Høgskolen i Molde
Vitenskapelig høgskole i logistikk

6. Statistisk inferens

Øving 6: Løsning

"*MAT110 Statistikk 1*", 2021

versjon 01

- Utlevering:
  - mandag 29. mars kl. 12:00
  - skriv ut oppgavene på papir, i farger, og ta med på øvingsdagen.

- Innlevering:
  - mandag 26. april kl. 12:00 mandag
  - dvs. man har 4 uker på å gjøre oppgavene
  - vi anbefaler at man prøver å bli ferdig etter 2 uker, siden ny øving kommer da

- FORarbeid:
  - Se videoene for uke 13, 14, 15 og 16
  - Les kompendiet.
  - Les gjennom øvingoppgave 6 så dere vet problemstillingene

- ETTERarbeid:
  - Gå gjennom løsning 6 som legges ut torsdag 8. april kl. 16:00.

Problem 6.1 — statistisk inferens - McDonald's

McDonald's i Molde ønsker å gjøre en analyse angående bemanning og jobbfordeling i rushperioden mellom 16:00 - 18:00 i ukedagene. De har oppdaget de smertefulle konsekvensene av å ha for få eller for mange ansatte i denne perioden:

- For få ansatte:
  gir lange køer med misfornøyde kunder som derfor heller drar til Burger King
- For mange ansatte:
  gir høye driftskostnader

For å kunne dimensjonere og fordele de ansatte i denne perioden, har de innsett at det er to typer data som i størst grad påvirker dimensjoneringen av bemanningen. [1]

- IAT - *Inter Arrival Time*:
  Tiden mellom to påfølgende kunder.
  "Kunde" betyr her en *gruppe* personer som kommer til McDonald's som en enhet. [2]

- Størrelsen på ordrene fra hele gruppen målt i bearbeidingstiden
  ordren krever fra de ansatte. [3]

Figure 6.1: McDonald's.

[1] "*Dimensjonering*" betyr her hvor *mange* ansatte de har.

[2] En familie er et typisk eksempel på en slik gruppe. McDonald's anser hele familien som "kunden". En vennegjeng er et annet eksempel på en slik gruppe.

[3] Dette er selve tiden det tar å lage ordren, dvs. eventuelle ventetider på grunn av køer ikke tatt med i disse tidene. Disse tidene er mao. kun beregnet fra selve ordren, ikke fra hvordan trykket var akkurat da ordren ble bestilt.

Før McDonald's starter selve analysen rundt dimensjoneringen av bemanningen, ønsker de å skaffe seg et godt statistisk utgangspunkt. De følger oppsettet:

1. Definere populasjonen.
2. Skaffe til veie data over en gitt periode. [4]
3. Beregne nøkkeltall, dvs. beskrivende statistikk for de to datasettene.
4. Formulere statistiske modeller for utvalgene og populasjonen.

a) Hva er populasjonsmengden i tilfellet hos McDonald's?

Løsning:

Populasjonsmengden er alle potensielle kunder hos McDonald's i Molde i hverdagene fra 16:00 til 18:00.

Du skal hjelpe McDonald's å samle inn forsøksdata. Dataene forekommer som *par*

$$z_i = (x_i, y_i) \tag{6.1}$$

hvor i står for kunde nr. i, hvor $i = 1, 2, 3, \ldots, n$.
For hver ankommet kunde i, måles altså tallene:

$$x_i = \text{tiden mellom kunde nr. } i \text{ og forrige kunde } i-1 \text{ (IAT) (antall minutter)} \tag{6.2}$$
$$y_i = \text{bearbeidingstiden for ordren til kunde nr. } i \text{ (antall minutter)} \tag{6.3}$$

b) Beskriv i korte trekk hvordan du vil gjennomføre datainnsamlingen slik at forsøksdataene $z_1, z_2, \ldots, z_n$ skal utgjøre et tilfeldig utvalg fra populasjonen.

[4] Målet er at dataene utgjør et tilfeldig utvalg.

Løsning:

Slik gjennomføres datainnsamlingen:

- observere påfølgende kunder mellom 16:00 og 18:00 på en gitt hverdag.
- Men hvilke hverdager? For at vi skal kunne få et *tilfeldig* utvalg må vi velge 100 hverdager, hvor hver hverdag har like stor sjanse for å bli valgt.
- I tillegg må vi luke ut spesielle dager hvor vi *vet* at etterspørselen enten er merkbart høyere eller lavere enn vanlig. Fridager eller kampanjedager er slike dager. Disse dagene er ikke representative for de dagene vi studerer.

Rent praktisk:

- Første ankommende kunde teller vi ikke, siden vi ikke har en tidligere kunde å regne IAT fra.
- Vi må notere ankomsttiden for neste ankommende gruppe. Vi kan da beregne IAT-tiden x_i for denne gruppen. Deretter må vi registrere hva *hele* gruppen bestilte, og beregne den totale bearbedingstiden y_i for alle ordrene.

