Gruvedrift

Notat for TMA4240/TMA4245 Statistikk*

Institutt for matematiske fag,
NTNU

I forbindelse med planlegging av gruvedrift i et område er det mange hensyn som må tas når en skal vurdere om prosjektet er lønnsomt. Blant annet er informasjon om den romlige fordelingen til malmen i berget nødvendig for effektivt å kunne utvinne mineralene. Vi vil her se nærmere på målinger av innholdet av et bestemt oksid i malmprøver hentet fra borehull. Målingene av oksidinnhold kan gjøres på to forskjellige måter:

- Røntgenfluorescensanalyse på et laboratorium (XRF); gir stor nøyaktighet, men er relativt kostbart og tidkrevende.
- Måling på stedet med et bærbart røntgenapparat (XMET); ikke like nøyaktig, men billigere og raskere.

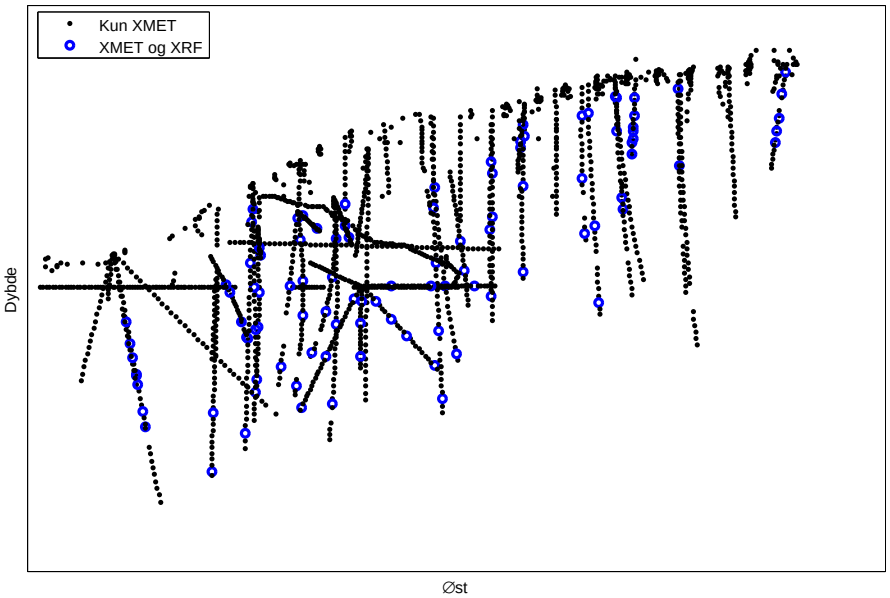
En nærliggende problemstilling er hvor stor unøyaktighet som introduseres i dataene når man bruker XMET i stedet for XRF.

Dataene i `malmdata.txt` består av 1871 observasjoner av oksidinnhold målt med XMET, hvorav 103 observasjoner også er målt med XRF. De syv kolonnene er, fra venstre: Observasjon nr., posisjon (øst), posisjon (nord), dybde, oksidinnhold (XRF), oksidinnhold (XMET) og mineraliseringsklasse (se nedenfor). Figur 1 viser hvor i gruva målingene er hentet fra. De første fem observasjonene er listet i tabell 1.

Tabell 1 – De første fem observasjonene i datasettet. Alle tilhører klasse 1, og det foreligger ikke XRF-målinger for noen av dem (i `.txt`-fila er dette markert ved at XRF-målingen er satt til 300).

Obs. nr.	Øst	Nord	Dybde	XRF	XMET	Klasse
1	−326.2	146.85	44.8	-	1.22	1
2	−323.48	138.37	40.57	-	0.97	1
3	−320.76	129.87	36.35	-	0.98	1
4	−317.93	121.37	32.15	-	0.82	1
5	−315.17	112.87	27.97	-	0.85	1

*Notatet er skrevet av Jacob Skauvold i samarbeid med Jo Eidsvik, og er basert på forskningsrapporten *The value of information in mineral exploration* `http://www.math.ntnu.no/preprint/statistics/2012/S2-2012.pdf` av Jo Eidsvik og Steinar L. Ellefmo. Dersom du finner feil eller har forslag til forbedringer, ta kontakt med Jo Eidsvik, `joeid@math.ntnu.no`



Figur 1 – Posisjonene i gruva hvor målingene er tatt. Den horisontale aksen er østlig lengde, mens den vertikale aksen er dybde. De fleste punktene ligger langs rette linjer som tilsvarer borehull. Observasjonene hvor det bare er gjort XMET-målinger er markert med sorte prikker, mens de hvor det finnes både XMET og XRF-målinger er markert med blå sirkler. Figuren er laget med scriptet i `plotPositions.m`. Koordinatene er tredimensjonale. Scriptet bør derfor kjøres, og plottet roteres, for bedre å gjengi informasjonen.

