UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST102 — Videregående kurs i statistikk
Eksamensdag:	Torsdag 29. mai 1997.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 4 sider.	
Vedlegg:	1) Data for 31 Black cherry trær, 2) Regresjonsanalyser av tredatene, 3) Residualplott for tredataene, 4) Tabeller over Students t fordeling og kji-kvadrat fordelingen.
Tillatte hjelpemidler:	Formelsamlinger i ST101 og ST102, Matematiske formelsamlinger (Rottmanns eller Jahren og Knutsens), lommeregner.

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

Oppgave 1.

I denne oppgaven skal vi se nærmere på den enkle lineære regresjonsmodellen uten konstantledd. Vi antar således at

$$Y_i = bx_i + \epsilon_i\,; \quad i = 1, 2, \ldots, n; \tag{1}$$

hvor x_i-ene er gitte størrelser, ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte, og b og σ^2 er ukjente parametere.

a) Vis at minste kvadraters estimator for b er

$$\widehat{B} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

b) Vis at $\widehat{B}$ er forventningsrett. Bestem $\mathrm{Var}(\widehat{B})$.

(Fortsettes side 2.)

c) En estimator for σ^2 er

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \hat{B}x_i\right)^2.$$

Det kan vises at $\hat{B}$ og S^2 er uavhengige og at $(n-1)S^2/\sigma^2$ er kjikvadrat fordelt med $n-1$ frihetsgrader. (Du skal ikke vise dette.) Forklar hvorfor

$$\frac{\hat{B}-b}{S}\sqrt{\sum_{i=1}^{n}x_i^2}$$

er Student t-fordelt med $n-1$ frihetsgrader.

d) Vis at

$$\sum_{i=1}^{n}(Y_i - \hat{B}x_i)^2 = \sum_{i=1}^{n}Y_i^2 - \hat{B}^2\sum_{i=1}^{n}x_i^2,$$

og bruk dette til å vise at S^2 er en forventningsrett estimator for σ^2. (Du får altså her ikke benytte resultatet gitt i punkt c.)

Oppgave 2.

Forstmenn har behov for å anslå mengden av tømmer i et gitt skogområde. Derfor trenger de en rask og enkel måte for å anslå volumet av et gitt tre. Det er vanskelig å anslå direkte volumet av et tre på rot. Men det er forholdsvis lett å måle dets høyde og enda enklere å måle dets diameter (ved bakken). Forstmenn har derfor behov for å utvikle en ligning eller en tabell som gjør det enkelt å anslå volumet av et tre ut fra dets diameter og/eller høyde.

I Vedlegg 1 er det gitt målinger av diameter (i tommer 4,5 fot over bakken), høyde (i fot) og volum (i kubikkfot) for et utvalg av trær ("black cherry trees") fra et bestemt skogsområde i Pennsylvania. Til hjelp nedenfor er også høyden multiplisert med kvadratet av diameteren oppgitt (`HoydeDiam2`).

Det er foretatt tre enkle regresjonsanalyser av disse dataene. I alle analysene er volum den avhengige variabelen (Y) mens den uavhengige variabelen (x) er henholdsvis diameter, høyde og høyde$\times$(diameter)^2. Et redigert utdrag av BLSS-kjøringer for disse tre regresjonsanalysene er gitt i Vedlegg 2 (bl.a. er P-verdiene fjernet). I Vedlegg 3 er det gitt plott hvor residualene for disse regresjonsanalysene er plottet mot de tilpassede verdiene ($\hat{Y}$). Skalaen langs y-aksen er med vilje fjernet fra disse plottene. Videre er plottene i Vedlegg 3 ikke (nødvendigvis) gitt i samme rekkefølge som regresjonsanalysene i Vedlegg 2.

(Fortsettes side 3.)

a) Identifiser hvilket residualplott som hører til hver av de tre regresjonsanalysene. Svaret skal begrunnes.

b) Diskuter for hver av de tre regresjonsanalysene om forutsetningene for å benytte lineær regresjon ser ut til å være oppfylt. Hvilken av regresjonsanalysene syns du gir den beste analysen av dataene? Begrunn ditt valg og gi en fortolkning av resultatene av den valgte regresjonsanalysen.

