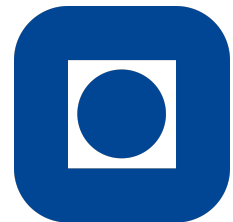


ISTx1003: Statistisk læring og data science

Zoom-forelesning 1

Mette Langaas, IMF/NTNU. 21.10.2020



NTNU

Praktisk om denne zoom-forelesningen

- Vanlig zoom-møte - ikke webinar
- Husk å ikke ha på mikrofonen (ha på mute)
- Du velger selv om du vil ha på kamera
- Forelesningen blir tatt opp og lenke blir lagt ut på Blackboard
- I dag: still spørsmål i chaten til alle (ikke bare til foreleser) - fordi studentassistent Simen hjelper til bare i dag
- Du kan stille spørsmål muntlig - hvis foreleser spør om noe er uklart, hvis du har noe å spørre om mens foreleser snakker må du si ifra i chaten først.
- Du kan selvfølgelig rekke opp handa, men det er det vanskelig å følge med på - bruk heller chat!

Plan

1. Hvem er vi?
2. Statistisk læring og data science
3. De tre temaene i modulen: regresjon, klassifikasjon og klyngeanalyse
4. Læringsmål
5. Læringsressurser og pensum
6. Prosjektoppgaven og Blackboard-informasjon
7. Ukens tema: regresjon - med enkel lineær regresjon
8. Forberedelser til neste uke

Avbrutt av en pause.

Hvem er vi?

Studentene:

- BIDATA i Trondheim, Gjøvik og Ålesund (totalt 163 studenter)

Fra nettsiden til studiet:

Dataingeniørstudiet er for deg som vil spesialisere deg i programmering. I tillegg vil du lære å bruke effektive utviklingsverktøy og metoder for å jobbe i team i større utviklingsprosjekter. Med en dataingeniørutdanning vil du være med på å forme fremtiden og digitalisere alle bransjer for alle formål. Den raske teknologiutviklingen medfører hele tiden nye og spennende produkter og tjenester. Som dataingeniør er du med på innovasjon og nyskaping hele tiden.

Vil du synes denne modulen er spennende og gøy? Kan denne modulen hjelpe deg til å bli en bedre dataingeniør?

Hvem er vi?

Fagteamet:

- Faglig ansvarlig for innholdet i modulen er Mette Langaas mette.langaas@ntnu.no.

Og i **veilederteamet** (for prosjektet) inngår i tillegg

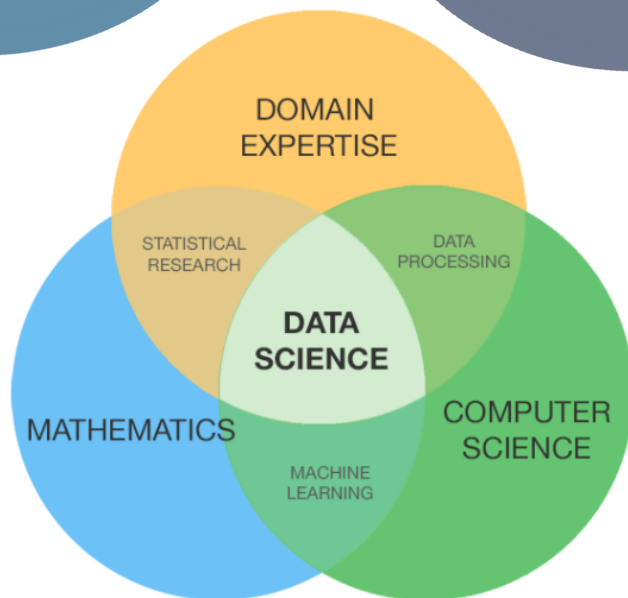
- Trondheim: studentassistentene Simen og Thomas, og øvingslærer Martin O. Berild
- Gjøvik: Charles Curry og Bernt Tore Jensen
- Ålesund: Siebe B. van Albada.

Statistisk læring og data science

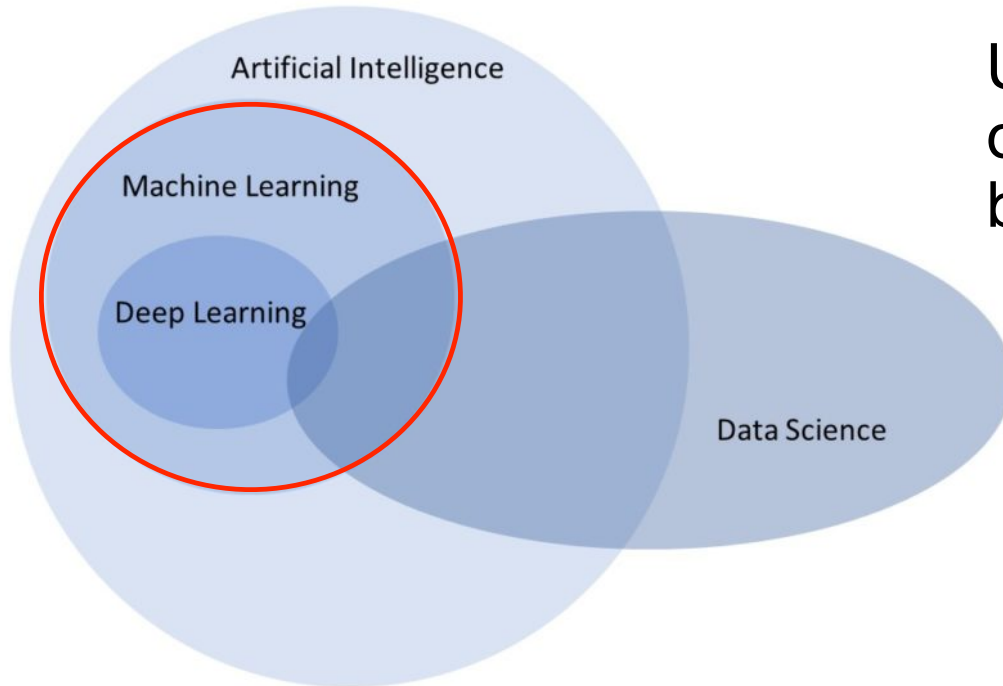
**STATISTISK
LÆRING**

**DATA
SCIENCE**

regresjon
klassifikasjon
klyngeanalyse



- Modell
- Metode
- **Algoritme**
- Resultat



Ustrukturerte
data: tekst og
bilder

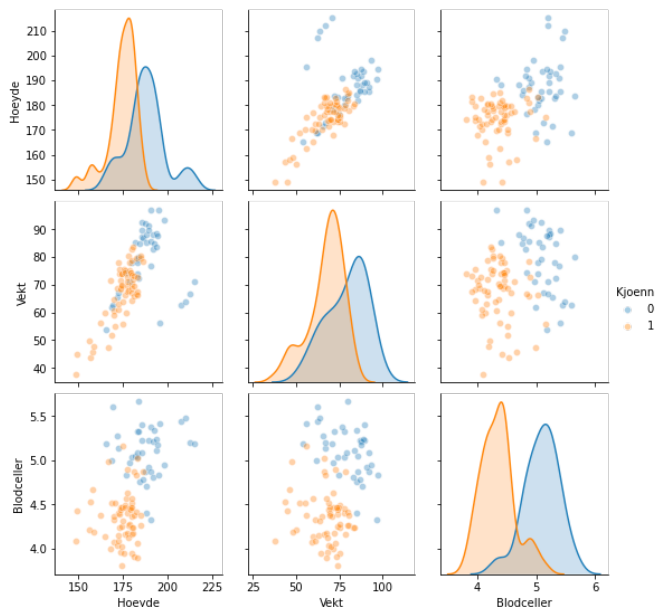
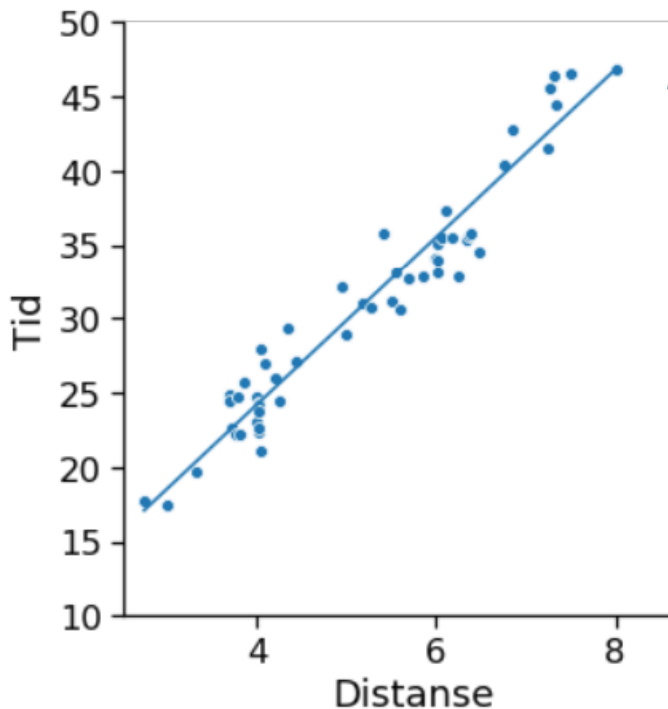
Regresjon

Veiledet læring - vi kjenner responsen

Fellesmodulen



Løpe: sammenheng
mellom distanse og tid



Prosjektet:

Hva er sammenhengen mellom
antall røde blodceller pr liter blod
- og høyde, vekt, kjønn og
idrettsgren - til idrettsutøvere?

