TENTAMEN: Statistisk modellering för I3, TMS161, lördagen den 22 Oktober kl 8.30-11.30 på V. **Jour:** John Gustafsson, ankn. 5316.
Hjälpmedel: På hemsidan tillgänglig ordlista och formelsamling med tabeller, BETA samt 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 efter tentamens slut ut på hemsidan:
http://www.math.chalmers.se/∼anders.sjogren/StatMod/

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 För att studera kostnaden för olika utbildningar åren 1988-1990, samlade man in data från 5 olika utbildningar under de tre åren. Vid en tvåsidig variansanalys av de observerade kostnaderna fick man följande ANOVA-tabell

Analysis of variance					
Source	DF	Sum of squares	Mean square	F Stat	Prob > F
År	2	9010	4505	81.0	0.000
Utbildn.	4	9349	2337	42.0	0.000
År×Utb.	8	1182	148	2.7	0.0127
Error	75	4173	56		
Total	89	23714			

Vid test på 5% signifikansnivå kan vi därmed dra följande slutsats:

(a) ☐ Både typen av utbildning och vilket år man undersöker har effekt på kostnaden, och någon eller några av åren har olika effekt på kostnaden beroende på vilken utbildning man tittar på.

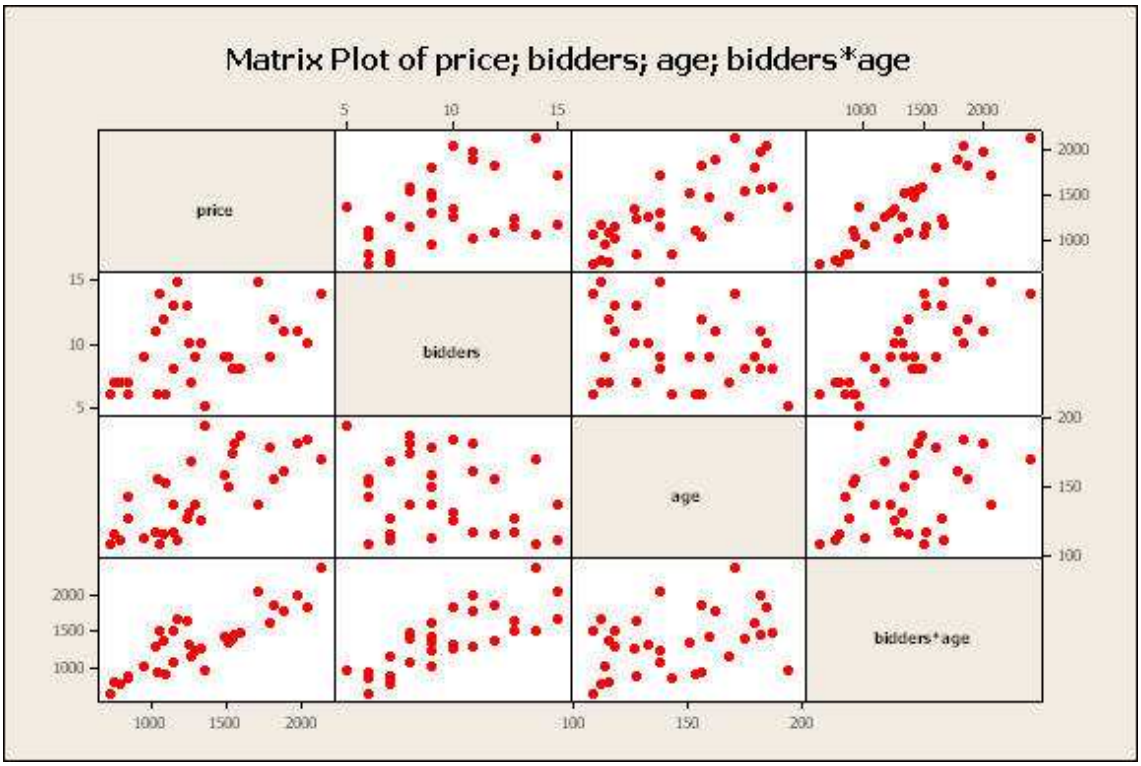
(b) ☐ Både typen av utbildning och vilket år man undersöker har betydelse, men det finns ingen signifikant skillnad mellan årskostnadseffekten för olika typer av utbildningar.