Figure 6.2: Datainnsamling.

I tabell 6.1 har McDonald's samlet inn data fra kundene mellom 16:00 - 18:00 i hverdagene og målt x_i (IAT) og y_i (bearbeidingstiden) for $n = 100$ kunder.

Antall minutter:

(0.85 , 4.94)	(0.30 , 5.50)	(0.21 , 5.78)	(0.40 , 3.84)	(0.17 , 7.62)
(0.31 , 6.61)	(0.08 , 7.83)	(0.69 , 4.45)	(0.29 , 8.50)	(0.02 , 8.41)
(0.61 , 7.04)	(1.06 , 6.27)	(0.36 , 9.73)	(0.58 , 3.04)	(0.09 , 2.52)
(0.45 , 4.98)	(0.33 , 5.70)	(0.06 , 6.83)	(0.25 , 3.41)	(0.22 , 7.32)
(0.52 , 6.84)	(0.30 , 8.74)	(0.64 , 10.59)	(1.07 , 5.23)	(1.38 , 6.50)
(0.06 , 7.88)	(0.07 , 9.00)	(0.31 , 8.44)	(1.09 , 11.86)	(0.28 , 9.32)
(0.09 , 4.40)	(1.65 , 8.07)	(0.03 , 10.09)	(0.29 , 7.97)	(0.27 , 4.88)
(0.41 , 3.22)	(0.25 , 7.29)	(0.21 , 5.85)	(0.10 , 4.89)	(0.44 , 6.42)
(0.09 , 6.41)	(0.06 , 5.88)	(0.78 , 4.01)	(1.60 , 6.97)	(0.08 , 6.48)
(0.09 , 4.99)	(0.20 , 5.30)	(0.40 , 4.58)	(0.39 , 10.04)	(0.75 , 5.27)
(0.57 , 4.92)	(1.22 , 5.07)	(0.50 , 6.36)	(0.59 , 5.88)	(0.84 , 7.53)
(0.09 , 6.25)	(0.04 , 10.54)	(0.51 , 7.41)	(0.98 , 8.12)	(0.78 , 6.65)
(0.28 , 8.40)	(0.33 , 7.38)	(0.08 , 6.13)	(0.07 , 6.72)	(0.30 , 5.07)
(0.10 , 6.97)	(0.29 , 3.36)	(0.16 , 5.22)	(1.60 , 8.56)	(0.37 , 5.76)
(0.48 , 5.29)	(0.25 , 7.45)	(0.29 , 2.90)	(1.59 , 6.44)	(0.13 , 1.72)
(0.31 , 4.87)	(0.07 , 4.36)	(1.29 , 4.29)	(0.47 , 6.70)	(0.12 , 10.40)
(0.24 , 5.27)	(1.07 , 8.33)	(0.05 , 9.12)	(0.19 , 4.32)	(0.17 , 5.03)
(0.00 , 4.78)	(1.03 , 7.81)	(0.71 , 3.75)	(0.12 , 5.33)	(0.28 , 5.36)
(0.03 , 6.26)	(1.28 , 7.42)	(1.18 , 2.28)	(0.20 , 5,29)	(0.01 , 3.44)
(0.22 , 5.48)	(0.61 , 6.61)	(0.02 , 6.77)	(0.49 , 7.29)	(0.12 , 5.62)

Table 6.1: Dataene (x_i, y_i), hvor $i = 1, 2, 3, \ldots, 100$.

c) Lag stolpediagrammer av relativfrekvensene f_r for x_i (IAT) og y_i (bearbeidingstiden). [5] Del opp verdiområdet til x_i og y_i i følgende intervaller: [6]

- x_i - start fra 0 og lag intervaller med lengde 0.2 frem til 3:

$$[0.0, 0.2\rangle$$
$$[0.2, 0.4\rangle$$
$$\vdots$$
$$[2.8, 3.0\rangle$$

- y_i - start fra 0 og lag intervaller med lengde 1 frem til 12:

$$[0, 1\rangle$$
$$[1, 2\rangle$$
$$\vdots$$
$$[12, 13\rangle$$

Løsning:

Regner relativ frekvens f_r for intervallet $[0.0, 0.2)$: [7]

$$f_r(n_1) = \frac{n_1}{n} = \frac{34}{100} = 0.34 \tag{6.4}$$

Tilsvarende for alle andre intervall.

[5] Altså ett stolpediagram for x_i og ett stolpediagram for y_i.

[6] Beregn relativ frekvensene f_r for hvert intervall og plott dem i et stolpediagram. Se kapittel 6 i boksen. Vi regnet der ut relativfrekvenser f_r for eksemplet med legemiddel.

