Innlevering 6

Oppgave 1

Anta at vi har observert observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ og at vi ønsker å tilpasse disse til en regresjonsmodell på formen

$Y_i = ax_i + \varepsilon_i$,

der $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ antas uavhengige og identisk normalfordelt med forventningsverdi lik null og varians lik σ^2. Vi har dermed at $Y_1, Y_2, \ldots, Y_n$ er uavhengige stokastiske variabler, og $Y_i \sim N(ax_i, \sigma^2)$.

Merk at vi altså betrakter de observerte verdiene $y_1, y_2, \ldots, y_n$ som realisasjoner av stokastiske variabler $Y_1, Y_2, \ldots, Y_n$, mens verdiene $x_1, x_2, \ldots, x_n$ betrakter vi som kjente tall.

Modellen har to parametre, a og σ^2, og vi ønsker å estimere verdien til disse fra de observerte parene $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Deloppgave a) **

Finn uttrykk for rimelighetsfunksjonen $L(a, \sigma^2)$ for situasjonen over.

Bruk $L(a, \sigma^2)$ til å finne uttrykk for log-rimelighetsfunksjonen $\ell(a, \sigma^2)$.

Finn sannsynlighetsmaksimeringsestimatorene for a og σ^2 og vis at disse kan skrives på formen

$$\hat{a} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{a}x_i)^2.$$

Deloppgave b) *

Finn forventningsverdi og varians for estimatoren $\hat{a}$. Du skal forenkle uttrykkene så mye det lar seg gjøre.

Er $\hat{a}$ forventningsrett? Begrunn svaret.

Her er deloppgave b) slutt.

Det kan vises (NB: du trenger ikkje vise det) at

$\sum_{i=1}^n \left(\frac{Y_i - \hat{a}x_i}{\sigma}\right)^2 \sim \chi^2_{n-1}$.

Dette resultatet kan du benytte til å besvare spørsmålene under.

Deloppgave c) *

Finn forventningsverdien til $\widehat{\sigma}^2$.

Forklar hvordan du kan se at $\widehat{\sigma}^2$ er forventningsskjev.

Foreslå en "korrigert" estimator for σ^2 (kall denne $\tilde{\sigma}^2$) som er forventningsrett.

Finn variansen til den forventningsrette estimatoren for σ^2.

$$Besvarelse$$

a) $Y_i \sim N(ax_i, \sigma^2)$

Rimlighetsfunksjon blir som følger

$$\begin{aligned}
L(a,\sigma^2) &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2}\frac{(Y_i-ax_i)^2}{\sigma^2}\right)\right)\\
l(a,\sigma^2) &= \sum_{i=1}^{n}\left(-\frac{1}{2}ln(2\pi)-\frac{1}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}(Y_i-ax_i)^2\right)\\
&= -\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-ax_i)^2
\end{aligned}$$

Vi må så pariellderivere for og løyse for når utrykka er 0 for å finne sannsynlighetsmaksimeringsestimatorene for a og σ^2.

$$\begin{aligned}
\frac{\partial l}{\partial a} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(Y_i-ax_i)(-1)\\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-ax_i)\\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}Y_i-\sum_{i=1}^{n}ax_i\right) \qquad (1)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i-ax_i)^2\\
&= \frac{1}{2\sigma^2}\left(-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-ax_i)^2\right) \qquad (2)
\end{aligned}$$

Vi må so sette (1) og (2) lik 0 for å finne sannsynlighetsmaksimeringsestimatorene.

$$
\begin{aligned}
a\sum x_i &= \sum Y_i \\
a &= \frac{\sum Y_i}{\sum x_i} \\
\Rightarrow \hat{a} &= \frac{\sum x_i Y_i}{\sum x_i^2}
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
\frac{1}{\sigma}^2 \sum_{i=1}^{n}(Y_i - ax_i)^2 &= n \\
\sum_{i=1}^{n}(Y_i - ax_i)^2 &= n\sigma^2 \\
\Rightarrow \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - ax_i)^2
\end{aligned}
\tag{2}
$$

b)

$$
\begin{aligned}
E[\hat{a}] &= E\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{1}{\sum x_i^2} E\left[\sum x_i Y_i\right] \\
&= \frac{\sum x_i}{\sum x_i^2} E\left[\sum Y_i\right] \\
&= \frac{1}{\sum x_i} a \sum x_i \\
&= a, \quad \text{forventningsrett Ü}
\end{aligned}
$$

$$
\begin{aligned}
Var[\hat{a}] &= Var\left[\frac{\sum Y_i}{\sum x_i}\right] \\
&= \frac{1}{\left(\sum x_i\right)^2} \sum Var\left[Y_i\right] \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}
$$

c) $\sum_{i=1}^{n} \left(\frac{Y_i - \hat{a}x_i}{\sigma}\right)^2 \sim \chi^2_{n-1}$.

$$E[\chi^2_{n-1}] = n-1$$

$$E\left[\sum_{i=1}^{n}\left(\frac{Y_i - \hat{a}x_i}{\sigma}\right)^2\right] = n-1$$

$$E\left[\frac{1}{\sigma^2}n\hat{\sigma}^2\right] = n-1 \quad \text{fordi } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a}x_i)^2.$$

$$E[\widehat{\sigma}^2] = \frac{\sigma^2(n-1)}{n}$$

Som vi ser er ikkje dette forventningsrett i hennhald til χ^2-fordeling Forå korigere dette definerer vi

$$\tilde{\sigma}^2 = \frac{n}{n-1}\widehat{\sigma}^2$$

forventningsverdien blir da

$$E\left[\tilde{\sigma}^2\right] = \frac{n}{n-2}E\left[\widehat{\sigma}^2\right] = \frac{n}{n-1}\frac{\sigma^2(n-1)}{n} = \sigma^2$$

Som er forventningsrett :)

For å finne variansen tar vi i bruk den nye estimatoren $\tilde{\sigma}^2$ Vi veit og at $Var[x] = 2\nu \Rightarrow Var\left[\chi^2_{n-1}\right] = \frac{1}{n^2}2(n-1)$

$$Var\left[\tilde{\sigma}^2\right] = Var\left[\frac{n}{n-2}\widehat{\sigma}^2\right]$$

$$= \frac{n^2}{(n-2)^2}Var\left[\frac{\sigma^2(n-1)}{n}\right]$$

$$= \frac{n^2}{(n-2)^2}\sigma^4 Var\left[\frac{\chi^2_{n-1}}{n}\right]$$

$$= \frac{\sigma^4 n^2 2(n-1)}{(n-2)^2}$$

$$Var\left[\tilde{\sigma}^2\right] = \frac{2\sigma^4}{n-1}$$

Oppgave 2 *

I denne oppgaven skal du benytte stokastisk simulering til å utforske hvordan et residualplott ser ut når modellen som antas i enkel lineær regresjon er korrekt og hvordan residualplott ser ut i noen tilfeller hvor den antatte modellen ikkje er korrekt.