Støy

Vi ønsker å anslå hvor nøyaktige XMET-målingene er i forhold til XRF-målingene. Betrakt de 103 observasjonene hvor oksidinnholdet er målt på begge måter. La $y_{11}, \ldots, y_{1103}$ være XRF-målingene, og la $y_{21}, \ldots, y_{2103}$ være XMET-målingene. Vi antar at y_{1i}, $i = 1, \ldots, 103$ er helt nøyaktige målinger dvs. den målte verdien er den riktige verdien. Vi antar videre at XMET-målingene er gjenstand for normalfordelt støy,

$$y_{2i} = y_{1i} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \tau^2), \qquad i = 1, \ldots, 103,$$

og vi vil derfor se nærmere på fordelingen til $\epsilon_i = y_{2i} - y_{1i}$, $i = 1, \ldots, 103$. Figur 2 viser hvordan observasjonene av ϵ ligger spredt om null. Figuren viser også et histogram og et Q-Q plott som kan brukes til å vurdere om antakelsen om normalfordeling holder.

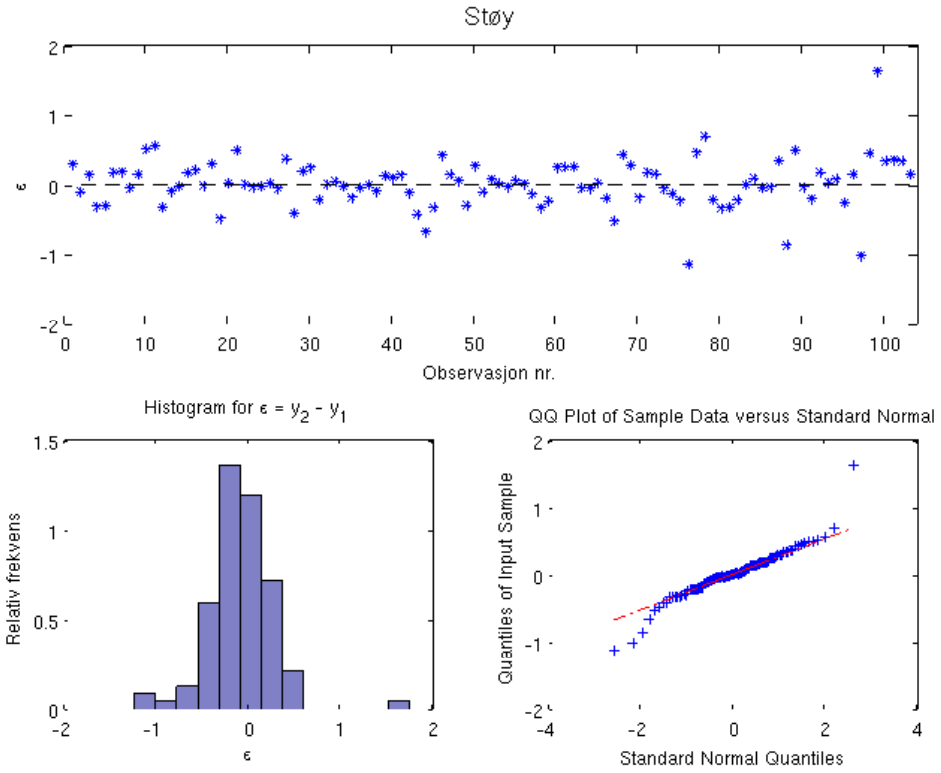
Som estimat for forventningsverdien μ_ϵ til ϵ får vi

$$\bar{\epsilon} = \frac{1}{103} \sum_{i=1}^{103} \epsilon_i = 0.0174.$$

Siden forventningsverdien per antakelse er null, er det to naturlige valg av estimator for variansen (se kap. 4.2 og 8.2 i W.M.M.Y.):

$$s^2 = \frac{1}{102} \sum_{i=1}^{103} (\epsilon_i - \bar{\epsilon})^2 = 0.3461^2 \quad \text{og} \quad \hat{\tau}^2 = \frac{1}{103} \sum_{i=1}^{103} \epsilon_i^2 = 0.3449^2.$$

I dette tilfellet spiller det liten rolle hvilken estimator vi velger. Vi noterer oss at $|\bar{\epsilon}|$ er liten sammenliknet med det estimerte standardavviket.