Klassifikasjon

Veiledet læring - vi kjenner svaret!

Hvilke tall er håndskrevet på brevet?

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9



Prosjektet:

Hvem vinner tennismatchen?
Data på serve-
ess, dobbeltfeil
og upressede feil,
og hvem som
vant!

treningssett

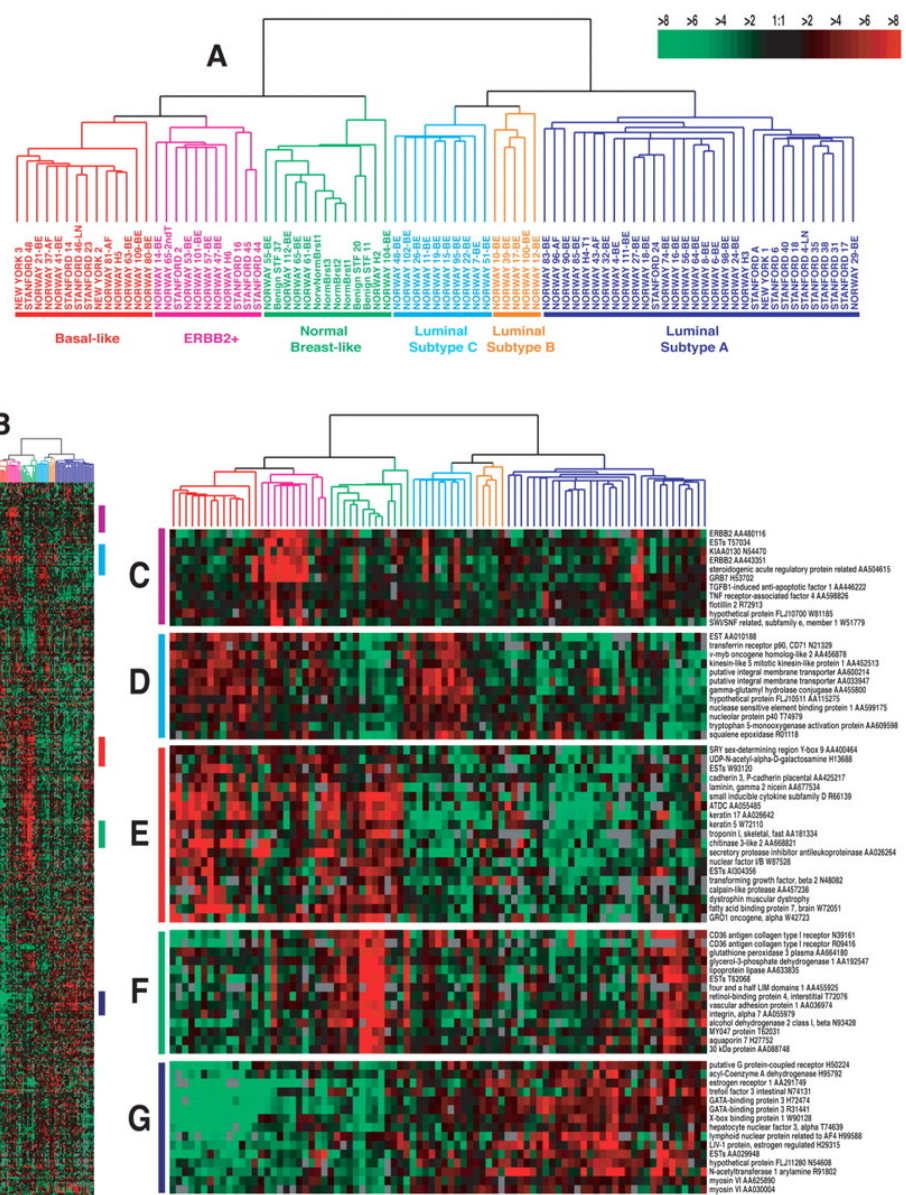
valideringssett

testsett

Klyngeanalyse

Ikke-veiledet læring ingen fasit kjent!

- Finn ukjente klynger i data.
- Observasjoner innen hver klynge er mer lik hverandre enn observasjoner fra ulike klynger

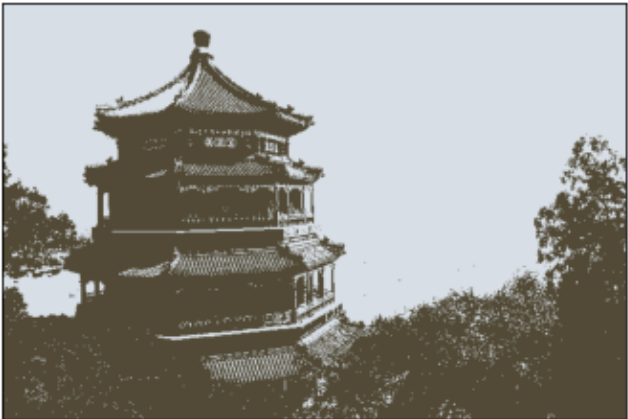


Klyngeanalyse

china.jpg

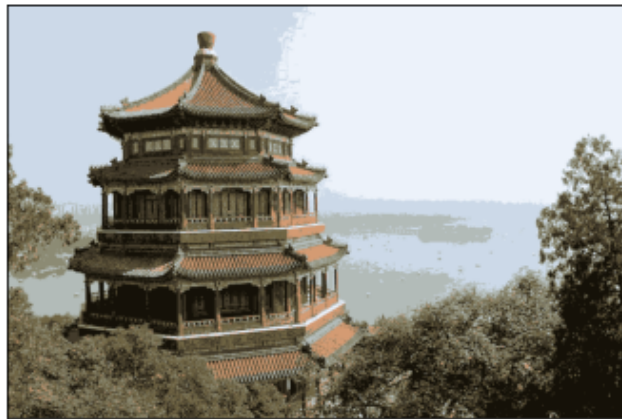


Bilde etter 2-gjennomsnitt klyngeanalyse



Prosjektet:

Kan vi bruke klyngeanalyse til å fjerne detaljer og støy - ved å dele pikslene inn i to eller flere klynger?



Plan

1. Hvem er vi?
2. Statistisk læring og data science
3. De tre temaene i modulen: regresjon, klassifikasjon og klyngeanalyse
4. Læringsmål
5. Læringsressurser og pensum
6. Prosjektoppgaven og Blackboard-informasjon
7. Ukens tema: regresjon - med enkel lineær regresjon
8. Forberedelser til neste uke

Avbrutt av en pause.

Læringsmål

Etter du har gjennomført denne modulen skal du kunne:

- forstå når du kan bruke regresjon, klassifikasjon og klyngedianalyse til å løse et problem
- kunne gjennomføre multivariat regresjon på et datasett
- bruke logistisk regresjon til å finne nærmeste nabo for å utføre en klassifikasjon
- bruke hierarkisk k-means klyngedianalyse på et datasett for å finne egnet avstandsmål
- kunne kommunisere resultatene fra regresjon/klassifikasjon/klyngedianalyse til medstudenter og ingeniører
- bli en kritisk leser av resultater fra statistikk/maskinlæring/statistisk læring/data science/kunstig intelligens når disse rapporteres i media, og forstå om resultatene er realistiske ut fra informasjonen som gis
- **kunne besvare prosjektoppgaven på en god måte!**

Pensum og læringsressurser

Pensum er definert som

**"svarene på det du blir spurt om på
prosjektoppgaven"**

og de kan du finne ved å bruke læringsressursene.

Det blir ikke spørsmål fra prosjektmodulen på den skriftlige eksamenen, men det du lærer i modulen kan gjøre at du får en god forståelse av temaene i fellesmodulen - spesielt inferensdelen av emnet (estimering, hypotesetesting, enkel lineær regresjon).

Pensum og læringsressurser

Læringsressurser

- tre kompendier,
- tilhørende videoer, og
- zoomforelesningene.

Læringsressursene legges ut under den relevante uken på Bb (men de to første finnes også utenfor Bb)

Snarvei til Bb 1003: <https://s.ntnu.no/ist1003>

Prosjektoppgaven

Vi ser hvor informasjonen ligger på Blackboard og hvordan melde seg på gruppe.

Snarvei til Bb 1003: <https://s.ntnu.no/ist1003>

Vi ser på prosjektoppgaven på <https://s.ntnu.no/isthub>

- Oppgave 1: Regresjon. 8 spørsmåls punkt.
- Oppgave 2: Klassifikasjon. 6 spørsmåls punkt.
- Oppgave 3: Klyngeanalyse. 6 spørsmåls punkt

Hvordan evalueres prosjektoppgaven?

