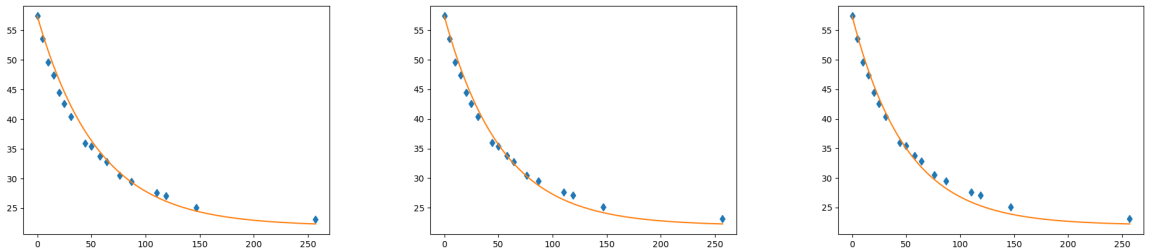
2 - 5 - REGRESJON

På begynnelsen av 1800-tallet oppdaget Guiseppe Piazzi asteroiden Ceres, men han ble syk, og så forsvant Ceres bak solen. Når man skulle lete den opp igjen på andre siden av solen litt senere, var det ikke så godt å vite den nøyaktige posisjonen, men Gauss kastet seg over problemet, og klarte å spå temmelig presist hvor den kom til å dukke opp. Han oppfant i den forbindelse **minste kvadraters metode**.[1] Dette er et av standardverktøyene i kofferten, og vi har allerede hatt bruk for det. Hvis du går helt tilbake til oppgave 15 i økt 1-1 i TMA4101, ser du et plot av noen målte temperaturer og en tilpasset kurve fra Newtons avkjølingslov. Her ser du det samme datasettet (diamanter) og Newtons avkjølingslov (heltrukken) for tre forskjellige α:



Hvilken α er "riktig"? Det er klart at vi ikke kan forvente perfekt match, for Newtons avkjølingslov er en grov modell. Hva som er "riktig" α kommer an på hva vi mener med "riktig", og en definisjon av "riktig", er gitt av minste kvadraters metode.

For å forstå minste kvadraters metode, må vi først forstå funksjoner av to variable. En funksjon av to variable skrives

$$f(\mathbf{x}) \qquad \text{eller} \qquad f(x_1, x_2)\,.$$

og du tenker på $\mathbf{x}$ som en gps-koordinat og på $f(\mathbf{x})$ som den korresponderende høyden over havet. Dette er enklest å komme igang med om du er vant til å lese kart. Et kart er en todimensjonal representasjon av et tredimensjonalt terreng, og med litt trening leser man dette omtrent like godt som om man så terrenget med sine egne øyne, eller faktisk bedre.

[1] https://en.wikipedia.org/wiki/Least_squares

De brune linjene på kartet kalles **ekvidistanselinjer** eller **høydekoter**. Disse forteller noe om høyden over havet; dersom du følger en ekvidistanselinje, går du hverken opp eller ned. Den matematiske ekvivalenten kalles **nivåkurve**. Disse er gitt ved

$$c = f(\mathbf{x}),$$

og forskjellige verdier for c gir forskjellige høyder over $\mathbf{x}$-planet.

1 Skisser nivåkurvene til

$$f(\mathbf{x}) = x_1 + 2x_2$$

Et polynom i to variable er gitt ved

$$p(\mathbf{x}) = \sum_{k,m} a_{km} x_1^k x_2^m$$

Dersom $k + m \leq n$ og $k + m = n$ for minst en kombinasjon av k og m, sier vi at polynomet har orden n. For eksempel er et generelt førsteordens polynom gitt ved

$$p(\mathbf{x}) = a_1 x_1 + a_2 x_2 + a_0.$$

Grafen til denne blir alltid et plan i $\mathbb{R}^3$. Et generelt andreordens polynom er gitt ved

$$p(\mathbf{x}) = a_{20} x_1^2 + a_{11} x_1 x_2 + a_{02} x_2^2 + a_1 x_1 + a_2 x_2 + a_0$$