Figur 2 – *Øverst:* Fordelingen av ϵ_i om null (stiplet linje) for $i = 1, \ldots, 103$, *Venstre:* Histogram for ϵ, *Høyre:* Q-Q plott av ϵ for vurdering av normalitet. Figuren er laget med `noise.m`

Hypotesetest: Er μ_ϵ forskjellig fra null?

Antar at ϵ er normalfordelt med forventningsverdi μ_ϵ og varians σ^2, stiller opp hypotesene

$$H_0 : \mu_\epsilon = 0 \quad \text{og} \quad H_1 : \mu_\epsilon \neq 0,$$

og bruker at under nullhypotesen er observatoren

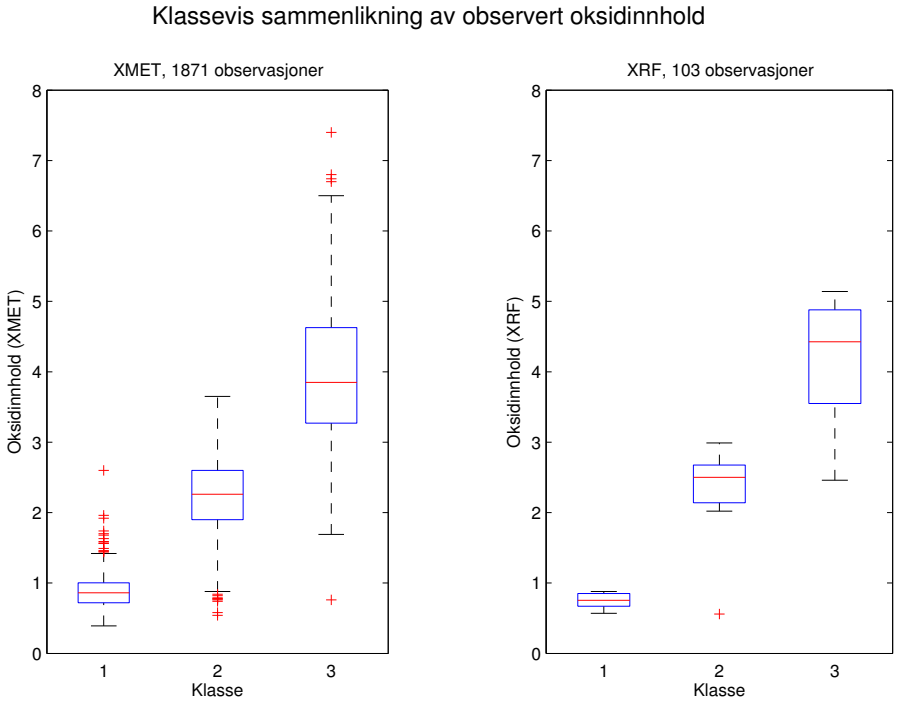
$$T = \frac{\bar{\epsilon}}{\sqrt{s^2/103}}$$

t-fordelt med 102 frihetsgrader (se kap. 10.4 i W.M.M.Y.). Velger signifikansnivå $\alpha = 0.05$, og finner de kritiske verdiene $t_{0.975,102} = -t_{0.025,102} = 1.9835$. Finner så den observerte verdien $T = 0.5095$. Vi konkluderer at μ_ϵ ikke er signifikant forskjellig fra null på signifikansnivå $\alpha = 0.05$.

Merk at for denne utvalgsstørrelsen er t-testen mer eller mindre identisk med en tilsvarende test med normalfordeling.

Klassifisering av malmprøver

Utfra kvalitative geologiske betraktninger kan malmprøvene deles inn i tre klasser; klasse 1, klasse 2 og klasse 3, med økende grad av mineralisering. Typisk kommer de høyeste målingene av oksidinnhold fra klasse 3-prøver. Box plot av oksidinnhold med én boks for hver klasse er vist i figur 3 både for XMET-data og for XRF-data.