- For hvert spørsmåls punkt: maksimalt 1 poeng
- NTNUs karaktergrenser planlegges brukt:
 - A: 17.8 - 20 poeng
 - B: 15.4 - 17.8 poeng
 - C: 13 - 15.4 poeng
 - D: 10.6 - 13 poeng
 - E: 8.2 - 10.6 poeng
 - F: under 8.2 poeng
- Du må ha minst E for å bestå, karakter oppgis i sammen med slutt karakter
- Endelig karakter i emnet:
 $0.2 \cdot \text{prosjekt} + 0.8 \cdot \text{skriftlig eksamen}$

Jobbe med prosjektoppgaven

Uke- tema- oppgave på prosjektet

Uke 43: Enkel lineær regresjon:

Oppgave 1: Q1.1-Q1.4

Uke 44: Multippel lineær regresjon:

Oppgave 1: Q1.5-Q1.8

Uke 45: Klassifikasjon

Oppgave 2

Uke 46: Klyngeanalyse

Oppgave 3

Tema: Regresjon

PAUSE TIL

9.15

I dag (denne uka): enkel lineær regresjon

- prosjektet oppg 1: Q1.1-Q1.4

Neste uke: multippel lineær regresjon

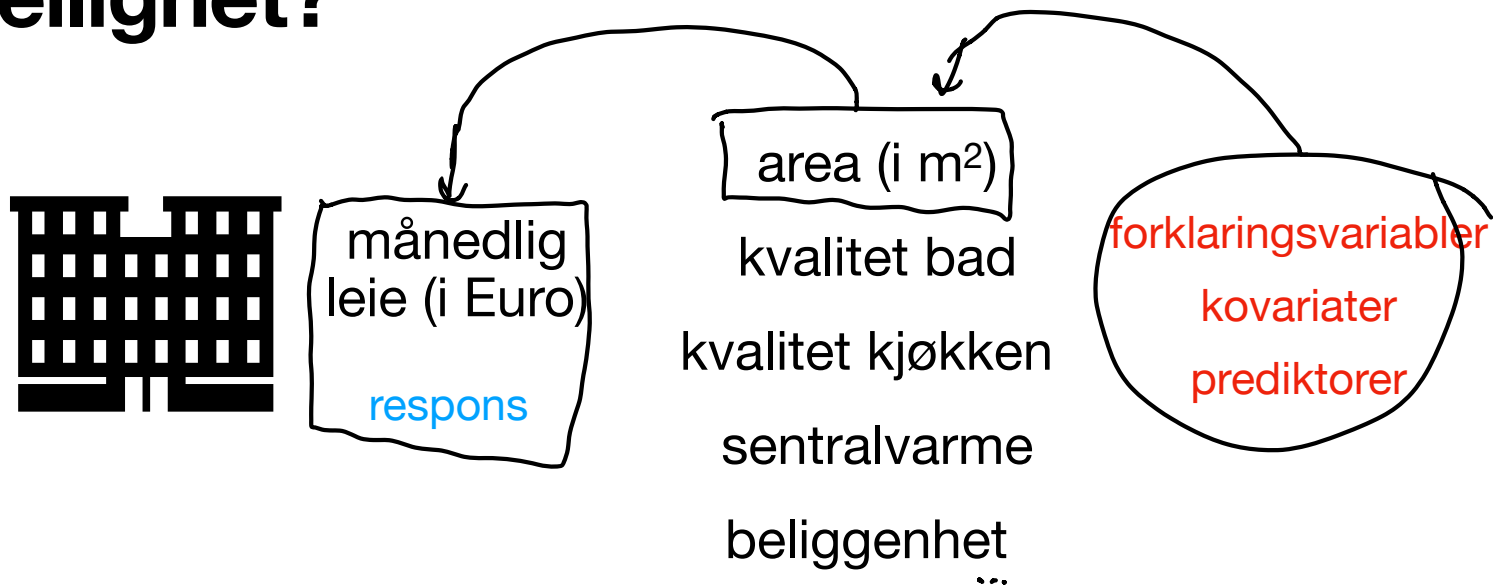
- prosjektet oppg 1: Q1.5-Q1.8

Målet med regresjon:

1. Vi vil lage en modell for å hjelpe oss med å **forstå sammenhengen** mellom *en respons* og *en eller flere forklaringsvariabler*.
2. Vi vil lage en modell for å **predikere en respons** fra *en eller flere forklaringsvariabler* (mer eller mindre sort boks).

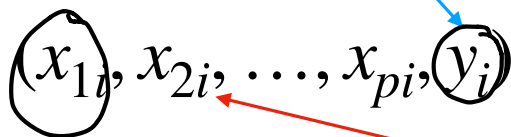
MEG! → Mette.hangaas@nfm.no

Hvor mye bør det koste å leie en leilighet?



kontinuerlig variabel

respons



forklaringsvariabler - kontinuerlige eller kategoriske

Mietspiegel für München 2019



Wohnfläche



Baujahr



Wohnlage



Gebäudetypen



Haustypen



Warmwasser

Berechnungsergebnisse

Mietspiegel für München 2019

Ihre Angaben:

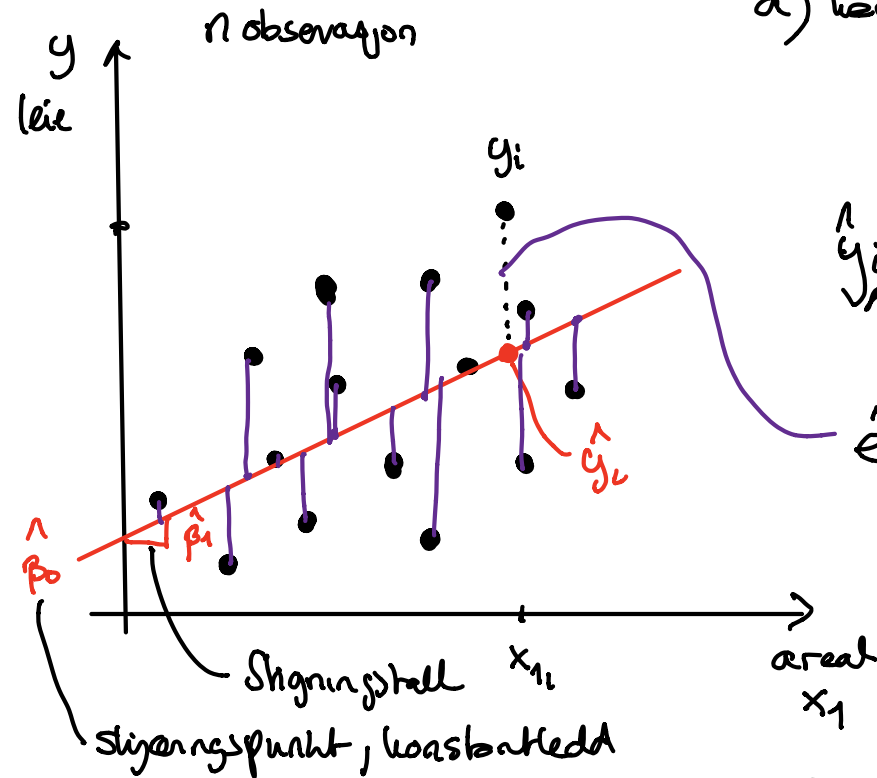
- Wohnfläche: 45 m²
- Baujahr: 2009-2012
- Wohnlage: Zentrale gute Lage
- Gebäudetypen: Wohnblock
- Haustypen: Anderer Haustyp
- Warmwasserversorgung:
Warmwasserversorgung ist in Küche und
Bad vorhanden.
- Heizung: Beheizungsmöglichkeit in allen
Wohnräumen vorhanden
- Fußbodenheizung: Ja
- Thermostatventile: Thermostatventile
und/oder Fußbodenheizung vorhanden
- Sanitärbereich: Besondere
Zusatzausstattung im Bad vorhanden
- Offene Küche: Nein

Berechnungsergebnisse:

- Durchschnittliche ortsübliche Miete:
13,76 Euro/m²/Monat
619,20 Euro/Monat
- Durchschnittliche ortsübliche Miete mit Spanne
nach unten:
12,08 Euro/m²/Monat
543,60 Euro/Monat
- Durchschnittliche ortsübliche Miete mit Spanne
nach oben:
15,99 Euro/m²/Monat
719,55 Euro/Monat
- Durchschnittliche ortsübliche Miete mit
begründeten Abweichungen:
13,76 Euro/m²/Monat
619,20 Euro/Monat

Enkel lineær regresjon

a) kan vi tilpasse en rett linje til dataene?



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

predikert respons i dataene våre

$$\hat{e}_i = y_i - \hat{y}_i \text{ residual}$$

$\hat{\beta}_0$ og $\hat{\beta}_1$ velges slik at

$$SSE = \sum_{i=1}^n \hat{e}_i^2 \text{ minimeres}$$

b) Kan vi tolke linja? Hvor sikker er jeg på $\hat{\beta}_1$ og linja?