Nivåkurvene til andreordens polynomer er de berømte kjeglesnittene.[2] Parabelen ($x_2 = x_1^2$) og sirkelen ($\|\mathbf{x}\| = 1$) kjenner du fra før. For en generasjon siden måtte sivingstudenter kunne alle disse på fingrene, men idag fokuserer vi mer på andre ting. Ellipsen [3] med halvakser a og b og sentrum i $\mathbf{y}$ kan være grei å kjenne til, den skal vi få litt bruk for:

$$\frac{(x_1 - y_1)^2}{a^2} + \frac{(x_2 - y_2)^2}{b^2} = 1$$

Finn og skisser nivåkurvene til flatene gitt ved

2 $z = 4x_1^2 + 5x_2^2$.

3 $z = 4x_1^2 + 8x_1 + 5x_2^2 - 10x_2 + 9$.

Et klassisk polynom av to variable er den ideelle gasslov. Du kan tenke på denne enten som en empirisk lov som sakte men sikkert ble oppdaget på 1600-tallet og utover via Boyles lov, Charles lov, Avogadros lov og Guy-Lussacs lov, eller som en utledet regel i statistisk mekanikk.[4] Den sier at i en ideell gass følger trykket p, temperaturen T og volumet V formelen

$$\frac{pV}{T} = nR = Nk,$$

der

$k = 1.380649 \cdot 10^{-23}$ J/K er Boltzmanns konstant[5]

$R = 8.31446261815324$ J/mol K er den ideelle gasskonstant

[2] https://en.wikipedia.org/wiki/Conic_section
[3] https://en.wikipedia.org/wiki/Ellipse
[4] https://en.wikipedia.org/wiki/Ideal_gas_law
[5] https://en.wikipedia.org/wiki/Boltzmann_constant

N er antall gassmolekyler

n er antall mol gassmolekyler
(husk at $N = N_A n$, der $N_A = 6.02214076 \cdot 10^{23}$/mol er Avogadros konstant)

Hvis man liker bedre statistisk mekanikk, kan man skrive

$$\frac{3}{2}\frac{pV}{N} = \frac{3}{2}kT = \frac{1}{2}mv^2$$

der v er forventningsverdien til den χ-fordelte partikkelhastigheten.[6] La oss for enkelhets skyld sette $Nk = 1$ og skrive den ideelle gasslov slik:

