Universitetet i Stavanger

EKSAMEN I: STA100 SANNSYNLIGHETSREGNING OG STATISTIKK

VARIGHET: 4 TIMER DATO: 11 SEPTEMBER, 2015

HJELPEMIDLER: Godkjent enkel kalkulator (HP30S, Casio FX82, TI-30, Citizen SR-270X , Texas BA II Plus eller HP17bII+).

EKSAMEN BESTÅR AV 5 OPPGAVER PÅ 6 SIDER OG 5 SIDER VEDLEGG, TOTALT 11 SIDER.

EMNEANSVARLIG: Jan Terje Kvaløy TELEFON: 51 83 22 55

Oppgave 1

En bedrift produserer rør som settes sammen til gassrørledninger. Bedriften har en maskin som produserer rør, som til en bestemt rørledning skal være omtrent 12 meter lange. Lengden til rør produsert av maskinen er normalfordelte. I første omgang (punkt a) og b)) antar vi at lengden til rørene har forventning $\mu = 12$ og standardavvik $\sigma = 0.1$ meter.

a) Forklar kort hva det i praksis betyr at lengden til rørene har forventing $\mu = 12$.
Hva er sannsynligheten for at et tilfeldig valgt rør er lengre enn 12.2 meter?
Hva er sannsynligheten for at lengden til et tilfeldig valgt rør er mellom 11.9 og 12.1 meter?

Bedriften har fått i oppdrag å produsere en 10.8 kilometer (dvs 10800 meter) lang rørledning. De har en feilmargin på $\pm$ 5 meter, det vil si at de kan levere en rørledning som er inntil 5 meter kortere eller 5 meter lengre enn den spesifiserte lengden på 10800 meter. Bedriften tenker å produsere 900 rør til rørledningen, og totallengden på rørledningen blir da summen av lengdene til disse rørene. Anta at hvert rør produsert av maskinen til bedriften har en lengde som er uavhengig av de andre rørene.

b) Hva er forventet totallengde på rørledningen?
Hva er variansen til totallengden?
Hva er sannsynligheten for at rørledningen blir for lang eller for kort (dvs får en lengde utenfor tillatt feilmargin)?
Dersom standardavviket til lengdene på rørene, σ, kunne justeres til en annen verdi, hvilken verdi måtte det justeres til for at sannsynligheten for at rørledningen blir for lang eller for kort skal bli 0.01?

Før produksjonen settes i gang for fullt insisterer en av ingeniørene ved bedriften på at de må undersøke om maskinen er korrekt justert slik at forventet lengde av rørene virkelig er 12 meter. For å undersøke dette produserer de 12 rør og måler lengden til disse. Resultatene av målingene er gitt under. Vi antar i resten av oppgaven at μ er ukjent, mens vi fremdeles har at $\sigma = 0.1$ og at lengdene på ulike rør er uavhengige.

Måleresultater:

11.87	11.82	11.99	12.01	11.93	11.98	12.08	12.11	11.92	11.79	12.02	12.07

c) Regn ut gjennomsnitt og median til dataene.
Finn et 95% konfidensintervall for μ.
Hvor mange målinger må man i denne situasjonen minst gjøre dersom man ønsker et 95% konfidensintervall for μ med lengde på maksimalt 0.02?

Oppgave 2

Et oljeserviceselskap deltar i konkurransen om to lignende oppdrag på to ulike felt på Norsk sokkel. Vi definerer følgende hendelser:

$$A = \text{firmaet får oppdraget på felt A.}$$
$$B = \text{firmaet får oppdraget på felt B.}$$

Ut fra opplysningene selskapet sitter inne med regner de med at $P(A) = 0.35$, $P(B) = 0.45$ og $P(A \cap B) = 0.30$.

a) Er hendelsene A og B uavhengige? Begrunn svaret.
Hva er sannsynligheten for at firmaet får minst ett av oppdragene?
Hva er sannsynligheten for at firmaet får ingen av oppdragene?

Selskapet regner med at oppdraget på felt A vil gi et overskudd på 10 millioner, mens oppdraget på felt B vil gi et overskudd på 20 millioner.

La O = totalt overskudd fra de to mulige oppdragene.

b) Vis at fordelingen til O blir:

o	0	10	20	30
$P(O = o)$	0.50	0.05	0.15	0.30

Regn ut forventa totalt overskudd.
Regn også ut standardavviket til det totale overskuddet.

Oppgave 3

En produsenten av bærebjelker har utviklet en helt ny type bjelker. Bæreevnen til bærebjelkene (målt i hvor mange tonn belastning de tåler) antas normalfordelt, men med ukjent forventningsverdi og ukjent standardavvik. Bæreevnen til ulike bjelker antas uavhengig. For å undersøke egenskapene til de nye bjelkene har de testet bæreevnen til ti slike bjelker. Resultatene ble:

10.3 10.6 10.3 11.1 10.1 10.8 10.5 10.1 10.3 10.6

Det oppgis at $\sum_{i=1}^{10} x_i = 104.7$ og $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 0.901$, der x_i er bæreevnen til bjelke i.

a) Regn ut et estimat for σ.
Finn et 95% konfidensintervall for σ.
Forklar kort hvorfor det for praktisk bruk av bærebjelkene er viktig å ha et godt estimat for σ.