↑
KFI
konfidensintervall

↖
Hypotesetest

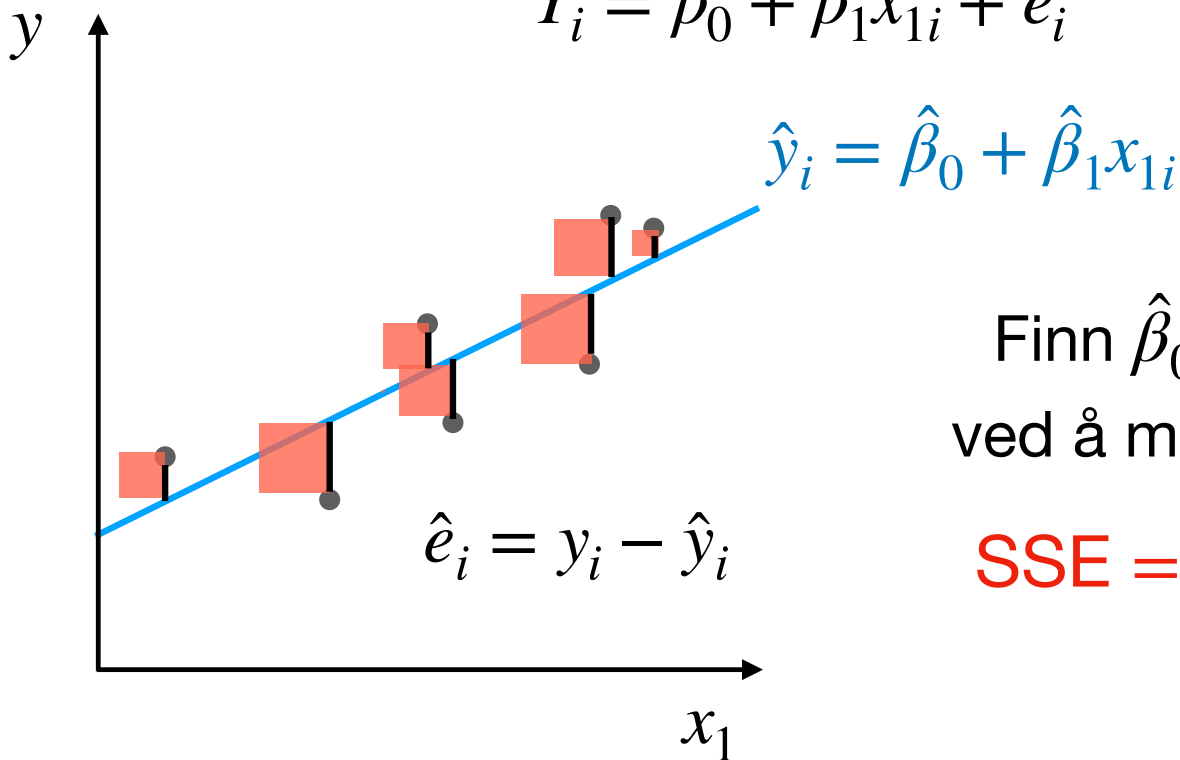
c) Fremtidige prediksjoner av $\hat{y}$

TRENGER
ANTAGELSER

Parameterestimering

Enkel lineær regresjon

$$Y_i = \beta_0 + \beta_1 x_{1i} + e_i$$



Finn $\hat{\beta}_0$ og $\hat{\beta}_1$
ved å minimere

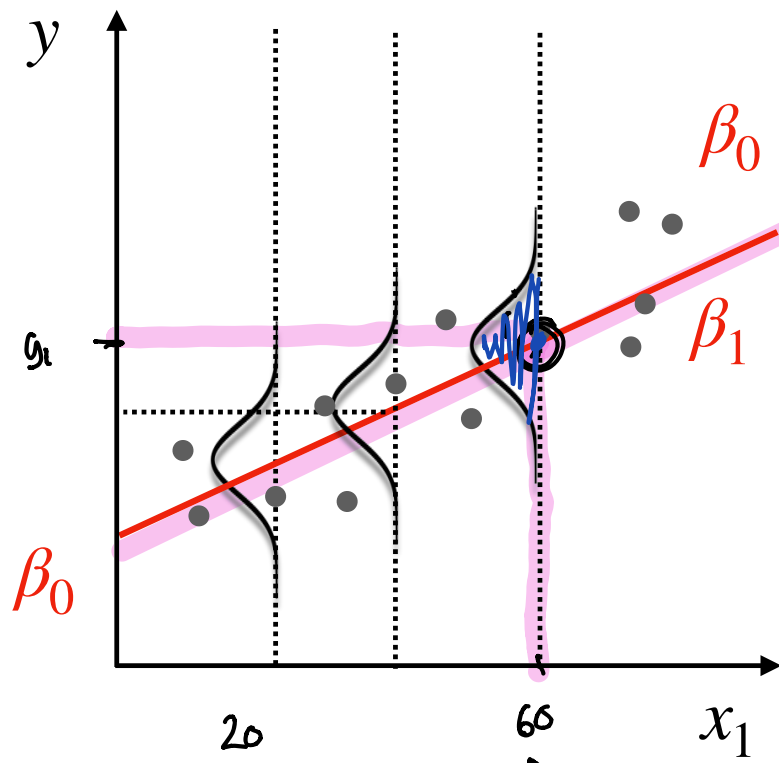
$$\text{SSE} = \sum_{i=1}^n \hat{e}_i^2$$

Enkel lineær regresjonsmodell

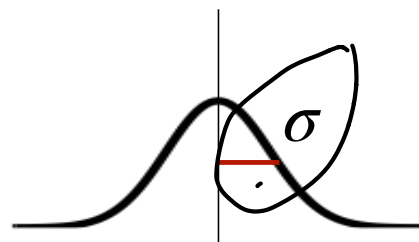
uavhengige
observasjonspar

$$(x_{1i}, y_i)$$

$$i = 1, \dots, n$$



$$Y_i = \boxed{\beta_0 + \beta_1 x_{1i}} + \underbrace{e_i}_{\text{residual}}$$



$$\underbrace{e_i}_{\text{residual}} \sim N(0, \sigma)$$

$$\hat{e}_i = y_i - \hat{y}_i \quad \boxed{\text{residual}}$$

Regresjonsanalyse i fem steg

Vi skal bruke `statmodels.api` og `statmodels.formula.api` for lineær regresjon

Steg 1: Bli kjent med dataene ved å se på oppsummeringsmål og ulike typer plott

Steg 2: Spesifiser en matematisk modell

Steg 3: Initialiser og tilpass modellen

Steg 4: Presenter resultater fra den tilpassede modellen

Steg 5: Evaluer om modellen passer til dataene

Steg 1: Bli kjent med dataene

X_1 : plote : histogram, bokuplott

y : 

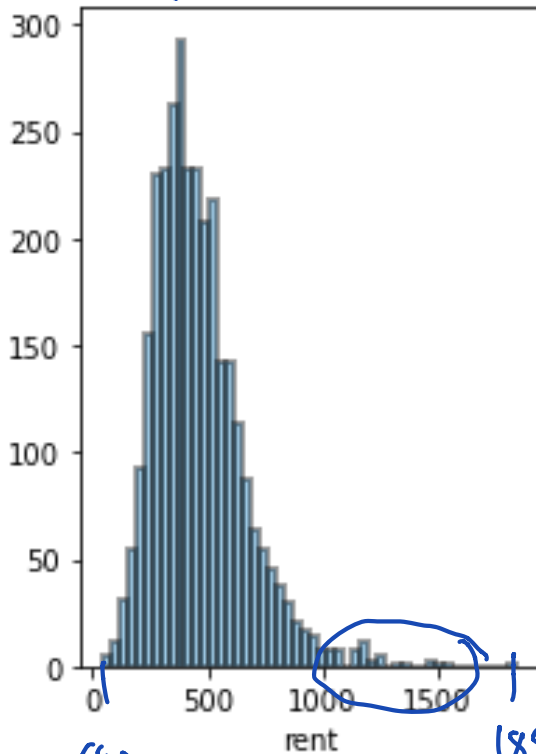
rekkevidde : min, max
 $\bar{x}$, sd

kryssplott X_1 mot $y \leftarrow$ er det lineær sammenheng

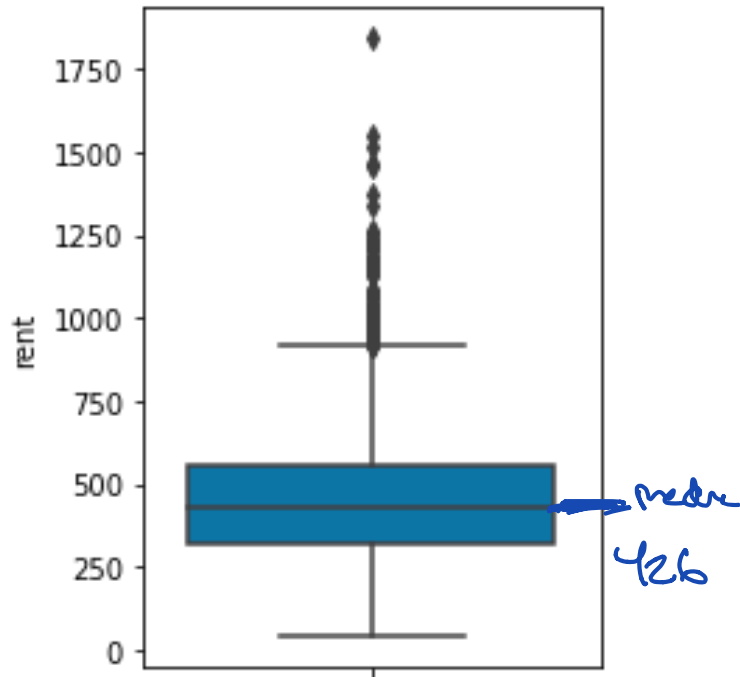
kan nye forklares?