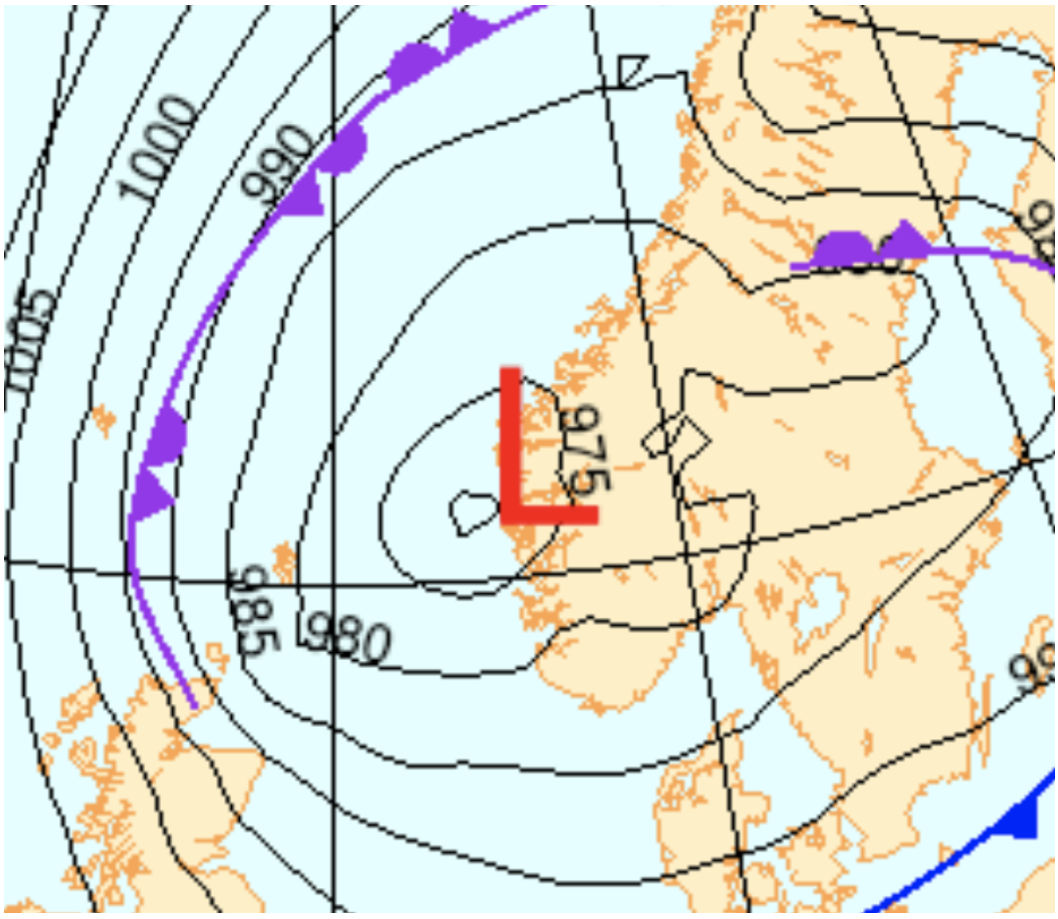
$$T(p, V) = pV$$

Siden både trykk og volum ikke kan være negative størrelser, er det naturlig å tenke at T er en funksjon fra $(0, \infty) \times (0, \infty)$ til $(0, \infty)$. I termodynamikk kalles nivåkurvene til T **isotermer** siden T er konstant på dem.

4 Skisser isotermene til den ideelle gasslov, og plot $T(p, V)$ i python.

Nivåkurvene til trykk kalles **isobarer**.

5 Hva er trykket på Steinkjer denne dagen?[7]



[6] https://en.wikipedia.org/wiki/Maxwell-Boltzmann_distribution
[7] https://www.met.no/vaer-og-klima/meteorologens-prognosekart

Vi har nå to uavhengige variable, og det er interessant å vite hvordan f endres med hensyn på begge. Derfor finnes det to deriverte. De er

$$\frac{\partial f}{\partial x_1} = \lim_{h\to 0} \frac{f(x_1+h, x_2) - f(x_1, x_2)}{h} \qquad \text{og} \qquad \frac{\partial f}{\partial x_2} = \lim_{h\to 0} \frac{f(x_1, x_2+h) - f(x_1, x_2)}{h}$$

og uttales henholdsvis "f derivert med hensyn på x_1 og x_2". Å partiellderivere er enkelt; man finner $\frac{\partial f}{\partial x_1}$ ved å betrakte x_2 som en konstant, og så derivere i vei med hensyn på x_1. Samme for x_2. Finn de partiellderiverte til

6 $f(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^2 + x_1 + x_2.$ 7 $T(p, V) = pV$

De partiellderiverte angir stigningen i koordinatretningene. Står du i et punkt og peker skiene rett øst er stigningen gitt ved $\frac{\partial f}{\partial x_1}$ og peker du dem rett nord er den gitt ved $\frac{\partial f}{\partial x_2}$. Vi setter de partiellderiverte sammen i en radvektor som kalles **gradienten**:

$$f' = \nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)$$

Dersom du peker skiene i en annen retning enn øst eller nord, finner du stigningen ved å prikke gradientvektoren med en enhetsvektor i retningen til skiene; dette kalles **retningsderivert**. Skjønner du retningsderivert og skalarproduktet, bør det være klart at gradienten også angir den bratteste retningen i terrenget, altså den retningen du må peke skiene hvis du skal gå rett oppover.

8 Finn gradientvektorenen til funksjonene i oppgave 6 og 7.

9 Finn stigningen på fjellsiden $f(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^2 + x_1 + x_2$ når du står i punktet $(1, 2)$ og skiene peker rett nordvest.