Oppgave 4

I et borettslag med mange omtrent like store leiligheter fikk ni leilighetene i løpet av sommeren 2014 installert varmepumpe. Varmepumpe er en energi-effektiv form for oppvarming som kan lede til lavere strømforbruk. For de ni leilighetene ble strømforbruket i januar 2014 (før installering av varmepumpe) og i januar 2015 (etter installering av varmepumpe) sammenlignet. Disse tallene samt differansen fra 2014 til 2015 er gitt i tabellen under.

leilighet i	1	2	3	4	5	6	7	8	9
forbruk 2014, x_i	1745	1236	1521	921	1681	1581	1291	1321	1568
forbruk 2015, y_i	1653	1106	1417	963	1589	1298	1167	1361	1235
differanse, d_i	92	130	104	-42	92	283	124	-40	333

Det oppgis at $\sum_{i=1}^{9} d_i = 1076$ og $\sum_{i=1}^{9}(d_i - \bar{d})^2 = 125720$.

a) Er forventet strømforbruk redusert fra januar 2014 til januar 2015? Formuler dette som en hypotesetest, og utfør testen. Bruk 5% signifikansnivå.
Hva betyr resultatet av testen i praksis? Kan vi fra dette konkludere med at installering av varmepumper gir redusert strømforbruk?
Hva kalles den typen forsøk som er utført her? Nevn en annen type fremgangsmåte som kunne vært brukt her for å evaluere effekten av installering av varmepumpe, og diskuter kort fordeler og ulemper med de to ulike fremgangsmåtene.

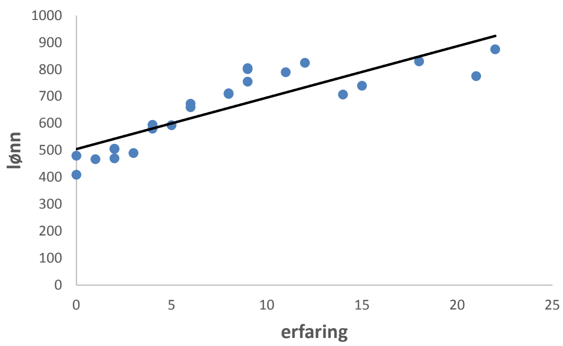
Oppgave 5

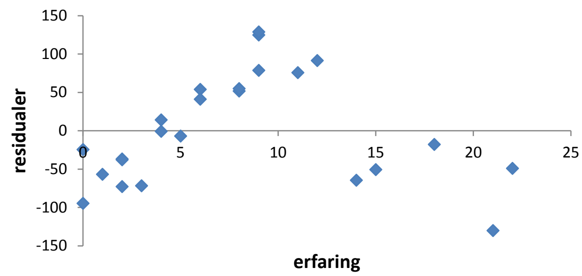
I en kartlegging av grunnlønn (i tusen kroner) og antall års erfaring for ansatte i en bestemt bransje med bachelorgrad i ingeniørfag fant man tallene gitt i tabellen under.

erfaring, x_i	1	0	4	22	8	2	15	0	2	18	4	2
lønn, y_i	467	410	595	875	709	506	740	480	505	830	580	470
erfaring, x_i	6	9	11	8	21	14	9	3	12	5	6	9
lønn, y_i	673	755	790	712	775	707	801	490	825	593	660	805

Det oppgis at $\sum_{i=1}^{24} x_i = 191$ og $\sum_{i=1}^{24} y_i = 15753$.

To plott knyttet til en enkel lineær regresjonsmodell for dataene er vist under.





Datautskriften vi får når vi bruker Excel til å tilpasse en enkel lineær regresjonsmodell til dataene på forrige side er vist under.

Regresjonsstatistikk	
Multippel R	0,86
R-kvadrat	0,75
Justert R-kvadrat	0,74
Standardfeil	72,4
Observasjoner	24

Variansanalyse

	fg	*SK*	*GK*	*F*	*Signifkans-F*
Regresjon	1	340608,4	340608,4	65,0	0,00000
Residualer	22	115369,2	5244,1		
Totalt	23	455977,6			

	Koeffisienter	*Standardfeil*	*t-Stat*	*P-verdi*	*Nederste 95%*	*Øverste 95%*
Skjæringspunkt	504,6	23,9	21,1	0,00000	455,0	554,3
erfaring	19,1	2,37	8,06	0,00000	14,2	24,0

a) Skriv ned den generelle enkle lineære regresjonsmodellen, og gi praktiske tolkninger av α og β-parametrene i regresjonsmodellen i denne situasjonen.
Hvorfor er det rimelig å bruke erfaring som x-variabelen og lønn som Y-variabelen når vi tilpasser en regresjonsmodell til disse dataene?
Hvordan er residualet til en observasjon definert? Regn ut residualet til den siste observasjonen ($x_{24} = 9,\ y_{24} = 805$).
Hvilke antagelser om modellen kan vi sjekke ut fra plottet av residualer gitt på forrige side? Ser disse antagelsene ut til å være oppfylte?