Vi skal starte med å anta følgende modell. For $i = 1, 2, \ldots, n$ la

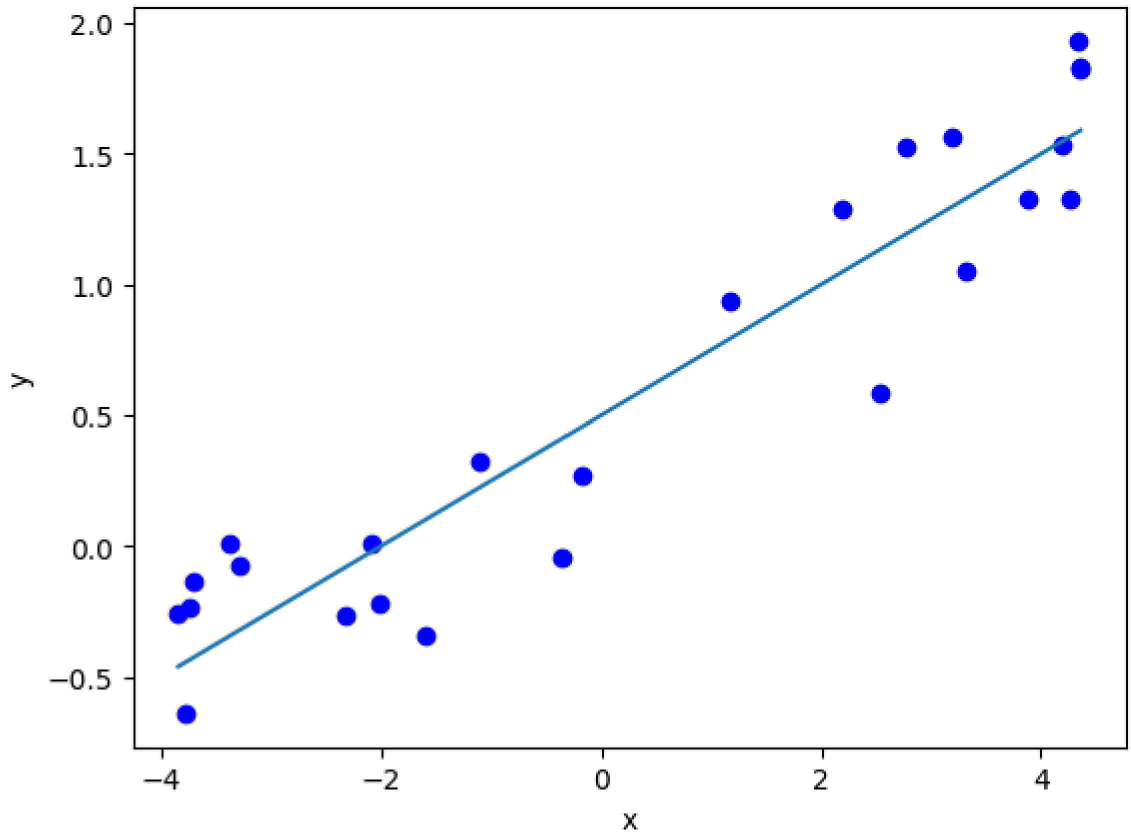
$Y_i = 0.5 + 0.25x_i + \varepsilon_i$,

der $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige og normalfordelte med forventningsverdi lik null og varians lik 0.25^2. Her er altså modellen som antas i enkel lineær regresjon korrekt, og parameterverdiene er $\alpha = 0.5$, $\beta = 0.25$ og $\sigma = 0.25$.

I python-koden under har du fått oppgitt verdier for x_i $i = 1, 2, \ldots, 25$. Deretter genereres tilhørende verdier for $y_1, y_2, \ldots, y_n$ ifølge modellen formulert ovenfor. De genererte verdiene visuliseres så i et spredningsplott.

In [ ]:

```
# Du trenger ikkje endre noe i denne koden!

import numpy as np
#from scipy.stats import norm
import matplotlib.pyplot as plt

#Initialisering av parameterverdier
n = 25
alpha = 0.5
beta = 0.25
sigma = 0.25

#Simulering av data etter modell
# x_1,x_2,...,x_n i intervallet [-5,5]
x = np.array([-3.842, -3.784, -3.745, -3.708, -3.37,  -3.288, -2.312, -2.078, -2.
              -1.595,-1.106, -0.352, -0.171,  1.166,  2.196,  2.538,  2.772,  3.18
              3.309,3.876, 4.2,    4.261,  4.337,  4.352,  4.359])
# genererer tilhørende verdier for y_1,y_2,...,y_n
y = alpha + beta * x + np.random.normal(loc=0,scale=sigma,size=n)
y_r = alpha + beta *x

#Visualiserer resultatet i et plott
plt.plot(x,y,'bo')
plt.xlabel('x')
plt.ylabel('y')

plt.plot(x,y_r)
plt.show()
```



Under er det gitt en python-funksjon som tar vektorer x og y som input og regner ut estimatene $\hat{\alpha}$, $\hat{\beta}$ og S^2 i en enkel lineær regresjonsmodell. Dette er tilsvarende metode som ble gjort i oppgave 1, bare nå med et konstantledd (SME for lineærregresjon, og forventningsrett estimator for variansen).

In [ ]:

```
# Du trenger ikkje endre noe i denne koden!

def estimerELR(x,y):
    #Beregner gjennomsnitt
    xStrek = np.mean(x)
    yStrek = np.mean(y)
    #Estimater for parametere
    betaHat = np.sum((x-xStrek)*y)/np.sum((x-xStrek)**2)
    alphaHat = yStrek - betaHat * xStrek
    S2 = np.sum((y-(alphaHat+betaHat*x))**2)/(len(x)-2)
    #Returnerer resultatet i en liste
    return [alphaHat,betaHat,S2]

paramHat = estimerELR(x,y)
print('alphaHat: ',paramHat[0])
print('betaHat: ',paramHat[1])
print('s2: ',paramHat[2])
```

```
alphaHat:  0.5135104002695219
betaHat:  0.2489039335412089
s2:  0.08084928463572486
```

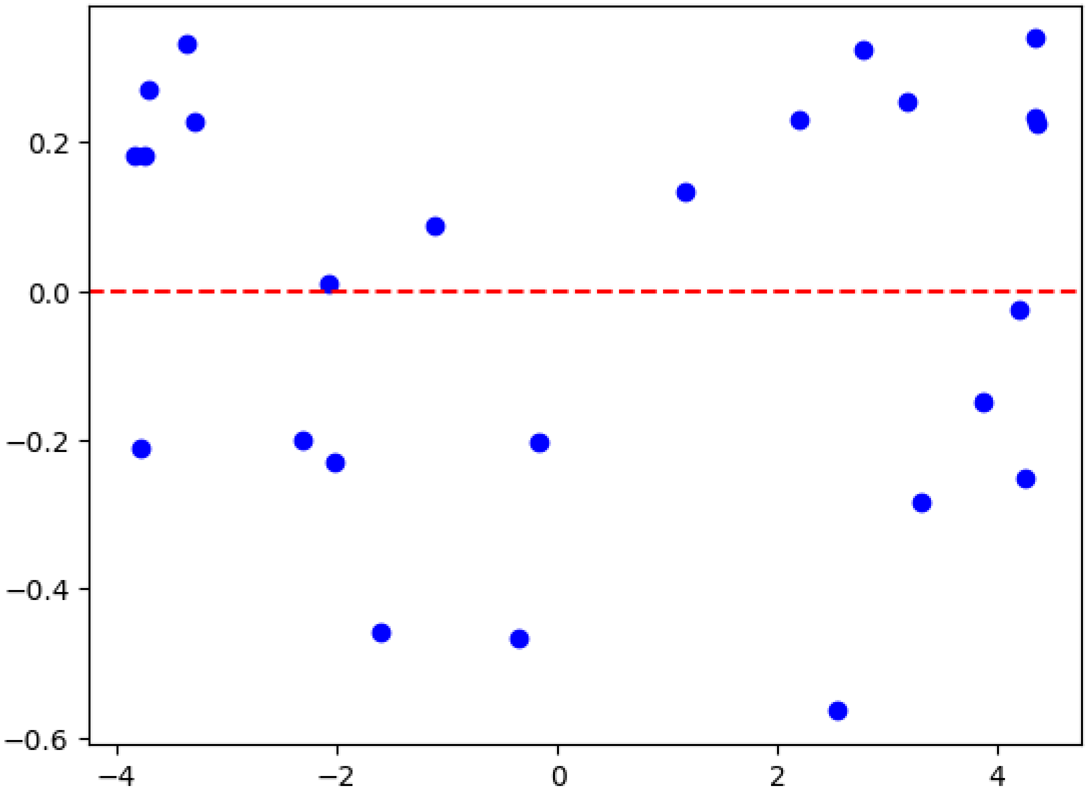
Deloppgave a)

Kjør de to bitene med python-kode gitt over. Betrakt nå de genererte x og y-verdiene som observerte verdier, og skriv under python-kode som regner ut de resulterende (estimerte) residualene. Lag også et residualplott hvor du plotter x_i-verdiene langs x-aksen og de (estimerte) residualene langs y-aksen.