10 Hvilken vei må du peke skiene dersom du vil gå langs med en ekvidistanselinje?

11 Hva om du vil kjøre rett utfor så bratt som mulig?

Hvis du skjønte alt dette, er det forhåpentligvis innlysende for deg at om du står på toppen av et fjell, er stigningen null uansett hvilken retning du peker skiene. Dersom fjellet er deriverbart, skjer dette kun dersom gradienten er nullvektoren:

$$f' = (0, 0)\,.$$

12 Finn toppunktet til fjellet gitt ved $z = 1 - x_1^2 - x_2^2 - x_1x_2 + x_1 + x_2$.

Det sies ellers at skisportens vugge er et eller annet sted i Telemark. Tull og tøys. Ski er visst en kinesisk oppfinnelse.[8]

[8] https://secretsoftheice.com/news/2018/10/04/skis/

Nå er vi i posisjon til å introdusere minste kvadraters metode. Studass Hausken gikk over en blomstereng en gang. Blomsterengen inneholdt n blomster med koordinater (x_k, y_k) der k løp fra 1 til n, og Hausken syntes det var praktisk å organisere disse koordinatene i to vektorer:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{og} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Blomstene duftet godt, så Hausken ønsket å gå gjennom engen på en rett linje slik at duftopplevelsen ble maksimert. Han tenkte det var best å summere opp kvadratet av den vertikale avstanden fra hver blomst til den rette linjen

$$y = \beta_1 x + \beta_0$$

og så minimere denne summen:

$$\begin{aligned} f(\beta_1, \beta_0) &= \sum_{k=1}^{n} (\beta_1 x_k + \beta_0 - y_k)^2 \\ &= \sum_{k=1}^{n} (\beta_1 x_k + \beta_0 - y_k)(\beta_1 x_k + \beta_0 - y_k) \\ &= \sum_{k=1}^{n} (\beta_1^2 x_k^2 + \beta_0^2 + y_k^2 + 2\beta_1\beta_0 x_k - 2\beta_1 x_k y_k - 2\beta_0 y_k) \\ &= \sum_{k=1}^{n} \beta_1^2 x_k^2 + \sum_{k=1}^{n} \beta_0^2 + \sum_{k=1}^{n} y_k^2 + \sum_{k=1}^{n} 2\beta_1\beta_0 x_k - \sum_{k=1}^{n} 2\beta_1 x_k y_k - \sum_{k=1}^{n} 2\beta_0 y_k \\ &= \beta_1^2 \|\mathbf{x}\|^2 + n\beta_0^2 + \|\mathbf{y}\|^2 + 2\beta_1\beta_0 n\overline{\mathbf{x}} - 2\beta_1 \mathbf{x}^T\mathbf{y} - 2\beta_0 n\overline{\mathbf{y}} \end{aligned}$$

Slik oppfant han den enkleste formen for **regresjon**.[9] Jeg har bevisst tatt med dette slik at du skal forstå at matrisemultiplikasjon vil spare deg for masse kronglete notasjon i TMA4245.

13 Siden f er en positiv kvadratisk funksjon av β_1 og β_0, er rimelig klart at det må finnes et bunnpunkt et eller annet sted. Sett gradienten til f lik null, og utled uttrykkene for **regresjonskoeffisientene**

$$\beta_1 = \frac{\mathbf{x}^T\mathbf{y} - n\bar{\mathbf{x}}\bar{\mathbf{y}}}{\|\mathbf{x}\|^2 - n(\bar{\mathbf{x}})^2} \qquad \text{og} \qquad \beta_0 = \overline{\mathbf{y}} - \beta_1\overline{\mathbf{x}}$$

[9] https://en.wikipedia.org/wiki/Linear_regression

Regresjon betyr å finne en kurve som "passer til" datasettet på en eller annen måte. Det alle lærer først er å finne det første ordens regresjonspolynomet ved minste kvadraters metode, fordi dette er enklest, og fordi man ofte lurer på om det finnes en rettlinjet avhengighet mellom to størrelser. Denne type regresjon er **lineær**, siden uttrykket over er lineært i β_1 og β_0. Variabelen x kalles **forklaringsvariabelen**, og y kalles **responsvariabelen**. Husk at korrelasjon ikke impliserer kausalitet![10]

