# *Er neste datapar ved kalibrering en ekstremverdi som skal forkastes?*

*v/Rune Øverland, Trainor Elsikkerhet AS*

## 1. Innledning

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

PI = Estimert Y ± [ usikkerhetsintervall ] [1]

Vi bruker dette verktøyet til å etablere et usikkerhetsintervall (for eksempel ved 95 %). Formålet er å vurdere om det neste datapunktet vi innhenter ved kalibrering ligger innenfor eller utenfor dette usikkerhetsintervallet. Ligger datapunktet utenfor usikkerhetsintervallet, kaller vi datapunktet for en ekstremverdi eller «uteligger», og skal forkastes. Usikkerhetsintervallet etableres slik at det er 95 prosent sannsynlig at det neste dataparet ligger innenfor usikkerhetsintervallet.

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

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

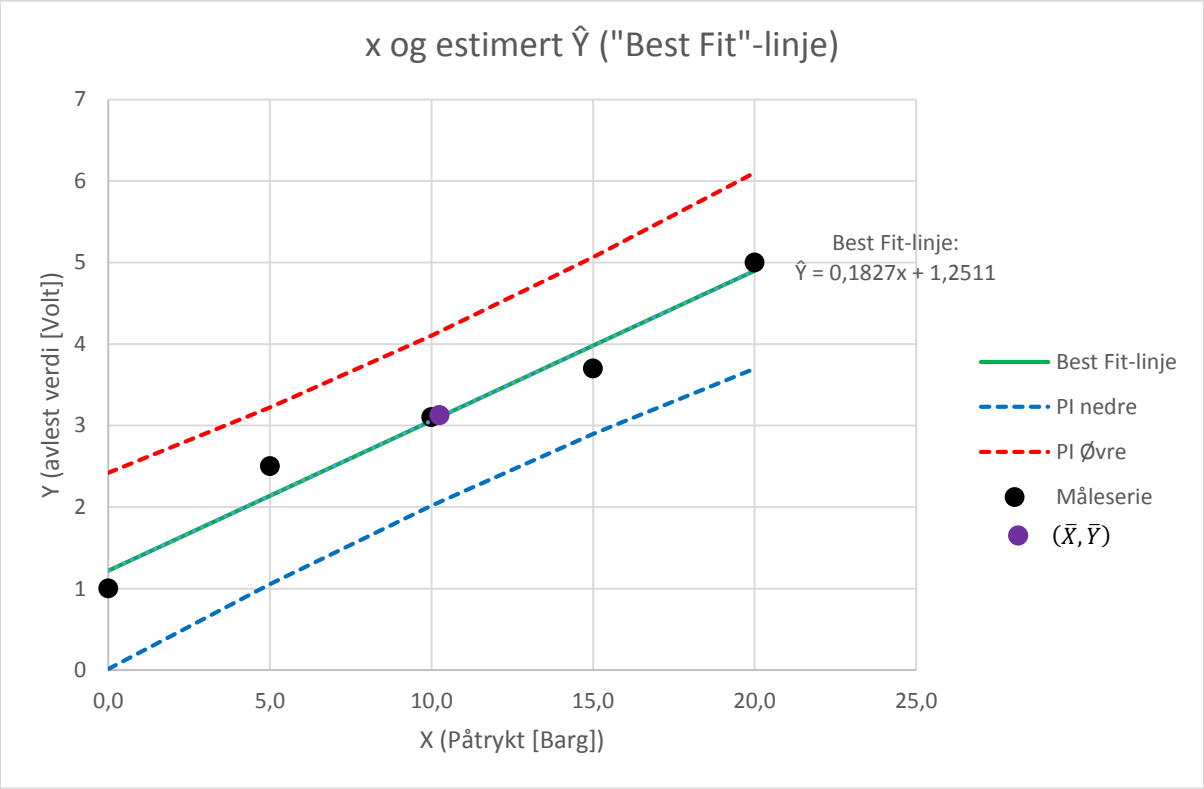
Uttrykket for PI er temmelig likt uttrykket for CI som jeg utledet i forrige artikkel. Forskjellen er at det har blitt lagt til **1** under kvadratrottegnet.

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 ($\hat{y}$ =bx+a). |
| 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 og det usikkerhetsintervallet (for eksempel 95%) vi er på jakt etter. |
| $\alpha$ | $\alpha$ forteller om usikkerhetsintervallet: 100(1- $\alpha$) 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 ($\hat{y}$ =bx+a). |
| $\bar{X}$ | 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 PI) 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, og grafisk presentert i figur 1.

