TENTAMEN: Statistisk modellering för I3, TMS161, måndagen den 9 januari 2006 kl 8.30-11:30 på V. **Jour:** Magnus Karlsson, tel: 772 42 91.
Hjälpmedel: Utdelad formelsamling med tabeller, BETA, på kursen använd ordlista och typgodkänd räknedosa.
Poängberäkning: Uppgifterna är av flervalstyp, där endast ett alternativ är rätt. Korrekt besvarad uppgift ger 2 poäng, obesvarad uppgift (vet inte eller alternativ f) ger 0 poäng och felaktigt besvarad uppgift ger -0.5 poäng (flera ifyllda alternativ ger automatiskt -1/2 poäng). Inlämnade lösningar kommer ej tas hänsyn till vid rättningen. Fyll i och lämna in denna sida.
Svar: Läggs ut på `www.math.chalmers.se/~anders.sjogren/StatMod/` efter tentamens slut.

Uppgift	a	b	c	d	e	f (vet ej)	Poäng
1	☐	☐	☐	☐	☐	☐	
2	☐	☐	☐	☐	☐	☐	
3	☐	☐	☐	☐	☐	☐	
4	☐	☐	☐	☐	☐	☐	
5	☐	☐	☐	☐	☐	☐	
6	☐	☐	☐	☐	☐	☐	
7	☐	☐	☐	☐	☐	☐	
8	☐	☐	☐	☐	☐	☐	
9	☐	☐	☐	☐	☐	☐	
10	☐	☐	☐	☐	☐	☐	
11	☐	☐	☐	☐	☐	☐	
12	☐	☐	☐	☐	☐	☐	
13	☐	☐	☐	☐	☐	☐	
14	☐	☐	☐	☐	☐	☐	
15	☐	☐	☐	☐	☐	☐	

1 I ett försök undersöks hur faktorerna magnesiumhalt och kalciumhalt i mat påverkar blodtrycket hos råttor. Vardera av faktorerna testas på tre olika nivåer: låg, mellan och hög. Man ansätter en generell modell:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Resultaten från försöken gav en samverkan som inte går att transformera bort.

Läs nu följande påståenden om den fortsatta analysen av data:

1: Det viktigaste är att testa om det finns någon skillnad på huvudeffekterna.

2: Parvisa jämförelser av effekter kan göras m.h.a. Tukeys metod.

3: Samverkan tyder på modellfel.

Vilket eller vilka av dessa påståenden är korrekt/korrekta?

(a) ☐ Påstående 1 är sant, men inte de andra.

(b) ☐ Påstående 1 och 2 är sanna, men inte 3.

(c) ☐ Påstående 2 är sant, men inte de andra.

(d) ☐ Påstående 3 är sant, men inte de andra.

(e) ☐ Inget påstående är sant.

(f) ☐ Vet ej.

2 Nedan visas en samverkansplot över ett två-faktorförsök, där vardera faktor hade tre nivåer. Man ansätter en generell modell:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Läs följande påståenden om figuren:

1: Faktorerna A och B verkar additiva.

2: Genom att transformera responsen kommer det att gå att få fram en additiv modell.

3: Figuren tyder på felaktiga modellantaganden.

Vilket eller vilka av dessa påståenden är korrekt/korrekta?

(a) ☐ Påstående 1 är sant, men inte de andra.

(b) ☐ Påstående 1 och 2 är sanna, men inte 3.

(c) ☐ Påstående 2 är sant, men inte de andra.

(d) ☐ Påstående 3 är sant, men inte de andra.

(e) ☐ Inget påstående är sant.

(f) ☐ Vet ej.

3 Vi undersöker effekten av en faktor med en envägs variansanalys med blockning. Blockfaktorn har 3 nivåer och den undersökta faktorn har 4 nivåer. För varje kombination av huvudfaktorn och blockfaktorn mäts vår svarsvariabel 5 gånger.

Ett giltigt villkor för att effekterna hos minst ett enskilt par av nivåer skiljer sig åt signifikant på en total signifikansnivå av 5% är att det minsta p-värdet bland alla parvisa jämförelser är mindre än $0.05/k$, för

(a) ☐ $k = 3$.

(b) ☐ $k = 4$.

(c) ☐ $k = 5$.

(d) ☐ $k = 6$.

