Norges miljø- og biovitenskapelige universitet

Fakultet: **KBM**

Eksamen i: **STAT100 STATISTIKK**

Tidspunkt: **Fredag 18. mai 2018** **14.00 – 17.30 (3.5 timer)**

Kursansvarlig: **Trygve Almøy 95141344**

Tillatte hjelpemidler: C3. Alle typer kalkulatorer, alle andre hjelpemidler.

Oppgavesettet er på**:** **9 sider inkludert svar-ark**

Oppgavesettet består av to deler: DEL A og DEL B

DEL A: Oppgave 1 og Oppgave 2 teller 25% hver. Totalt teller DEL A 50% av eksamensbesvarelsen. Alle delspørsmål i Oppgave 2 vektlegges likt.

DEL B: Flervalgsspørsmål – teller 50% av eksamen. Alle flervalgsspørsmål teller likt. Alle flervalgsspørsmål har 6 svaralternativer (A-F). Merk: Svar-ark på side 9 brukes til å besvare DEL B**. Husk å levere dette arket!**

Lykke til!

Oppgave 1
Lam på beite kan i enkelte deler av landet være plaget av innvollssnyltere. Siden medisinen er svært dyr vil en ikke anbefale den brukt før en er rimelig sikker på at den virker.
For å prøve å finne ut noe om effekten av medisiner mot dette ble det utført et forsøk på to følgende to måter:

Forsøk 1:
En trakk tilfeldig 12 lam som alle var infisert av snyltere, og delte disse inn i to like store grupper ved loddtrekning.
Den ene fikk injeksjon av et middel mot snyltere, den andre ble ikke behandlet.
Deretter ble lammene sluppet på beite, der de gikk sammen med sine 12 mødre. Om høsten ble lammene slaktet, og antallet innvollsormer talt opp.

Forsøk 2:
Fordi beiteområdet kan være infisert i forskjellig grad, ble det foreslått et alternativt forsøk: 6 søskenpar der alle var infisert ble trukket tilfeldig ut, i hvert par ble det ene lammet behandlet det andre ikke. Så ble søskenparet sluppet på beite, der de gikk sammen med mora. Om høsten ble lammene slaktet, og antallet innvollsormer talt opp.

Siden de ansatte på forsøksstasjonen var godt skolert i forsøksplanlegging, men ikke i analyse av data fra forsøket, ble det kjørt diverse analyser (noe korrekt og noe galt). Data og resultater fra analysen er vedlagt.

Forsøk 1

ubehandlet	40	54	26	63	21	37
behandlet	18	43	50	28	16	32

Forsøk 2

Søsken nr.	1	2	3	4	5	6
ubehandlet	42	28	29	21	62	55
behandlet	27	20	18	18	50	62

Tabell 1. Data fra begge forsøk

Skriv en kort rapport over begge forsøkene (ta med modeller, antagelser, parametere med tolkning, estimater med usikkerhet, hva slags tester som brukes), hvilke analysemetoder som er korrekte og hvilke som er feilaktige. Gi konkrete forklaringer på resultatene av forsøkene, og gi en vurdering av hvilket av de to forsøkene som er å foretrekke.

```
      Two Sample t-test
t = 1.0511, df = 10, p-value = 0.159
alternative hypothesis: true difference in means is greater than 0
sample estimates:
       mean of x        mean of y      pooled std.dev.
        40.16         31.16                14.83
------------------------------------------------------------------
      Paired t-test
t = 1.1106, df = 5, p-value = 0.1586
alternative hypothesis: true difference in means is greater than 0
sample estimates:
     mean of the differences      std.dev. of the differences
                     9.00000            19.84
```

Tabell 2. Diverse resultater fra forsøk 1.

```
      Two Sample t-test
t = 0.6863, df = 10, p-value = 0.2541
alternative hypothesis: true difference in means is greater than 0
sample estimates:
       mean of x        mean of y  pooled std.dev.
        39.50000         32.50000         17.66635
------------------------------------------------------------------------
      Paired t-test
t = 2.15, df = 5, p-value = 0.04212
alternative hypothesis: true difference in means is greater than 0
sample estimates:
     mean of the differences         std.dev. of the differences
                    7.000000                            7.974961
```

Tabell 3: Diverse resultater fra forsøk 2.