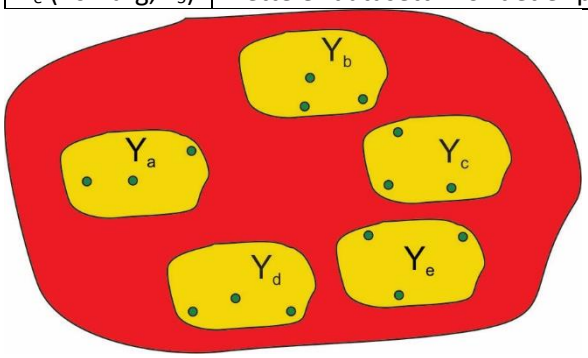
| X (Barg) | Y (Volt) | Datapar |
|---|---|---|
| 0,0 | 1,0 | $(x_1,y_1)$ |
| 5,0 | 2,5 | $(x_2,y_2)$ |
| 10,0 | 3,1 | $(x_3,y_3)$ |
| 15,0 | 3,7 | $(x_4,y_4)$ |
| 20,0 | 5,0 | $(x_5,y_5)$ |

*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 er påtrykt 0 Barg, med avleste sensorverdier $Y_1$. |
| $Y_b$ (5 Barg, $Y_2$) | Dette er datasett hvor det er påtrykt 5 Barg, med avleste sensorverdier $Y_2$. |
| $Y_c$ (10 Barg, $Y_3$) | Dette er datasett hvor det er påtrykt 10 Barg, med avleste sensorverdier $Y_3$. |
| $Y_d$ (15 Barg, $Y_4$) | Dette er datasett hvor det er påtrykt 15 Barg, med avleste sensorverdier $Y_4$. |
| $Y_e$ (20 Barg, $Y_5$) | Dette er datasett hvor det er 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) inn i 5 sub-populasjoner/utvalg (oransje områder. Sub-populasjonene kaller vi $Y_a$, $Y_b$, $Y_c$, $Y_d$, og $Y_e$.*

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

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

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 ($N \approx \infty$) 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 $Y_a$:

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

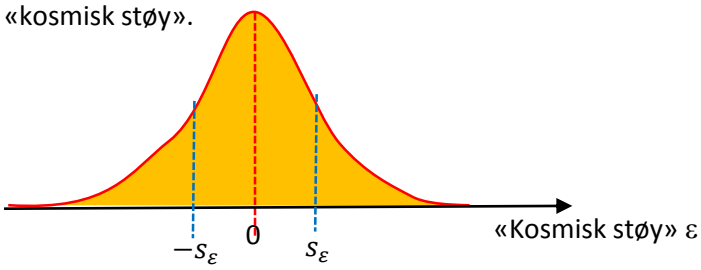
Uansett, vi plotter sentralverdier for våre sentralverdier (enten de er beregnede ($\overline{U_a}$) eller estimerte ($\mathrm{E}(\overline{U_a})$) inn i et diagram (figur 3). Vi trekker en rett linje gjennom sentralverdiene.

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 $s_\varepsilon$ for den «kosmiske» støyen. Variasjonen er symmetrisk rundt den lokale 0-verdien.*

«Kosmisk støy» er normalfordelt, med sentralverdi = 0

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

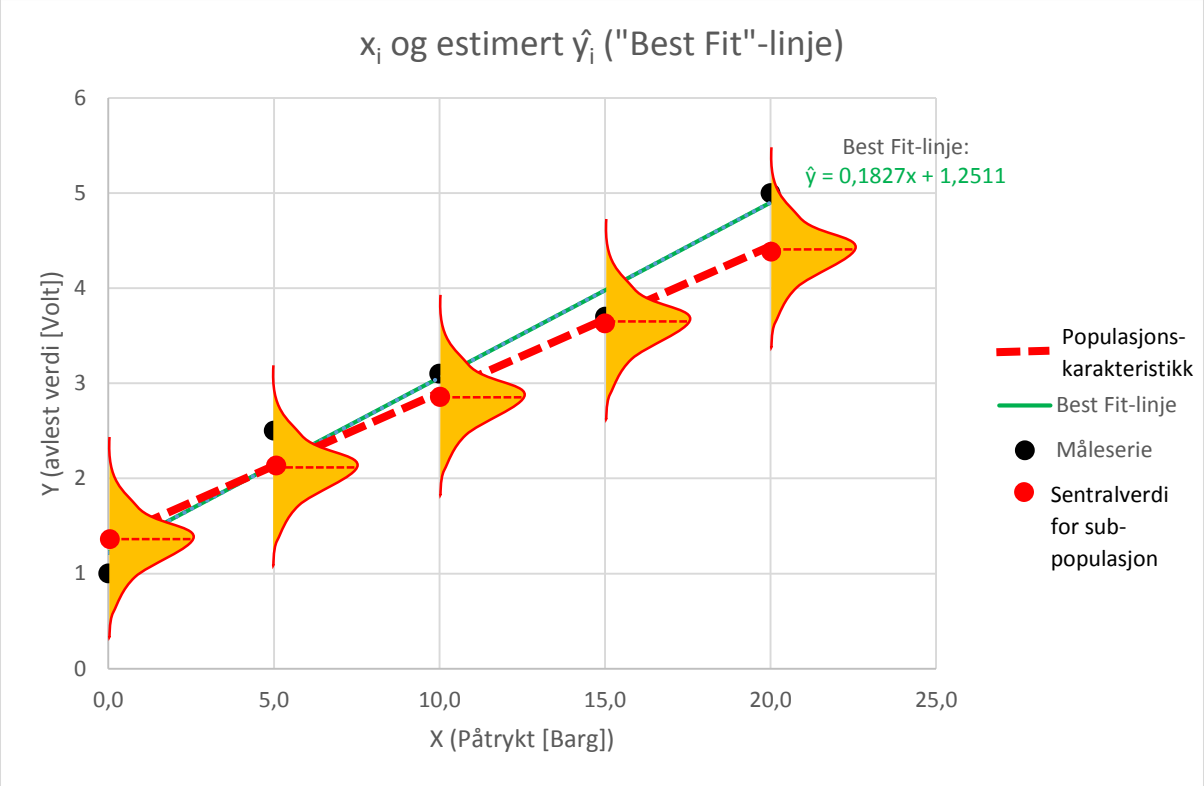
Variansen for den «kosmiske» støyen er:

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

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 (se figur 4): 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 at populasjonen har en lineær karakteristikk som kan beskrives slik:

$$U_{Y|x}=\beta x+\alpha \qquad [8]$$

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.

$$U_{Y|x}=\beta x+\alpha + \varepsilon \qquad [9]$$

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 assosieres som variasjon i områdene $Y_a$, $Y_b$, $Y_c$, $Y_d$, og $Y_e$ i figur 2.

**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.

| | Eksempel 1 | Eksempel 2 | Eksempel 3 | Eksempel 4 | Eksempel 5 |
|---|---|---|---|---|---|
| x (Barg) | y (volt) | y (volt) | y (volt) | y (volt) | y (volt) |
| 0 | 1,0 | 1,1 | 1,1 | 1,3 | 0,8 |
| 5 | 2,5 | 2,5 | 2,3 | 1,5 | 1,5 |
| 10 | 3,1 | 3,1 | 3,1 | 3,2 | 3,2 |
| 15 | 3,7 | 3,7 | 3,9 | 4,3 | 4,3 |
| 20 | 5,0 | 4,9 | 4,9 | 5,0 | 5,5 |
| $\hat{y}$ =bx+a | b = 0,184<br>a = 1,22 | b = 0,176<br>a = 1,30 | b = 0,184<br>a = 1,22 | b = 0,204<br>a = 1,02 | b = 0,244<br>a = 0,62 |

*Tabell 3: Eksempler på ulike responser (y-verdier) på grunn av «kosmisk støy». Dette gir noe forskjellige estimater på koeffisientene b og a. Alle eksemplene har det samme rotasjonspunktet* $(\bar{X}, \bar{Y})$ $\bar{X} = 10{,}0$ *og* $\bar{Y} = 3{,}06$.

I og med «kosmisk støy» (ε), må vi tilsvarende konstruere et usikkerhetsintervall (PI – Prediction Interval). Dette intervallet skal fortelle oss at vi med 95 prosent sikkerhet forventer å det neste datapunktet.

Det neste datapunktet har to usikkerheter knyttet til seg (se figur 5 og figur 6)

*Forklarlig avvik basert på regresjonsmodellen*

Den første usikkerheten er hvor regresjonslinjen befinner seg (figur 5). Den responsverdi (y-verdi) som best representerer datasettet er sentralverdien for datasettet ($\bar{Y}$).

Basert på regresjonsmodellen [$\mathrm{bx} + \mathrm{a}$ ], for en gitt neste x-verdi, forventer vi at den neste responsverdien skal ha verdien $(\hat{y}_i = \bar{Y} + Y_i)$.

$Y_i$ kalles ofte for det 'forklarlige' avviket.