Figur 3 – Venstre: XMET-målinger av oksidinnhold fordelt på klasse 1 (249 observasjoner), klasse 2 (479 observasjoner) og klasse 3 (1043 observasjoner). Høyre: XRF-målinger av oksidinnhold fordelt på klasse 1 (6 observasjoner), klasse 2 (11 observasjoner) og klasse 3 (86 observasjoner). Klasse 3 har de høyeste målingene, og størst forventningsverdi, men også større spredning enn klasse 1 og 2. Figuren er laget med `classboxplot.m`.

Estimering av forventingsverdi

La μ_1, μ_2 og μ_3 være forventet oksidinnhold for en malmprøve av henholdsvis klasse 1, 2 og 3. Vi regner ut estimat og konfidensintervall for μ_i, $i = 1, 2, 3$. Som estimator for μ_i bruker vi

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

hvor vi summerer over alle klasse i-observasjonene, $n_1 + n_2 + n_3 = n$, og n er 1871 eller 103 avhengig av om vi ser på XMET- eller XRF-data. For å finne konfidensintervaller antar vi at variabelen

$$T = \frac{\bar{y}_i - \mu_i}{\sqrt{s_i^2/n_i}} \qquad \text{med} \qquad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

er t-fordelt med $n_i - 1$ frihetsgrader (W.M.M.Y. kap. 9.4). Ved å bruke `meanCI.m` finner vi estimatene og 95%-konfidensintervallene i tabell 2.

Lineær regresjon på klasse

Det ser ut til å være en signifikant sammenheng mellom mineraliseringsklasse og oksidinnhold. For å undersøke denne sammenhengen nærmere utfører vi enkel lineær regresjon av

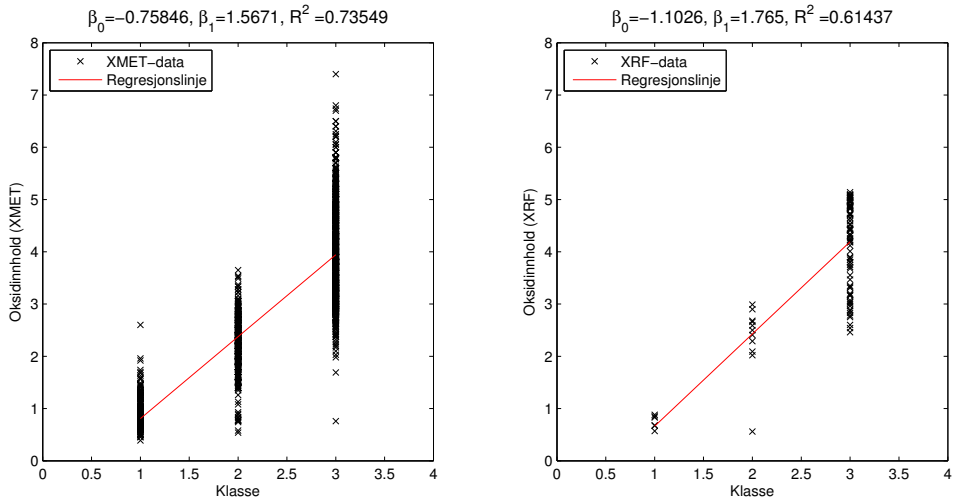
Tabell 2 – Estimat og 95% konfidensintervall for μ_1, μ_2 og μ_3, beregnet med `meanCI.m`.

Datagrunnlag	Forventningsverdi	Estimat	95% konfidensintervall	Antall frihetsgrader
XMET	μ_1	0.9113	(0.8821, 0.9405)	348
	μ_2	2.2263	(2.1795, 2.2731)	478
	μ_3	3.9773	(3.9231, 4.0315)	1042
XRF	μ_1	0.7467	(0.6165, 0.8768)	5
	μ_2	2.3355	(1.8907, 2.7802)	10
	μ_3	4.1983	(4.0277, 4.3688)	85

målt oksidinnhold, y, på mineraliseringsklasse $x \in \{1, 2, 3\}$ og en konstant. Vi skriver altså

$$y = \beta_0 + \beta_1 x + \epsilon', \quad \text{med} \quad \epsilon' \sim N(0, \sigma^2).$$

Vi utfører først regresjonen for alle 1871 observasjonene, og lar y være oksidinnhold målt med XMET. Deretter gjør vi det samme for de 103 XRF-observasjonene, og lar y være oksidinnholdet målt med XRF. Resultatet er vist i figur 4. Siden x bare kan ha heltalls-



Figur 4 – Regresjon av oksidinnhold på mineraliseringsklasse (1, 2, 3), og en konstant. Venstre: data og regresjonslinje for 1871 XMET-målinger. Høyre: Data og regresjonslinje for 103 XRF-målinger. Figuren er laget med `linRegresjon.m`.

verdiene 1, 2 og 3, har ikke regresjonslinja $\hat{y} = \beta_0 + \beta_1 x$ noen tolkning for andre verdier av x. Et plott av linja er likevel tatt med for referanse.