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej

4 I en undersökning ville man avgöra hur åldern på chefen för småföretag påverkade vilken lön han/hon fick. Man samlade därför in 59 datavärden från olika småföretag i USA. Låt X_i och Y_i vara åldern respektive lönen för chef nr i. (Åldern mäts här i år och lönen i tusentals dollar.) För de data som man samlat in går det bra att anpassa en enkel linjär regressionsmodell. Skattningen av medelresponsen för denna modell blir:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X = 242.70 + 3.13 \cdot X$$

p-värdena för test om β_0 och β_1 är noll är 0.012 respektive 0.342. Läs nu följande påståenden:

1: Modellen är ej relevant för X nära 0.

2: Det finns en stark relation mellan X och Y eftersom β_1 har ett relativt högt p-värde.

3: Pearsons korrelationskoefficient för X och Y bör ligga nära 1.

Vilket eller vilka av dessa påståenden är korrekt/korrekta?

(a) ☐ Påstående 1 är sant, men inte 2 och 3.

(b) ☐ Påstående 2 är sant, men inte 1 och 3.

(c) ☐ Påstående 2 och 3 är sanna, men inte 1.

(d) ☐ Påstående 1 och 3 är sanna, men inte 2.

(e) ☐ Inget påstående är sant.

(f) ☐ Vet ej.

5 I en tvåvägstabell med c kolumner och r rader har teststatistikan X^2 approximativt en χ^2-fördelning med $(c-1)(r-1)$ frihetsgrader.

Män och kvinnor fick svara på om de föredrar McDonald's eller Burger King. Svaren presenteras i tabellen nedan. (Fingerade data)

	McDonald's	Burger King
Man	51	42
Kvinna	41	40

X^2 beräknas till 0.310. Vilken är den starkaste slutsatsen vi kan dra från denna undersökning?

(a) ☐ På signifikansnivå 0.10 (men inte på 0.05) kan vi säga att kvinnor fördrar Burger King framför McDonald's i större utsträckning än män

(b) ☐ På signifikansnivå 0.05 (men inte på 0.025) kan vi säga att kvinnor fördrar Burger King framför McDonald's i större utsträckning än män

(c) ☐ På signifikansnivå 0.025 (men inte på 0.1) kan vi säga att kvinnor fördrar Burger King framför McDonald's i större utsträckning än män

(d) ☐ På signifikansnivå 0.1 kan vi säga att kvinnor fördrar Burger King framför McDonald's i större utsträckning än män

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej

6 Tabellen nedan visar ANOVA-tabellen för en tvåsidig variansanalys.

Analysis of variance				
Source	DF	Sum of squares	Mean square	F Stat
A	2	512.9	*	F_a
B	*	449.5	*	F_b
A×B	*	143.1	*	F_{ab}
Error	15	136.0	*	
Total	29	*		

Värdena i några av fälten saknas och är markerade med (*). Från de givna siffrorna kan man ändå beräkna F-statistikorna F_a, F_b och F_{ab}. De är:

(a) ☐ F_a=18.86, F_b=9.92, F_{ab}=1.58

(b) ☐ F_a=56.57, F_b=24.79, F_{ab}=0.26

(c) ☐ F_a=225.45, F_b=98.79, F_{ab}=3.95

(d) ☐ F_a=28.28, F_b=12.39, F_{ab}=1.97

(e) ☐ Inget av ovanstående.

(f) ☐ vet inte.

7 En racing-fantast vill undersöka om valet bland tre möjliga motormodifikationer påverkar hans tid på kvartsmilen. Eftersom han är statistiskt sinnad urformar han en försöksdesign där han kör en gång med varje modifikation under måndag till fredag under en vecka. Den inbördes ordningen bland de tre loppen under samma dag slumpas fram.

Fantasten är osäker på om veckodagen har en systematisk effekt, varför han gör envägs variansanalys både med och utan blockning. Resultaten av hans analyser syns i de två ANOVA-tabellerna nedan.

Analysis of variance					
Source	DF	Sum of squares	Mean square	F Stat	p-value
A	2	2.86	1.43	5.13	0.0245
Error	12	3.34	0.279		

Analysis of variance					
Source	DF	Sum of squares	Mean square	F Stat	p-value
A	2	2.86	1.43	4.21	0.0563
B	4	0.629	0.157	0.463	0.761
Error	8	2.71	0.339		

Läs följande påståenden:

i Valet av modifikation spelar signifikant roll på 5%-nivå i analysen utan blockning.

ii Valet av modifikation spelar signifikant roll på 5%-nivå i analysen med blockning.

iii Blockningen verkar inte vara nödvändig.

