UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Fredag 5. juni 2015
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 0 sider.	
Vedlegg:	Tabell over normal-, t-, χ^2 og F-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

Utskriften nedenfor viser en en-faktor variansanalyse av testresultater for femteklassinger for de fire skolene A, B, C og D. På hver skole ble 10 elever testet. Gjennomsnittet for skolene A, B, C og D er henholdsvis 600.1, 688.6, 667.4 og 487.1

```
Response: score
          Df Sum Sq Mean Sq F value    Pr(>F)
skole      3 246726   82242  26.529 3.119e-09
Residuals 36 111601    3100
```

a) Formuler modellen som brukes for en-faktor variansanalyse. Redegjør for forutsetningene. Forklar hva de enkelte størrelesene i utskriften betyr. Hvorfor må en hypotese om at det ikke er noen forskjell på resultatene mellom skolene forkastes?

b) Skolen D har de svakeste resultatene, og et spørsmål er om den skiller seg fra gjennomsnittet av de andre. Formuler en hypotese for å teste om denne forskjellen er signifikant. Utfør testen og konkluder.

Et spørsmål er om det er noen forskjell på resulatene for jenter og gutter. Det var like mange jenter som gutter som ble testet på hver skole. Utskriften nedenfor viser en to-faktor variansanalyse hvor også kjønn er tatt med som forklaringsfaktor.

```
Response: score
             Df Sum Sq Mean Sq F value    Pr(>F)
skole         3 246726   82242 27.7535 4.945e-09 ***
kjoenn        1   6200    6200  2.0923    0.1578
skole:kjoenn  3  10575    3525  1.1895    0.3293
Residuals    32  94826    2963
```

(Fortsettes på side 2.)

c) Forklar hvordan opplysningene i utskriften kan brukes til å teste om kjønn har betydning. Hva er konklusjonen din?

d) Gjennomsnittet for jenter i skolene A, B, C og D er 616.4, 725.8, 668.8 og 482.0. Gjennomsnittet for gutter skolene A, B, C og D er 583.8, 651.4, 666.0 og 492.2. Skisser et samspillsplott (interaksjonsplott) og forklar hvordan det støtter opp under konklusjonen fra punkt c).

Oppgave 2

Figuren nedenfor viser sju målinger for kvinner og syv målinger for menn av skrittlengde og hastighet på en tredemølle. De fylte sirklene viser målingene for kvinner og trekantene viser målingene for menn. Hastigheten kan innstilles og er derfor den uavhengige forklaringsvariabelen mens skrittlengde er respons. Som det framgår av plottet, er de valgte hastighetene de samme for kvinner og menn, og betegnes med $x_1, \cdots, x_7$. Dette betyr at dataene kan uttrykkes som parene $(y_1, x_1), \cdots, (y_{14}, x_{14})$ der de første sju parene er for kvinnene og $x_1 = x_8, \cdots, x_7 = x_{14}$.

Det er tilpasset to regresjonsmodeller. R-utskriftene er gjengitt nedenfor og regresjonslinjene er gjengitt i figuren. I den ene modellen er det tilpasset en felles regresjonslinje for kvinner og menn, i den andre separate linjer for de to gruppene, men med felles stigningskoeffesient. Den nederste linja svarer til kvinner. For faktoren som angir kjønn er cornerpoint/treatment parameteriseringen benyttet med kvinner som referansekategori.

a) Formuler de to regresjonsmodellene som er benyttet og gjør rede for forutsetningene.

b) Alle de tre linjene i figuren har formen $y = a + bx$. Bruk modellutskriftene til å angi de numeriske verdiene til størrelsene a og b for de tre linjene.

c) Angi hva designmatrisene $\mathbf{X}_1$ og $\mathbf{X}_2$ for de to tilpassede modellene er.

d) Forklar hvordan modellen med separate regresjonslinjer kan reparameteriseres slik at uttrykket $\mathbf{X}_3'\mathbf{X}_3$ har formen

$$\mathbf{X}_3'\mathbf{X}_3 = \begin{pmatrix} 7 & 0 & 7\bar{x} \\ 0 & 7 & 7\bar{x} \\ 7\bar{x} & 7\bar{x} & 2\sum_{j=1}^{7} x_j^2 \end{pmatrix}$$

der $x_1, \ldots, x_7$ er de valgte hastighetene og $\mathbf{X}_3$ er designmatrisen i denne parameteriseringen.

e) Av utskriftene for de to tilpassede modellene framgår det at stigningsforholdet for de tre linjene har samme numeriske verdier. Forklar hvorfor dette er tilfelle.