Konfidensintervaller for regresjonskoeffisientene

Vi regner ut 95%-konfidensintervaller for β_0 og β_1. Ifølge kap. 11.5 i W.M.M.Y. har både

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{s^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2}} \quad \text{og} \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2/S_{xx}}}$$

t-fordelinger med $n-2$ frihetsgrader. Her er

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

den forventningsrette estimatoren for variansen, σ^2, til residualene og $\hat{\beta}_i$ er estimatoren for β_i, $i = 0, 1$. Videre er $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, og i dette tilfellet er n enten lik 1871 eller 103 avhengig av hvilke data vi ser på. Bruker scriptet `coeffCI.m` for å regne ut konfidensintervallene for begge parametrene med XMET-dataene og med XRF-dataene. Resultatene er oppsummert i tabell 3.

Tabell 3 – Estimat og 95% konfidensintervall for β_0 og β_1, fra regresjonsmodell.

Datagrunnlag	Koeffisient	Estimat	95% konfidensintervall
XMET	β_0	-0.7585	$(-0.8649, -0.6521)$
	β_1	1.5671	$(1.5245, 1.6098)$
XRF	β_0	-1.1026	$(-1.8833, -0.3219)$
	β_1	1.7650	$(1.4890, 2.0410)$

Regresjon med XRF-dataene gir brattere stigning og lavere skjæringspunkt med y-aksen enn regresjon med XMET-dataene. I tillegg er konfidensintervallene for XRF-dataene bredere, hvilket er rimelig, siden det er langt færre XRF-observasjoner enn XMET-observasjoner.

Forventet respons

Vi antar som før at responsen y har forventningsverdi μ_1, μ_2 og μ_3 for henholdsvis klasse 1, klasse 2 og klasse 3. Regresjonsmodellen gir estimat for disse, nemlig $\hat{y}_i$, $i = 1, 2, 3$. med ugangspunkt i modellen kan vi konstruere konfidensintervaller for disse estimatene også. Tar da utgangspunkt i den tilfeldige variabelen

$$T = \frac{\hat{y_x} - \mu_x}{\sqrt{s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}}$$

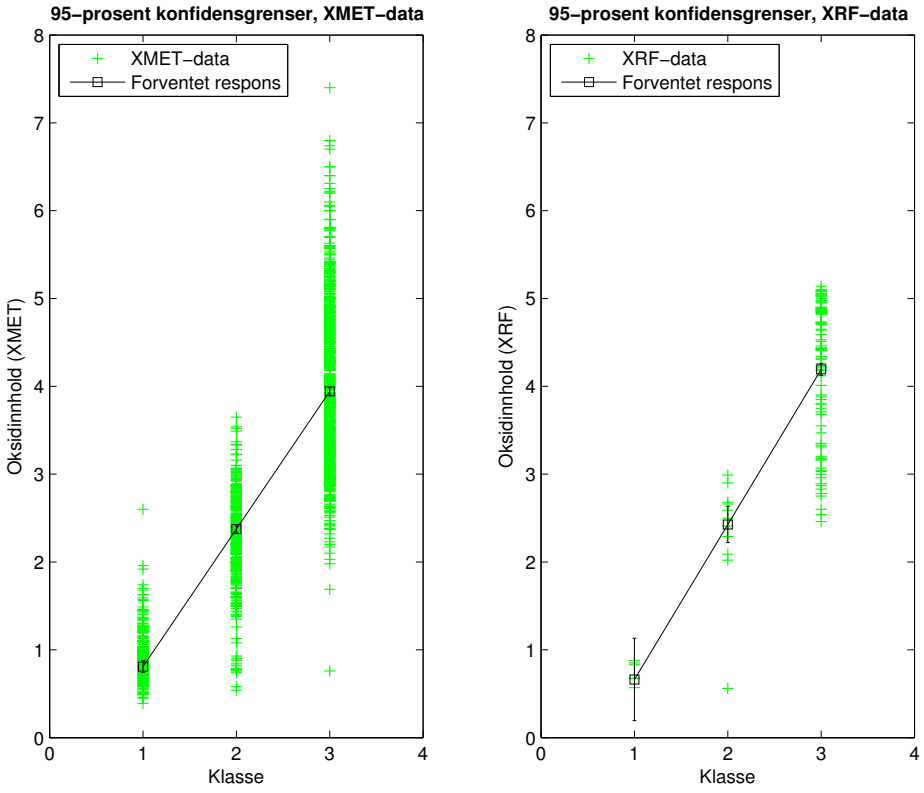
som er t-fordelt med $n-2$ frihetsgrader, ifølge kap. 11.6 i W.M.M.Y. Bruker `responseCI.m` til å regne ut 95%-konfidensintervall. Resultatene er gitt i tabell 4.