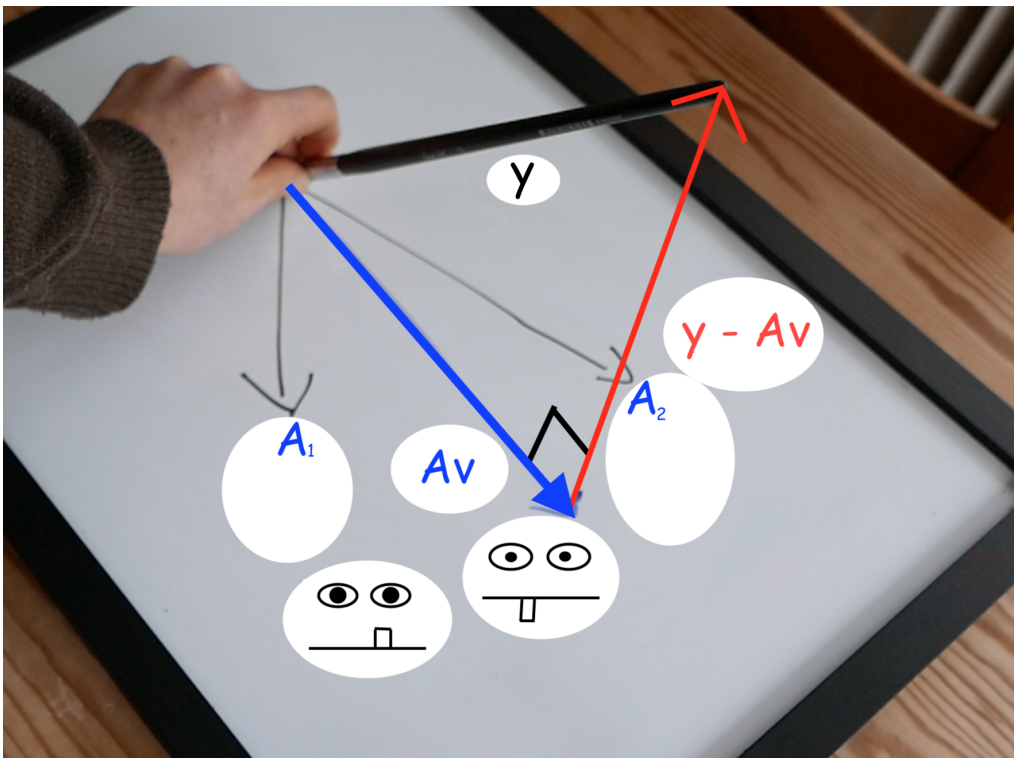
Uttrykket for regresjonen over kan utledes på en annen måte. Du husker kanskje hvordan vi polynominterpolerte i TMA4101. Hvis vi krever at det første ordens polynomet skal reise gjennom alle datapunktene, får vi n likninger:

$$\begin{aligned} y_1 &= \beta_1 x_1 + \beta_0 \\ y_2 &= \beta_1 x_2 + \beta_0 \\ &\vdots \\ y_n &= \beta_1 x_n + \beta_0 \end{aligned}$$

Det sier seg selv at dette ikke har noen løsning med mindre korrelasjonen er perfekt, altså at punktene virkelig ligger på en rett linje. På matriseform blir likningssystemet $A\mathbf{v} = \mathbf{y}$, der

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \qquad \mathbf{v} = \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}$$

14 Gang likningssystemet med A^T fra venstre og løs for $\mathbf{v}$. Hvis du regner riktig, skal du få det samme som i forrige oppgave.
(Se figur for geometrisk tolkning. Kolonnene i A kalles A_1 og A_2.)



