i)}{\sum_{i=1}^{5}(X_i^2)}=\frac{Kovarians\,(X,Y)}{Varians(X)} \quad [32]$$

Stigningstallet b for den lineære regresjonslinjen forholdet mellom funksjonen Kovarians for dataparene og variansen av den påtrykte verdien.

A11 =KOVARIANS.S(A6:A10;B6:B10)

	A	B	C	D	E
1	**Kalibreringskurver; på jakt etter statistisk signifikan**				
2					
3	Påtrykt	Avlest	Estimert		
4	X	Y	Ŷ	Confidence (CI)	
5	(BarG)	(Volt)	(Ŷ = bx +a)	Nedre	Øvre
6	0,0	1,0	1,22	0,48321041	1,95678959
7	5,0	2,5	2,14	1,61901109	2,66098891
8	10,0	3,1	3,06	2,63461433	3,48538567
9	15,0	3,7	3,98	3,45901109	4,50098891
10	20,0	5,0	4,9	4,16321041	5,63678959
11	11,5	62,5	0,184		

Figur 7: Skjermdump fra Microsoft Excel

I Microsoft Excel kan vi bruke funksjonen =KOVARIANS.S(A6:B10) for våre datapar (celle A11). Og, tilsvarende =VARIANS.S(A6:A10) for å bestemme variansen (celle B11).

7. Bestemme usikkerhetsintervall (standard avvik) for uttrykket b

Siden $b=\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n}(X_i^2)}$ kan vi kalkulere variansen av b slik:

$$Var\,(b)=Var\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n}(X_i^2)}\right) \quad [33]$$

Siden vi antar at alle x-verdier ikke er beheftet med variasjoner, kan vi derfor uttrykke disse som konstanter. Vi kan derfor omskrive utrykket over til å bli slik:

$$Var\,(b)=Var\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n}(X_i^2)}\right)=\frac{1}{(\sum_{i=1}^{n}(X_i^2))^2}\cdot\sum_{i=1}^{n}(X_iY_i) \quad [34]$$

Og husk at funksjonen Var er å ta kvadratet av uttrykket, så når vi trekker ut konstanter ($\sum_{i=1}^{n}(X_i{}^2)$ fra uttrykket, skal uttrykket ($\sum_{i=1}^{n} X_i{}^2$) kvadreres!

Videre, vi har at uttrykket Var ($\sum_{i=1}^{n}(X_iY_i)$) kan skrives slik:

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \text{Var}\left(X_1Y_1 + X_2Y_2 + \ldots . + X_nY_n\right) \quad [35]$$

Igjen betrakter vi x-verdier som konstanter og kan trekkes ut av parentesen. Og husk at funksjonen Var er å ta kvadratet av uttrykket, så når vi trekker ut konstanter ut av uttrykket, skal disse kvadreres!

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \text{Var}\left(X_1Y_1 + X_2Y_2 + \ldots . + X_nY_n\right) = X_1{}^2\text{Var}(Y_1) + X_2{}^2\text{Var}(Y_2) + \ldots + X_n{}^2\text{Var}(Y_n) \quad [36]$$

Fra figur 4, som tok for seg «Kosmisk støy», sa vi at variasjonsområdet for ε gjelder for alle x-verdier. I dette tilfellet kan vi derfor sette:

$$\text{Var}(Y_1) = \text{Var}(Y_2) = \ldots = \text{Var}(Y_n) = \text{Variasjon av «kosmisk støy»} = (\sigma_\varepsilon)^2 = (\sigma_{Residual})^2 \quad [37]$$

Vi har med andre ord én felles varians for residual y, og vi erstatter det ovennevnte uttrykket med $(\sigma_\varepsilon)^2 = (\sigma_{Residual})^2$.

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = X_1{}^2\text{Var}(Y_1) + X_2{}^2\text{Var}(Y_2) + \ldots + X_n{}^2\text{Var}(Y_n) = (X_1{}^2 + X_2{}^2 \ldots . + X_n{}^2) \cdot (\sigma_\varepsilon)^2 \quad [38]$$

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \left(\sum_{i=1}^{n} X_i{}^2\right) \cdot (\sigma_\varepsilon)^2 \quad [39]$$

Vi går tilbake til utrykket for Var(b), og får

$$\text{Var (b)} = \text{Var}\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n} X_i{}^2}\right) = \frac{1}{(\sum_{i=1}^{n}(X_i{}^2))^2} \cdot \text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \frac{1}{(\sum_{i=1}^{n}(X_i{}^2))^2} \cdot \left(\sum_{i=1}^{n} X_i{}^2\right) \cdot (\sigma_\varepsilon)^2 \quad [40]$$

$$\text{Var (b)} = \frac{\left(\sum_{i=1}^{n} X_i{}^2\right)}{\left(\sum_{i=1}^{n} X_i{}^2\right)} \cdot (\sigma_\varepsilon)^2 = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n} X_i{}^2} \quad [41]$$