Tabell 4 – Estimat og 95% konfidensintervall for μ_1, μ_2 og μ_3, jf. tabell 2.

Datagrunnlag	Forventningsverdi	Estimat	95% konfidensintervall
XMET	$\mu_1 = \hat{y}_1$	0.8087	(0.7415, 0.8759)
	$\mu_2 = \hat{y}_2$	2.3758	(0.7719, 0.8455)
	$\mu_3 = \hat{y}_3$	3.9430	(0.7660, 0.8514)
XRF	$\mu_1 = \hat{y}_1$	0.6624	(0.1943, 1.1305)
	$\mu_2 = \hat{y}_2$	2.4274	(0.4556, 0.8692)
	$\mu_3 = \hat{y}_3$	4.1924	(0.5950, 0.7298)

Konfidensintervallene er illustrert, sammen med dataene, i figur 5. Siden det er en overvekt av klasse 3-observasjoner, får vi smalere konfidensintervaller for klasse 2 og 3 enn for

klasse 1. Dette står i kontrast til figur 3, som viser at klasse 1 har klart minst spredning i dataene. Uoverensstemmelsen kan forklares delvis med at antakelsene for enkel lineær regresjon ikke er oppfylt. Spesielt er det antakelsen om lik varians for alle residualene som ikke holder.