Kjør gjerne (alle de tre) pythonkodebitene flere ganger slik at du får et inntrykk av hvordan residualplottet varierer for ulike datasett (generert fra den spesifiserte regresjonsmodellen). Diskuter kort hva du ser (eller ikkje kan se) i residualplottene.

In [ ]:
```
# Her kan du skrive din python-kode

estimerte_residualer = y - (paramHat[0] + paramHat[1] * x)
print(estimerte_residualer)

plt.plot(x,estimerte_residualer, 'bo')
plt.axhline(0, color='r', linestyle='--')

plt.show()
```

```
[ 0.18173765 -0.21116009  0.18167724  0.27190378  0.33333653  0.22784627
 -0.19945446  0.01099212 -0.2292419  -0.45718748  0.08693704 -0.46582623
 -0.2031644   0.13505881  0.23129434 -0.56395447  0.32375455  0.25413538
 -0.28277587 -0.14895582 -0.02431736 -0.25013816  0.33946555  0.23195423
  0.22608276]
```



Her er deloppgave a) slutt.

Du skal så utforske hvordan et residualplott kan bli seende ut når modellen som antas i enkel lineær regresjon ikkje er korrekt. For å gjøre dette skal du først generere x-verdier ved å trekke verdier fra samme fordeling som gjort over. Deretter skal du generere y-verdier ifølge

$$Y_i = 0.5 + 0.25x_i + 0.02x_i^2 + \varepsilon_i,$$

der $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige og normalfordelte med forventningsverdi null og varians lik 0.10^2.

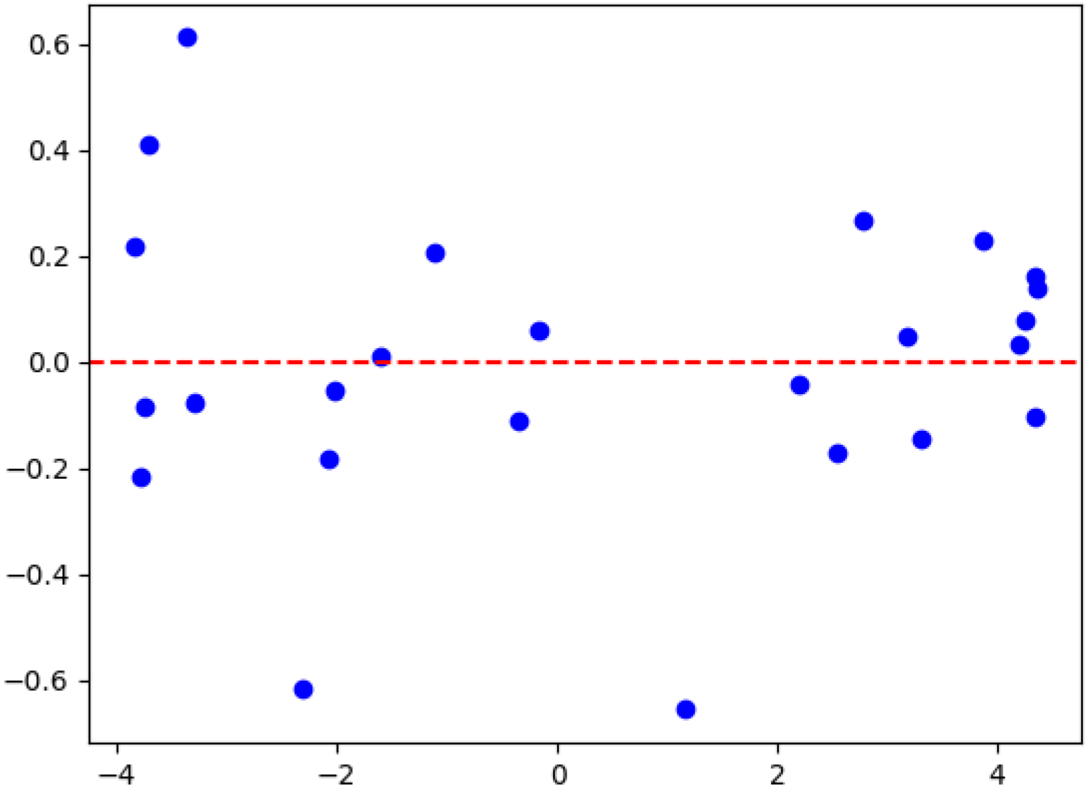
Deloppgave b)

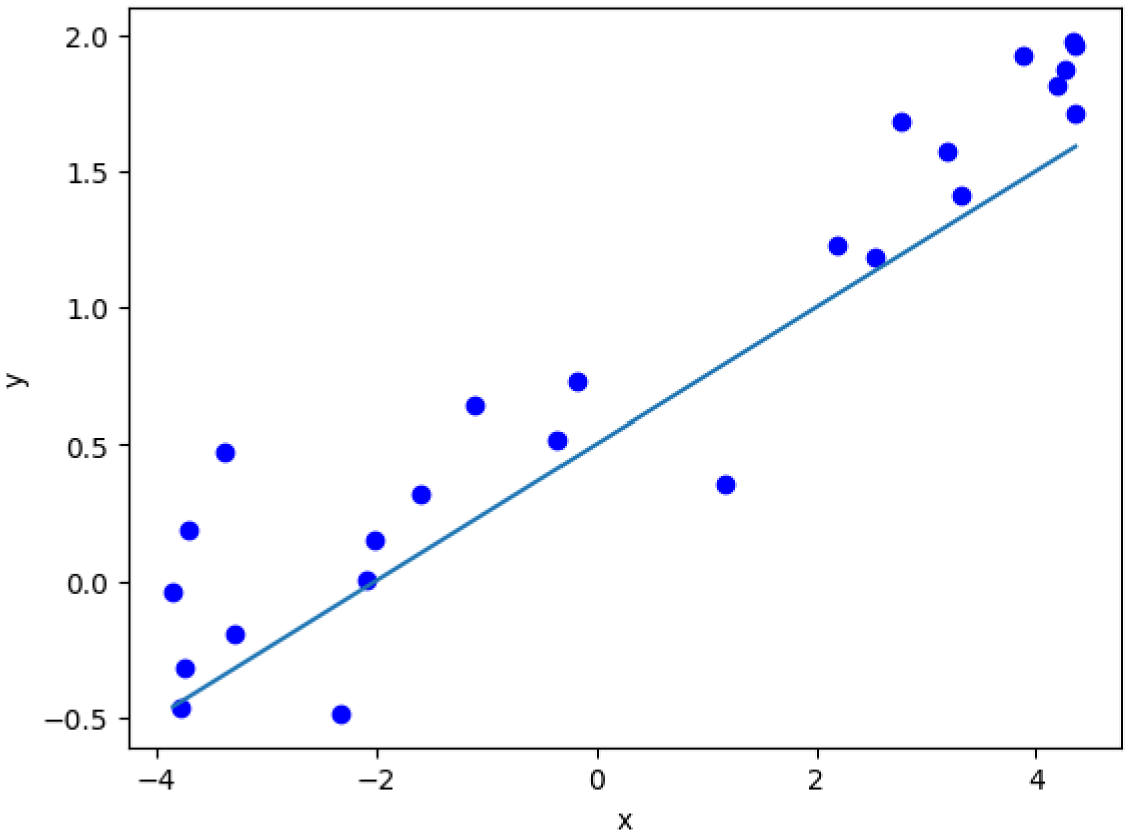
Skriv python-kode som genererer $n = 25$ par (x_i, Y_i) som beskrevet over. Betrakt så disse simulerte dataene som observerte data og tilpass en enkel lineær regresjonsmodell ved å kalle python-funksjonen estimerELR gitt over. Regn så ut (estimerte) residualer og generer residualplott.