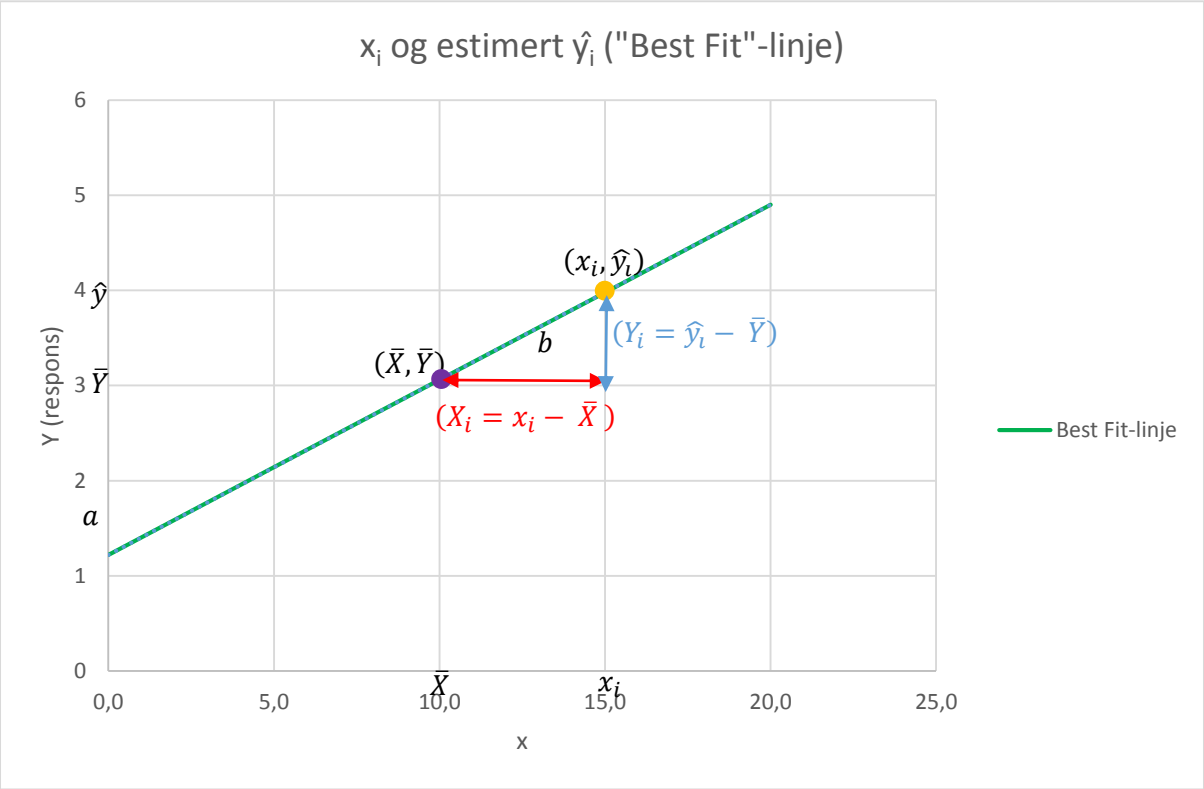
Koeffisientene b og a har et usikkerhetsintervall knyttet til seg. I denne artikkelen skal jeg nå vise uttrykket for disse.

Oppsummert: Vi skal bestemme et variasjonsuttrykk Var($\hat{y}_i$). [10]

*Uforklarlig avvik basert på regresjonsmodellen*

Kikk på figur 6. 'Kosmisk' støy kan ikke forklares direkte av regresjonsmodellen, og kalles derfor den uforklarlige bidraget.

Oppsummert: Vi skal bestemme et variasjonsuttrykk Var(ε). [11]




*Figur 5: Grafisk fremstilling av rotasjonspunktet $(\bar{X}, \bar{Y})$, og koeffisientene b og a for den rette regresjonslinjen* $\hat{y}$ =bx+a. [12]

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.

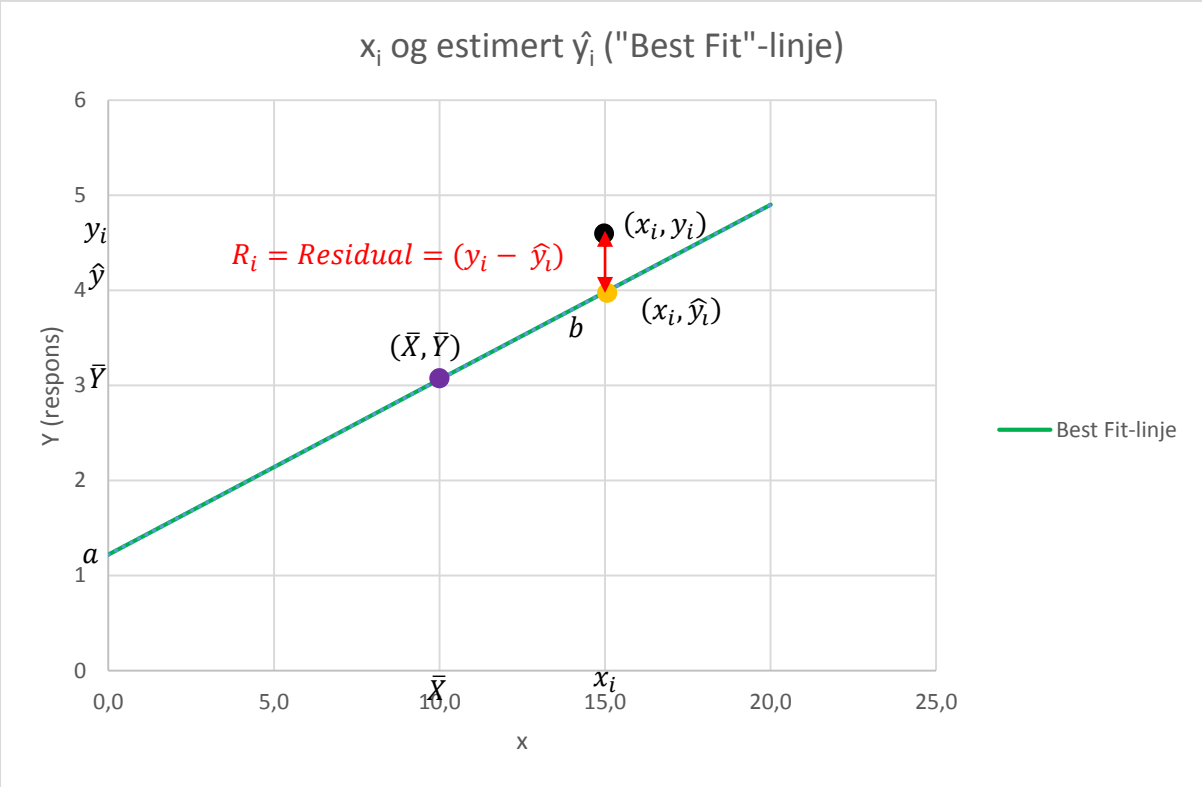
Men, denne gangen tar vi utgangspunkt i rotasjonspunktet for linjen. Vi har nå alternativ måte å beregne den forventede responsverdien (for det oransje punktet):