Responsen : leie (rent)

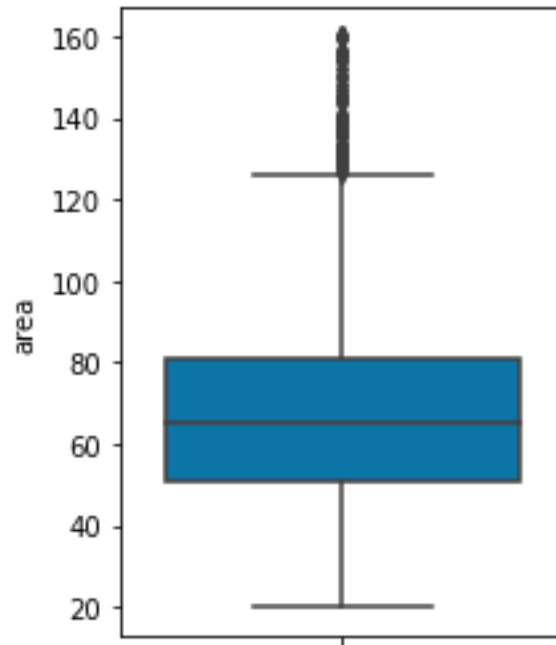
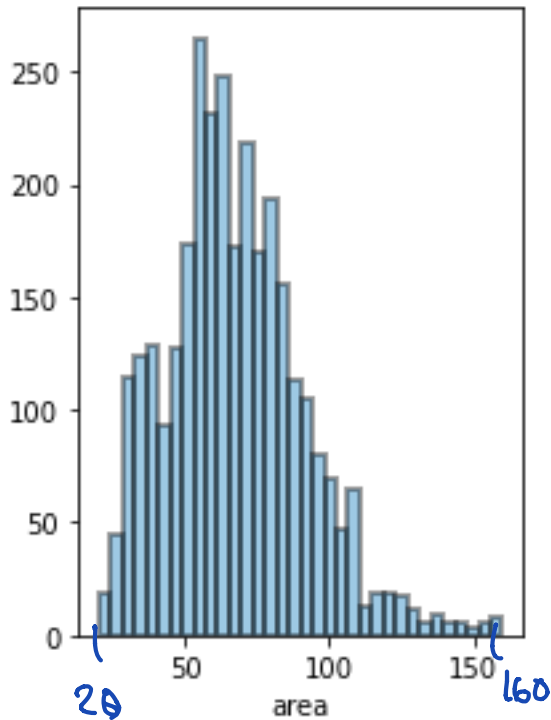
histogram



boxplott

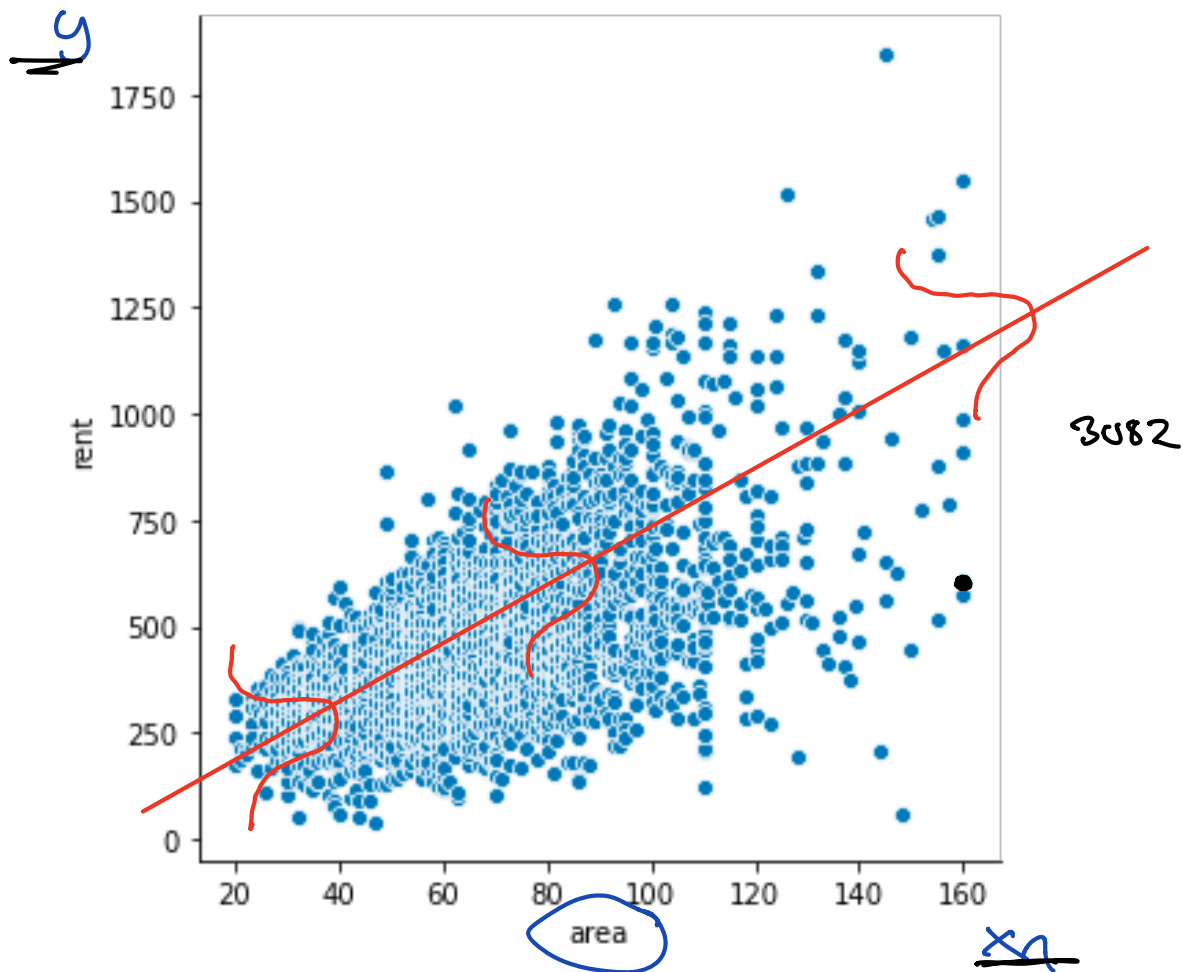


Kovariaten

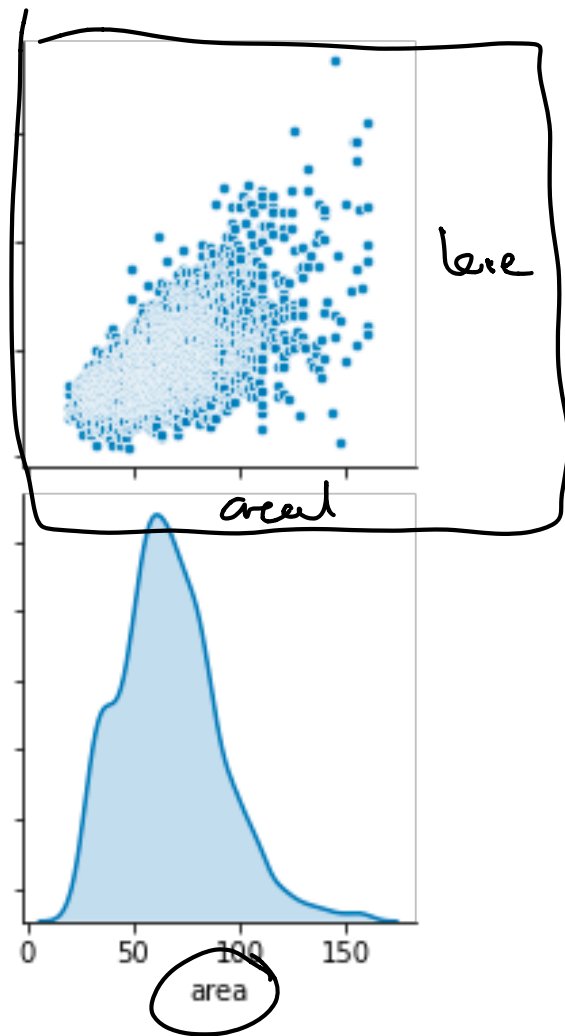
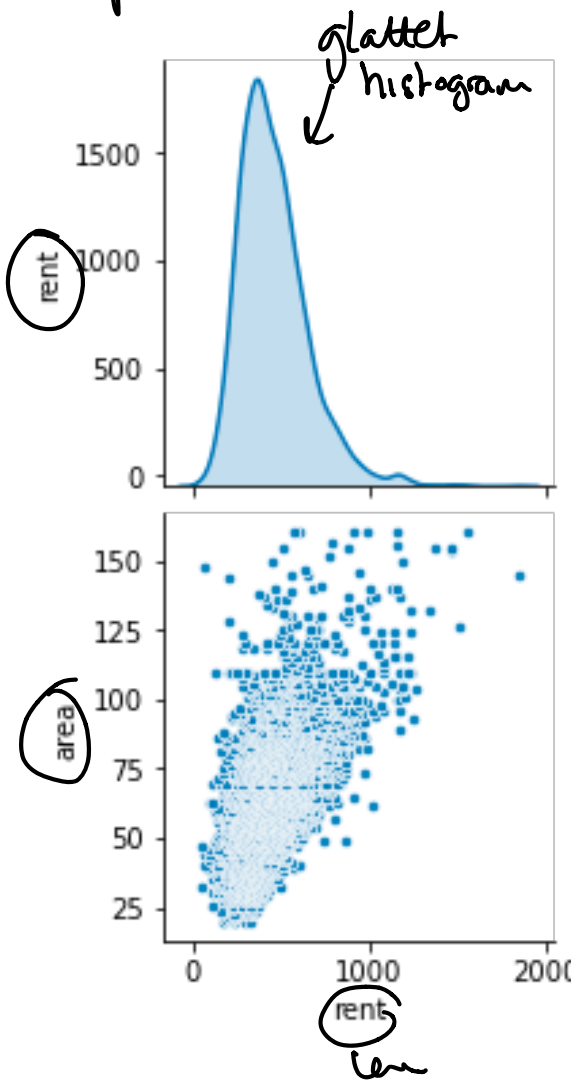