Kjør gjerne python-koden flere ganger slik at du får et inntrykk av hvordan residualplottet varierer for ulike datasett (generert fra den spesifiserte modellen). Diskuter kort hva du ser (eller ikkje kan se) i residualplottene.

In [ ]:

```
# Her kan du skrive din python-kode
# Du trenger ikkje endre noe i denne koden!

import numpy as np
#from scipy.stats import norm
import matplotlib.pyplot as plt

#Initialisering av parameterverdier
n = 25
alpha = 0.5
beta = 0.25
gamma = 0.02
sigma = 0.25

y = alpha + beta * x + gamma * x**2+ np.random.normal(loc=0,scale=sigma,size=n)


paramHat = estimerELR(x,y)
# print('alphaHat: ',paramHat[0])
# print('betaHat: ',paramHat[1])
# print('s2: ',paramHat[2])
# Her kan du skrive din python-kode

estimerte_residualer = y - (paramHat[0] + paramHat[1] * x)

plt.plot(x,estimerte_residualer, 'bo')
plt.axhline(0, color='r', linestyle='--')

plt.show()

#Visualiserer resultatet i et plott
plt.plot(x,y,'bo')
plt.xlabel('x')
plt.ylabel('y')

plt.plot(x,y_r)
plt.show()
```





Det er ingen tydlig mønster i residualplottene samtidig som det helder seg rimelig nære den estimerte linja. Det kan tyde på at det kan vere fornuftig å nytte lineærregresjonsmodell til oppgava.

Her er deloppgave b) slutt.

Du skal så utforske hvordan residualplottet blir seende ut for en annen modell som avviker fra hva som antas i en enkel lineær regresjonsmodell. Genererer igjen x_i-verdier på samme måte som over. Genererer deretter Y_i-verdier ifølge

$$Y_i = 0.5 + 0.25x_i + \varepsilon_i,$$

der $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige og $\varepsilon_i \sim N(0, 0.10^2 \cdot (0.1 + x_i^2))$

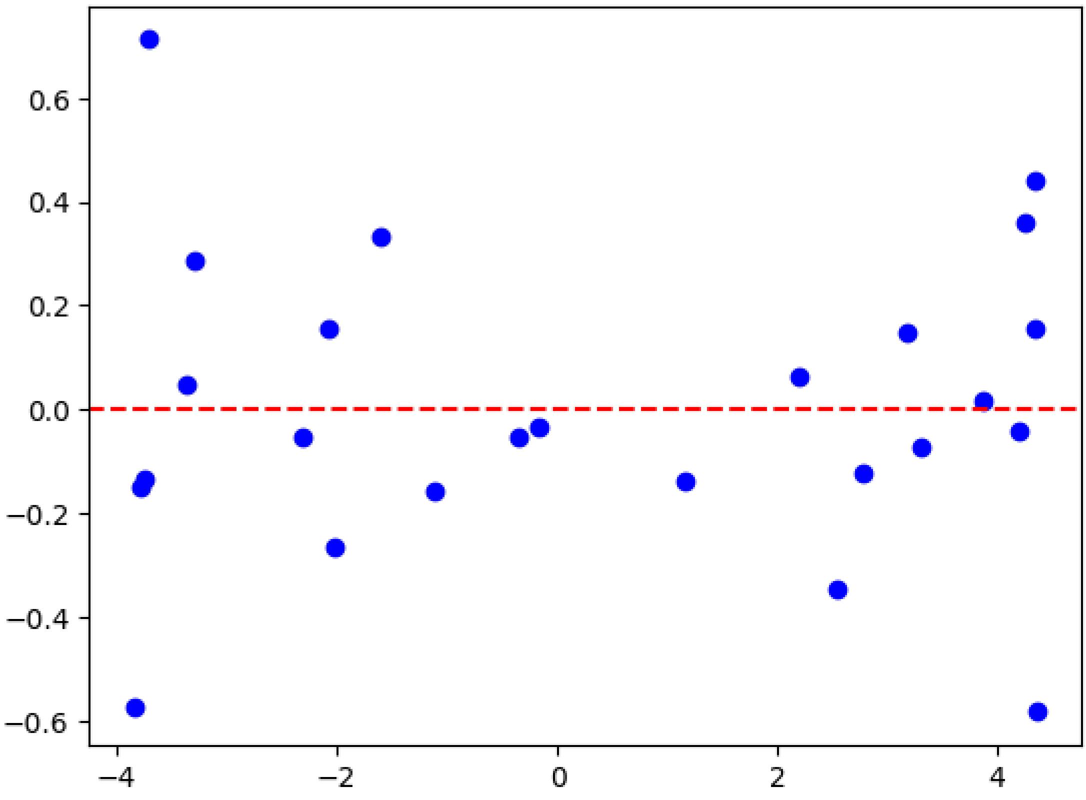
Deloppgave c)

Skriv og kjørpython-kode som simulerer x- og y-verdier som beskrevet over, og så bruker disse tilsvarende som i oppgave b) over til å generere tilhørende residualplott.

Kjør gjerne python-koden flere ganger slik at du får et inntrykk av hvordan residualplottet varierer for ulike datasett (generert fra den spesifiserte modellen). Diskuter kort hva du ser (eller ikkje kan se) i residualplottene.

In [ ]:

```
# Her kan du skrive din python-kode
epsilon_varians = 0.10**2 * (0.1 + x**2)
y = alpha + beta * x + np.random.normal(loc=0, scale=np.sqrt(epsilon_varians), siz
paramHat = estimerELR(x,y)
estimerte_residualer = y - (paramHat[0] + paramHat[1] * x)

plt.plot(x,estimerte_residualer, 'bo')
plt.axhline(0, color='r', linestyle='--')

plt.show()
```



Her og er det ingen tydlig mønster i residualplottene, det kan tyde på at det kan vere fornuftig å nytte lineærregresjonsmodell til oppgava.

Oppgave 3 *

Vi skal i denne oppgaven anta at bremselengden, Y , målt i meter for en bil som kjører x km/time antas å være normalfordelt med forventningsverdi βx^2 og standardavvik σx. En bil som for eksempel kjører i 50 km/time vil dermed ha en bremselengde som er normalfordelt med forventningsverdi 2500β og standardavvik 50σ. Modellen har to parametre, β og σ^2, og disse vil avhenge av forsøksbetingelsene, som for eksempel dekkenes egenskaper, veidekke og vær- og føreforhold.

Anta nå at verdiene til β og σ^2 er ukjent og skal estimeres. For å estimere disse parametrene gjøres n bremseprøver med ulike hastigheter, men forøvrig under identiske forsøksbetingelser. La x_i betegne hastigheten benyttet ved bremseprøve nummer i, og la Y_i være tilhørende bremselengde. Vi skal anta at bremseprøvene utføres på en slik måte at det er rimelig å betrakte $Y_1, Y_2, \ldots, Y_n$ som uavhengige stokastiske variabler. Vi lar som vanlig $y_1, y_2, \ldots, y_n$ betegne de målte bremselengdene.