Hint: Du kan ha nytte av formelen

$$\begin{pmatrix} 7 & 0 & c \\ 0 & 7 & c \\ c & c & d \end{pmatrix}^{-1} = \begin{pmatrix} 7d - c^2 & c^2 & -7c \\ c^2 & 7d - c^2 & -7c \\ -7c & -7c & 49 \end{pmatrix} / (49d - 14c^2).$$

(Fortsettes på side 3.)

Modell 1:

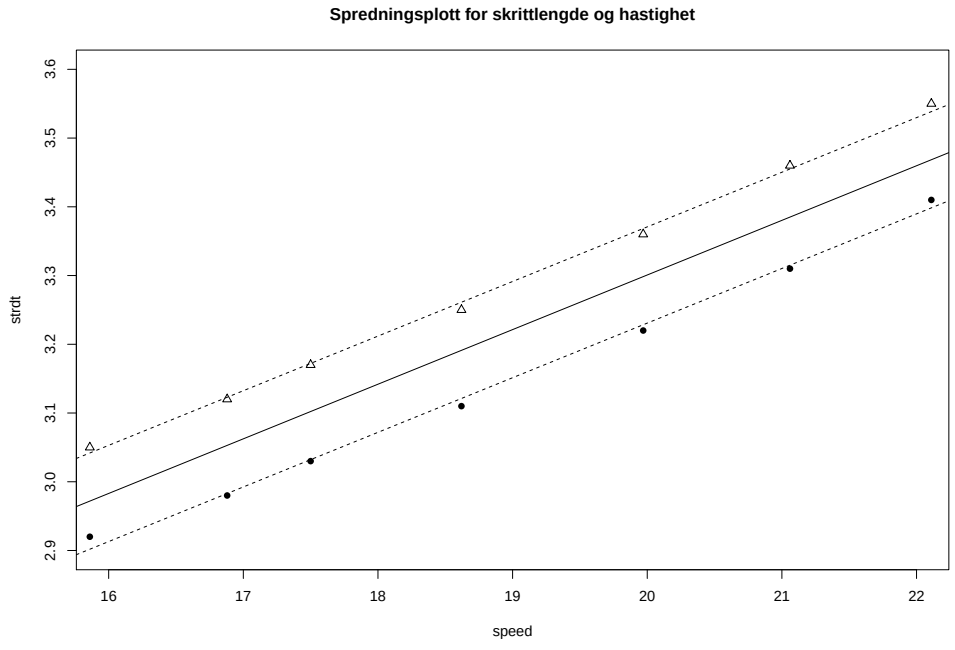
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71166    0.18235   9.387 7.07e-07 ***
speed        0.07946    0.00961   8.269 2.68e-06 ***
---

Residual standard error: 0.07623 on 12 degrees of freedom
Multiple R-squared:  0.8507,    Adjusted R-squared:  0.8382
F-statistic: 68.37 on 1 and 12 DF,  p-value: 2.677e-06
```

Modell 2:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.641660   0.024385   67.32 9.64e-16 ***
speed           0.079457   0.001277   62.22 2.29e-15 ***
factor(gender)2 0.140000   0.005415   25.85 3.35e-11 ***
---

Residual standard error: 0.01013 on 11 degrees of freedom
Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9971
F-statistic:  2270 on 2 and 11 DF,  p-value: 4.059e-15
```



(Fortsettes på side 4.)

Oppgave 3

La $X_1, \cdots, X_n$ være uavhengige og identisk eksponetialfordelte tilfeldige variable med tetthet

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\exp(-\frac{x}{\theta}), & x > 0 \\ 0 & \text{ellers} \end{cases}$$

der $\theta > 0$ er en parameter.

a) Sett opp et uttrykk for likelihooden. Finn scorefunksjonen og bruk denne til å bestemme sannsynlighetsmaksimeringsestimatoren (SME) $\hat{\theta}$ for θ. Hva er Fisher-informasjonen?

b) Bruk Cramer-Rao's ulikhet til å bestemme en nedre grense for variansen til forventningsrette estimatorer for θ. Hva betyr dette for variansen til SME $\hat{\theta}$?

c) Hva er den tilnærmede fordelingen til $\hat{\theta}$ når n er stor? Bruk dette resulatet til å angi et tilnærmet konfidensintervall for θ med konfidenskoeffesient $1 - \alpha$ og til å utlede en test for nullhypotesen $H_0 : \theta = \theta_0$ mot alternativet $H_a : \theta \neq \theta_0$ med tilnærmet nivå α.

d) Utled også sannsynlighetskvotetesten for nullhypotesen og alternativet fra punkt c). Hva er den tilnærmede fordelingen for testoservatoren?

SLUTT

(Fortsettes på side 5.)