Fra tidligere hadde vi $X_i = [x_i - \bar{X}]$. Alternativ måte i beskrive variansen til b på er slik:

$$\text{Var (b)} = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n} X_i{}^2} = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \quad [42]$$

Estimatet for standardavviket for stigningskoeffisienten b ($s_{\hat{b}}$) i en lineær regresjon er kvadratroten av variansuttrykket:

$$s_{\hat{b}} = \sqrt{Var(b)} = \sqrt{\frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{\sqrt{(\sigma_\varepsilon)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{Standardusikkerhet\ for\ Residual}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} \quad [43]$$

Vi trenger med andre ord å beregne standard usikkerhet for residual, og dividere med kvadratroten av summen av den kvadrerte x-differansen.

8. Bestemme standard usikkerhet for $\hat{y}$.

Usikkerhetsintervallet for estimert gjennomsnitt kan nå beregnes. Usikkerheten kan beregnes med Var($\hat{y}$).

Vi vet at estimert y ved x-verdien kan uttrykkes slik:

$$\hat{y} = \bar{Y} + b(x - \bar{X}) \quad [44]$$

Vi brukes samme varians-teorem som tidligere.

Var ($\hat{y}$) = Var ($\overline{Y}$ + b(x - $\bar{X}$)) [45]

Vi betrakter (x - $\bar{X}$) som en konstant. Vi løser opp Var-parentesen

Var ($\hat{y}$) = Var ($\overline{Y}$) + [$(x - \bar{X})^2$ · Var(b)] [46]

Det første deluttrykket i Var ($\hat{y}$) er Var ($\overline{Y}$). Vi har at sentralverdien av y er summen av alle y_i delt på antall y-verdier.

$$\overline{Y} = \frac{\sum_{i=1}^{n} yi}{n} \quad [47]$$

Vi har nå

$$\text{Var}(\overline{Y}) = \text{Var}\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n^2} \cdot \text{Var}\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2} \cdot (n \cdot (\sigma_\varepsilon)^2) = \frac{(\sigma_\varepsilon)^2}{n} \quad [48]$$

Fra før har vi at:

$$\text{Var (b)} = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \quad [49]$$

Nå har vi:

$$\text{Var}(\hat{y}) = \text{Var}(\overline{Y}) + [(x - \bar{X})^2 \cdot \text{Var(b)}] = \frac{(\sigma_\varepsilon)^2}{n} + [(x - \bar{X})^2 \cdot \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}] \quad [50]$$

Vi rydder i ligningen:

$$\text{Var}(\hat{y}) = (\sigma_\varepsilon)^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right) \quad [51]$$

Standard usikkerhet for estimert y ($\hat{y}$) er kvadratroten av varians-uttrykket:

$$s_{\hat{y}} = \sqrt{Var(\hat{y})} = \sqrt{(\sigma_\varepsilon)^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right)} = s_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} \quad [52]$$

9. CI (Confidence Interval)

Ved forsøk som inkluderer 30 eller færre datapar, bør vi bruke Student t-faktor for å ta hensynet til små sub-populasjoner. I vårt tilfelle har vi fem sub-populasjoner.

T-verdien kan vi finne i tabeller, eller la Excel beregne denne for oss.

$T_{(\alpha/2,v)}$ [53]

Konfidensintervallet er 100(1- α). α har verdien (0,05) = 5 prosent, i og med vi er på jakt etter usikkerhetsintervallet 95 %. Usikkerheten 5 % skal fordeles på to sider (/2 →2 «tail»). Parameteren 'v' er «Degrees of Freedom).

v = n – 2 [54]

Vi har n=5 datapar, og subtraherer med verdien 2 i og med vi har 2 variabler (x- og y-parametere).

