UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Torsdag 7. juni 2012.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 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.

Lav bentetthet øker sjansen for benbrudd. En tilstand med svært lav bentetthet kalles osteoporose. Vi skal se på en gruppe på 190 kvinner som alle har brukket hofteskålen. På grunnlag av bentettheten kan kvinnene deles inn i tre kategorier: normal, osteopeni, osteoporose. Osteopeni er et forstadium til osteoporose, dvs. at bentettheten er nedsatt, men ikke tilstrekkelig til å få diagnosen osteoporose. De 190 kvinnene fordeler seg over de tre kategoriene og over tre aldersgrupper som spesifisert i tabellen. Forskerne ønsker å belyse om det er noen forskjell mellom aldersgruppene når det gjelder bentetthet for hoftebruddspasienter.

	normal	osteopeni	osteoporose	sum
51-57	7	11	22	40
58-64	8	17	36	61
≥ 65	9	24	56	89
sum	24	52	114	190

Formuler nullhypotese og gjennomfør en passende test. Du får oppgitt at testobservator $\chi^2 = 1.5232$.

Oppgave 2.

Mange kvinner får problemer med osteoporose etter overgangsalderen. Forskere ved Rikshospitalet studerer hvordan genetiske faktorer bidrar til en slik tilstand. Vi skal studere et datasett der bentetthet er målt for 84 kvinner etter overgangsalder. For de

(Fortsettes på side 2.)

samme kvinnene har man målt gen-ekspresjon (et mål på hvor aktivt et gen er) i benceller for 20.000 gener. Vi skal ikke analysere et slikt høydimensjonalt datasett her, da det ville kreve spesielle metoder. Vi skal nøye oss med å se hvordan gen-ekspresjon for fire utvalgte gener kan brukes som forklaringsvariable i en multippel regresjonsanalyse med bentetthet som responsvariabel.
I utskriften nedenfor heter variabelen bentetthet 'bone' og de fire genene hhv. 'gene1', 'gene2', 'gene3' og 'gene4'. Data ligger i matrisen 'd'.

```
Call:
lm(formula = bone ~ gene1 + gene2 + gene3 + gene4, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97654 -0.91338  0.08307  0.82191  2.18842

Coefficients:
                Estimate  Std. Error t value Pr(>|t|)
(Intercept)      2.0388     3.7747   0.540 0.590640
gene1           -1.4291     0.3667  -3.898 0.000202 ***
gene2           -2.3275     0.5931  -3.924 0.000185 ***
gene3            2.7178     0.5550   4.897 5.06e-06 ***
gene4            1.6393     0.2501   6.555 5.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 0.9973 on 79 degrees of freedom
Multiple R-squared: 0.6293,     Adjusted R-squared: 0.6106
F-statistic: 33.53 on 4 and 79 DF,  p-value: 2.436e-16
```

a) Sett opp modellen som ligger til grunn for analysen på matriseform. Bruk notasjon $\mathbf{Y}$ for responsvektor, $\mathbf{X}$ for designmatrisen og $\boldsymbol{\beta}$ for parametervektoren. Angi dimensjonen til alle vektorer/matriser. Formuler de vanlige forutsetningene vi gjør for en slik modell. Vis at minste kvadraters estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ er forventningsrett for $\boldsymbol{\beta}$. Forklar kort hva de estimerte koeffisientene sier om sammenhengen mellom bentetthet og gen-ekspresjonene.

b) Begrunn kort hvorfor $\hat{\beta}_j$-ene blir normalfordelte. Utled et uttrykk for kovariansmatrisen til $\hat{\boldsymbol{\beta}}$, $\text{Cov}(\hat{\boldsymbol{\beta}})$. Du kan bruke at dersom $\mathbf{A}$ er en matrise med konstanter og $\mathbf{V} = \mathbf{AU}$, så er $\text{Cov}(\mathbf{V}) = \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{A}^T$. Beregn et 95% konfidensintervall for β_2, koeffisienten for gen 2.

La oss i resten av oppgaven se på en multippel regresjonsmodell med k forklaringsvariable og n observasjoner. For å teste nytten av modellen (model utility) utfører de fleste programvarer en F-test for å teste $\text{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. La predikerte verdier være $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Hvis vi definerer SSE= $\sum_i (Y_i - \hat{Y}_i)^2$, SST= $\sum_i (Y_i - \bar{Y})^2$ og SSR= $\sum_i (\hat{Y}_i - \bar{Y})^2$, kan du bruke uten å vise det at SST=SSE+SSR og at SSE og SSR er uavhengige.