$\hat{y}_i = \bar{Y} + b(x_i - \bar{X})$ [13]

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 ($y_i$). Avviket mellom den observerte responsen og regresjonspunktet [$y_i - \hat{y}_i$] er en av de viktigste størrelsene i denne artikkelen!*

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:

$$S = \sum_{i=1}^{5}(Residual_i)^2 = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 \qquad [14]$$

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

$$S = \sum_{i=1}^{5}(y_i - [\bar{Y} + b(x_i - \bar{X}])^2 \qquad [15]$$

Vi rydder litt i likningen, og får

$$S = \sum_{i=1}^{5}([y_i - \bar{Y}] + b[x_i - \bar{X}])^2 \qquad [16]$$

Vi setter

| | | |
|---|---|---|
| $y_i - \bar{Y} = Y_i$ | $Y_i$ er total residual for et punkt. Avvik mellom observasjon ($y_i$) og sentralverdien ($\bar{Y}$). | [17] |
| $x_i - \bar{X} = X_i$ | $X_i$ er endringen som gjøres fra påvirkningspunktet ($x_i$) og sentralverdien ($\bar{X}$). | [18] |

$$Y_i = \text{Total Residual} = \text{Forklarlig Residual} + \text{Uforklarlig Residual} = (\hat{y}_i - \bar{Y}) + (y_i - \hat{y}_i) = y_i - \bar{Y} \qquad [19]$$

Vi skal nå beskrive variansutrykket slik:

$$S = \sum_{i=1}^{5}([y_i - \bar{Y}] - b[x_i - \bar{X}])^2 \quad [20]$$

$$S = \sum_{i=1}^{5}(Y_i - bX_i)^2 \quad [21]$$

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 Var(ε).

Bestemme uttrykket Var(ε) er ganske enkelt.

Kikk på figur 3! Variasjon av det 'uforklarlige' er variasjonen av 'kosmisk støy'.

$$Var(\varepsilon) = (\sigma_\varepsilon)^2 \quad [22]$$

## 7. Bestemme utrykk for Var($\hat{y}_i$).

Vi har en flertrinnsvei på gå for å bestemme Var($\hat{y}_i$).

Først må vi bestemme verdien på koeffisienten b som inngår i regresjonslikningen. Dette gjøres i punkt 7a.

Deretter må vi bestemme et usikkerhetsintervall for denne koeffisienten. Dette gjøres i punkt 7b.

Til slutt bestemmer vi Var($\hat{y}_i$). Dette gjøres i punkt 7c.

### a. Bestemme utrykk koeffisienten b

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

$$\frac{dS}{db} = \frac{dS}{dK} \cdot \frac{dK}{db} \quad [23]$$

Vi definerer kjerneuttrykket:

$$K = Y_i - bX_i \quad [24]$$

Vi kan nå skrive variasjonsuttrykket [19] slik:

$$S = \sum_{i=1}^{5}(K)^2 \quad [25]$$

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:

$$\frac{dS}{dK} = 2\sum_{i=1}^{5}(K)^{2-1} = 2\sum_{i=1}^{5}(K)^1 = 2\sum_{i=1}^{5}K \quad [26]$$

Vi deriverer kjernen med hensyn til b:

$$\frac{dK}{db} = \frac{d(Y_i - \mathrm{b}X_i)}{db} = 0 - 1X_i b^{1-1} = -X_i b^0 = -X_i 1 = -X_i \qquad [27]$$

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

$$\frac{dS}{db} = \frac{dS}{dK} \cdot \frac{dK}{db} = [2\sum_{i=1}^{5} K] \cdot [-X_i] = 2\sum_{i=1}^{5}(Y_i - \mathrm{b}X_i) \cdot [-X_i] = 2\sum_{i=1}^{5}(-X_i Y_i + \mathrm{b}X_i^2) \qquad [28]$$

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

$$\frac{dS}{db} = 0 \qquad [29]$$

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

Vi dividerer venstre og høyre side med 2.

$$\frac{2\sum_{i=1}^{5}(-X_i Y_i + \mathrm{b}X_i^2)}{2} = \frac{0}{2} \qquad [31]$$

$$\sum_{i=1}^{5}(-X_i Y_i + \mathrm{b}X_i^2) = 0 \qquad [32]$$

Vi løser opp parentesen, og får:

$$\sum_{i=1}^{5}(-X_i Y_i) + \mathrm{b}\sum_{i=1}^{5}(X_i^2) = 0 \qquad [33]$$

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

$$\mathrm{b}\sum_{i=1}^{5}(X_i^2) = \sum_{i=1}^{5}(X_i Y_i) \qquad [34]$$

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

$$\mathrm{b} = \frac{\sum_{i=1}^{5}(X_i Y_i)}{\sum_{i=1}^{5}(X_i^2)} = \frac{Kovarians\ (X,Y)}{Varians(X)} \qquad [35]$$

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).