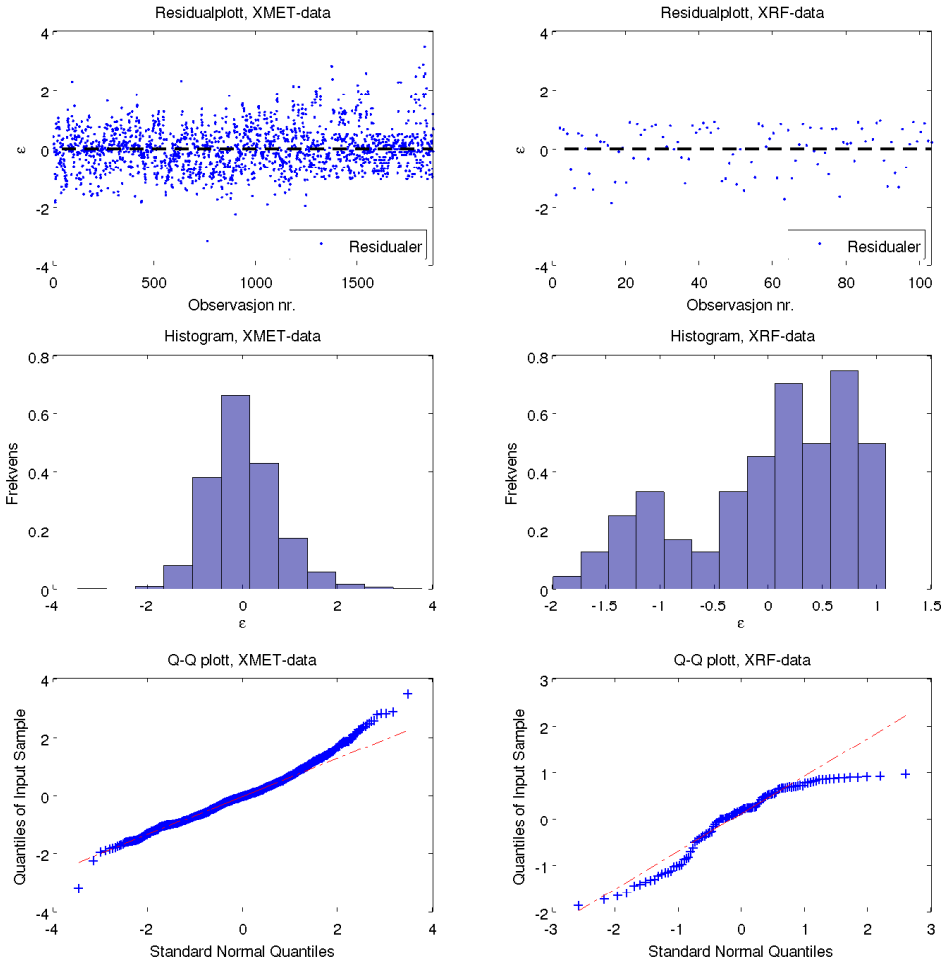


Figur 5 – Forventet respons og konfidensintervaller for klasse 1, 2 og 3. Plottet illustrerer effekten av antall observasjoner på bredden av konfidensintervallene: til venstre ses konfidensintervallene for XMET-data (1871 observasjoner), og til høyre intervallene for XRF-data (103 observasjoner). Konfidensintervallene for XMET-målingene er små og syns derfor dårlig. Figuren er laget med `responseCI.m`.

Vi vil undersøke i hvilken grad residualene $\epsilon' = y - (\beta_0 + \beta_1 x)$ oppfyller *antakelsene for enkel lineær regresjon*:

- Residualene er uavhengige av hverandre.
- Residualene er normalfordelte med forventningsverdi 0 og varians σ^2.
- Variansen σ^2 er den samme for alle residualene.

Vi plotter derfor i figur 6 residualplott, histogram og Q-Q plott for ϵ'. Residualene fra XRF-dataene ser ikke ut til å oppfylle antakelsene om normalfordelte residualer. Videre viser figur 4 tydelig at antakelsen om lik varians for alle residualene ikke holder for noen av dataene.



Figur 6 – Venstre: residualplott(øverst), histogram (midten) og Q-Q plott (nederst) for residualene fra regresjon med XMET-data. Høyre: tilsvarende for residualer etter regresjon med XRF-data. Figuren er laget med `linRegresjon.m`.

Utvidet modell

Siden antakelsen om at residualene har lik varians for alle verdier av x, dvs. for alle tre klassene, byr på problemer, kan vi prøve å unngå den ved i stedet å anta at variansen til residualene avhenger av klasse, men er den samme innenfor hver klasse. La ϵ_{ij} være residual nr. j fra klasse i. Antar da at ϵ_{ij} er normalfordelt med forventningsverdi null og varians σ_i.

$$\epsilon_{ij} \sim N(0, \sigma_i^2), \;\; j = 1, \ldots, n_i \;\; i = 1, 2, 3.$$

Vi vil fortsatt bruke den lineære modellen

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \;\; j = 1, \ldots, n_i \;\; i = 1, 2, 3,$$

men siden vi ikke lenger gjør de nødvendige antakelsene, kan vi ikke bruke minste kvadraters metode for å estimere β_0 og β_1. I stedet kan vi bruke sannsynlighetsmaksimering for å estimere vektoren $\theta = (\beta_0, \beta_1, \sigma_1, \sigma_2, \sigma_3)$, som inneholder de fem parametrene i modellen.

Betrakt den simultane fordelingsfunksjonen til residualene $\epsilon = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{3,n_3})$,

$$f(\epsilon|\theta) = \prod_{j=1}^{n_1} n(\epsilon_{1j}|0, \sigma_1^2) \cdot \prod_{k=1}^{n_2} n(\epsilon_{2k}|0, \sigma_2^2) \cdot \prod_{l=1}^{n_3} n(\epsilon_{3l}|0, \sigma_3^2).$$