(Fortsettes på side 3.)

c) Vis først hvorfor SST/σ^2 er χ^2-fordelt med $n-1$ frihetsgrader når H_0 er riktig (her er σ^2 variansen til støyleddene). Bruk dette til å finne fordelingen til SSR/σ^2 under H_0.

d) F-testen baserer seg på testobservator

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-(k+1))}.$$

Vis hvordan du kan finne fordelingen til F under H_0. Forklar hvorfor det er naturlig å forkaste H_0 når F er stor. Formuler en konklusjon for model utility testen for data i oppgaven.
(Tips: Husk at $V_1 \sim \chi^2_{\nu_1}$, $V_2 \sim \chi^2_{\nu_2}$, V_1, V_2 uavhengige, gir $\frac{V_1/\nu_1}{V_2/\nu_2} \sim F$-fordelt med ν_1 og ν_2 frihetsgrader)

e) Multippel R^2 er definert ved SSR/SST. Forklar hvordan den skal tolkes, og hvordan den henger sammen med F-observator for model utility-testen.

Oppgave 3.

a) Formuler modellen og vanlige forutsetninger for å sammenligne forventningene i I populasjoner basert på J observasjoner fra hver av dem. Sett opp hypotesene som testes.

b) Gitt at nullhypotesen i situasjonen over er forkastet, kan vi bruke Tukeys prosedyre for å finne ut hvilke forventninger som skiller seg fra de andre. Vis hvordan man kan finne simultane konfidensintervaller for alle differanser $\mu_{i_1} - \mu_{i_2}$ (der $i_1, i_2 = 1, ..., I$) og skisser hvordan man bruker slike intervaller til å finne signifikant forskjellige forventninger. Du kan bruke resultatet om studentifisert variasjonsbredde fra formelsamlingen uten å vise dette.

Oppgave 4.
En fordeling som er mye brukt for å modellere økonomiske data er Pareto-fordelingen. Vi skal se på hvordan parametrene i en Pareto-fordeling kan estimeres fra data ved hjelp av Maximum Likelihood metoden (sannsynlighetsmaksimeringsmetoden).
Når en kontinuerlig variabel X er Pareto-fordelt med parametre $x_m > 0$ og $\theta > 0$, er tettheten gitt ved

$$\begin{aligned} f(x|x_m,\theta) &= \theta x_m^{\theta} x^{-\theta-1} && \text{når} \quad x \geq x_m \\ &\quad 0 && \text{ellers} \end{aligned} \tag{1}$$

a) Anta først at lokasjonsparameteren x_m er gitt. Basert på uavhengige, identisk Pareto-fordelte variable $X_1, X_2, ..., X_n$, finn maksimum likelihood estimator (sannsynlighetsmak-

(Fortsettes på side 4.)

simeringsestimator) for formparameteren θ. Finn Fisher-informasjonen og angi fordelingen til $\hat{\theta}$ når n er stor.

b) Vi har $n = 30$ observasjoner fra en Pareto-fordeling med $x_m = 1$. Maksimum likelihood estimatet for θ beregnes til 2.1 fra disse observasjonene. Konstruer et tilnærmet 95% konfidensintervall for θ.

c) I stedet for normaltilnærming kunne vi ha brukt bootstrap-metoden til å beregne et konfidensintervall for θ. Basert på simuleringer fra Pareto-modellen med $x_m = 1$ og $\theta = \hat{\theta} = 2.1$, har vi beregnet $B = 1000$ bootstrap-replikasjoner $\hat{\theta}^*$. Fordelingen for disse kan oppsummeres ved

Min	0.5%	2.5%	25%	Median	Mean	75%	97.5%	99.5%	Max
1.192	1.386	1.531	1.897	2.159	2.206	2.445	3.172	3.574	5.114

I tabellen finner du bl.a. kvantilene i den empiriske fordelingen. Konstruer et 95% standard bootstrap konfidensintervall for θ og sammenlign med intervallet ovenfor. Kommenter kort.

d) Anta til sist at du også skal estimere lokasjonsparameteren x_m. Finn maksimum likelihood estimator for x_m.

Takk for innsatsen!