[7] At det er 34 kunder som har AIT-tid i intervallet $[0.0, 0.2)$ finner man ved å telle. Men bruk gjerne Excel-filen som ligger på Canvas. Sorter, og deretter tell. Eller få Excel til å gjøre tellingen for dere.

Intervall	Relativ frekvens f_r
$[0.0,\ 0.2\rangle$	0.34
$[0.2,\ 0.4\rangle$	0.28
$[0.4,\ 0.6\rangle$	0.13
$[0.6,\ 0.8\rangle$	0.08
$[0,8,\ 1.0\rangle$	0.03
$[1.0,\ 1.2\rangle$	0.06
$[1.2,\ 1.4\rangle$	0.04
$[1.4,\ 1.6\rangle$	0.02
$[1.6,\ 1.8\rangle$	0.02
$[1.8,\ 2.0\rangle$	0.00
$[2.0,\ 2.2\rangle$	0.00
$[2.2,\ 2.4\rangle$	0.00
$[2.4,\ 2.6\rangle$	0.00
$[2.6,\ 2.8\rangle$	0.00
$[2.8,\ 3.0]$	0.00

Table 6.2: f_r for x_i-ene (AIT).

Intervall	Relativ frekvens f_r
$[0,\ 1\rangle$	0.00
$[1,\ 2\rangle$	0.01
$[2,\ 3\rangle$	0.03
$[3,\ 4\rangle$	0.07
$[4,\ 5\rangle$	0.15
$[5,\ 6\rangle$	0.21
$[6,\ 7\rangle$	0.20
$[7,\ 8\rangle$	0.14
$[8,\ 9\rangle$	0.10
$[9,\ 10\rangle$	0.03
$[10,\ 11\rangle$	0.05
$[11,\ 12\rangle$	0.01
$[12,\ 13\rangle$	0.00

Table 6.3: f_r for y_i-ene (bearbeidingstid).

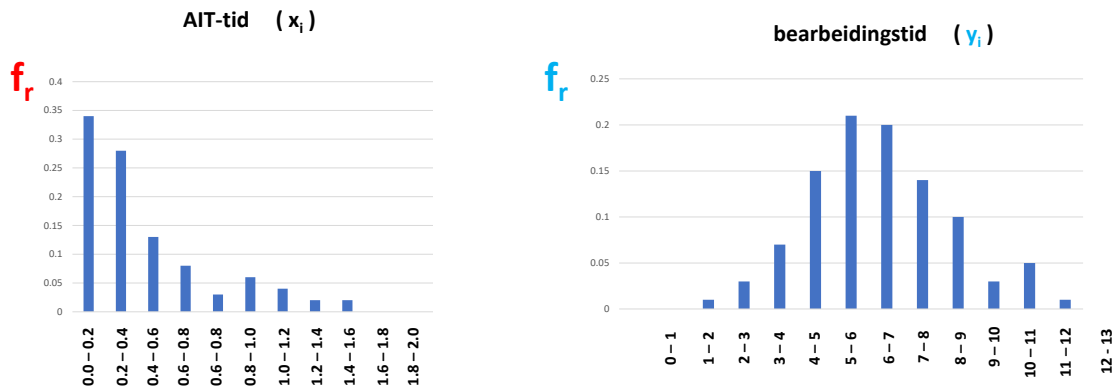


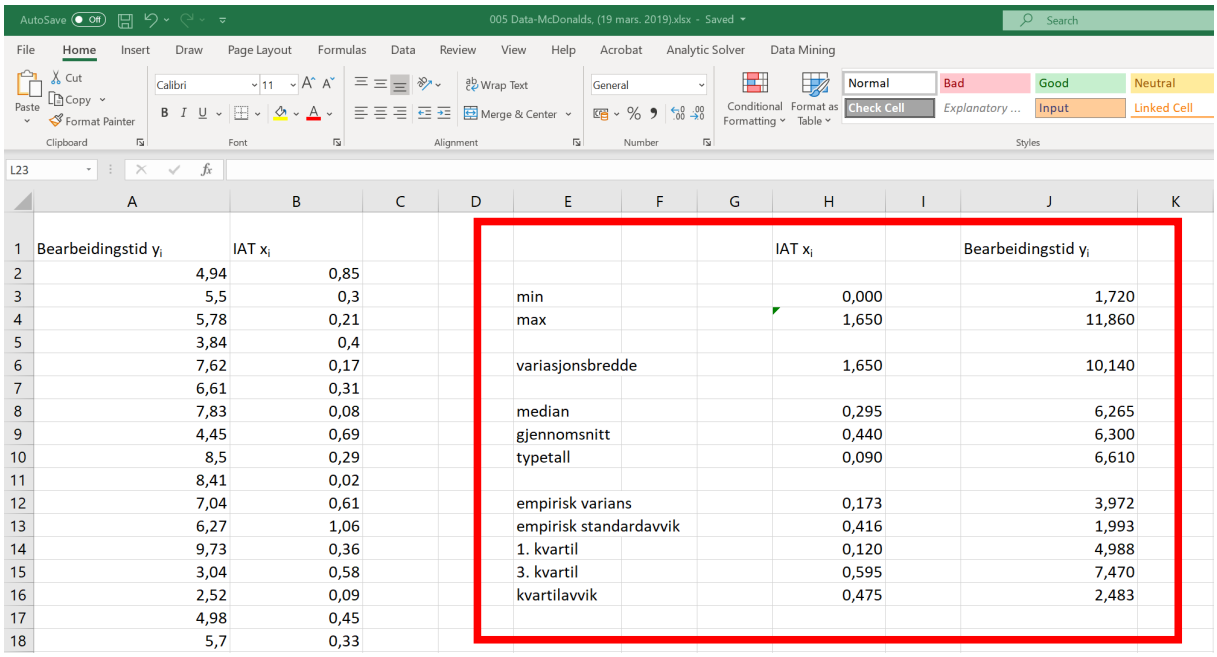
Figure 6.3: f_r-stolpediagram for IAT-tidene (x_i) og bearbeidingstidene (y_i).

