UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Tirsdag 3. juni 2008.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabeller over normal-, t-, F- og χ^2-fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

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

Oppgave 1.

(a) Anta at $X_1, X_2, ..., X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2. Angi Student's t-test for nullhypotesen $\mathrm{H}_0 : \mu = 0$ mot $\mathrm{H}_1 : \mu \neq 0$.

Diskuter i hvilken grad testen er gyldig når antagelsen om normalfordeling ikke holder.

(b) Anta istedet at X_i-ene er logistisk fordelt med tetthet

$$f(x|\mu,\tau) = \frac{1}{\tau}\frac{\exp((x-\mu)/\tau)}{[1+\exp((x-\mu)/\tau)]^2} \text{ for } x \in \Re.$$

Tegn en grov skisse av $f(x|0,1) = e^x/(1+e^x)^2$.

Vis at $f(x|\mu,\tau)$ er symmetrisk om $x = \mu$.

(Det følger at $\mathrm{E}[X_i] = \mu$. Det kan også vises at $\mathrm{Var}[X_i] = \frac{\pi^2}{3}\tau^2$, men det skal du ikke gjøre).

(Fortsettes side 2.)

(c) Utled at log-likelihood basert på $X_1, X_2, ..., X_n$ uavhengige og logistisk fordelte blir

$$l(\mu, \tau) = -n \log(\tau) + \sum_{i=1}^{n} \frac{X_i - \mu}{\tau} - 2 \sum_{i=1}^{n} \log[1 + \exp((X_i - \mu)/\tau)]$$

Finn også eksplisitte uttrykk for den tilhørende score-funksjonen.

Forklar hva den observert informasjonsmatrisen $\bar{J}(\mu, \tau)$ er (eksplisitte uttrykk for den aktuelle modellen er ikke nødvendig).

(d) Gi en kort beskrivelse av algoritmer for iterasjon mot MLE $(\hat{\mu}, \hat{\tau})$.

Hvilken algoritme er enklest å implementere for det aktuelle problemet?

Foreslå også fornuftige startverdier for iterasjonen.

(e) Basert på $n = 20$ observasjoner gjengitt i tabellen under ble gjennomsnittet $\bar{X} = 1.0495$ og empirisk varians $s^2 = 3.057$. Videre ble MLE $(\hat{\mu}, \hat{\tau}) = (1.0413, 0.9453)$, mens invers observert informasjonsmatrise evaluert i MLE er

$$\bar{J}(\hat{\mu}, \hat{\tau})^{-1} = \begin{bmatrix} 0.1310 & 0.0005 \\ 0.0005 & 0.0324 \end{bmatrix}$$

(beregnet under logistisk modell). Angi tilnærmet fordeling for $\hat{\mu}$ og gjennomfør på denne bakgrunn en test for $H_0 : \mu = 0$ mot $H_1 : \mu \neq 0$.

Gjør også en t-test for hypotesen. Sammenlign resultatene og kommenter.

Tabell over $X_1, ..., X_{20}$ (sortert)

−2.27	−2.13	−1.03	−0.15	0.24	0.36	0.49	0.51	0.97	0.98
1.09	1.13	1.28	1.65	1.74	1.78	2.84	2.92	3.95	4.64

(f) Forklar hva den generaliserte likelihood ratio testen (GLRT) er.

Angi eksplisitt testobservator for GLRT for modellen i denne oppgaven under nullhypotesen $H_0 : \mu = 0$. Angi også testobservatorens tilnærmede fordeling under nullhypotesen.

Under nullhypotesen ble MLE for parameteren τ lik $\tau^\star = 1.1425$, $\sum_{i=1}^{n} \log[1 + \exp((X_i - \hat{\mu})/\hat{\tau})] = 20.153$ og $\sum_{i=1}^{n} \log[1 + \exp(X_i/\tau^\star)] = 29.074$. Benytt dette til å gjennomføre en GLRT av $H_0 : \mu = 0$.

Sammenlign med resultatene i punkt (e).

Oppgave 2.