Excel-funksjon:

=T.INV.2T(0,05;3) [55]

Setter vi det hele sammen

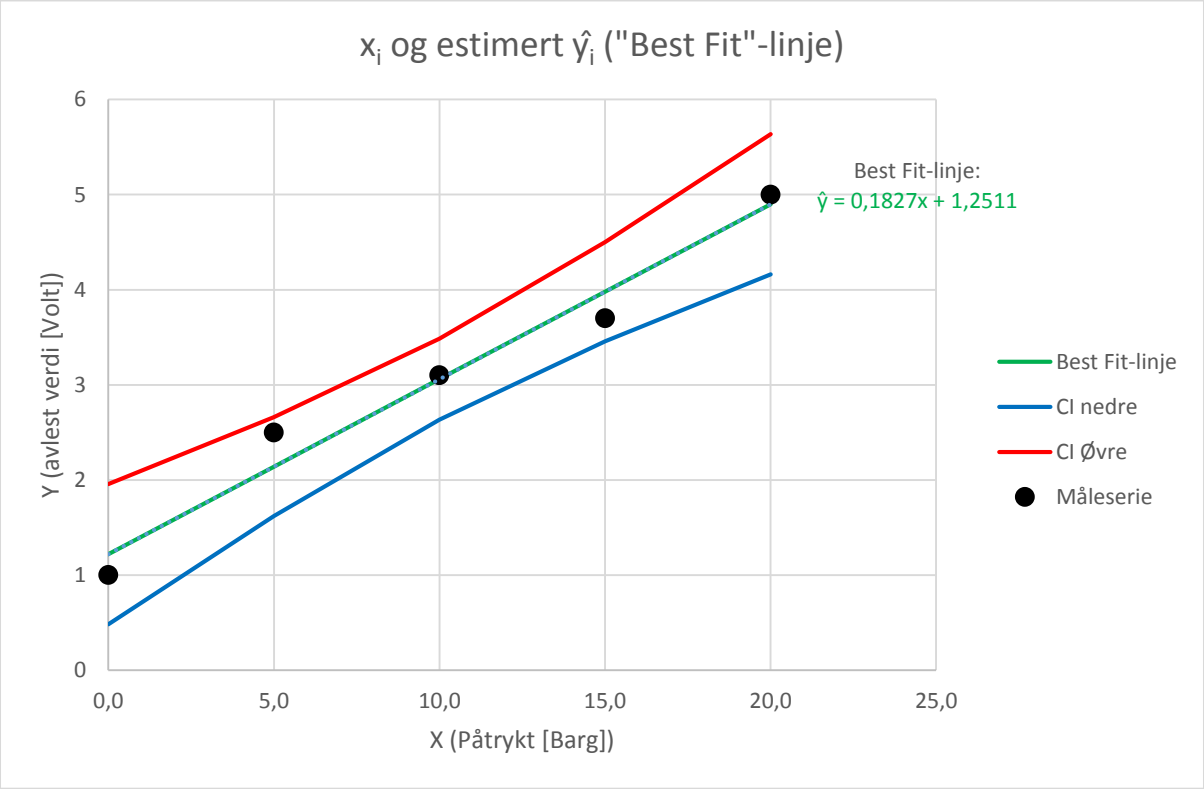
CI = [«Best Fit»-linje] ± [usikkerhetsintervall] [56]

$$CI = \hat{y} \pm \left[T_{(\alpha/2,v)} \cdot \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-2)}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right] \quad [57]$$

Vårt datagrunnlag for lineær regresjon:

Påtrykt	Avlest	Datapar
X	Y	
(BarG)	(Volt)	
0,0	1,0	(x_1,y_1)
5,0	2,5	(x_2,y_2)
10,0	3,1	(x_3,y_3)
15,0	3,7	(x_4,y_4)
20,0	5,0	(x_5,y_5)

Tabell 3: Datapar vi fikk ved kalibrering av trykksensor



Figur 8: Grafisk presentasjon fra Microsoft Excel på datapar, «Best Fit»-linje med usikkerhetsintervall.

Neste kalibreringskurve («Best Fit»-linje) skal med 95 prosent sannsynlighet ligge innenfor nevnte «Confidence Intervall». Det er 5 prosent sjanse for at den neste «Best Fit»-linje vil ligge utenfor «Confidence Interval».

Da det er liten sannsynlighet for at neste «Best Fit»-linje skal ligge utenfor «Confidence Interval», og vi bør være varsom med å godkjenne/ta i bruk en slik «Best Fit»-linje. Iallfall må vi nøye undersøke hvorfor vår neste «Best Fit»-linje havnet på utsiden av «Confidence Interval». Det kan være at en

eller flere av dataparene innehar feil knyttet til seg, noe som gjør «Best Fit»-linjen til en ekstremverdi/»uteligger».

10. Koeffisienten a

Basert på et datasett (x_i, y_i) skal vi beregne en Best Fit»-linje basert på lineær regresjon

$\hat{y}_i = \mathrm{b}x_i + \mathrm{a}$ [58]

slik at summen av kvadratet av residualene

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

blir minst mulig.