Vilket eller vilka av påståendena är sanna?

(a) ☐ Endast i.

(b) ☐ Endast ii.

(c) ☐ Endast i & iii.

(d) ☐ Endast ii & iii.

(e) ☐ Inget av ovanstående alternativ.

(f) ☐ Vet ej

8 Chefen på ett företag vill analysera om andelen producerade enheter som är defekta skiljer sig mellan olika produktionsavdelningar. På varje avdelning tas ett stickprov. Stickprovsstorleken och antalet defekta enheter inrapporteras. Hur ska chefen analysera dessa data på bästa sätt?

(a) ☐ Med χ^2-test för kategoriska data.

(b) ☐ Med ensidig variansanalys utan blockning.

(c) ☐ Med ensidig variansanalys med blockning.

(d) ☐ Med tvåsidig variansanalys

(e) ☐ Med regressionsanalys

(f) ☐ Vet ej

9 Beakta följande påståenden, i sammanhanget linjär regression:

i Externt studentiserade residualer har alltid större absolutvärde än motsvarande internt studentiserade residualer.

ii Ett symmetriskt konfidensintervall är alltid större än motsvarande symmetriska prediktionsintervall.

iii Internt studentiserade residualer är bättre lämpade för att upptäcka outliers än motsvarande externt studentiserade residualer.

Vilket eller vilka av påståendena är sanna?

(a) ☐ Endast i.

(b) ☐ Endast ii.

(c) ☐ Endast iii.

(d) ☐ Endast i & ii.

(e) ☐ Inget av ovanstående alternativ.

(f) ☐ Vet ej

10 Betrakta följande påståenden.

1 Hög multilkolinearitet indikerar linjärt beroende mellan regressorerna.

2 Multikolinearitet mäter icke-linjärt samband mellan svarsvariabeln och regressorerna.

3 Ett problem med hög multikollinearitet är att den gör att det kan vara svårt att skilja effekterna av olika regressorer från varandra.

Vilket eller vilka av påståendena är korrekta?

(a) ☐ Endast 1 är korrekt.

(b) ☐ Endast 2 är korrekt.

(c) ☐ Endast 3 är korrekt.

(d) ☐ 1 och 2 är korrekta. 3 är falskt.

(e) ☐ 1 och 3 är korrekta. 2 är falskt.

(f) ☐ Vet ej

11 Data har insamlats där man har mätt verkningsgraden i en förbränningsprocess mot temperaturen vid vilken förbränningen sker. Man försöker hitta en bra modell för sina data och testar då följande två modeller:
Modell 1: $Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$ där $\epsilon_i \in N(0, \sigma^2)$ och oberoende.
Modell 2: $Y_i = \beta_2 + \beta_3 \cdot X_i^2 + \delta_i$ där $\delta_i \in N(0, \sigma^2)$ och oberoende.

På nästa sida visas två grafer för modell 1 och två grafer för modell 2, en qq-plot och en plot av studentiserade residualer mot de anpassade värdena. Vad kan man utifrån dessa grafer säga om de båda modellerna?

(a) ☐ Modell 2 är bättre eftersom en kvadratisk modell aldrig är känslig för i vilken ordningen försöken har gjorts.

(b) ☐ Modell 1 är bättre eftersom den är studentiserad.

(c) ☐ Modell 2 bättre eftersom den bättre uppfyller modellantagandena.

(d) ☐ Modell 1 är bättre eftersom den bättre uppfyller modellantagandena.

(e) ☐ Modell 1 är bättre eftersom en linjär modell aldrig är känslig för i vilken ordning försöken har gjorts.

(f) ☐ Vet ej.

12 För att studera effekten av två olika mediciner har man genomfört ett två-faktorförsök. Faktorerna var medicin A och medicin B och nivåerna för dessa var 0 eller 1, där 0 motsvarar att man inte delar ut någon medicin och 1 motsvarar att man ger medicin. Vid en tvåsidig variansanalys av de observerade effekterna fick man följande ANOVA-tabell

Analysis of variance					
Source	DF	Sum of squares	Mean square	F Stat	Prob > F
Medicin A	1	153.27	153.27	17.67	0.000
Medicin B	1	120.06	120.06	13.84	0.000
A×B	1	0.06	0.06	0.01	0.921
Error	36	312.31	8.68		
Total	39	585.70			

Läs följande påståenden om ANOVA-tabellen:

1: Totalt har 40 mätningar genomförts.