[10] https://www.tylervigen.com/spurious-correlations

Hvis du tar et overbestemt likningssystem og ganger med den transponerte av systemmatrisen fra venstre og løser, løser du noe som kalles **normallikningene**. Grunnen er enkel. Systemet

$$A^T A\mathbf{v} = A^T\mathbf{y}$$

kan like gjerne skrives

$$A^T(A\mathbf{v} - \mathbf{y}) = \mathbf{0}$$

og dette forklarer geometrisk hva som skjedde i forrige oppgave.

15 Hvis du skjønte transponeringsoperasjonen og hvordan vi skriver skalarprodukt, bør det nå være klart at likningssystemet over krever at $A\mathbf{v}-\mathbf{y}$ står ortogonalt på alle kolonner i A. Dette betyr at vi velger $\mathbf{v}$ slik at avstanden mellom $A\mathbf{v}$ og $\mathbf{y}$ blir minimert; se figuren på forrige side.

Oppgaven over gjør i bunn og grunn det samme som oppgave 14 i forrige uke. Forskjellen er at kolonnene i A ikke er ortogonale; er de det kan vi regne ut $\mathbf{v}$ slik:

$$\mathbf{v} = \frac{\mathbf{y}^T A_1}{A_1^T A_1} A_1 + \frac{\mathbf{y}^T A_2}{A_2^T A_2} A_2$$

Fordelen med å forstå den geometriske ideen bak normallikningene, er at man nå kan forstå hvordan man kjører regresjon basert på andre ting enn rette linjer. Man setter opp et lineært likningssystem, og så løser man normallikningene. Et kvadratisk regresjonspolynom skrives for eksempel

$$y = \beta_2 x^2 + \beta_1 x + \beta_0.$$

Merk at denne regresjonen også er lineær, siden uttrykket for det kvadratiske regresjonspolynomet er lineært i β_2, β_1 og β_0.

16 Finn det kvadratiske regresjonspolynomet til dette datasettet:

x	y
1	2
2	3
3	4
4	5
5	1

Det går fint å regregere med andre typer funksjoner enn polynomer. Teknikken er det samme, minimer summen av kvadratene av avstandene mellom responsvariabelens datapunktener og regresjonskurven evaluert i de korresponderende datapunktene for forklaringsvariabelen. Jeg var på en hytte i Kragerø en gang og badet i en sirkulær badestamp med et sirkulært høl i bånn. Når jeg skulle tømme den, målte jeg vannstanden etterhvert som stampen tømtes, og fikk tabellen i margen. Hvis man ikke visste bedre, ville det kanskje være plausibelt å anta at utstrømningen er proporsjonal med vannhøyden, siden vanntrykk er proporsjonalt med vanndybde:

t (s)	h (cm)
0	48
10	38
20	28
30	20
40	14
50	8

$$\dot{h} = ah.$$

17 Finn h og estimér a. Hva med integrasjonskonstanten C?

Ifølge fysikkboken er det Torricellis lov som gjelder:

$$h(t) = h_0\left(1 - \frac{t}{T}\right)^2 \qquad T = \frac{V_0}{A}\sqrt{\frac{2}{gh_0}}$$

der g er tyngdeakselerasjonen, V_0 er vannvolumet ved t_0 og A er arealet til tappehullet.[11]

18 Radien til stampen var om lag en meter, men jeg målte aldri tappehullet. Hvor stort var det?

19 Torricellis lov sier jo at h skal være en parabel. Kjør vanlig andreordens polynomregresjon på datasettet (slik som i oppgave 16), og sammenlikne med forrige oppgave.

I eksmeplet denne økten startet med, er temperaturmålingene
np.array([57.4,53.6,49.6,47.4,44.5,42.6,40.4,36.0,35.4,33.8,32.8,30.5,29.5,27.6,27.1,25.1,23.1])
og måletidspunktene (i minutter etter start)
np.array([0,5,10,15,20,25,31,44,50,58,64,76,87,110,119,147,257])

