UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2 - FASIT
Eksamensdag:	Mandag 30. mai 2005.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabeller over normal-, 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) Modell:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

der ε_{ijk}-ene er uavhengige og normalfordelte med forventning 0 og felles varians σ^2. Da modellen er overparametrisert, må en legge inn begrensningene

$$\sum_i \alpha_i = \sum_j \beta_i = \sum_i \delta_{ij} = \sum_j \delta_{ij} = 0.$$

Full anova tabell:

Source	df	SS	MS	F
Kjonn	1	2.133	2.133	0.6410
Signaltype	2	97.267	48.634	14.6119
Interaksjon	2	23.267	11.634	3.4952
Error	24	79.880	3.328	
Total	29	202.547		

(Fortsettes side 2.)

(b) For å teste om `Interaksjon` er en signifikant faktor, tar vi utgangspunkt i null-hypotesen

$H_{AB} : \delta_{ij} = 0$ for alle i, j.

Under H_{AB} er $F_{AB} = \frac{SS_{AB}/(I-1)(J-1)}{SS_E/IJ(K-1)}$ Fisher-fordelt med $((I-1)/J-1), IJ(K-1)) = (2, 24)$ frihetsgrader. En kan da forkaste H_{AB} på α nivå hvis $F_{AB} > F_{1-\alpha,2,24}$.

For å teste om `Interaksjon` er signifikant, ser vi at $F_{AB} = 3.4952$ som er større enn $F_{0.95,2,24} = 3.40$ men mindre enn $F_{0.975,2,24} = 4.32$ noe som betyr at vi ville forkaste $H_{AB} : \delta_{ij} = 0$ på 0.05 nivå men ikke på 0.025 nivå.

For å teste om `Kjonn` er signifikant, ser vi at $F_A = 0.6410$ som er mindre enn $F_{0.9,1,24} = 2.93$, som medfører at det ikke er grunnlag for å forkaste H_A. Merk dog at dette *kan* endre seg hvis vi velger å ta bort interaksjonsleddet.

For å teste om `Signaltype` er signifikant, ser vi at $F_B = 14.6119$ som er større enn $F_{0.99,2,24} = 5.61$ noe som betyr at vi vil forkaste $H_B : \beta_1 = \beta_2 = \beta_3 = 0$ på $\alpha = 0.01$ nivå.

Merk dog at de to siste tester ikke er så aktuelle å utføre hvis ikke interaksjonsleddet er fjernet først.

Oppgave 2.

(a) Regresjonsmodell:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + e_i$$

der y_i er pris på bolig i mens $x_{i,1}$ og $x_{i,2}$ er tilhørende kvm og takst. Det antas her e_i-ene er uavhengige og normalfordelte med forventning 0 og felles varians σ^2.

Første kollonne er estimat på β-ene, 2. kollonne er standard feil, 3. kollonne er testobservator for hypotesen $H_0 : \beta_j = 0$, som er estimat delt på standard feil mens siste kollonne er tilhørende p-verdi.

Residual standard error: $\hat{\sigma}$.

Multiple R-squared: R^2, som beskriver forklaringsgraden til modellen.

(b) Vi har

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * 105 + \hat{\beta}_2 * 23 = 24.0$$

(Fortsettes side 3.)

Videre har vi at

$$\begin{aligned} var[y-\hat{y}] =& var[\beta_0+\beta_1 x_1+\beta_2 x_2+e_i-\hat{\beta}_0-\hat{\beta}_1 x_1-\hat{\beta}_2 x_2] \\ =& \sigma^2(1+var[\beta_0]+x_1^2 var[\beta_1]+x_2^2 var[\beta_2]+2x_1 cov[\beta_0,\beta_1]+2x_2 cov[\beta_0,\beta_2]+2x_1x_2c \end{aligned}$$

Et 95% prediksjonsintervall kan vi få ved

$$\hat{y} \pm t_{0.025,23} s_{\hat{y}-y} = [23.42, 24.58]$$

(c) Det er en sterk korrelasjon mellom de to forklaringsvariablene, noe som gjør at de prøver å forklare det samme.

Oppgave 3.

(a) Vi har

$$L(\boldsymbol{\beta}) = \prod_{c=1}^{C} [1-e^{-(\beta_0+\beta_1 x+\beta_2 x^2)}]^{y_c} [e^{-(\beta_0+\beta_1 x+\beta_2 x^2)}]^{n_c-y_c}$$

som gir log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{c=1}^{C} [y_c \log[1-e^{-(\beta_0+\beta_1 x+\beta_2 x^2)}] - (n_c-y_c)(\beta_0+\beta_1 x+\beta_2 x^2)]$$

og dermed

$$\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum_{c=1}^{C} x_c^j [y_c \frac{e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}}{1-e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}} - (n_c-y_c)], \quad j=0,1,2$$

Videre har vi at

$$\begin{aligned} E[\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta})] =& \sum_{c=1}^{C} x_c^j [E[y_c] \frac{e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}}{1-e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}} - (n_c - E[y_c])] \\ =& \sum_{c=1}^{C} x_c^j [n_c e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)} - n_c(1-(1-e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)})] \\ =& 0 \end{aligned}$$

som svarer til hva den generelle teorien sier.

(b) En har at under generelle antagelser så vil maksimum likelihood estimater være tilnærmet forventningsrette og normalfordelte. Dette gir tilnærmede konfidensintervaller

$$\hat{\beta}_1 \pm z_{\alpha/2} s_{\hat{\beta}_1} = 0.1627 \pm 1.96\sqrt{13.2389/1000} = [-0.0628, 0.3882]$$

(Fortsettes side 4.)

(c) Standard bootstrap konfidensintervall:

$[2\hat{\beta}_1 - \bar{\beta}_1^*, 2\hat{\beta}_1 - \underline{\beta}_1^*] = [-0.0698, 0.3848]$

Alternativt, $\underline{\delta} = -0.0594 - 0.1627 = -0.2221, \overline{\delta} = 0.2325$ som gir intervall

$[\hat{\beta}_1 - \overline{\delta}, \hat{\beta}_1 - \overline{\delta}] = [-0.0698, 0.3848]$

Ganske likt normaltilnærming som er naturlig da plottene viser at $\hat{\beta}_1$ er rimelig normalfordelt.

Oppgave 4.

Vi kan teste på uavhengighet ved å bruke

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

som testobservator. Under $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ er X^2 tilnærmet kji-kvadrat fordelt med $(I-1)(J-1) = 2$ frihetsgrader.
Forventninger:

	Alder		
Opinion	18-29 år gamle	30-49 år gamle	50 år eller eldre
For	163.65	303.92	275.43
Imot	60.45	112.08	101.57

som gir $X^2 = 6.6814$. Denne verdien er mindre enn $\chi^2_{0.975,2} = 7.38$ men større enn $\chi^2_{0.95,2} = 5.99$ som viser at data gir grunnlag for å forkaste H_0 på et $\alpha = 0.0.05$ nivå (P-verdien ligger mellom 0.025 og 0.05).

SLUTT