Oppgave 2

Får å få en oversikt over hvordan inntekt endrer seg undersøkte vi 52 personer med mastergrad og noterte inntekten (som årsinntekt i 1000 kroner) og hvor lang tid (målt i antall år) det var gått siden avlagt mastergrad. Se også figur 1 og figur 2. Gjennomsnittlig tid fra avlagt mastergrad var omtrent 16 år. Vi tilpasset data til en lineær regresjonsmodell med inntekt som respons, og det ga følgende utskrift i R.

```
Coefficients:
             Estimate Std. Error t value
(Intercept)  420.054     27.593  15.223
Tid            9.375      1.450   6.466
s: 105.8 on 50 degrees of freedom   Multiple R-squared: 0.4554

fit         lwr.CI  upr.CI   lwr.PI   upr.PI
420.054    364.63  475.47   200.357  639.7513
```

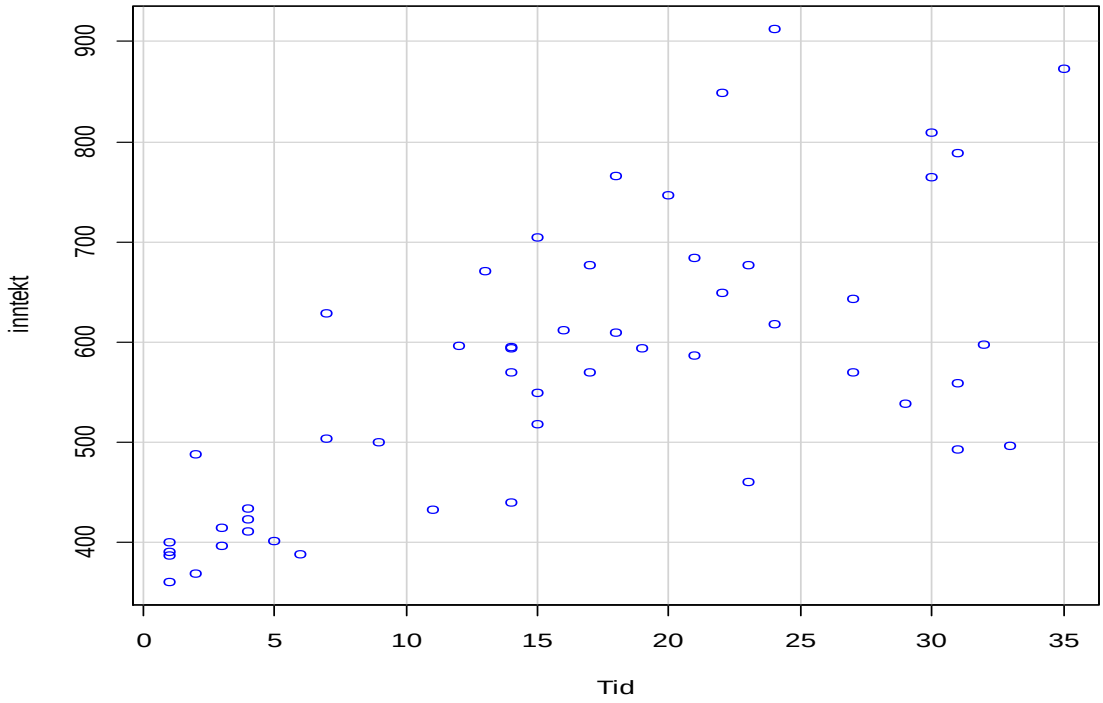
Tabell: 4 resultater fra regresjonsanalysen

a) Skriv ned modellen som er brukt.
Gi en tolkning av alle parametere i modellen.
Hvordan vil du estimere forventet årlig lønnsvekst?
Hvordan vil du estimere forventet begynnerlønn?