20 Estimer α ved minste kvadraters metode. Temperaturen i omgivelsene var 22 grader.

[11] https://en.wikipedia.org/wiki/Torricelli's_law

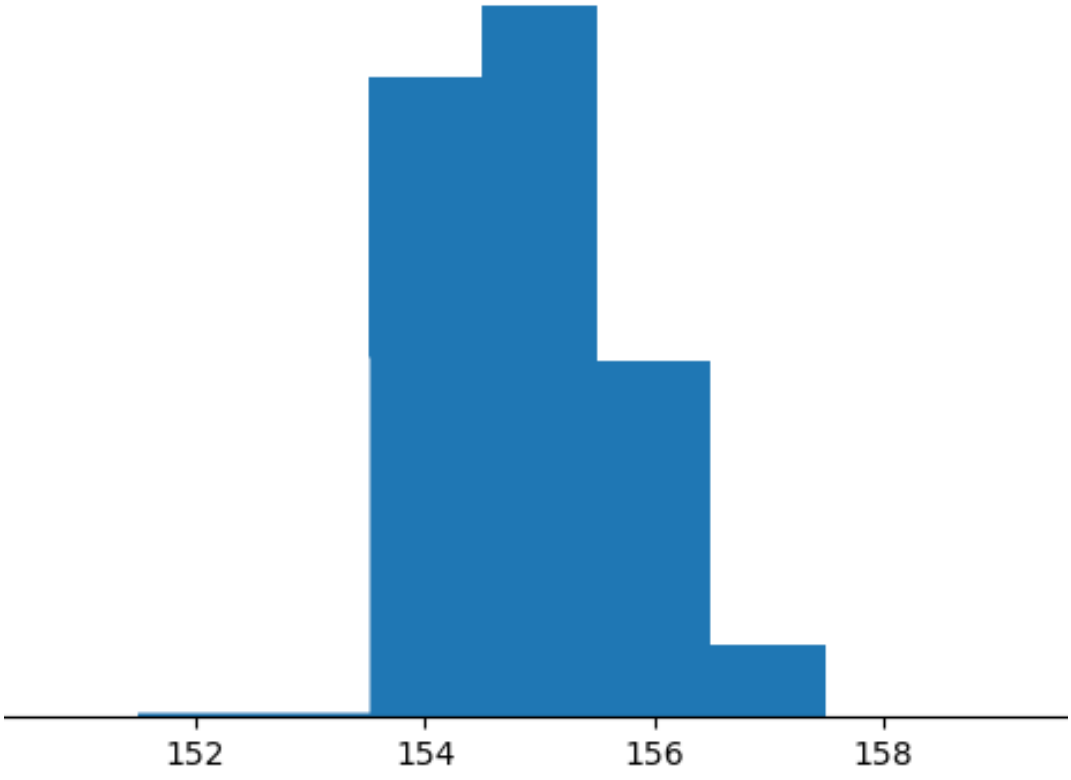
UKENS NØTTER

Forestill deg at du har et måleapparat og at du måler en normalfordelt størrelse, men så vet du at den ene halen i normalfordelingen er under deteksjonsgrensen for måleapparatet.[a] Hvis du nå ønsker å estimere μ og σ, kan du ikke ta gjennomsnitt og empirisk standardavvik, for gjennomsnittet til treffe for høyt og det empiriske standardavviket for lavt.

1 I datasettet over bredden på kaibordene fra integraløkten i TMA4101 har jeg her tatt ut de laveste målingene (tre bord på 153 og et på 152). Finn estimater for μ og σ i normalfordelingen

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

ved å kjøre minste kvadraters metode. Det blir et ikkelineært problem; det enkleste er nok å implementere noe som heter "steepest descent method", eller en av de andre liknende variantene som leter etter minimumspunkter til flervariable funksjoner.
(På springer finner du optimeringsboken til Nocedal og Wright. Den er grei å ha.)



[a] For eksempel ved at du prøver å finne høyden på den gjennomsnittlige nordmann med en målestokk som ikke kan måle kortere lengder enn 170cm.

1	154
2	154
3	
4	156
5	155
6	
7	154
8	155
9	
10	154
11	155
12	155
13	154
14	156
15	155
16	154
17	155
18	154
19	154
20	156
21	155
22	154
23	155
24	157
25	
26	155
27	156
28	155
29	156