UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST-IN 216 — Sentrale statistiske modeller og metoder.
Eksamensdag:	Torsdag 14. desember 2000.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabeller over standardnormalfordeling, χ^2- og F-fordelinger.
Tillatte hjelpemidler:	Rottmans formelsamling, Formelsamlingene for ST101 og ST102, lommekalkulator.

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

Oppgave 1.

Anta følgende situasjon: Vi skal studere relasjonen mellom kornavling og gjødselmengde for to forskjellige gjødseltyper. Vi samler data ved å måle verdiene av avling for 3 forskjellige nivåer for hver av de to gjødseltypene (nivåene kan godt være forskjellige for de to gjødseltypene). Vi måler to uavhengige verdier av avling (replikater) for hvert nivå av hver type gjødsel (på for eksempel to forskjellige jordstykker). La

$$y_{ijk} = \mu + \alpha_i + \gamma x_{ij} + e_{ijk}$$

være modellen for denne situasjonen. Her er y_{ijk} responsen for k-te replikat for gjødseltype i og gjødselnivå j. Konstantene α_j angir nivåforskjellene for de to gjødseltypene, x_{ij} representerer gjødselmengde for j'te nivå av gjødseltype i, γ er regresjonskoeffisienten for x_{ij} og e_{ijk} er et tilfeldig feilledd. Indeksen i antar verdiene 1 eller 2 avhengig av gjødseltype, j antar verdien 1, 2 eller 3 avhengig av gjødselnivå og k antar verdien 1 eller 2 avhengig av måling (replikat).

a) Sett opp dette som en regresjonsmodell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$. Hva blir $\boldsymbol{X}$ og hva blir $\boldsymbol{b}$? Hvilken antagelse gjør vi her om forskjellen/likheten av effekten av gjødselnivå for de to gjødseltypene?

(Fortsettes side 2.)

b) Vi antar at vi har samlet inn tall y_{ijk}. Hvor mange målinger av y svarer dette oppsettet til? I tabellen presenteres variansanalysetabellen for denne modellen. Fyll ut de stedene som mangler verdi (de samme reglene som i vanlig ANOVA gjelder også her for utregning av MS og F). Hva vil du si om betydningen av mengden gjødsel? Virker den inn på avlingen? Skriv også opp nullhypotesene til testene som svarer til de to p-verdiene.

Effekt	SS	DF	MS	F	p-verdi
Gjødseltype (α_i)	1.54	1			0.0059
Gjødselmengde (γ)	10.35	1			0.0001
Error (residual)	1.077	9			
Total (SYY)	12.97				

c) R^2 for disse dataene var lik 0.92. Hva betyr det? Kombiner noen av de oppgitte tallene til å regne ut regresjonskvadratsummen (SS_{reg}) for denne modellen?

d) En måte å få entydige estimater av regresjonskoeffisientene på i slike situasjoner er å legge restriksjoner på dem. En vanlig slik restriksjon er å anta at $\alpha_1 + \alpha_2 = 0$. Under denne restriksjonen blir parameterne i modellen lik $\hat{\mu} = 0.32$, $\hat{\alpha}_1 = 0.36$, $\hat{\alpha}_2 = -0.36$, $\hat{\gamma} = 1.14$. Tegn opp regresjonslinjene for de to gjødseltypene i området $3 \leq x_{ij} \leq 10$.

Oppgave 2.

Anta at vi har 33 observasjoner av 10 x-variable. I tabellen under angis egenverdier for kovariansmatrisen til disse. `PRIN1` betyr prinsipal komponent nr. 1, osv.

a) Fyll ut de verdiene som mangler i tabellen. Hvor mange prosent av variasjonen beskriver de to første komponentene.

	Eigenvalue	Proportion	Cumulative
PRIN1	.	.	0.70978
PRIN2	4.1356	0.174259	.
PRIN3	1.8702	0.078802	0.96284
PRIN4	0.4270	0.017992	.
PRIN5	0.2292	0.009659	0.99049
PRIN6	0.0990	0.004172	0.99466
PRIN7	0.0470	0.001981	0.99664
PRIN8	0.0375	0.001579	0.99822
PRIN9	0.0231	0.000972	0.99919
PRIN10	0.0192	0.000808	1.00000

(Fortsettes side 3.)

b) Lag et todimensjonalt plott av de to første ladningsvektorene (svarende til `PRIN1` og `PRIN2` i tabellen under). Kan du fra plottet si noe om korrelasjoner mellom variablene?