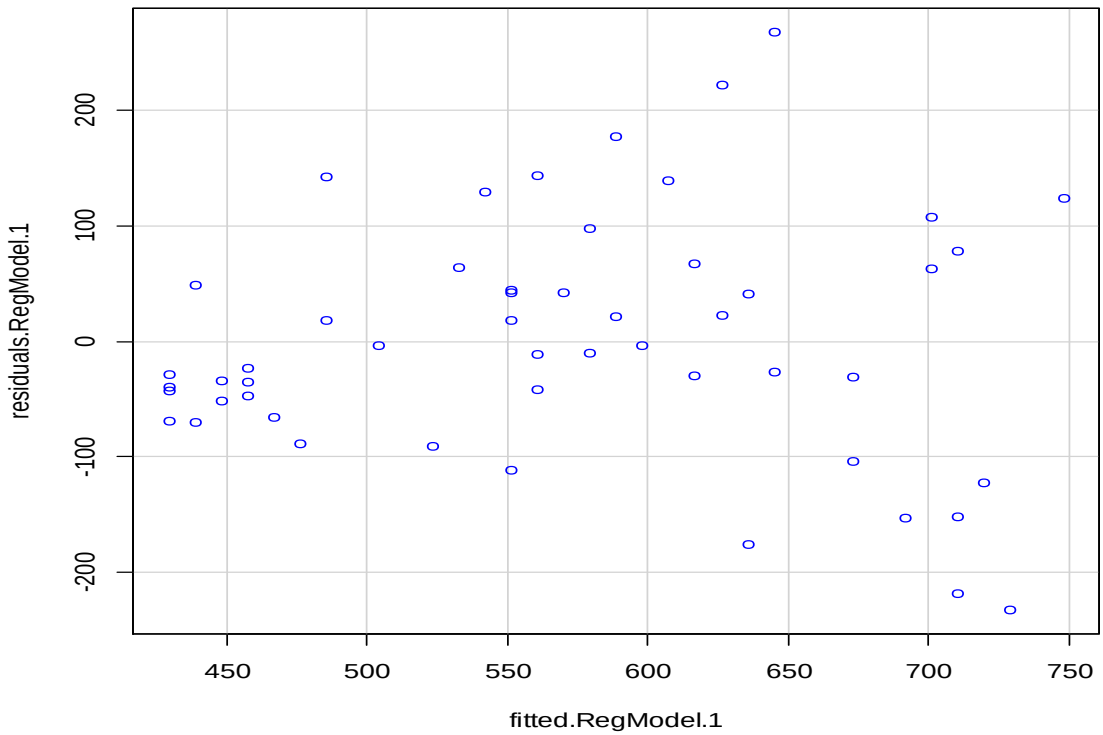
b) Lag et 95 % konfidensintervall for stigningskoeffisienten (β).
Gi en forklaring på hva dette intervallet sier.

c) Se plot av residualer mot tilpassede verdier (figur 2).
Kan du se noen svakheter ved modellen?
Har du mulige forslag til forbedringer?
En av de personene som ble observert hadde en inntekt på 642 000 kroner og det var 27 år siden vedkommende var ferdig med mastergrad.
Hva blir residualet for denne personen?

d) Bruke denne analysen til å anslå (predikere) din inntekt rett etter at du er ferdig med mastergrad (tid er null).
Hva vil du anslå inntekten til?
Skriv ned et 95 % prediksjonsintervall for denne framtidige inntekten.
Gi en forklaring på hva dette intervallet sier.
Hvorfor tror du det blir så bredt?



Figur1: Inntekt mot tid siden avlagt mastergrad



Figur 2. Residualer mot tilpassede verdier

Flervalg

For oppgave F1 til og med F4
Ei bedrift ønsker å satse på blåskjellproduksjon langs norskekysten. Det er slik at utbyttet (målt som spiselig mat etter koking i prosent av brutto vekt) kan variere langs kysten. Bedriften utførte et forsøk ved at de valgte ut tre lokaliteter A, B og C, deretter samla de 5 prøver på hver lokalitet. Dette gav følgende resultat:

Lokalitet A	*Lokalitet B*	*Lokalitet C*
19	*17*	*20*
22	*21*	*22*
19	*18*	*23*
20	*23*	*23*
22	*22*	*21*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lokalitet		7.6	3.800		0,375
Residuals		42.8	3.567		

Tabell 5: Data og analyseresultater F1-F4

La Y_{ij} være utbytte av skjell nummer j fra lokalitet i.

F1
Hvilken modell ble brukt til analysen?
A) $Y_{ij} = \mu + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma)$ B) $Y_{ij} = \mu + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma_i)$
C) $Y_{ij} = \mu_j + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma)$ D) $Y_{ij} = \mu_i + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma)$
E) $Y_{ij} = \mu_j + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma_j)$ F) $Y_{ij} = \mu_j + \varepsilon_{ij}$, der $\varepsilon_{ij} \sim N(0, \sigma_j)$

F2
Hvor mange frihetsgrader tilhører henholdsvis lokalitet og residuals (Error)
A) 2 og 12 B) 2 og 14 C) 14 og 2 D) 3 og 12 E) 3 og 14 F) 3 og 15