d) Beregn beskrivende nøkkeltall for både IAT og bearbeidingstiden. Inkluder følgende størrelser: [8]

- min
- maks
- variasjonsbredde
- median
- gjennomsnitt
- typetall
- empirisk varians
- empirisk standardavvik
- 1. kvartil (25%)
- 3. kvartil (75%)
- kvartilavvik

Løsning:

Bruker Excel-filen som ligger på Canvas. Da kan man regne ut st o rrelsene som oppgaven ber om: [9]



	A	B	C	D	E	F	G	H	I	J	K
1	Bearbeidingstid y_i	IAT x_i						IAT x_i		Bearbeidingstid y_i	
2	4,94	0,85									
3	5,5	0,3			min			0,000		1,720	
4	5,78	0,21			max			1,650		11,860	
5	3,84	0,4									
6	7,62	0,17			variasjonsbredde			1,650		10,140	
7	6,61	0,31									
8	7,83	0,08			median			0,295		6,265	
9	4,45	0,69			gjennomsnitt			0,440		6,300	
10	8,5	0,29			typetall			0,090		6,610	
11	8,41	0,02									
12	7,04	0,61			empirisk varians			0,173		3,972	
13	6,27	1,06			empirisk standardavvik			0,416		1,993	
14	9,73	0,36			1. kvartil			0,120		4,988	
15	3,04	0,58			3. kvartil			0,595		7,470	
16	2,52	0,09			kvartilavvik			0,475		2,483	
17	4,98	0,45									
18	5,7	0,33									

Figure 6.4: Størrelser som beskriver nøkkeltall.

[8] Bruk gjerne Excel for å regne ut disse nøkkeltallene. En Excel-fil med tallene fra tabell 6.1 ligger på Canvas. Da blir det ikke så mye arbeid. Oppgi tallene med 3 desimalers nøyaktighet.

[9] I dette utklippet fra Excel så er det konsekvent valgt 3 desimaler. Dette kan kan justeres i Excel.

Definer populasjonsvariablene:

$$X = \text{tiden mellom to påfølgende kunder i tidsrommet 16:00-18:00 på en gitt hverdag} \tag{6.5}$$

$$Y = \text{bearbeidingstiden til en kunde i i tidsrommet 16:00-18:00 på en gitt hverdag} \tag{6.6}$$

Definer tilhørende forsøksvariabler:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \tag{6.7}$$

e) Er

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X \tag{6.8}$$

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \tag{6.9}$$

rimelig å anta?
Gi en kort begrunnelse.

Løsning:

Identiske fordelte variabler:

At X_i-ene er identisk fordelt lik populasjonsvariabelen X er en grei antagelse å ta siden vi har luket ut alle dagene fra trekningen som *ikke* er representative for de dagene vi ønsker å analysere.

Tilsvarende for Y_i-ene og populasjonsvariabelen Y.

Uavhengige forsøksvariabler X_i og Y_i:

At X_i-ene er gjensidig *uavhengige* betyr at to påfølgende kunder *ikke* er koblet mot hverandre. Et brudd på denne antagelsen fås f.eks. dersom det kommer en vennegjeng. Siden vennene følger hverandre så er inter-arrival tidene den samme, dvs. inter-arrival tidene deres være avhengig av hverandre. Men personer som kommer samlet som *en* kunde. Dermed er uavhengigheten til X_i-ene rimelig å anta.

Tilsvarende for bearbeidingstidene Y_i.

f) Er det rimelig å anta at X_i-ene er gjensidig uavhengige av Y_i-ene?
Gi en kort begrunnelse.