Deloppgave a)

Utled estimatorer for β og σ^2 ved å benytte sannsynlighetsmaksimeringsprinsippet. Vis at estimatorene kan skrives på formen

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2},$$
$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\beta} x_i^2}{x_i} \right)^2.$$

Deloppgave b)

Finn forventningsverdi og varians for estimatoren $\hat{\beta}$.

Hvilken sannsynlighetsfordeling har $\hat{\beta}$? Begrunn svaret.

Deloppgave c)

Bruk antagelsene gjort i oppgaveteksten over og sammenhenger mellom fordelinger som vi har diskutert tidligere i kurset til å vise at

$$\sum_{i=1}^{n} \left(\frac{Y_i - \beta x_i^2}{\sigma x_i} \right)^2$$

er χ^2-fordelt med n frihetsgrader. Merk at det i uttrykket over står den (ukjente) sanne verdien β.

Her er deloppgave c) slutt.

Videre i oppgaven kan du uten bevis benytte at dersom man i uttrykket gitt i deloppgave c) erstatter den (ukjente) sanne verdien β med estimatoren $\hat{\beta}$ vil størrelsen fremdeles være χ^2-fordelt, men antall frihetsgrader vil reduseres med en. Man har altså at

$$V = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\beta} x_i^2}{\sigma x_i} \right)^2$$

er χ^2-fordelt med $n-1$ frihetsgrader. Du kan dessuten uten bevis benytte at $\hat{\beta}$ og V er uavhengige stokastiske variabler.

Deloppgave d)

Identifiser en pivotal som kan brukes til å utlede en konfidensintervall for β. Vis hvilken sannsynlighetsfordeling pivotalen har, og bruk så dette til å utlede et $(1-\alpha) \cdot 100\%$-konfidensintervall for β.

$$Besvarelse$$

Benytter notasjonen
$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, $S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{y})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{y})$

a)

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi x_i^2 \sigma^2}} exp\left(-\frac{1}{2} \frac{(Y_i - \beta x_i^2)^2}{x_i^2 \sigma^2} \right) \right)$$

$$l(\beta, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{1}{2} ln(2\pi) - \frac{1}{2} ln(x_i^2 \sigma^2) - \frac{2}{2x_i^2 \sigma^2} (Y_i - \beta x_i^2)^2 \right)$$

$$= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \sum_{i=1}^{n} ln(x_i) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(Y_i - \beta x_i^2)^2}{x_i^2}$$

Vi må så pariellderivere for og løyse for når utrykka er 0 for å finne sannsynlighetsmaksimeringsestimatorene for β og σ^2.

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(Y_i - \beta x_i^2) \frac{-x_i^2}{x_i^2}$$

$$\Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta x_i^2) = 0$$

$$= \sum_{i=1}^{n} \beta x_i^2 = \sum_{i=1}^{n} Y_i$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum \frac{(Y_i - \beta x_i^2)^2}{x_i^2}$$

$$\frac{1}{\sigma^2}\sum \frac{(Y_i - \beta x_i^2)^2}{x_i^2} = n$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \beta x_i^2)^2}{x_i^2}$$

b) Forbentnigs og varians

$$E[\hat{\beta}] = E\left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{\sum_{i=1}^n \beta x_i^2}{\sum_{i=1}^n x_i^2} = \beta$$

$$Var[\hat{\beta}] = Var\left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\sum_{i=1}^{n} \sigma^2 x_i^2 = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

Sidan Y_i er normalfordelt, og $\hat{\beta}$ er ein lineærkombinasjon av Y_i vil også $\hat{\beta}$ vere normalfordelt med forventningverdi β og varians $\frac{\sigma^2}{\sum_{i=1}^n x_i^2}$

c) dersom $X_1, X_2, \ldots, X_n$ er uavhengige og normal fordelt med forventnigsverdi βx^2 og varians $x^2\sigma^2$ har vi at

$\sum \frac{(Y_i - x_i^2\beta)^2}{x_i^2\sigma^2} \sim \chi_n^2$

og vi kan skrive $\sum \frac{(Y_i - x_i^2\beta)^2}{x_i^2\sigma^2} = \sum \left(\frac{(Y_i - x_i^2\beta)}{x_i\sigma}\right)^2$

d)

$$V = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\beta}x_i^2}{\sigma x_i}\right)^2$$

bruker estimatorane $\hat{\beta} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2}$ og $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(Y_i - \beta x_i^2)^2}{x_i^2}$

Visst $Z \sim N(0,1)$ så er $T = \frac{Z}{\sqrt{V/\nu}} \sim t_n$

og β og V er uavhengige variabla, og vi kan normaliser Z

$$Z = \frac{\hat{\beta} - E[\hat{\beta}]}{\sqrt{Var[\hat{\hat{\beta}}]}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

så har vi at $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(Y_i - \beta x_i^2)^2}{x_i^2}$ og V kan skrivast $V = \frac{1}{\sigma^2}\sum_{i=1}^n \left(\frac{Y_i - \hat{\beta}x_i^2}{x_i}\right)^2 \sim_{n-1}^2$

Fra ditta ser vi at $\hat{\sigma}^2 = \frac{1}{n}\sigma^2 V \Rightarrow V = \frac{\hat{\sigma}^2}{\sigma^2}n$ der $\nu = 1 - n$

sette deet inn for T

$$T = Z\sqrt{\frac{\nu}{V}}$$

$$T = (\hat{\beta} - \beta)\sqrt{\frac{S_{xx}}{\sigma^2}}\sqrt{\frac{1-n}{\hat{\sigma}^2 n}\sigma^2}$$

$$= (\hat{\beta} - \beta)\sqrt{\frac{S_{xx}(1-n)}{\hat{\sigma}^2 n}}$$

$$\Rightarrow \beta \pm t_{\alpha/2,n-1}\sqrt{\frac{\hat{\sigma}^2 n}{s_{xx}(1-n)}}$$

Oppgave 4

I denne oppgaven skal vi tilpasse en enkel lineær regresjonsmodell til et datasett hvor x_i-ene er målt tetthet til australsk tømmer, mens tilhørende Y_i er målt verdi for den såkalte Janka-hardheten til det samme tømmeret. En grundigere presentasjon og diskusjon av datasettet finnes i 'E.J. Williams. Regression analysis. John Wiley & Sons Inc., New York, 1959; Tabell 3.1, side 43'.

Deloppgave a) *

In [ ]:
```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

x = np.array([24.7,24.8,27.3,28.4,28.4,29.0,30.3,32.7,35.6,38.5,38.8,39.3,39.4,39.
              42.9,45.8,46.9,48.2,51.5,51.5,53.4,56.0,56.5,57.3,57.6,59.2,59.8,66.
y = np.array([484,427,413,517,549,648,587,704,979,914,1070,1020,1210,989,1160,101(
              1400,1760,1710,2010,1880,1980,1820,2020,1980,2310,1940,3260,2700,28(
```

Visualiser dataene i et spredningsplott.

Anta så en enkel lineær regresjonsmodell for dataene og estimer parametrene α, β og σ^2 basert på sannsynlighetsmaksimeringsprinsippet (dvs. regn ut estimater for de tre parametrene). Legg til den estimerte linja $y = \hat{\alpha} + \hat{\beta}x$ i spredningsplottet.