(c) ☐ De olika åren har olika effekt på kostnaden, men kostnaden skiljer sig inte signfikant åt för olika utbildningar.

(d) ☐ De olika utbildningarna har olika effekt på kostnaden, men kostnaden skiljer sig inte signfikant åt för olika år.

(e) ☐ Inget av ovanstående.

(f) ☐ vet inte.



2 En auktionsfirma vill beskriva hur åldern och antalet budgivare på en klocka påverkar priset på klockan. Man bestämmer sig för följande modell.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, där ϵ är $N(0, \sigma)$.

I denna modell är

$X_1 = Z_1$
$X_2 = Z_2$
$X_3 = Z_1 Z_2$

där Z_1 är antalet budgivare och Z_2 är klockans ålder. Ovan syns en scatterplot över resultaten. En Minitab-utskrift och frågeställningen följer på nästa sida. (Bidders står för budgivare, age står för ålder och price för pris.)

```
Regression Analysis: price versus age; bidders; bidders*age

The regression equation is
price = 320 + 0.88 age - 93.3 bidders + 1.30 bidders*age


Predictor      Coef  SE Coef      T      P   VIF
Constant      320.5    295.1   1.09  0.287
age           0.878    2.032   0.43  0.669  12.2
bidders      -93.26    29.89  -3.12  0.004  28.3
bidders*age  1.2978   0.2123   6.11  0.000  30.5


S = 88.9145   R-Sq = 95.4%   R-Sq(adj) = 94.9%


Analysis of Variance

Source          DF       SS       MS       F      P
Regression       3  4578427  1526142  193.04  0.000
Residual Error  28   221362     7906
Total           31  4799790
```

Betrakta följande påståenden:

1 I exemplet ovan verkar multikolinearitet vara ett problem.

2 Multikolinearitet innebär att det finns (starka) korrelationer mellan två eller flera av regressorerna.

3 Multikolinearitet innebär att responsvariabeln är korrelerad med regressorerna.

(a) ☐ Endast 1 är korrekt.

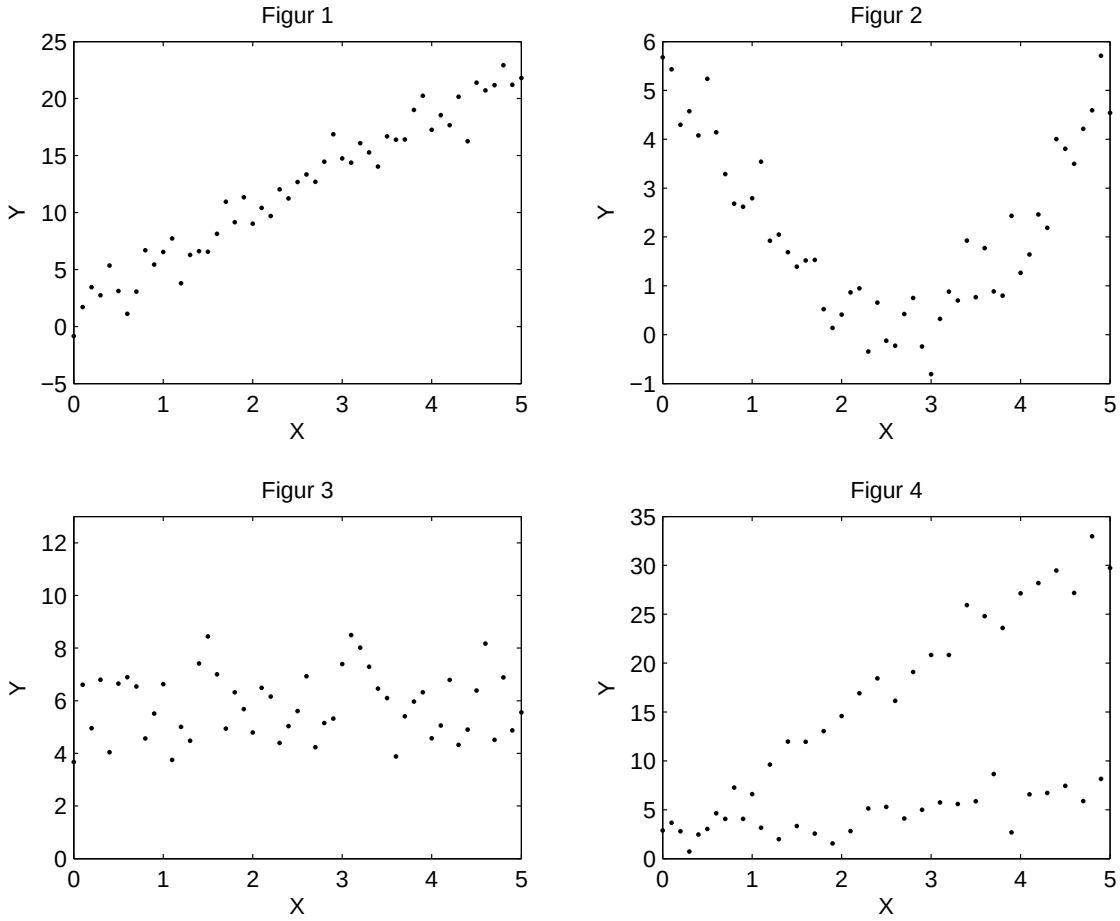
(b) ☐ Endast 2 är korrekt.