Lungefunksjon måles ofte ved "forced expiratory volume in one second" (FEV1), som er antall liter luft et individ er istand til å blåse ut i løpet av ett sekund. Vi skal i denne oppgaven se på hvordan responsen $Y_i = \text{FEV1}$

(Fortsettes side 3.)

avhenger av forklaringsvariable x_{i1} = høyde (målt i cm), x_{i2} = kjønn (kodet som 1 for menn og 2 for kvinner) og x_{i3} = bmi ("body mass index", som er vekt i kg delt på kvadratet av høyde i meter) i et datasett over 154 seksti år gamle nordtrøndere.

(a) I R-utskriften under er det gjengitt resultatet av en enkel lineær regresjon av FEV1 kun mot høyde.

Forklar hva utskriften forteller om sammenhengen mellom høyde og FEV1. Finn spesielt (den empiriske) korrelasjonen mellom disse variablene basert på utskriften.

Beregn også et 95% konfidensintervall for stigningskoeffisenten β_1 og test på denne bakgrunn nullhypotesen at en 1 cm forskjell i høyde tilsvarer en forskjell i FEV1 på 0.04 liter.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.105926   0.863665  -8.228 7.97e-14 ***
hoyde        0.058514   0.005124  11.419  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5247 on 152 degrees of freedom
Multiple R-Squared: 0.4618,     Adjusted R-squared: 0.4582
F-statistic: 130.4 on 1 and 152 DF,  p-value: < 2.2e-16
```

(b) Under er det gjengitt resultatet av en Student t-test for forskjell i FEV1 mellom kvinner og menn. Det er også gjengitt resultatet av en regresjonsanalyse med respons FEV1 og forklaringsvariable høyde og kjønn (se neste side). Konkluder om sammenheng mellom FEV1 og kjønn basert på begge analyser.

Forklar hvorfor testene gir forskjellig resultat.

```
> t.test(fev1~kjonn,var.equal=T)

        Two Sample t-test

data:  fev1 by kjonn
t = 8.6033, df = 152, p-value = 8.962e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.637081 1.016907
sample estimates:
mean in group 1 mean in group 2
       3.233302        2.406308
```

(Fortsettes side 4.)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.802614   1.588610  -3.653 0.000357 ***
hoyde        0.052067   0.008352   6.234 4.33e-09 ***
kjonn       -0.137023   0.140170  -0.978 0.329861
```

(c) I en neste regresjon er bmi og kvadratet av bmi lagt inn i regresjonen. Resultatene er gjengitt i R-utskriften under. Beskriv den estimerte sammenhengen mellom FEV1 og bmi ved en skisse.

Estimer den bmi-verdien som (under modellen) gir det høyeste nivået for FEV1 i dette utvalget.

Det viser seg at bmi ikke viser signifikant effekt hvis kvadratleddet tas ut av modellen. Kan du gi noen rimelig forklaring på dette, for eksempel i lys av fordelingen til bmi i dette utvalget (gitt i R-utskriften under)?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.880367   1.362471  -7.252 2.04e-11 ***
hoyde        0.057232   0.005144  11.125  < 2e-16 ***
bmi          0.203648   0.077461   2.629  0.00945 **
I(bmi^2)    -0.003373   0.001308  -2.580  0.01084 *
---

> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.50   25.10   27.35   28.05   30.37   42.20
```

(d) Tilslutt et teoretisk problem. Anta en generell multippel lineær regresjonsituasjon med designmatrise $\mathbf{X}$ og vektor av responser $\mathbf{Y}$. Vi gjør de vanlige antagelsene ved multippel linær regresjon: Komponentene Y_i i $\mathbf{Y}, i = 1, \ldots, n$, er uavhengige med samme varians σ^2 og forventningen til Y_i er lik $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$. La $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})^\top$.

Vis at minste kvadraters estimatoren for β, dvs. $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, har følgende egenskaper:

(i) $\hat{\beta}$ er forventningsrett, $\mathrm{E}[\hat{\beta}] = \beta$

(ii) $\hat{\beta}$ har kovariansmatrise $\Sigma(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$

Hint: Vis først at $\mathrm{E}[\mathbf{X}^\top \mathbf{Y}] = \mathbf{X}^\top \mathbf{X} \beta$ og at kovariansmatrisen for $\mathbf{X}^\top \mathbf{Y}$ er lik $\Sigma(\mathbf{X}^\top \mathbf{Y}) = \mathbf{X}^\top \mathbf{X} \sigma^2$.

SLUTT