Regn ut de (estimerte) residualene og visualiser disse i et residualplott. Diskuter det du ser i spredningsplottet og i residualplottet. Tyder plottene på at en enkel lineær regresjonsmodell passer for dette datasettet?

Her er deloppgave a) slutt.

Uansett hva du konkluderte med i deloppgave a) skal du videre i oppgaven gi svar basert på en enkel lineær regresjonsmodell. Merk dessuten at du i resten av denne oppgaven kan benytte resultater som er utledet i læreboka/introvideoer/forelesninger, men må passe på at forutsetningene for resultatene du benytter er oppfylt.

Deloppgave b) *

Benytt datasettet til å gjennomføre en hypotesetest hvor du tester $H_0 : \alpha = 0$ mot $H_1 : \alpha \neq 0$. Dvs. spesifiser hvilken testobservator du vil benytte, angi hvilken sannsynlighetsfordeling testobservatoren har når H_0 er sann, finn en beslutningsregel slik at testen får signifikansnivå lik 0.10, og benytt de observerte data til å bestemme om man skal forkaste H_0 eller ikkje.

Deloppgave c)

Du skal så finne et 90%-prediksjonsintervall for Janka-hardheten, Y_0 i en trestamme hvor tettheten i trestammen er målt til $x = x_0$. Angi svaret som et intervall hvor nedre og øvre grense i intervallet er en funksjon av x_0. Plott opp nedre og øvre grense av prediksjonsintervallet sammen med spredningsplottet for $x_0 \in [24, 70]$.

Det å måle hardheten, altså x, i en trestamme kan gjøres raskt, mens det å måle Janka-hardheten, altså Y, er en mer arbeidskrevende prosess. Basert på dine resultater i denne oppgaven, vil du si at det er fornuftig erstatte en måling av Janka-hardheten med prediksjonen $y = \hat{\alpha} + \hat{\beta}x$ der x er målt hardhet? Begrunn svaret ditt.

$$Besvarelse$$

a) vi veit at for ein enkel lineær regresjonsmodell vil dei estimerte parametera for α, βog σ hennholdsvis vere $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$, $\hat{\beta} = \frac{\S_{xy}}{S_{xx}}$, og $\hat{\sigma}^2 = \frac{1}{n}(Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$

I spreiingsplottet ser vi ein generell positiv trend, noko som tydar at hardheten aukar proposjonalt med tettheten.

Residualplotter viser forskjell mellom de observerte verdiane og den predikert verdien. Ideelt ønska vi ein tilfeldig fordeling av punkat i plottet, og ikkje noko mønster rundt linja $y = 0$. I plottet ser det ut til å ikkje vere noko mømnster. Observasjonen tyder på at lineær regresjon er rimelig.

b) $H_0 : \alpha = 0$ mot $H_1 : \alpha \neq 0$ og vi benytter ei student-t test.

Testobservator $T = \frac{\hat{\alpha} - \alpha_0}{\sqrt{Var[\hat{\alpha}]}}$ der $\alpha_0 = 0$, og $Var[\hat{\alpha}] = \sigma^2 \frac{\sum x_i^2}{nS_{xx}}$ med $\hat{\sigma}^2 = \frac{1}{n}S_{yy}$

Men vi ønsker å korrigere for forventingsskjev estimatoren til $\hat{\sigma}^2$ og erstatter den med den forventnigsrette estimatoren $S^2 = \frac{n}{n-2}\hat{\sigma}^2$.

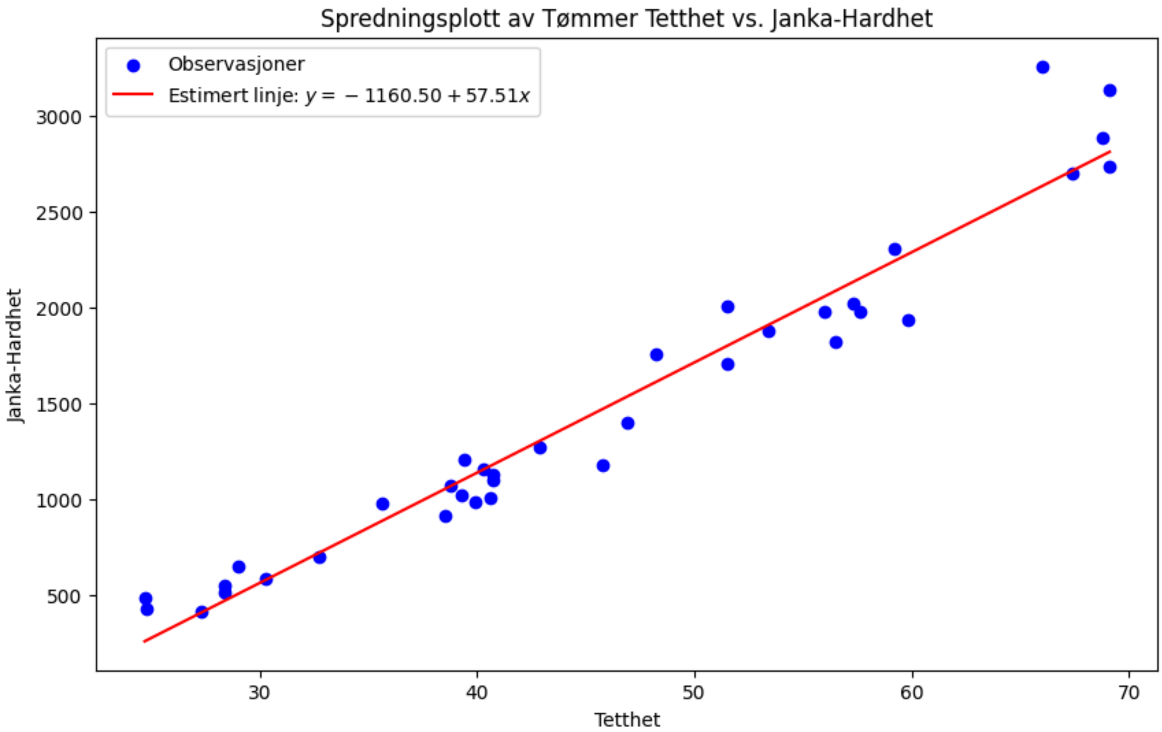
Når H_0 er sann vil T følgje ein student t-fordeling med n-2 frihetsgrader fordi det er estimert med 2 parameter frå datasettet. for å kunne fastsette ein besultningsregel med signifikansnivå på 0.10 finner vi dei kritiske verdiane for ein tosidig student t-fordeling med $\nu = n - 2$ altså vil $t_{\alpha/2, n-2}$ og $t_{1-\alpha/2, n-2}$ avvisningsområdet.