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	0.066837	0.017481	-.098841	0.042169	0.933173
X2	-.123512	0.402030	-.253089	-.152552	-.062140
X3	-.003288	-.008179	-.082976	-.018520	0.079650
X4	0.237902	-.555921	0.622191	-.148393	0.067659
X5	0.000912	0.688978	0.692613	0.156225	0.074986
X6	0.569792	0.046254	-.111791	0.404738	-.217089
X7	-.372591	-.145508	0.050723	0.523682	0.093982
X8	0.568314	0.072007	-.152926	0.349297	0.102988
X9	-.360983	-.160261	0.037121	0.606391	-.082692
X10	-.083022	-.008941	-.112349	0.020138	0.188089

c) Hvilke variable har størst og minst bidrag til den første komponenten? Er det noe i de ladningene du ser over som tilsier at man kanskje burde vektet/standardisert disse variablene.

d) Beskriv forskjellen mellom scoreplot og ladningsplot og forklar hvordan de brukes.

Oppgave 3.

a) I en studie av forekomst av Type II diabetes fant man diabetes hos 97 av 21779 "normalvektige" og hos 623 av 16897 "overvektige".

Beregn odds-ratioen for diabetes mellom overvektige og normalvektige. Forklar hva odds-ratioen estimerer og hvordan den fortolkes for disse dataene. (Med overvektig menes Body Mass Indeks = BMI = (vekt i kg)/(høyde i m)$^2 \geq 25$ og normalvektig BMI< 25.)

b) La $x_i' = (x_{i1}, \ldots, x_{ip})$ være en kovariat, $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ en regresjonsparameter og $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ en lineær prediktor for individ nr. i. Med $Y_i =$ indikator for at individ i har diabetes antas den logistiske regresjonsmodellen

$$P(Y_i = 1) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \, .$$

Forklar hvorfor e^{β_j} kan fortolkes som en odds-ratio for $j = 1, \ldots, p$.

c) På neste side er det angitt resultater fra diabetesstudien mot kovariatene x_{i1} = alder (i år) og x_{i2} = BMI (målt kontinuerlig). Beregn odds-ratioen for diabetes

1) mellom 50-åringer og 25 åringer
2) mellom individer med BMI = 20 og BMI = 30
3) mellom en 50-åring med BMI = 30 og en 25-åring med BMI = 20.

(Fortsettes side 4.)

```
> summary(glm(diab~alder+bmi,family=binomial))

                  Value  Std. Error   t value
(Intercept) -11.63876355 0.267396487 -43.52624
      alder   0.04117991 0.002989308  13.77573
        bmi   0.20752482 0.007905826  26.24961
    Null Deviance: 7107.495 on 38675 degrees of freedom

Residual Deviance: 6080.94 on 38673 degrees of freedom

Correlation of Coefficients:
      (Intercept)      alder
alder -0.5108159
  bmi -0.7726740  -0.1316488
```

d) Finn 95% konfidensintervall for odds-ratioene du beregnet i forrige punkt. Finn også et 95% konfidensintervall for sannsynligheten for at en 50-åring med BMI = 30 skal få diabetes.

e) Under er angitt resultater fra logistiske regresjonsanalyser bare mot alder og bare mot BMI.
Gi en forklaring på fenomenet som opptrer. Korrelasjonen mellom alder og BMI var lik 0.32 i denne studien.

```
> summary(glm(diab~alder,family=binomial))

                  Value  Std. Error   t value
(Intercept) -6.74306128 0.158562830 -42.52612
      alder  0.05467347 0.002768061  19.75154

    Null Deviance: 7107.495 on 38675 degrees of freedom

Residual Deviance: 6689.684 on 38674 degrees of freedom

> summary(glm(diab~bmi,family=binomial))

                  Value  Std. Error   t value
(Intercept) -10.0706151 0.218932856 -45.99865
        bmi   0.2280191 0.007493923  30.42720

    Null Deviance: 7107.495 on 38675 degrees of freedom

Residual Deviance: 6282.455 on 38674 degrees of freedom
```

Oppgave 4.

a) I levetidsanalyse er sensurering et sentralt begrep. Forklar hva sensurerte levetidsdata er og hvorfor man trenger å utvikle egne metoder for denne type data.

b) La T være en levetid og $S(t) = P(T > t)$ levetidsfunksjonen for T. Denne estimeres ofte ved Kaplan-Meier estimatoren

(Fortsettes side 5.)