Løsning:

X_i-ene er gjensidig uavhengige av Y_i:

Kunden endrer ikke bestillingen etter *når* forrige kunde ankom.
Antagelsen om uavhengighet er dermed rimelig.

g) På bakgrunn av kunnskapen vi har om McDonald's-eksemplet samt de beskrivende nøkkeltallene for observasjonene, formuler to statistiske modeller:

(a) Statistisk modell for IAT - Inter-Arrival Times (X_i-ene)
(b) Statistisk modell for bearbeidingstiden (Y_i-ene)

som du mener representerer situasjonen godt. [10]

Løsning:

Statistisk modell for X_i-ene: (inter-arrivel tidene)

Stolpediagrammet for x_i-ene i figur 6.3 indikerer at tetthetsfunksjonen synker raskt mot null for voksende inter-arrival tider. Vi kan spekulere i at den synker *eksponentielt*.

[10] Se på stolpediagrammene og se om du gjenkjenner noen kjente sannsynlighetsfordelinger. Søk også på internett for å se hva andre har gjort for IAT - Inter Arrival Times.

Vi søker på nettet på ordene ”exponential’ og ”inter arrival time” Her er en link til Wikipedia om eksponentialfordelingen. Vi innser fra nettsiden at eksponential fordelingen er en ofte brukt fordeling for å modellere IAT, hvor kundene ankommer uavhengig av hverandre. Vi foreslår derfor følgende modell: [11]

Den statistiske modellen for de stokastiske forsøksvariablene $X_1, X_2, \ldots, X_{100}$ og populasjonsvariabelen X er gitt ved:

$$\boxed{X_1, X_2, X_3, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X \sim \text{Exp}[\lambda]_{\lambda>0}} \qquad (6.10)$$

hvor

$$(0,\infty) = \text{verdimengden } V \text{ til de stokastiske variablene} \qquad (6.11)$$

$$\text{Exp}[\lambda]_{\lambda>0} = \text{familien av eksponential-fordelinger med parameter } \theta = \lambda \qquad (6.12)$$

$$(0,\infty) = \text{parametermengden } \Theta, \qquad (6.13)$$
$$\text{dvs. de mulige verdiene for parameteren } \theta = \lambda \text{ som er } \lambda > 0$$

[11] Dette med statistisk modell er litt teknisk og litt vanskelig. I kompenidet er det utarbeidet eksempler som er analoge til oppgavene i denne øvingen. Derfor er det både viktig og lurt å gå gjennom teorien først, og deretter gjøre øvingen.

Statistisk modell for Y_i-ene: (bearbeidingstidene)

Stolpediagrammet for y_i-ene i figur 6.3 indikerer at tetthetsfunksjonen følger en *normalfordeling*.
Vi foreslår derfor følgende modell:

Den statistiske modellen for de stokastiske forsøksvariablene $Y_1, Y_2, \ldots, Y_{100}$ og populasjonsvariabelen Y er gitt ved:

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Y \sim N[\mu, \sigma]_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} \tag{6.14}$$

hvor [12]

$$\overbrace{\mathbb{R}}^{\text{reelle tall}} = \text{verdimengden } V \text{ til de stokastiske variablene} \tag{6.15}$$

$$N[\mu, \sigma]_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} = \text{familien av normalfordelinger med parametre } \mu \text{ og } \sigma \tag{6.16}$$

$$\mathbb{R}\times\mathbb{R}_+ = \text{parametermengden } \Theta, \text{ dvs. de mulige verdiene for parameteren } \mu \text{ og } \sigma \tag{6.17}$$

■

[12] $\mathbb{R}_+$ = alle positive reelle tall.
Elementene i $\mathbb{R}\times\mathbb{R}_+$ er alle par (μ, σ), hvor $-\infty < \mu < \infty$ og $\sigma > 0$.

Høgskolen i Molde
Vitenskapelig høgskole i logistikk

11. Regresjonsanalyse

Problem 11.1 — regresjon

a) Hva er regresjonsanalyse? [1]

Løsning:

Regresjonsanalyse er teori og metoder for å analysere og utnytte samvariasjon mellom variable.

b) Hva er formålet med regresjonsanalyse? [2]

Løsning:

Formålet med regresjonsanalyse er å konstruere modeller som kan brukes til å anslå verdien ("prediksjon/forutsi") av en variabel Y ved hjelp av informasjon om en annen variabel X.

[1] Her trengs kun et *kort* svar. Bruk internett, boken eller noe annet du synes er egnet å slå opp i dersom du ikke har det i hodet.

[2] Se boken eller andre se steder. I dette kurset ser vi kun på tilfellet når vi har én uavhengig variabel.

c) Hvilke to typer regresjon skiller man ofte mellom?

Løsning:

Man skiller ofte mellom lineær regresjon og ikke-lineær regresjon.