I stedet for den enkle lineære modellen skal vi nå prøve en modell med et andregradspolynom:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

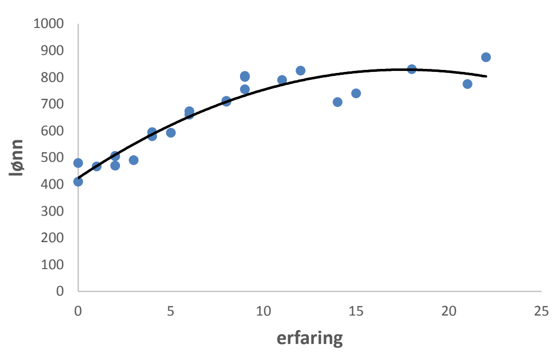
der $e \sim N(0, \sigma)$ og vi antar at $e_1, \ldots, e_n$ er uavhengige. Datautskrift og plott er vist under.

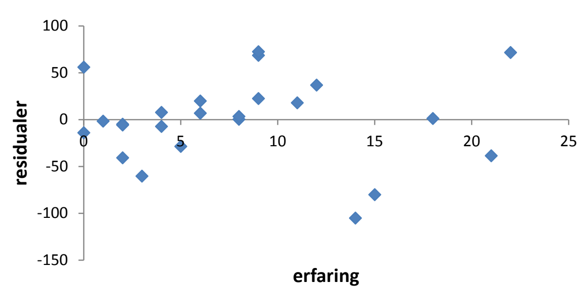
Regresjonsstatistikk	
Multippel R	0,95
R-kvadrat	0,90
Justert R-kvadrat	0,89
Standardfeil	47,0
Observasjoner	24

Variansanalyse

	fg	*SK*	*GK*	*F*	*Signifkans-F*
Regresjon	2	409632,0	204816,0	92,8	0,00000
Residualer	21	46345,7	2206,9		
Totalt	23	455977,6			

	Koeffisienter	*Standardfeil*	*t-Stat*	*P-verdi*	*Nederste 95%*	*Øverste 95%*
Skjæringspunkt	423,9	21,2	20,0	0,00000	379,8	468,0
erfaring	46,1	5,07	9,09	0,00000	35,5	56,6
erfaring2	-1,31	0,23	-5,59	0,00002	-1,80	-0,82





b) Tyder plottene på forrige side på at modellen med andregradspolynom er bedre enn den enkle lineære modellen? Forklar.
Hvilke andre plott av residualene bør vi lage, og hvilke antagelser sjekker man med disse plottene?
Regn ut estimert forventet lønn for en person med $x = 10$ års erfaring.
Er det grunn til å tro at den estimerte modellen med andregradspolynom vil være en god modell for å estimere forventet lønn for personer med mer enn 25 års erfaring? Forklar kort.

Til slutt prøver vi å tilpasse en modell med tredjegradspolynom:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

Datautskrift for denne modellen er vist under.

Regresjonsstatistikk	
Multippel R	0,95
R-kvadrat	0,91
Justert R-kvadrat	0,89
Standardfeil	46,3
Observasjoner	24

Variansanalyse

	fg	*SK*	*GK*	*F*	*Signifkans-F*
Regresjon	3	413165,1	137721,7	64,3	0,00000
Residualer	20	42812,5	2140,6		
Totalt	23	455977,6			

	Koeffisienter	*Standardfeil*	*t-Stat*	*P-verdi*	*Nederste 95%*	*Øverste 95%*
Skjæringspunkt	405,0	25,5	15,9	0,00000	351,7	458,3
erfaring	59,7	11,73	5,09	0,00006	35,2	84,2
erfaring2	-3,06	1,38	-2,22	0,039	-5,94	-0,18
erfaring3	0,055	0,043	1,28	0,214	-0,03	0,14

c) Skriv ned den observerte verdien på R^2 og forklar kort hva denne forteller oss.
Kan vi bruke R^2 til å sammenligne de tre forskjellige modellene? Hvorfor/hvorfor ikke?
Hvilket annet mål kan vi alternativt bruke til å sammenligne modellene? Hvilken modell ser ut fra dette målet ut til å være best?

Har variablene x, x^2 og x^3 (erfaring, erfaring^2 og erfaring^3) samlet sett innvirkning på Y-variabelen (lønn)? Formuler dette som en hypotesetest og gi resultatet av testen.

Trenger vi å ha med tredjegradsleddet (x^3) i modellen? Formuler dette som en hypotesetest og gi resultatet av testen.

Hvilken modell er best?