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{Y(t_j)}\right)$$

der $0 < t_1 < t_2 < \cdots$ er tidene for dødsfall, d_j = antall som dør ved t_j og $Y(t_j)$ = antall som fortsatt er under observasjon like før tid t_j.
Gi en begrunnelse for denne estimatoren.

c) Intensiteten til T er definert ved

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t < T \le t + \Delta | T \ge t)$$

og den kumulative intensiteten til T ved $\Lambda(t) = \int_0^t \lambda(s)ds$.

Utled sammenhengen $S(t) = e^{-\Lambda(t)}$ og foreslå to estimatorer for $\Lambda(t)$.

d) La $\hat{S}_0(t)$ og $\hat{S}_1(t)$ være Kaplan-Meier estimatorene i to ulike grupper. Foreslå og begrunn en grafisk metode for å sjekke om dødeligheten i de to gruppene kan beskrives ved en proporsjonal intensitetsmodell $\lambda_1(t) = R\lambda_0(t)$ der $\lambda_j(t)$ = intensiteten i gruppe j.

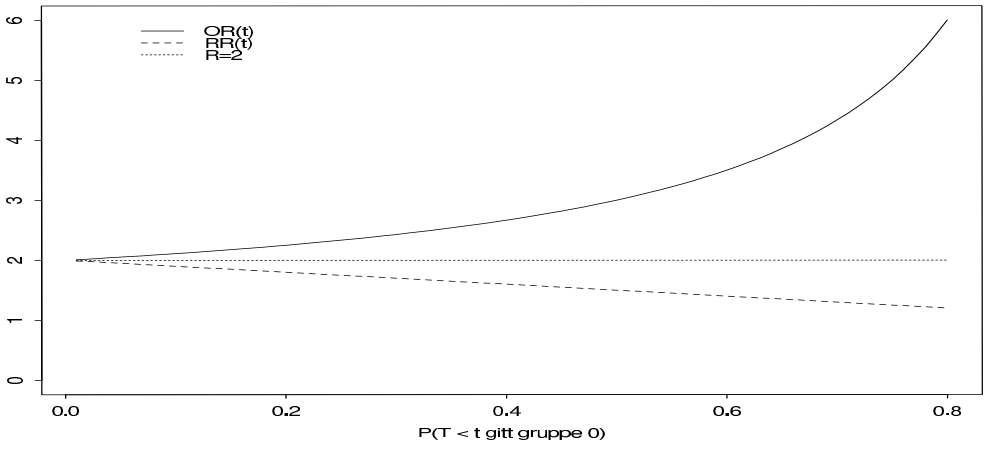
e) Anta nå at den proporsjonale intensitetsmodellen $\lambda_1(t) = R\lambda_0(t)$ holder. Finn et uttrykk for den relative risken

$$RR(t) = \frac{P(T \le t \mid \text{Gruppe 1})}{P(T \le t \mid \text{Gruppe 0})},$$

(dvs. $RR(t)$ er forholdet mellom sannsynlighetene for død i de to gruppene før t) ved hjelp av intensitetsratioen R og overlevelsesfunksjonen $S_0(t)$ i gruppe 0. Sett også opp et lignende uttrykk for odds-ratioen for dødelighet mellom de to gruppene inntil tid t

$$OR(t) = \frac{P(T \le t \mid \text{Gruppe 1})}{P(T > t \mid \text{Gruppe 1})} \Big/ \frac{P(T \le t \mid \text{Gruppe 0})}{P(T > t \mid \text{Gruppe 0})}.$$

I figuren under er $OR(t)$ og $RR(t)$ plottet mot $F_0(t) = 1 - S_0(t)$ for intensitetsratio $R = 2$. Diskuter på bakgrunn av figuren likhet og forskjell mellom begrepene intensitets-ratio, relativ risk og odds-ratio.



SLUTT