### b. 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:

$$\text{Var (b)} = \text{Var}\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n}(X_i{}^2)}\right) \qquad [36]$$

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:

$$\text{Var (b)} = \text{Var}\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n}(X_i{}^2)}\right) = \frac{1}{(\sum_{i=1}^{n}(X_i{}^2))^2} \cdot \sum_{i=1}^{n}(X_iY_i) \qquad [37]$$

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:

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \text{Var}\left(X_1Y_1 + X_2Y_2 + \ldots + X_nY_n\right) \qquad [38]$$

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!

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \text{Var}\left(X_1Y_1 + X_2Y_2 + \ldots + X_nY_n\right) = X_1{}^2\text{Var}(Y_1) + X_2{}^2\text{Var}(Y_2) + \ldots + X_n{}^2\text{Var}(Y_n) \qquad [39]$$

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:

$$\text{Var}(Y_1) = \text{Var}(Y_2) = \ldots = \text{Var}(Y_n) = \text{Variasjon av «kosmisk støy»} = (\sigma_\varepsilon)^2 = (\sigma_{Residual})^2 \qquad [40]$$

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

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = X_1{}^2\text{Var}(Y_1) + X_2{}^2\text{Var}(Y_2) + \ldots + X_n{}^2\text{Var}(Y_n) = (X_1{}^2 + X_2{}^2 \ldots + X_n{}^2) \cdot (\sigma_\varepsilon)^2 \qquad [41]$$

$$\text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \left(\sum_{i=1}^{n} X_i{}^2\right) \cdot (\sigma_\varepsilon)^2 \qquad [42]$$

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

$$\text{Var (b)} = \text{Var}\left(\frac{\sum_{i=1}^{n}(X_iY_i)}{\sum_{i=1}^{n} X_i{}^2}\right) = \frac{1}{(\sum_{i=1}^{n}(X_i{}^2))^2} \cdot \text{Var}\left(\sum_{i=1}^{n}(X_iY_i)\right) = \frac{1}{(\sum_{i=1}^{n}(X_i{}^2))^2} \cdot \left(\sum_{i=1}^{n} X_i{}^2\right) \cdot (\sigma_\varepsilon)^2 \qquad [43]$$

$$\text{Var (b)} = \frac{(\sum_{i=1}^{n} X_i{}^2)}{(\sum_{i=1}^{n} X_i{}^2)} \cdot (\sigma_\varepsilon)^2 = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n} X_i{}^2} \qquad [44]$$

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

[45]

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

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

$$s_{\hat{b}} = \sqrt{Var(b)} = \sqrt{\frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{\sqrt{(\sigma_\varepsilon)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} = \frac{Standardusikkerhet\ for\ Residual}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2}} \quad [46]$$

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

### c. Bestemme Var($\hat{y}_i$).

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:

$$\hat{y} = \bar{Y} + b(x - \bar{X}) \quad [47]$$

Vi brukes samme varians-teorem som tidligere.

$$Var(\hat{y}) = Var(\bar{Y} + b(x - \bar{X})) \quad [48]$$

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

$$Var(\hat{y}) = Var(\bar{Y}) + [(x - \bar{X})^2 \cdot Var(b)] \quad [49]$$

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

$$\bar{Y} = \frac{\sum_{i=1}^{n} yi}{n} \quad [50]$$

Vi har nå

$$Var(\bar{Y}) = Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n^2} \cdot Var\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2} \cdot (n \cdot (\sigma_\varepsilon)^2) = \frac{(\sigma_\varepsilon)^2}{n} \quad [51]$$

Fra før har vi at:

$$Var(b) = \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \quad [52]$$

Nå har vi:

$$Var(\hat{y}) = Var(\bar{Y}) + [(x - \bar{X})^2 \cdot Var(b)] = \frac{(\sigma_\varepsilon)^2}{n} + [(x - \bar{X})^2 \cdot \frac{(\sigma_\varepsilon)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}] \quad [53]$$

Vi rydder i ligningen:

$$Var(\hat{y}) = (\sigma_\varepsilon)^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right) \quad [54]$$

## 8. PI (Prediction Interval)

### *a. Standard usikkert*

Variasjonsintervallet for den neste y, Var($y_p$), er:

Total forventet variasjon = Forventet uforklarlig variasjon + forventet forklarlig variasjon

$$Var(y_p) = Var(\varepsilon) + Var(\hat{y}_i) \quad [55]$$

$$Var(y_p) = (\sigma_\varepsilon)^2 + (\sigma_\varepsilon)^2 \cdot \left(\frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right) \quad [56]$$

$$Var(y_p) = (\sigma_\varepsilon)^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right) \quad [57]$$

Standard usikkerhet for Var($y_p$) er:

$$S_e = \sqrt{Var(y_p)} = \sqrt{(\sigma_\varepsilon)^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}\right)} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}} \quad [58]$$

### *b. Utvidet standard usikkerhet*

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.

$$T_{(\alpha/2,v)} \quad [59]$$

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).

$$v = n - 2 \quad [60]$$

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) [61]

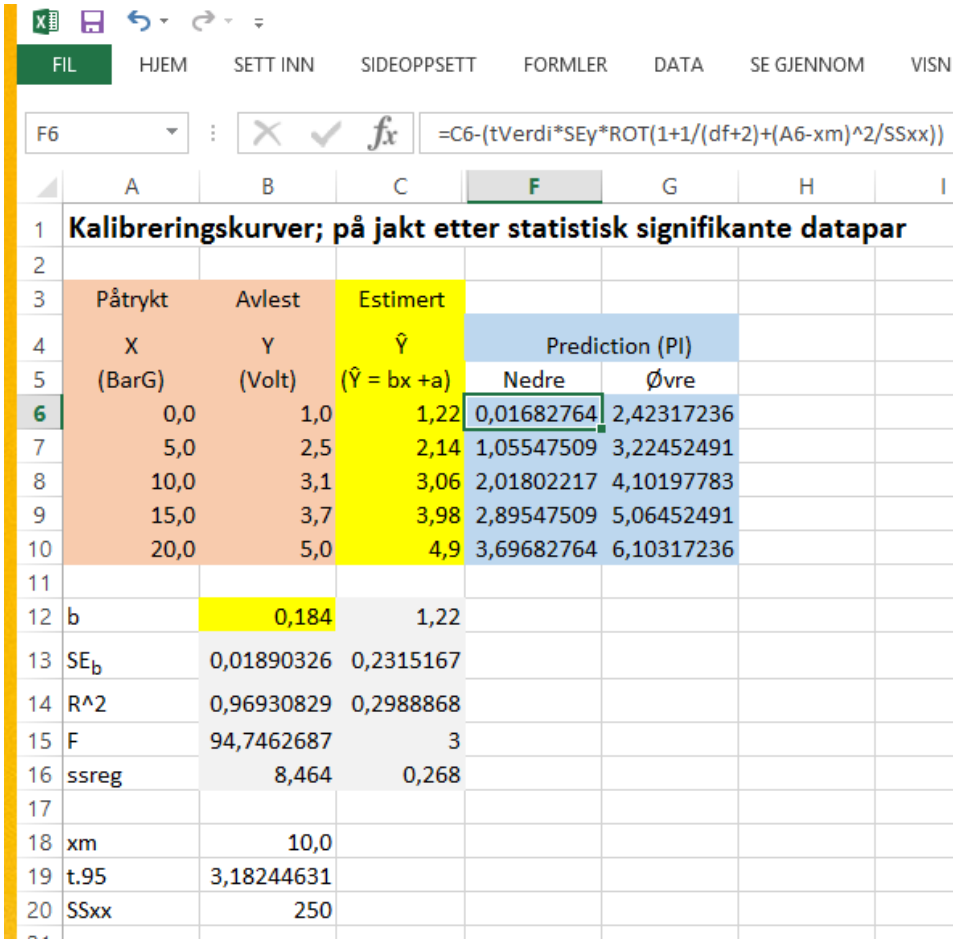
### c. *Øvre og nedre grenseverdi for «Prediction»-intervall*

Setter vi det hele sammen

PI = [«Best Fit»-linje] ± [ usikkerhetsintervall ] [62]

$$PI = [bx + a] \pm \left[ T_{(\alpha/2,v)} \cdot \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{(n-2)}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \right] \quad [63]$$

I Microsoft Excel kan det se slik ut for å beregne $PI_{nedre}$ og $PI_{øvre}$.