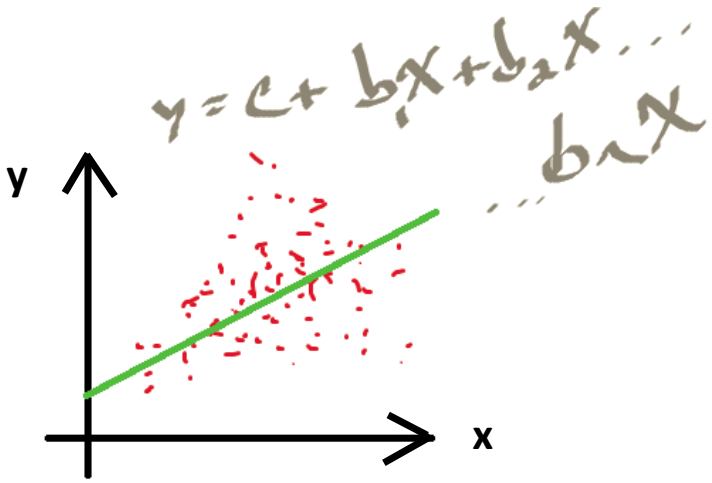


Figure 11.1: Regresjon.

■

Problem 11.2 — økonomi - regresjonsanalyse

EiendomsMegler 1 Midt-Norge ønsker å se nærmere på sammenhengen mellom areal x og pris y på leiligheter. De gjør observasjoner av x og tilhørende y:

$$x = \text{areal} \tag{11.1}$$
$$y = \text{pris} \quad (\text{i } 1000 \text{ NOK}) \tag{11.2}$$

Eiendomsfirmaet solgte $n = 6$ leiligheter i mars 2018. Areal og pris for disse leilighetene er oppsummert i følgende tabell:

x (areal i m^2)	43	60	75	80	95	105
y pris (i 1000 NOK)	2100	2850	3050	3800	4525	4500

Figure 11.2: Areal x og pris y.

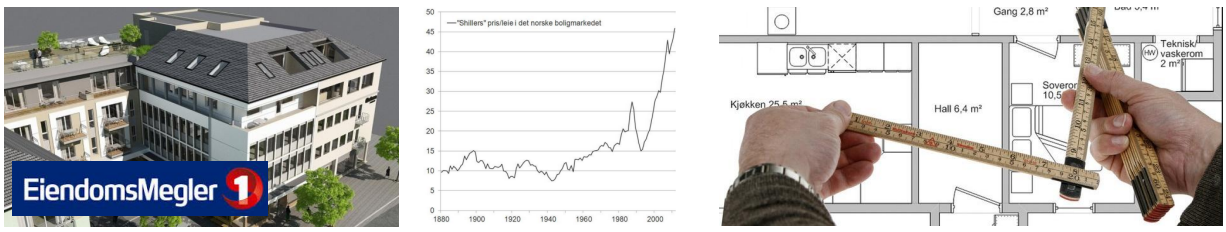


Figure 11.3: Areal og pris.

PS:
Man kan fint løse denne oppgaven kun ved bruk av vanlig kalkulator. Uten bruk av Excel. Men denne oppgaven er et eksempel på at et dataprogram som f.eks. Excel kan brukes i statistikksammenheng. Derfor er det laget en Excel-fil:

006 Øving 6, oppgave 11.2, pris-areal, (29. mars 2021).xlsx

Denne Excel-filen ligger på Canvas. Direktelink finner du her. Selv om det er frivillig å bruke Excel så kan det være en fin måte å sjekke dine svar på i denne oppgaven. I tillegg så illustrerer det, til en viss grad, nytten av dataprogrammer innen statistikk.

a) Hva er gjennomsnittlig areal $\bar{x}$ av leilighetene? Og gjennomsnittspris $\bar{y}$?

Løsning:

Gjennomsnittlig areal $\bar{x}$ av leilightene:

$$\bar{x} \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i \tag{11.3}$$

$$= \frac{1}{6}\Big(43+60+75+80+95+105\Big)\ \text{m}^2 = \ 76.33\ \ \text{m}^2 \tag{11.4}$$

Gjennomsnittlig pris $\bar{y}$ på leilightene: (i 1000 NOK)

$$\bar{y} \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i \tag{11.5}$$

$$= \frac{1}{6}\Big(2100+2850+3050+3800+4525+4500\Big)\ \text{NOK} \tag{11.6}$$

$$= 3470.83\ \ \text{NOK} \tag{11.7}$$

b) Hva er den empiriske variansen for arealet x, dvs. hva er S_x^2? Og for prisen, S_y^2?