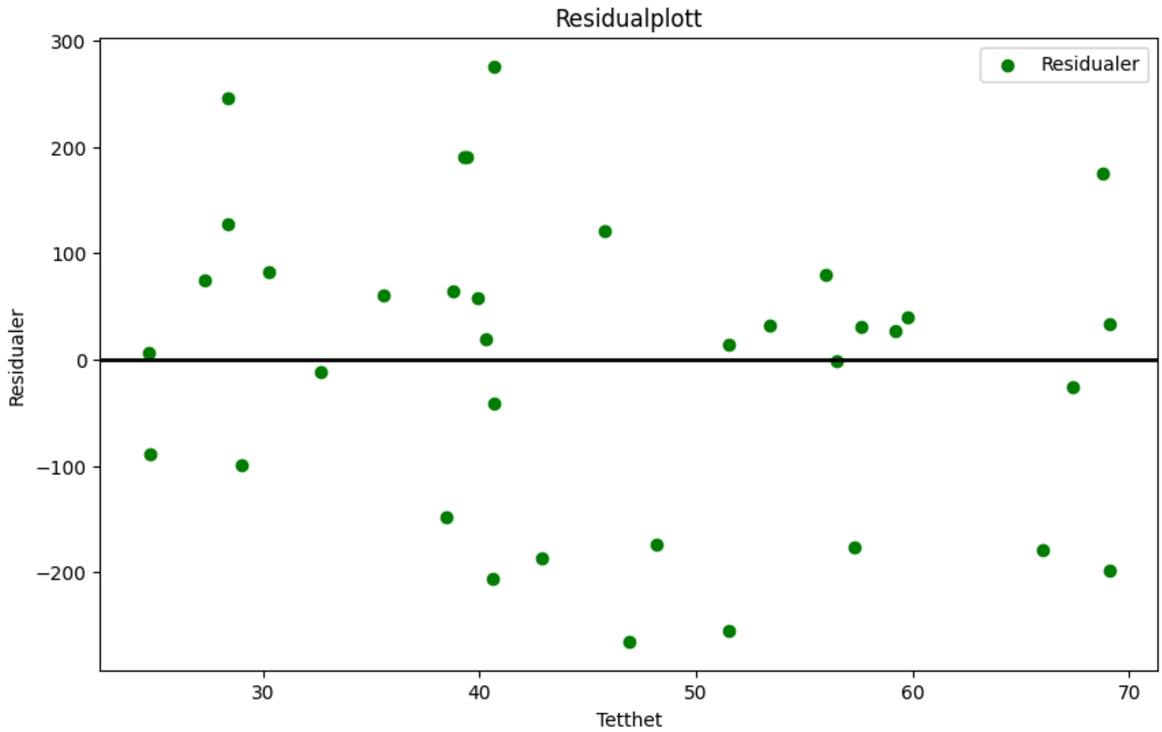
In [ ]:

```
# Her kan du skrive python-kode for å gjøre beregningene du trenger for å besvare
from scipy import stats
def estimerParam(x,y):
    xSnitt = np.mean(x)
    ySnitt = np.mean(y)
```

```python
    n = len(x)

    betaHat  = np.sum((x-xSnitt)*(y-ySnitt))/np.sum((x-xSnitt)**2)
    alphaHat = ySnitt-betaHat*xSnitt
    sigma2 = np.sum((y-(alphaHat+betaHat*x))**2)/(n)
    s2 = n/(n-2)*sigma2
    VarAlpha = s2*np.sum(x**2)/(n*np.sum((x-xSnitt)**2))
    T = alphaHat/np.sqrt(VarAlpha)
    return[alphaHat,betaHat,sigma2,s2],T

param,T = estimerParam(x,y)
y_hat = param[0] + param[1]*x
residuals = np.random.normal(loc=0,scale=np.sqrt(param[2]),size=len(x))

print('alphaHat: ',param[0])
print('betaHat: ',param[1])
print('s2: ',param[2])

print('T: ',T)
print('Kritsik verdi', stats.t.ppf(alpha/2, n-2))

# Visualiser dataene i et spredningsplott
plt.figure(figsize=(10, 6))
plt.scatter(x, y, color='blue', label='Observasjoner')
plt.title('Spredningsplott av Tømmer Tetthet vs. Janka-Hardhet')
plt.xlabel('Tetthet')
plt.ylabel('Janka-Hardhet')
plt.plot(x, y_hat, color='red', label=f'Estimert linje: $y={param[0]:.2f} + {parar

plt.legend()
plt.show()

# Visualisere residualene i et residualplott
plt.figure(figsize=(10, 6))
plt.scatter(x, residuals, color='green', label='Residualer')
plt.axhline(0, color='black', lw=2)  # Linje for residual = 0
plt.title('Residualplott')
plt.xlabel('Tetthet')
plt.ylabel('Residualer')
plt.legend()
plt.show()
```

```
alphaHat:  -1160.499703659406
betaHat:  57.50667476417555
s2:  31649.06878660552
T:  -10.688008138743822
Kritsik verdi -0.685306279212829
```





b) vi ser at vi ligger langt utanfor den nedtre grensa, altså skal vi avvise H_0

c) Ein liten samling av det vi har funne ut.

- $y = \alpha + \beta x + \epsilon \Rightarrow \hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0 + \epsilon$
- $\epsilon \sim N(0, \sigma^2)$
- $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}}) = \frac{S_x y}{S_{xx}}$
- $\hat{\alpha} \sim N(\alpha, \sigma^2 \frac{\sum x_i^2}{n S_{xx}}) = \overline{y} - \hat{\beta}\overline{x}$
- $\frac{\sum x_i^2}{n S_{xx}} = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}$ fordi $\sum x_i^2 = n\overline{x}^2 + S_{xx}$

- Forventningsrett estimator for σ^2 er $S^2 = \frac{n\hat{\sigma^2}}{n-2} = \frac{1}{n-2}\sum(Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$, S^2 er uavhenging av $\hat{\alpha}$ og $\hat{\beta}$
- $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$

Vidare finner ci at vi kan skrive

$$\begin{aligned}\hat{y}_0 &= \hat{\alpha} + \hat{\beta}x_0 \\ Y_0 - \hat{y}_0 &= Y_0 - \hat{\alpha} - \hat{\beta}x_0\end{aligned}$$

Sidan $\hat{alpha}$ og $\hat{\beta}$ er lineørfunksjoner av $Y_1, Y_2, .., Y_n$ bir differansen $Y_0 - \hat{Y}_0$ også ein lineærfunksjon av $Y_1, Y_2, .., Y_n$ og Y_0 og Y_i er uavhengig og normalfirdelt.

Vidaare har vi då

$$\begin{aligned}E[Y_0 - \hat{Y}_0] &= E[Y_0 - \hat{\alpha} - \hat{\beta}x_0] \\ &= E[Y_0] - E[\hat{\alpha}] - x_0E[\hat{\beta}] \\ &= \alpha + x_0\beta - \alpha - x_0\beta = 0\end{aligned}$$

Altså er modellen forventnigsrett.

Vidare no bruker vi at $\hat{\alpha} = \overline{Y} - \overline{x}\hat{\beta}$

$$\begin{aligned}Var[Y_0 - \hat{Y_0}] &= Var[Y_0 - \overline{Y} + \hat{\beta}\overline{x} - \hat{\beta}x_0] \\ &= Var[Y_0] + Var[\overline{Y}] + (x_0 - \overline{x})^2 Var[\hat{\beta}] \\ &= \sigma^2 + \frac{\sigma^2}{n} + (x_0 - \overline{x})^2\frac{\sigma^2}{S_{xx}} \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\end{aligned}$$

Vi har no vist at vi kan sei at $Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\right)$

Den tilnærma standariseringa blir da

$$Z = \frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}} \sim N(0,1)$$

σ^2 er ukjent, så vi bytter den til den forventnigsrette estimatoren S^2, det betyr at vi mister 2 frihetsgrader.