2: Medicin A verkar inte påverka effekten av medicin B.

3: Medicin B tycks inte ge någon effekt.

4: Ett rimligt nästa steg i analysen är att anpassa en additiv modell.

Vilket eller vilka av dessa påståenden är korrekt/korrekta?

(a) ☐ Påstående 2 är sant, men inte 1, 3 och 4.

(b) ☐ Påstående 3 och 4 är sanna, men inte 1 och 2.

(c) ☐ Påstående 1, 2 och 4 är sanna, men inte 3.

(d) ☐ Påstående 1, 2 och 3 är sanna, men inte 4.

(e) ☐ Inget påstående är sant.

(f) ☐ Vet ej.

13 Du arbetar som marknadsföringsansvarig för en butik och sitter och väljer mellan att satsa på en marknadsföringskampanj eller att locka kunder med nedsatta priser. För att få råd analyserar du gamla data, där du har försäljningssiffror under perioder med olika kombinationer av de båda åtgärderna. Du har tre olika nivåer på varje åtgärd: ingen, lite och mycket. I nuläget passar det bara att antingen ha lite marknadsföring, eller ha lite nedsatta priser.

Du drar slumpmässigt ur dina gamla data, så att du får försäljningssiffran från 5 veckor för varje av de nio kombinationerna av faktor-nivåer (totalt 45 mätvärden).

För att undersöka skillnaden mellan kombinationerna lite marknadsföring/ ingen prisnedsättning och ingen marknadsföring/lite prisnedsättning använder du en tvåvägs faktormodell där du undersöker den aktuella kontrasten. Du erinrar dig dock att det kan vara problem med multipla jämförelser, och i detta fall finns ju 36 sådana möjliga parvisa jämförelser av kombinationer av marknadsföringsnivå och prisnedsättningsnivå.

Vilket av följande alternativt är i första hand giltigt och i andra hand det mest kraftfulla sättet att analysera den aktuella kontrasten, bland följande alternativ:

(a) ☐ Ett t-test utan korrektion för de 36 möjliga jämförelserna.

(b) ☐ Ett t-test med Scheffés korrektion för de 36 möjliga jämförelserna.

(c) ☐ Ett t-test med Tukey's korrektion för de 36 möjliga jämförelserna.

(d) ☐ Ett t-test med Bonferroni's korrektion för de 36 möjliga jämförelserna.

(e) ☐ Yate's viktade kvadratmedelvärdesteknik.

(f) ☐ Vet ej

14 Kalle har kommit över ett intressant dataset med löner för ett tvärsnitt av befolkningen. Datasetet innehåller uppgifter om varje individs årslön Y (i dollar), hur många års utbildning individen har X_1 och antalet år individen arbetat i samma företag X_2. Han vill passa datan till en multipel regressionsmodell. För att tillgodose kravet på normalfördelade residualer visar det sig lämpligt att som svarsvariabel använda logaritmen av årslönen.

$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, där ϵ är $N(0, \sigma)$.

Anpassning av data till modellen ger: $\hat{\beta}_0 = 10.2$ $\hat{\beta}_1 = 0.03$ $\hat{\beta}_2 = 0.04$

Kalle har studerat i 12 år. Han funderar på att läsa vidare på ett treårigt program. Kalle vill räkna ut hur mycket högre årslön han kan förvänta sig enligt modellen efter fem år i ett företag om han väljer att studera 3 år till.

Hur mycket högre årslön kan Kalle förvänta sig enligt modellen efter fem år i ett företag om han väljer att studera 3 år till?

(a) ☐ 8397.5 dollar

(b) ☐ 32859.5 dollar

(c) ☐ 3343.5 dollar

(d) ☐ 1094 dollar

(e) ☐ 4435.5 dollar

(f) ☐ Vet ej

15 Betrakta figurerna 1-3. I vilken figur eller vilka figurer verkar det rimligt att anpassa en linjär regressionsmodell utan att transformera vare sig X eller Y?

(a) ☐ I samtliga figurer.

(b) ☐ I figurerna 2 och 3, men inte 1.

(c) ☐ I figurerna 1 och 2 men inte 3.

(d) ☐ Endast i figur 2.

(e) ☐ I ingen av figurerna.

(f) ☐ Vet ej.