Kryssplott

(x_i, y_i)



Parplott



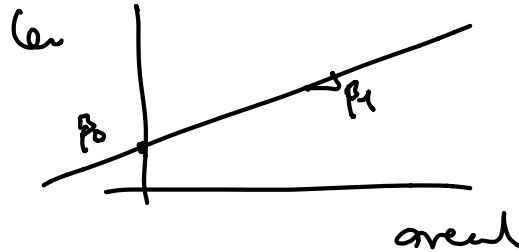
Steg 2: Matematisk modell

response $\sim$ hvariat
y $\sim$ x₁

leie areal
'rent' 'area'

Steg 3: Initier og tilpass

→ her estimerer vi parametrene $\beta_0, \beta_1 \Rightarrow \hat{\beta}_0, \hat{\beta}_1$



Steg 4: Resultat

$$y_i = \beta_0 + \beta_1 x_{1i}$$

↑
real

OLS Regression Results

```

=====
Dep. Variable:          rent      R-squared:          0.342
Model:                  OLS      Adj. R-squared:       0.341
Method:                 Least Squares      F-statistic:      1599.
Date:                   Tue, 20 Oct 2020    Prob (F-statistic): 6.13e-282
Time:                   17:39:44          Log-Likelihood:    -19990.
No. Observations:      3082          AIC:              3.998e+04
Df Residuals:          3080          BIC:              4.000e+04
Df Model:               1
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	134.5922	8.613	15.626	0.000	117.703	151.481
area	4.8215	0.121	39.982	0.000	4.585	5.058