F3
Hva blir F-verdien i tabell 5?
A) 0,18 B) 5,63 C) 1,07 D) 0,94 E) 0,375 F) 3,8

F4
Dersom du tester en nullhypotese om at det ikke er effekt av lokalitet, hva er rett utsagn:
A) Kan ikke forkaste nullhypotese på 5% signifikansnivå.
B) Kan forkaste nullhypotese på 5% signifikansnivå
C) Kan forkaste nullhypotese på alle mulige signifikansnivå.
D) Kan påstå at det er effekt av lokalitet på 10 % signifikansnivå.
E) Kan forkaste nullhypotesen på 1% signifikansnivå fordi p-verdien er større enn 0,01.
F) Modellen passer ikke til å teste hypoteser.

For oppgave F5 til og med F9
Siden lokalitet (undersøkt i F1 - F4), hadde relativt beskjeden effekt, ville en heller se etter andre faktorer som kunne påvirke utbyttet. Skjell ble sortert etter størrelse (lite eller stort) og etter den årstid (vår eller høst) som prøvene ble tatt. For hver kombinasjon årstid/størrelse ble det tatt 6 prøver. Vi velger å kalle de 4 kombinasjonene for:

1: høst/liten
2: høst/stor
3: vår/liten
4: vår/stor

	Df	Sum Sq	Mean Sq	F value
Årstid/størrelse	3	684	228.0	95
Residuals	20	48	2.4	

Kombinasjon	mean	sd	data:n
1	24	1.41	6
2	13	1.54	6
3	24	1.78	6
4	27	1.41	6

Tabell 6: Data og resultat for F5- F9

F5
Et stort skjell tatt om høsten ga utbytte på 11. Hva blir residualet til dette skjellet?
A) 13 B) 684 C) -2 D) 2 E) -11 F) 11

F6
Hvordan vil du estimere standardavviket i utbytte for skjell innen samme årstid/størrelse kombinasjon?
A) 1,41 B) 6,93 C) 1,53 D) 95 E) 1,55 F) 1,78

For oppgavene F7 til og med F9
La μ_i være forventet utbytte for kombinasjon nummer i. (i =1, 2, 3, eller 4).

F7
Hvilken kontrast ser på effekt av å ta skjell om våren framfor om høsten?
A) $\theta = \frac{1}{2}(\mu_3 + \mu_4) - \frac{1}{2}(\mu_1 + \mu_2)$
B) $\theta = \frac{1}{2}(\mu_3 + \mu_1) - \frac{1}{2}(\mu_4 + \mu_2)$
C) $\theta = (\mu_3 - \mu_4)$
D) $\theta = \frac{1}{2}(\mu_3 + \mu_4) - \frac{1}{4}(\mu_1 + \mu_2)$
E) $\theta = \frac{1}{2}(\mu_3 + \mu_4) + \frac{1}{2}(\mu_1 + \mu_2)$
F) $\theta = \frac{1}{2}(\mu_3 + \mu_4) - \frac{1}{2}(\mu_1)$

F8
En bestemmer seg for å ta skjell om våren og se på kontrasten: $\theta = \mu_4 - \mu_3$
Hva blir estimatet til denne kontrasten.
A) 24 B) 27 C) 11 D) 2 E) -11 F) 3

F9
En bestemmer seg for å ta skjell om våren og se på kontrasten $\theta = \mu_4 - \mu_3$
Hva blir standardfeilen til estimatet for denne kontrasten?
A) 0,8 B) 0,89 C) 0 D) 2,19 E)0,81 F) 2,4

F10
Dersom to begivenheter er uavhengige, hva er da P(A|B)?
A) 1 B) P(A) C) P(B) D) 0 E) P(B|A) F) $P(A \cap B)$

F11
La en tilfeldig variabel være normalfordelt med ukjent forventning og kjent standardavvik (σ). Dersom du planlegger å lage et 95 % konfidensintervall for forventningen, men vil ha dette smalt, kan dette gjøres ved:
A) Gjøre utvalgsgjennomsnittet stort.
B) Gjøre utvalgsgjennomsnittet lite
C) Øke σ
D) Gjøre forventningen liten
E) Redusere σ
F) Øke antall observasjoner

F12
I en stor forsamling ble det registrert kjønn og røykevaner, noe som ga følgende tabell:

	Røyker	Røyker ikke
Kvinne	0,2	0,1
Mann	0,3	0,4

Tallene gir sannsynlighet, for eksempel er sannsynligheten for å trekke ut en kvinnelig røyker lik 0,2. Hvor stor andel av mennene røyker ikke?
A) 0,4 B) 0,57 C) 0,6 D) 0,5 E) 0,67 F) 0,8

F13
La R være begivenheten at en person røyker og U være begivenheten at en person har lang utdanning. En undersøkelse ga at $2*P(R|U) = P(R|\bar{U})$. Hvordan kan dette resultatet tolkes?
A) Det er dobbelt så vanlig å røyke som ikke å røyke.
B) Det er dobbelt så mange røykere som ikke-røykere.
C) Det er dobbelt så vanlig at røykere har kort utdanning i forhold til ikke-røykere.
D) Det er dobbelt så mange som både har kort utdanning og som røyker i forhold til de som både har lang utdanning og som røyker.
E) Det er dobbelt så vanlig å røyke blant de med kort utdanning i forhold til de med lang utdanning.
F) Halvparten av de som røyker har kort utdanning

For oppgavene F14 – F16
BMI (Body mass index) er vekt i kg dividert med kvadratet av kroppshøyden i meter. For menn mellom 20 og 30 år er BMI tilnærmet normalfordelt med forventning 25 og standardavvik 5.

F14
Hvor stor andel av menn mellom 20 og 30 år har en BMI større enn 30?
A) 0,3 B) 0,5 C) 0,16 D) 0,02 E) 0,62 F) 0,84

F15
Hvis vi har et tilfeldig utvalg på 4 menn, hva er sannsynligheten for at alle har en BMI over 25?
A) 0,412 B) 0,5 C) 0,16 D) 0,02 E) 0,063 F) 0,84

F16
Hvis vi har et tilfeldig utvalg på 4 menn, hva er sannsynligheten for at gjennomsnittlig BMI for disse er over 30.
A) 0,3 B) 0,5 C) 0,16 D) 0,02 E) 0,063 F) 0,84

For oppgavene F17 og F18. I en valgundersøkelse fant vi følgende.

	Menn	Kvinner	sum
Sosialistisk	450	520	970
Borgerlig	485	420	905
Stemte ikke	100	140	240
Sum	1035	1080	2115

Følgende utskrift i R kan du bruke

```
Expected counts:
                  Menn        Kvinner
Sosialistisk      474         495
Borgerlig         442         462
Stemte ikke       117         122

Chi-square components:
                  Menn        Kvinner
Sosialistisk      1.28        1.23
Borgerlig         4.01        3.84
Stemte ikke       2.59        2.48
```

F17
La H_0 være at kjønn og stemmegivning er uavhengige begivenheter. Hva er blir testobservator (kalt Q i læreboka), og hva blir konklusjon?
A) Q = 15,4, ikke forkast H_0 B) Q = 12,4, ikke forkast H_0 C) Q= 12,4, forkast H_0
D) Q= 2, ikke forkast H_0 E) Q= 15,4, forkast H_0 F) Q = 2112, forkast H_0

F18.
Ta for deg kun kvinner som stemmer ved valg. Et 95 % konfidensintervall for andelen kvinner som stemmer sosialistisk er gitt ved:
A) (0,52; 0,58) B) (0,55; 0,58) C) (0,50; 0,55)
D (0,40; 0,65) E) (0,5; 0,5) F) (0,553; 0,554)

F19
Anta du vil estimere forventet høyde i en populasjon og måler høyden på 4 menn. Fra før vet vi at standardavviket er 10 cm. Det viser seg å være to tvillingpar du har undersøkt. Korrelasjon i høyde mellom tvillinger er 0,9. Da er standardavviket til estimatoren ($\bar{X}$) lik
A) 5 B) 10 C) 13,8 D) 2,5 E) 47,5 F) 6,9

F20
En kvinne tar mammografiundersøkelse. La S = {kvinnen har brystkreft} og la M = {mammogrammet viser kreft}.
Anta at $P(M|S) = 0.95,\ P(M|\bar{S}) = 0.035$ og $P(S) = 0.007$. Hva blir sannsynligheten for at en kvinne har brystkreft hvis mammogrammet viser kreft?

A) 0.007 B) 0.062 C: 0.16 D) 0.04 E) 0.002 F) 0,95

Riv ut arket og levere dette sammen med besvarelsen.
Bare ett kryss i hver rute.

Oppgave	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						