Vi erstatter $\hat{y}_i$ med $[\mathrm{b}x_i + \mathrm{a}]$ og får

$S = \sum_{i=1}^{n}(y_i - [\mathrm{b}x_i + \mathrm{a}])^2$ [59]

Vi løser opp firkantparentesen:

$S = \sum_{i=1}^{n}(y_i - \mathrm{b}x_i - \mathrm{a})^2$

For å lettere komme frem til det deriverte uttrykket av S med hensyn til a, bruker vi kjerneregelen.

$\frac{dS}{da} = \frac{dS}{dK} \cdot \frac{dK}{da}$ [60]

Vi definerer kjerneuttrykket:

$K = y_i - \mathrm{b}x_i - \mathrm{a}$ [61]

Vi kan nå skrive S slik:

$S = \sum_{i=1}^{n}(K)^2$, når $K = y_i - \mathrm{b}x_i - \mathrm{a}$ [62]

Kjerneregelen sier at vi skal derivere funksjonen S med hensyn til kjernen K; $(\frac{dS}{dK})$, og multiplisere dette med den deriverte av kjernen K med hensyn til a-variabelen; $(\frac{dK}{da})$.

Vi gjør altså derivasjonen av S i to etapper.

Vi deriverer funksjonen med hensyn til kjernen:

$\frac{dS}{dK} = 2\sum_{i=1}^{n}(K)^{2-1} = 2\sum_{i=1}^{n}(K)^1 = 2\sum_{i=1}^{n}K$ [63]

Vi deriverer kjernen $(y_i - \mathrm{b}x_i - \mathrm{a})$ med hensyn til a:

$\frac{dK}{da} = \frac{d(y_i - \mathrm{b}x_i - \mathrm{a})}{da} = 0 - 0 - 1a^{1-1} = -\mathrm{a}^0 = -1$ [64]

x_i og y_i betraktes som konstanter når vi gjør partiell derivasjon med hensyn til a. Og, de deriverte av x_i og y_i blir derfor 0.

Vi setter sammen uttrykkene, slik at vi kan bestemme den deriverte av S med hensyn på a, slik

$\frac{dS}{da} = \frac{dS}{dK} \cdot \frac{dK}{da} = [2\sum_{i=1}^{n}K] \cdot [-1] = [2\sum_{i=1}^{n}(y_i - \mathrm{b}x_i - \mathrm{a})] \cdot [-1] = -2 \cdot \sum_{i=1}^{n}(y_i - \mathrm{b}x_i - \mathrm{a})$ [65]

Vi løser opp parentesen

$\frac{dS}{da}$ = -2 $\sum_{i=1}^{n}(y_i - \mathrm{b}x_i - \mathrm{a})$ = -2$\sum_{i=1}^{n} y_i$ +2b$\sum_{i=1}^{n} x_i$ +2$\sum_{i=1}^{n} a$ [66]

Vi kan erstatte deler av de to første uttrykkene, da summen av alle y er lik antall n multiplisert med sentralverdien $\bar{Y}$ [67]. Tilsvarende; summen av alle x er lik antall n multiplisert med sentralverdien $\bar{X}$ [68].

$\sum_{i=1}^{n} y_i$ = $n\bar{Y}$ [67]

$\sum_{i=1}^{n} x_i$ = n$\bar{X}$ [68]

Derivasjonslikningen [66] kan nå skrives på formen

$\frac{dS}{da}$ = -2$n\bar{Y}$ +2b$n\bar{X}$ +2na [69]

Vi er interessert å bestemme den a-verdi som gjør det deriverte uttrykket [69] lik 0.

$\frac{dS}{da} = 0$ [70]

-2$n\bar{Y}$ +2b$n\bar{X}$ +2na = 0 [71]

(-2n) er felles for alle ledd i [71], og vi setter -2n utenfor parentesen

-2n($\bar{Y}$ -b$\bar{X}$ - a) = 0 [72]

Vi dividerer venstre og høyre side [72] med -2n.

$\frac{-2\mathrm{n}(\bar{Y} - \mathrm{b}\bar{X} - a)}{-2n}$ = $\frac{0}{-2n}$ [73]

$\bar{Y} - \mathrm{b}\bar{X} - a$ = 0 [74]

Vi rydder litt til, og kommer frem til vårt sluttuttrykk for a: [75]

a = $\bar{Y}$ - b$\bar{X}$

Denne likningen forteller oss at konstanten a settes slik at regresjonslinjen *må* gå gjennom sentralverdiene for $\bar{X}$ og $\bar{Y}$. Dette gir mening da dette punktet ($\bar{X}$,$\bar{Y}$) utgjør *det* single dataparet som best representerer *alle* datapar som vi har «trukket» ut av sub-populasjonene. [76]

Med vennlig hilsen

Trainor Elsikkerhet AS

Rune Øverland
Senioringeniør

Tønsberg april 2015