```

=====
Omnibus:                209.552      Durbin-Watson:          2.048
Prob(Omnibus):           0.000      Jarque-Bera (JB):       442.485
Skew:                    0.448      Prob(JB):               8.23e-97
Kurtosis:                4.625      Cond. No.               215.
=====

```

Warnings:

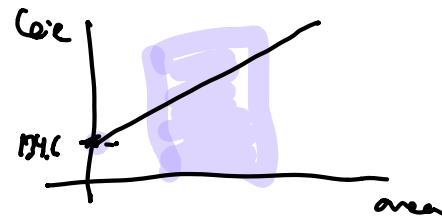
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

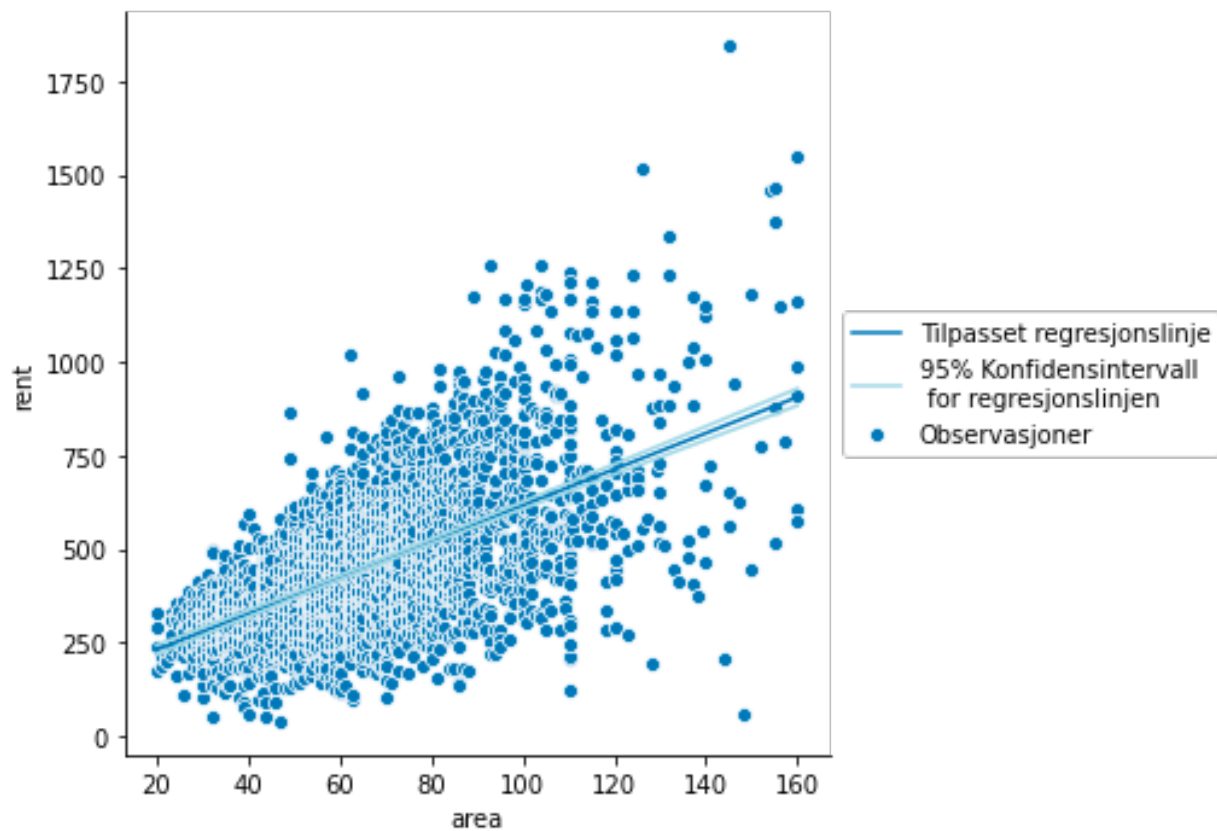
$\hat{\beta}_1$: hvor mye øker leie i gjennomsnitt når areal øker med 1 m²

$\hat{\beta}_0$

$H_0: \beta_1 = 0 \rightarrow p\text{-verdi}$

4.82



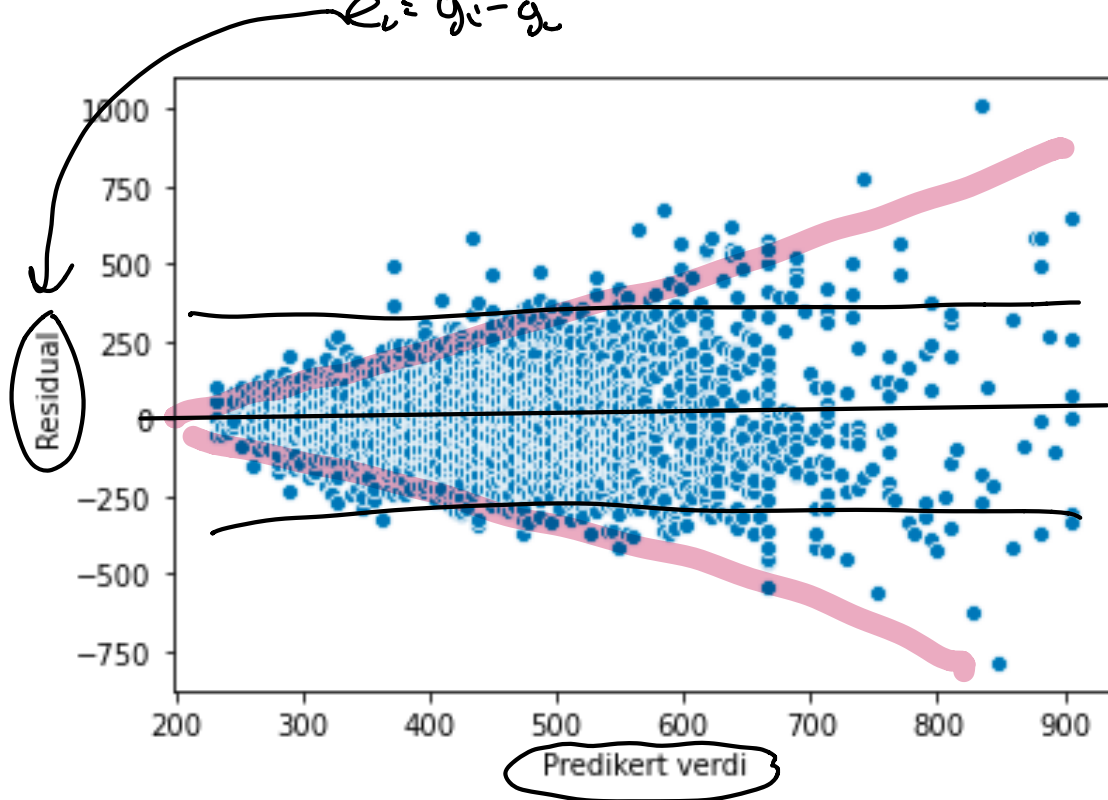