Løsning:

Empirisk varians for arealet x:

$$S_x^2 \overset{\text{def.}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{11.8}$$

$$= \frac{1}{6-1}\Big((43-76.33)^2+(60-76.33)^2+(75-76.33)^2 + (80-76.33)^2+(95-76.33)^2+(105-76.33)^2\Big)\ (\text{m}^2)^2$$

$$= 512.67\ \ (\text{m}^2)^2 \tag{11.9}$$

Empirisk varians for prisen y: (i 1000 NOK)

$$S_y^2 \overset{\text{def.}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{11.10}$$

$$= \frac{1}{6-1}\Big((2100-3470.83)^2+(2850-3470.83)^2+(3050-3470.83)^2 + (3800-3470.83)^2+(4525-3470.83)^2+(4500-3470.83)^2\Big)\ \text{NOK}^2$$

$$= 944\,104.17\ \ \text{NOK}^2 \tag{11.11}$$

c) Hva er den empiriske kovariansen (samvariasjon) mellom x og y, dvs. hva er S_{xy}?

Løsning:

Empirisk kovarians (samvariasjon) mellom x og y: (i 1000 NOK)

$$S_{xy} \overset{\text{def.}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{11.12}$$

$$= \frac{1}{6-1}\Big((43-76.33)(2100-3470.83)+(60-76.33)(2850-3470.83) + (75-76.33)(3050-3470.83)+(80-76.33)(3800-3470.83) + (95-76.33)(4525-3470.83)+(105-76.33)(4500-3470.83)\Big)\ \text{m}^2\,\text{NOK}$$

$$= 21\,356.67\ \ \text{m}^2\,\text{NOK} \tag{11.13}$$

d) Finn minste kvadraters **regresjonslinje** for x og y. [3]

Løsning:

Vi bruker minste kvadraters **regresjonslinje** for x og y.
Parametrene $\hat{\beta}$ og $\hat{\alpha}$ er da: (dropper benevning her)

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} = \frac{21\,356.67}{512.67} \approx 41.66 \tag{11.14}$$

$$\hat{\alpha} = \bar{y} - \beta\bar{x} = 3470.83 - 41.66\cdot 76.33 \approx 290.94 \tag{11.15}$$

Minste kvadraters lineære regresjonslinje $\hat{y} = \hat{\alpha} + \hat{\beta}\,x$ blir dermed: (se boken)

$$\hat{y} = 290.94 + 41.66x \tag{11.16}$$

[3] Bruk boken.

e) Den største leiligheten i tabellen i figur 11.2 er "bare" 105 m^2.
Dersom eiendomsmeglerfirmaet ønsker å *estimere* hvor mye en leilighet på f.eks. 140 m^2 vil koste basert på slagstallene fra mars så kan de bruke den estimerte modellen fra oppgave 11.2 **d**), altså regresjonslinjen.

Hvor mye predikerer regresjonslinjen at en leilighet på 140 m^2 vil koste?

Løsning:

Regresjonslinjen i lign.(11.16) predikerer at en leilighet på 140 m^2 vil koste:
(i 1000 NOK)

$$\hat{y}(140) = \Big(290.94 + 41.66 \cdot 140\Big) \text{ NOK} = 6123.34 \text{ NOK} \qquad (11.17)$$

altså litt over 6.1 mill. NOK.

f) Finn forklaringsstyrken R^2 uten å gjøre noe regning "for hånd".
Bare les av fra Excel-utskriften i figur 11.4. [4]

Løsning:

Forklaringskraften R^2 kan leses direkte fra Excel-utskriften:
(Se cellen som heter "*R Square*" i Excel-utskriften): [5]

$$R^2 = 0.9423 \qquad (11.18)$$

[4] Hint: Se i boken.

[5] Man kan også regne ut forklaringskraften R^2 "for hånd" via definisjonen $R^2 = 1 - SSE/SST$. Men det er mye mer arbeidskrevende. Fint at dataprogrammer (som f.eks. Excel) kan hjelpe oss med slikt.

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,970746809							
R Square	0,942349368							
Adjusted R Square	-1,5							
Standard Error	260,8356815							
Observations	1							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	6	4448379,822	741396,637	65,38345404	#NUM!			
Residual	4	272141,0111	68035,25276					
Total	10	4720520,833						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
X Variable 5	290,9395319	407,4209876	0,714100502	0,514621858	-840,2424749	1422,121539	-840,2424749	1422,121539
X Variable 6	41,6579974	5,151864817	8,086003589	0,001271107	27,35412755	55,96186725	27,35412755	55,96186725

Figure 11.4: Utskrift fra Excel.

g) Kommenter svaret i oppgave 11.2 **f**). [6]

Løsning:

Kommentar:

At $R^2 = 0.9423$ betyr at for ei leilighet med et gitt areal x så kan vi "i stor grad", tilsvarende 94.23 %, forutsi/predikere prisen. Vi sier at modellen har stor forklaringskraft.

h) Bruk **Excel** til å plotte **regresjonslinjen** fra oppgave 11.2 **d**).