Her er $\epsilon_{ij} = y_{ij} - (\beta_0 + \beta_1 x_{ij})$, slik at $f(\epsilon|\theta)$ egentlig er en funksjon av β_0, β_1, σ_i, x_{ij} og y_{ij} for $j = 1, \ldots, n_i$, $i = 1, 2, 3$. Bytter vi plass på argumentene og parametrene, får vi likelihood-funksjonen $\mathcal{L}(\theta|\epsilon) = f(\epsilon|\theta)$, dvs. $\mathcal{L}$ har samme form som simultanfordelingen, men nå er det θ som er den uavhengige variabelen, og ϵ som er parameter. La nå $\tilde{\ell} = -2 \ln \mathcal{L}$, slik at vi får

$$\tilde{\ell}(\theta|\epsilon) = n_1 \ln(2\pi\sigma_1^2) + n_2 \ln(2\pi\sigma_2^2) + n_3 \ln \sigma_3^2 + \sum_{j=1}^{n_1} \left(\frac{\epsilon_{1j}}{\sigma_1}\right)^2 + \sum_{k=1}^{n_2} \left(\frac{\epsilon_{2k}}{\sigma_2}\right)^2 + \sum_{l=1}^{n_3} \left(\frac{\epsilon_{3l}}{\sigma_3}\right)^2.$$

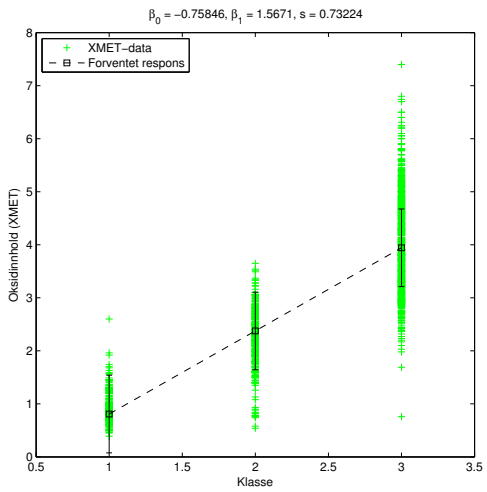
Sannsynlighetsmaksimeringsestimatet for θ er den verdien som maksimerer $\mathcal{L}(\theta|\epsilon)$ med de gitte dataene x og y. Dette er, siden $\tilde{\ell} = -2 \ln \mathcal{L}$, den samme verdien av θ som minimerer $\tilde{\ell}(\theta|\epsilon)$, altså

$$\arg\max_{\theta} \mathcal{L}(\theta|\epsilon) = \arg\min_{\theta} \tilde{\ell}(\theta|\epsilon).$$

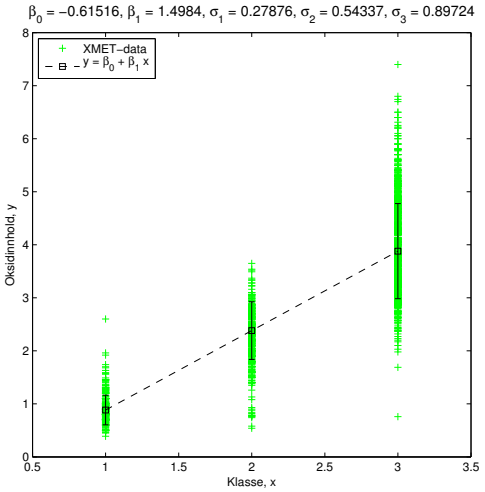
Denne verdien kan finnes numerisk ved f.eks. å bruke funksjonen `fminunc` fra Matlab Optimization Toolbox, slik det er gjort i `likelihood.m`. Når scriptet kjøres med XMET-dataene fås estimatene i tabell 5. I figur 7b er forventet respons og estimerte standardavvik illustrert med errorbars. Sammenliknet med figur 7a, som viser tilsvarende for den opprinnelige modellen, er forholdet mellom usikkerhetene for klasse 1, 2 og 3 her mer i tråd med tendensen en ser i figur 3.

Tabell 5 – Sannsynlighetsmaksimeringsestimater (SME) for parametrene i den utvidete modellen, beregnet numerisk med `likelihood.m`. Sammenliknet med verdiene for XMET-dataene i tabell 3 har β_0 blitt større, mens β_1 har blitt noe mindre.

Parameter	SME
β_0	-0.6152
β_1	1.4984
σ_1	0.2788
σ_2	0.5434
σ_3	0.8972



(a) Enkel lineær regresjonsmodell, β_0 og β_1 estimert med minste kvadraters metode, s^2 er den forventningsrette estimatoren for variansen til residualene. Figuren er laget med `responseCI.m`.



(b) Alternativ modell, β_0, β_1, σ_1, σ_2 og σ_3 estimert med sannsynlighetsmaksimering. Figuren er laget med `likelihood.m`.

Figur 7 – Forventet respons og standardavvik beregnet utfra den enkle lineære regresjonsmodellen, og den alternative modellen med ulik varians.