I en kjegle er volumet proporsjonalt med høyden multiplisert med kvadratet av diameteren. Hvis dette er en god modell for sammenhengen mellom volum, høyde og diameter av et tre, vil konstantleddet i regresjonsmodellen være null når høyde$\times$(diameter)^2 benyttes som uavhengig variabel.

c) Hvordan vil du benytte resultatene i Vedlegg 2 til å teste nullhypotesen om at konstantleddet er lik null i regresjonsmodellen der høyde$\times$(diameter)^2 benyttes som uavhengig variabel? Utfør testen og kommenter resultatet.

Vi vil til slutt benytte regresjonsmodellen (1) uten konstantledd fra forrige oppgave til å analysere tredataene når den uavhengige variabelen er høyde$\times$(diameter)^2. Du kan benytte resultatene fra Oppgave 1 selv om du ikke klarte å vise disse.

d) Estimer b i regresjonsmodellen (1). Estimer også residualvariansen σ^2. For å lette beregningene oppgis det at med y_i =volum av tre nummer i og x_i =høyde$\times$(diameter)^2 av tre nummer i, blir $\sum_{i=1}^{31} y_i^2 = 3,63250{\cdot}10^4$, $\sum_{i=1}^{31} x_i^2 = 8,13308{\cdot}10^9$ og $\sum_{i=1}^{31} x_i y_i = 1,71454{\cdot}10^7$.

e) Bestem et 95% konfidensintervall for b.

Oppgave 3.

Eksponensialfordelingen brukes ofte til å beskrive levetiden til tekniske komponenter. Vi skal i denne oppgaven se nærmere på noen statistiske metoder for eksponensialfordelte data.

Vi antar derfor at $Y_1, Y_2, \ldots, Y_n$ er uavhengige og identisk fordelte med sannsynlighetstetthet

$$f_Y(y) = \lambda e^{-\lambda y};\ y > 0,$$

hvor $\lambda > 0$.

a) Sett opp likelihoodfunksjonen og vis at sannsynlighetsmaksimeringsestimatoren (SME) for λ blir $\widehat{\lambda} = n/\sum_{i=1}^{n} Y_i$.

b) Bestem den tilnærmete fordelingen til $\widehat{\lambda}$ når n er stor.

(Fortsettes side 4.)

c) Vis at $2\lambda \sum_{i=1}^{n} Y_i$ er kji-kvadratfordelt med $2n$ frihetsgrader.

d) Utled en test med 5% nivå for testing av nullhypotesen $H_0 : \lambda \leq \lambda_0$ mot alternativet $H_1 : \lambda > \lambda_0$, hvor λ_0 er et gitt tall.

e) Utled styrkefunksjonen til testen. Hvor stor må n (minst) være for at teststyrken skal være 90% når $\lambda = 2\lambda_0$?

Oppgave 4.

I perioden 1950-85 ble det registrert 58626 firebarnsmødrene i Norge. Av disse hadde 4699 fire gutter, 15251 hadde en jente og tre gutter, 20806 hadde to jenter og to gutter, 13901 hadde tre jenter og en gutt, mens 3969 hadde fire jenter (*Tidsskrift for Den Norske Lægeforening*, nr. 14, 1987)

a) Det antas ofte at antall jenter i en firebarnsfamilie er binomisk fordelt med $n = 4$ og p sannsynligheten for at et tilfeldig valgt barn skal være en jente. Bestem sannsynlighetsmaksimeringsestimatet for p i en slik binomisk modell ut fra dataene over. (Svar: $\widehat{p} = 0,488$.)

b) Hvordan vil du gå fram for å vurdere om den binomiske modellen gir en rimelig beskrivelse av virkeligheten? Utfør din analyse og kommenter resultatet.

SLUTT