Det kkan vi sjå ved at T kan skivas på formen $\frac{Z}{\sqrt{V/\nu}}, \nu = n - 2$

$$T = \frac{\frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}}}{\sqrt{\frac{\frac{(n-2)s^2}{\sigma^2}}{(n-2)}}}, \text{ der } \frac{(n-2)s^2}{\sigma^2} = V \sim \chi^2_{n-2}$$

Så ergo er $T \sim t_{n-2}$ vidare definerer vi dei kritiske verdiane i testen. $t_{\alpha/2,n-2}$ og $t_{1-\alpha/2,n-2}$ med $\alpha = 1 - 0.9 = 0.1$ og siden student t-fordelinga er symetrisk utnyttar vi at $t_{1-\alpha/2,n-2} = -t_{\alpha/2,n-2}$ Og vi kan da løyse for ulikheten

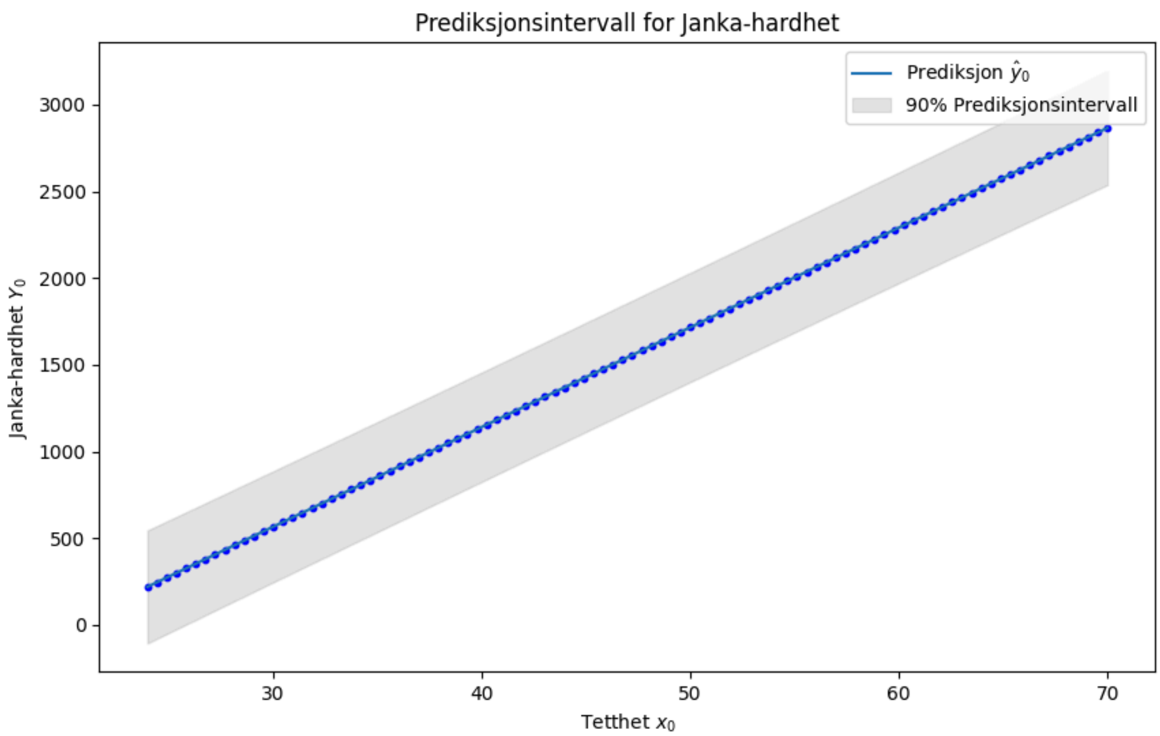
$$P\left(-t_{\alpha/2,n-2} \leq T \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$
$$\Rightarrow P\left(\hat{Y}_0 - t_{\alpha/2,n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2,n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \right.}\right.$$

Bruker så at $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$ og $S^2 = \frac{n\hat{\sigma}^2}{n-2}$

Då blir prediksjonsinteravallet

$$\hat{\alpha} +\hat{\beta}x_0 \pm t_{\alpha/2, n-2}\sqrt{\frac{n\hat{\sigma}^2}{n-2}\left(1 + \frac{1}{n} +\frac{(x_0-\overline{x})^2}{S_{xx}} \right)}

```
In [ ]: # Her kan du skrive python-kode for å gjøre beregningene du trenger for å besvare
from scipy import stats
hat_alpha = param[0]
hat_beta = param[1]
s2 = param[3]
n = len(x)
x_bar = np.mean(x)
S_xx = np.sum((x-x_bar)**2)
alpha = 0.10
t_critical = stats.t.ppf(alpha/2, n-2)  # t-verdien for 90% CI og n-2 frihetsgrade

x0 = np.linspace(24, 70, 100)
y_pred = hat_alpha + hat_beta * x0
SE = np.sqrt(s2 * (1 + 1/n + ((x0 - x_bar)**2) / S_xx))
lower_bound = y_pred - t_critical * SE
upper_bound = y_pred + t_critical * SE

plt.figure(figsize=(10, 6))
plt.plot(x0, y_pred, label='Prediksjon $\hat{y}_0$')
plt.fill_between(x0, lower_bound, upper_bound, color='gray', alpha=0.2, label='90%
plt.scatter(x0, y_pred, color='blue', s=10)  # Fiktivt spredningsplott
plt.title('Prediksjonsintervall for Janka-hardhet')
plt.xlabel('Tetthet $x_0$')
plt.ylabel('Janka-hardhet $Y_0$')
plt.legend()
plt.show()
abs(np.mean(SE))
```




```
Out[ ]: 188.09259483898578
```

Basert på prediksjonsintervallets bredde kan man vurdere om det er fornuftig å erstatte direkte målinger av Janka-hardheten med prediksjonene. Hvis intervallet er smalt nok gjennom hele området av interesser (for eksempel [24, 70] for tettheten), kan dette indikere at prediksjonen er nøyaktig nok for praktiske formål. Eit bredt interval kan derimot indikere at det er for mye usikkjerhet knyttet til prediksjonene, og direkte målinger vil være å foretrekke. Avgjørelsen avhenger av den akseptable feilmarginen for den spesifikkje anvendelsen. her har vi ein absoluttverdi på standaravviket på kring 188 ein eller anna enhet, som er nok så voldsom avvik fra den sanne verdien og direkte målinger bør være å foretrekke.

Fasit

- 1a: $L(a, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{(Y_i - a x_i)^2}{\sigma^2}\right\}$
- 1b: $\hat{a}$ er forventningsrett. $\mathrm{Var}[\hat{a}] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$
- 1c: $E[\widehat{\sigma}^2] = \frac{n-1}{n}\sigma^2$, $\mathrm{Var}[\tilde{\sigma}^2] = \frac{2}{n-1}(\sigma^2)^2$
- 3b: $E[\hat{\beta}] = \beta$, $\mathrm{Var}[\hat{\beta}] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$
- 3d: $\left[\hat{\beta} \pm t_{n-1,\alpha/2}\sqrt{\frac{n}{n-1}\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n} x_i^2}}\right]$
- 4a: $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$, $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\hat{\alpha} + \hat{\beta}x_i))^2$
- 4b: Forkast H_0.
- 4c: $\left[\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\alpha/2}\sqrt{\frac{n\widehat{\sigma}^2}{n-2}\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}\right]$

```
In [ ]: # Define a simple array of x values
x_values = np.array([24.7,24.8,27.3,28.4,28.4,29.0,30.3,32.7,35.6,38.5,38.8,39.3,
```

```python
                42.9,45.8,46.9,48.2,51.5,51.5,53.4,56.0,56.5,57.3,57.6,59.2,59.8,66.
n_values = len(x_values)
x_mean = np.mean(x_values)

# Calculate each term
sum_x_i_squared = np.sum(x_values**2)
sum_x_i_minus_x_mean_squared = np.sum((x_values - x_mean)**2)

# Calculate the first expression
first_expression = sum_x_i_squared / (n_values * sum_x_i_minus_x_mean_squared)

# Calculate the second expression
second_expression = (1/n_values) + (x_mean**2 / sum_x_i_minus_x_mean_squared)

# Check if they are equal
are_equal = np.isclose(first_expression, second_expression)

first_expression, second_expression, are_equal
```

Out[ ]: (0.35181309145468376, 0.35181309145468365, True)