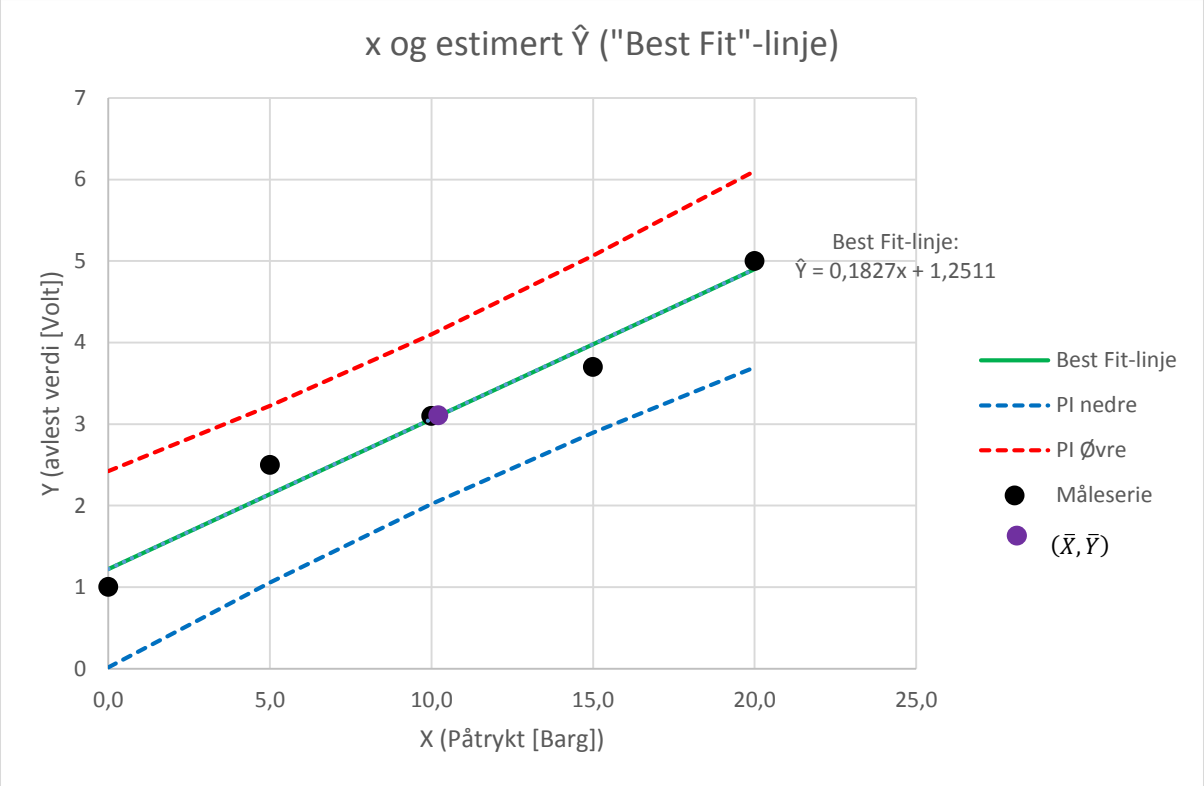
FIL HJEM SETT INN SIDEOPPSETT FORMLER DATA SE GJENNOM VISN

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

| | A | B | C | F | G | H | I |
|---|---|---|---|---|---|---|---|
| 1 | **Kalibreringskurver; på jakt etter statistisk signifikante datapar** | | | | | | |
| 2 | | | | | | | |
| 3 | Påtrykt | Avlest | Estimert | | | | |
| 4 | X | Y | Ŷ | Prediction (PI) | | | |
| 5 | (BarG) | (Volt) | (Ŷ = bx +a) | Nedre | Øvre | | |
| 6 | 0,0 | 1,0 | 1,22 | 0,01682764 | 2,42317236 | | |
| 7 | 5,0 | 2,5 | 2,14 | 1,05547509 | 3,22452491 | | |
| 8 | 10,0 | 3,1 | 3,06 | 2,01802217 | 4,10197783 | | |
| 9 | 15,0 | 3,7 | 3,98 | 2,89547509 | 5,06452491 | | |
| 10 | 20,0 | 5,0 | 4,9 | 3,69682764 | 6,10317236 | | |
| 11 | | | | | | | |
| 12 | b | 0,184 | 1,22 | | | | |
| 13 | $SE_b$ | 0,01890326 | 0,2315167 | | | | |
| 14 | R^2 | 0,96930829 | 0,2988868 | | | | |
| 15 | F | 94,7462687 | 3 | | | | |
| 16 | ssreg | 8,464 | 0,268 | | | | |
| 17 | | | | | | | |
| 18 | xm | 10,0 | | | | | |
| 19 | t.95 | 3,18244631 | | | | | |
| 20 | SSxx | 250 | | | | | |

*Figur 8: Skjermdump fra Microsoft Excel*

Vi kan la Excel tegne opp verdiene for oss (figur 8).




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

Neste datapunkt skal med 95 prosent sannsynlighet ligge innenfor det nevnte «Prediction» intervallet. Det er 5 prosent sjanse for at det neste datapunktet vil ligge utenfor «Prediction» intervallet.

Da det er liten sannsynlighet for at det neste datapunktet skal ligge utenfor «Prediction» intervallet, og vi bør være varsom med å godkjenne/ta i bruk et slikt datapunkt.

Iallfall må vi nøye undersøke hvorfor vårt neste datapar havnet på utsiden av «Prediction» intervallet. Det kan være at dataparet har en eller flere av feil knyttet til seg, noe som gjør dataparet til en ekstremverdi/»uteligger».

Med vennlig hilsen

Trainor Elsikkerhet AS

Rune Øverland
Senioringeniør

Tønsberg april 2015