Skriv ut ditt Excel-plott og legg ved i din innlevering. Din Excel-utskrift skal se omtrent ut som figur 11.6, men med regresjonslinjen fra oppgave 11.2 **d**) i tillegg. [7]

PS:
Usikker på hvordan man lager regresjonsplott i Excel? Se video HER.

[6] For et gitt areal, vi du si at regresjonslinjen predikerer prisen i stor eller liten grad? Med stort eller lite presisjonsnivå?

[7] Tittel på x-aksen kan f.eks. være "Areal (i m^2)". Tittel på y-aksen kan f.eks. være "Pris (i 1000 NOK)".

Løsning:

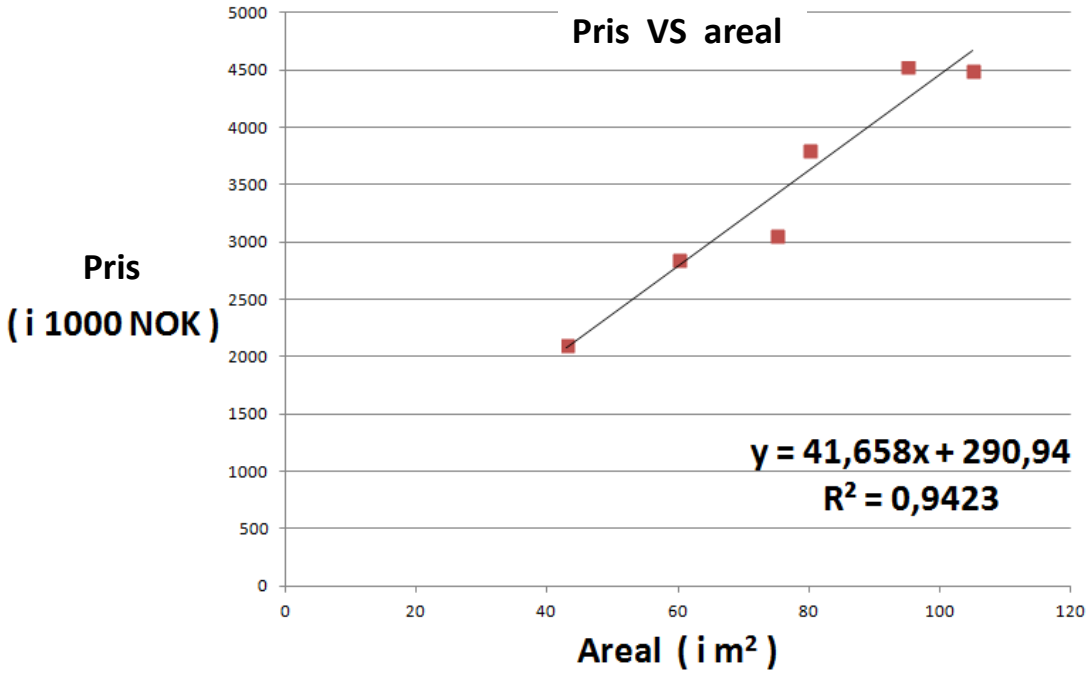


Figure 11.5: Utskrift fra Excel.

■

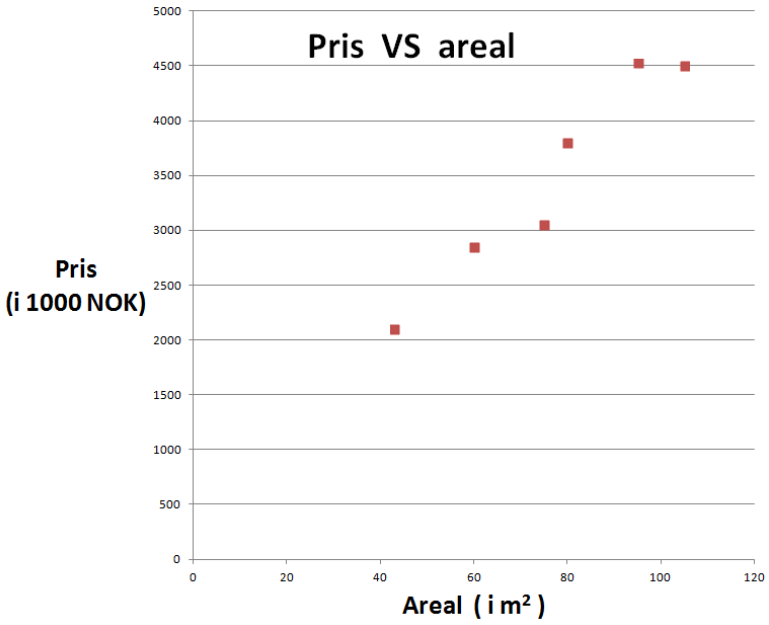


Figure 11.6: Excel.