Steg 5: Passer modellen?

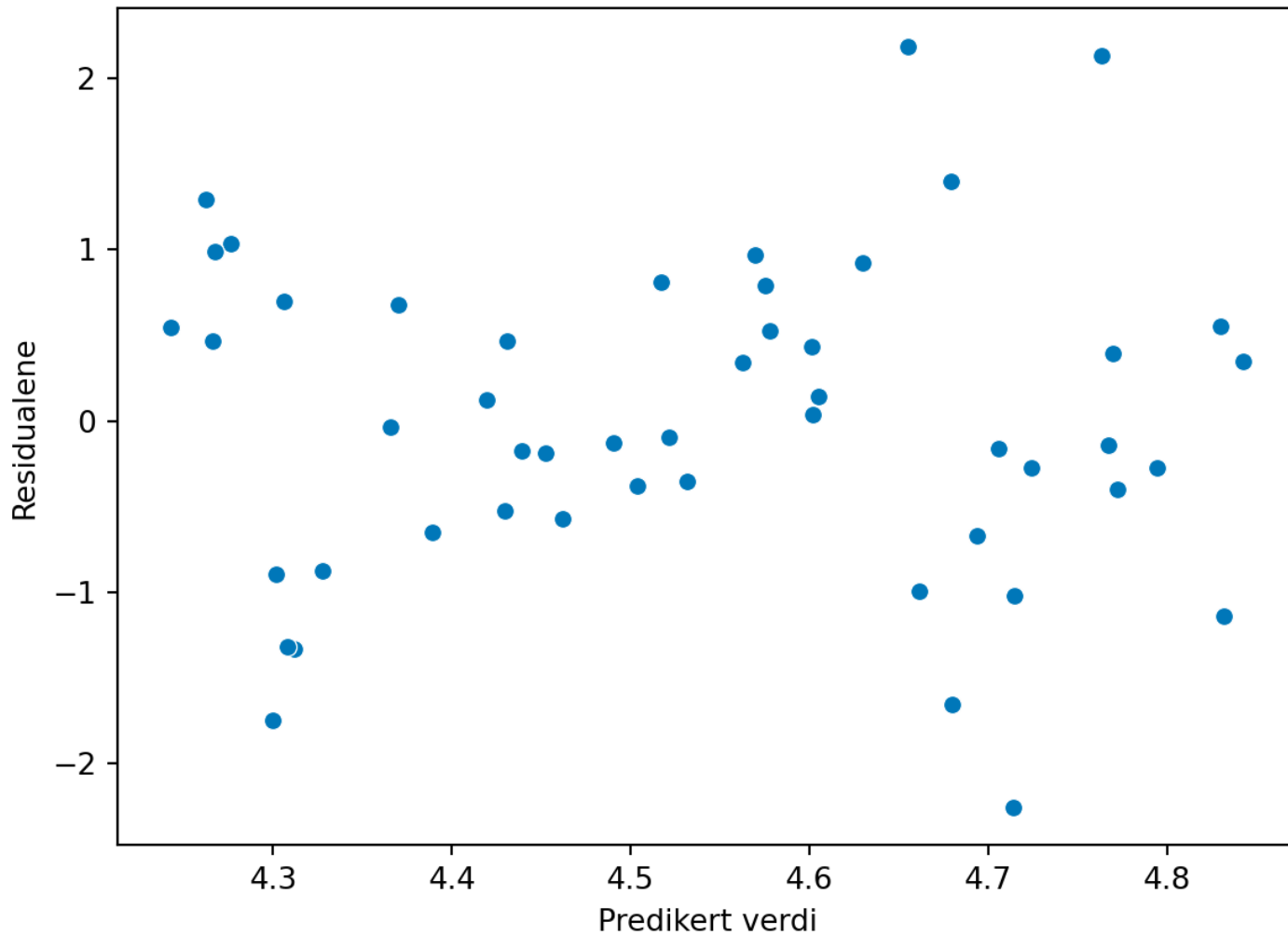
$\hat{y}_i$ predikert verd

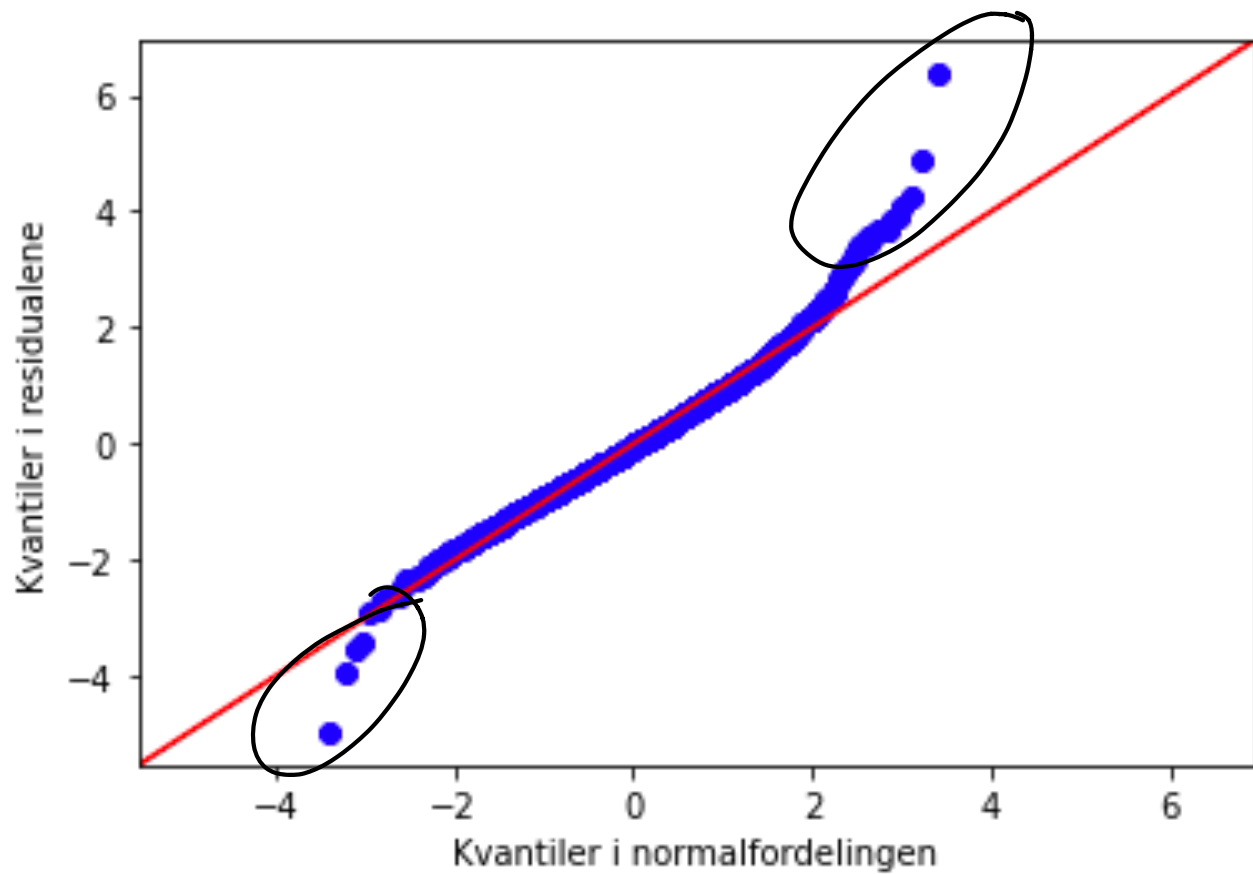
y_i obs

$$\hat{e}_i = y_i - \hat{y}_i$$

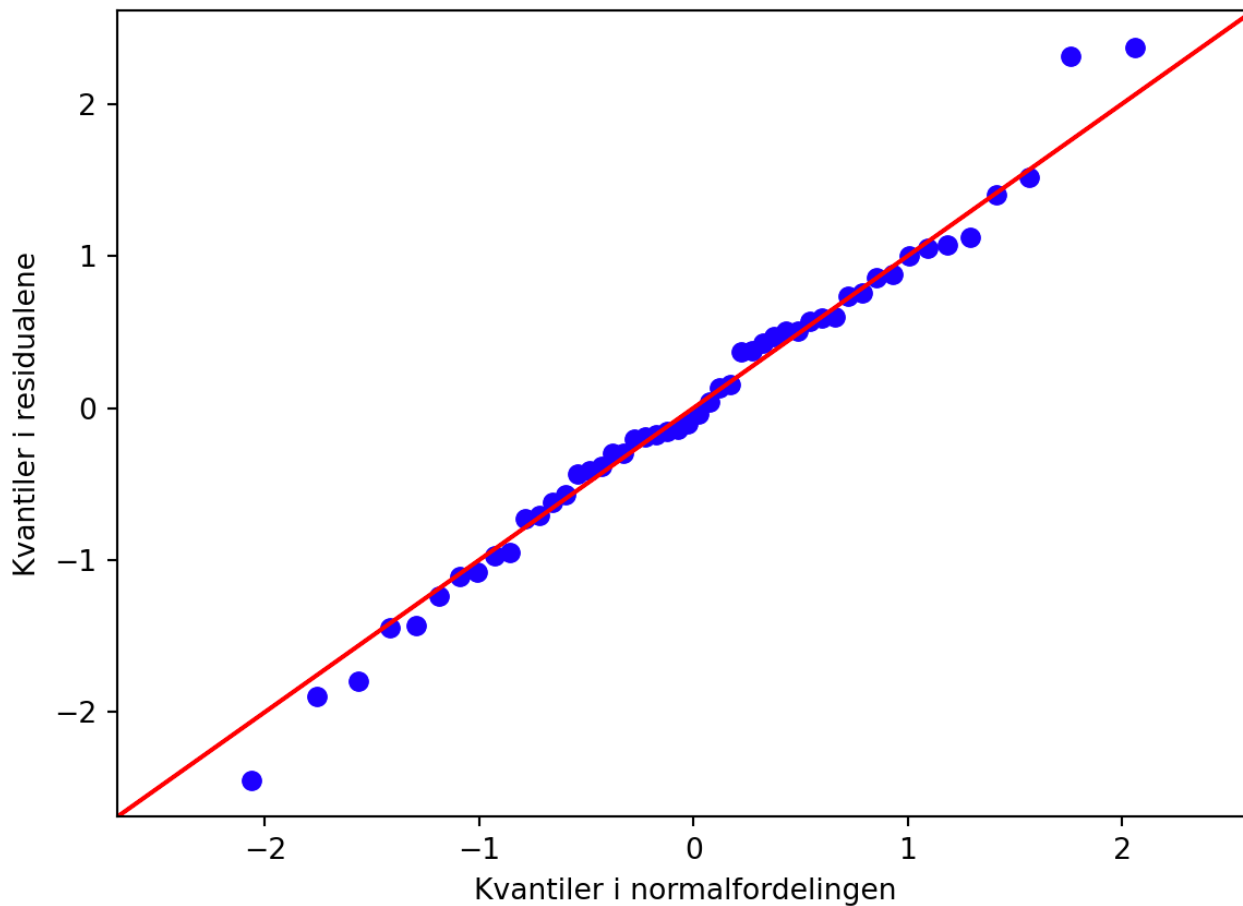


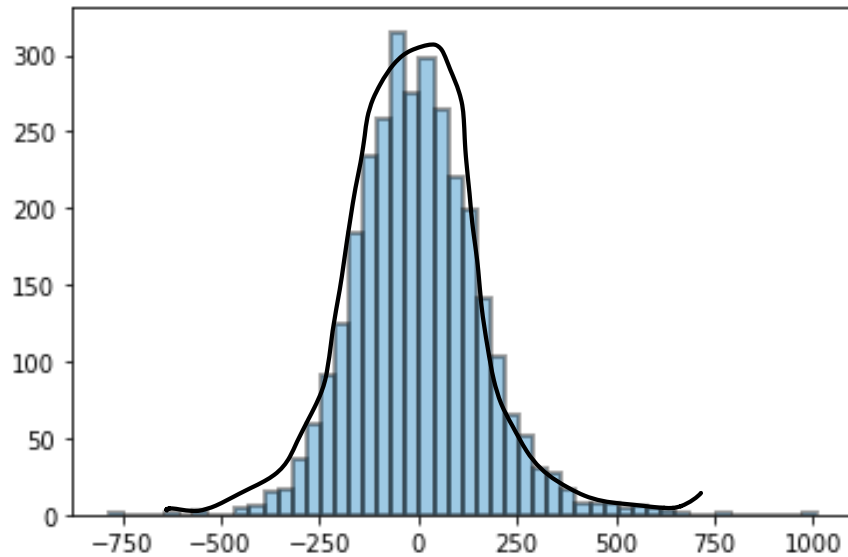
Residualplott når alt er bra!



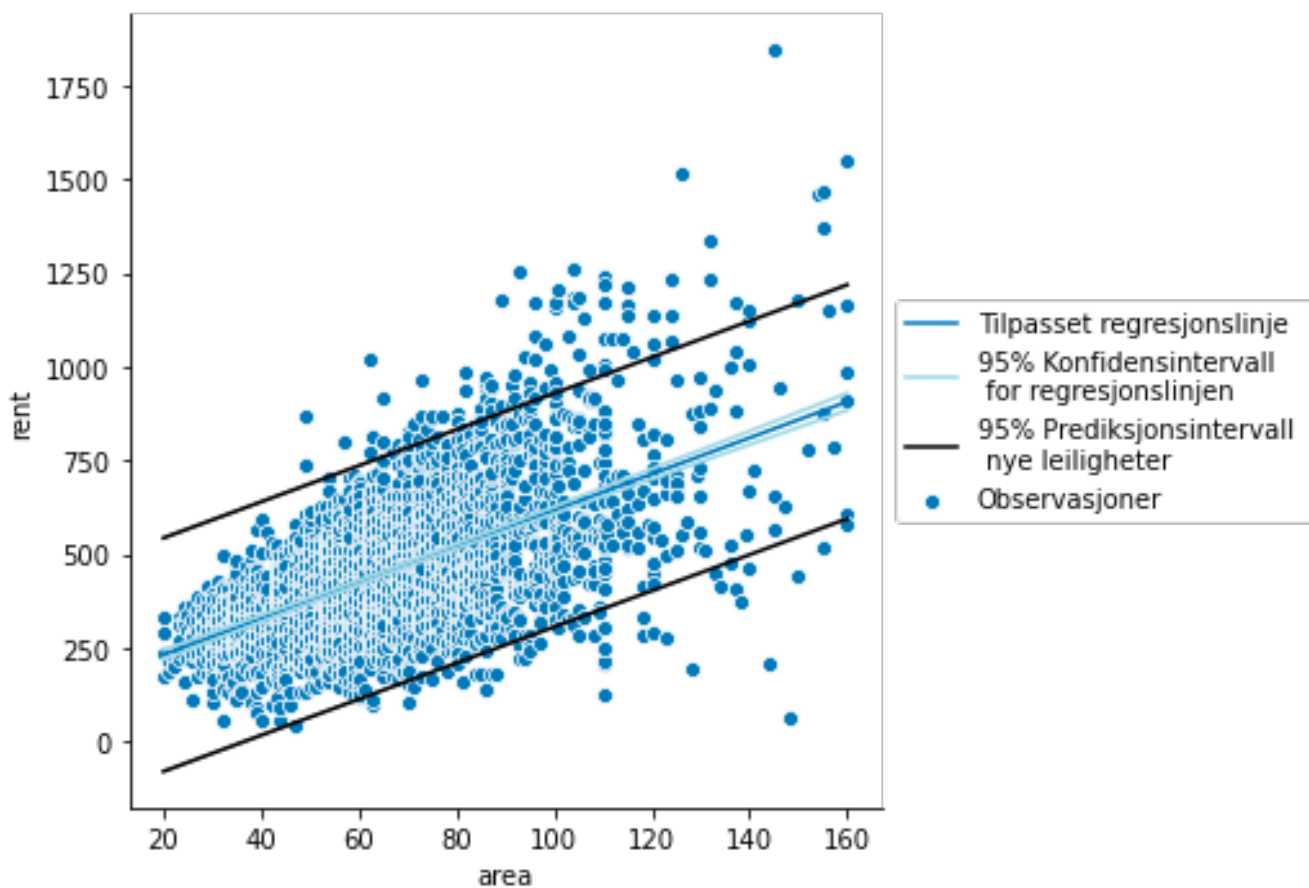


QQ-plott når alt er bra!





Vår modell passe ikke veldig godt i de delene.



Forberedelser til neste ukes zoom-time

- **Tema: Multippel lineær regresjon (MLR)**
- Se på videoene og les i kompendiet, og
- titt på Prosjektoppgaven Oppgave 1 Q1.5-Q1.8.