(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



3 Ovan visas fyra olika spridningsdiagram (scatter plots). I vilken eller vilka av dessa ger Pearsons korrelationskoefficient ett bra mått på associationen mellan variablerna X och Y?

(a) ☐ Endast Figur 1.

(b) ☐ Endast Figur 1 och Figur 3.

(c) ☐ Endast Figur 1 och Figur 4.

(d) ☐ Endast Figur 1, Figur 3 och Figur 4.

(e) ☐ Endast Figur 2.

(f) ☐ Vet ej.

4 Man vill avgöra om män i högre utsträckning än kvinnor röstar borgligt. Man tog ett stickprov av röstberättigade män och kvinnor och frågade dem: ”Röstar du borgligt?” Hur ska man på bästa sätt analysera resultatet?

(a) ☐ Med χ^2-test för oberoende i tvåsidig tabell för kategorisk data.

(b) ☐ Med ensidig variansanalys utan blockning.

(c) ☐ Med ensidig variansanalys med blockning.

(d) ☐ Med tvåsidig variansanalys.

(e) ☐ Med regressionsanalys.

(f) ☐ Vet ej

5 Du vill undersöka vem av fem tyngdlyftare som är starkast, genom att observera den tyngsta vikt de kan lyfta i sex olika "grenar". Du har fått reda på att skillnaden mellan olika lyftare kan anses vara multiplikativ, vilket t.ex. innebär att en lyftare i grunden kan lyfta 10% mer än en annan, oberoende av vilken gren det gäller. Vi låter Y_{ij} vara vikten lyftare i lyfter i gren j och vi vill analysera försöket med ensidig variansanalys med blockning.

Om ovanstående information tyder på att transformation av Y_{ij} bör göras innan vidare analys utförs, vilken transformation är det då?

(a) ☐ Analysera $\sqrt{x}$

(b) ☐ Analysera x^2

(c) ☐ Analysera e^x

(d) ☐ Analysera $\log(x)$

(e) ☐ Informationen ovan tyder inte på att någon transformation behövs.

(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	265.4	*
B	*	449.5	*	*
A×B	*	143.1	17.9	*
Error	15	136.0	9.1	
Total	29	*		

Värdena i några av fälten saknas och är markerade med (*). Från de givna siffrorna kan man ändå beräkna hur många nivåer som man använt sig av på varje faktor och hur många observationer som har gjorts per cell, d.v.s. per kombination av nivåerna i A och B. De är:

(a) ☐ Faktor A: 2 nivåer, faktor B: 4 nivåer, antal observationer per cell: 3.

(b) ☐ Faktor A: 3 nivåer, faktor B: 5 nivåer, antal observationer per cell: 15.

(c) ☐ Faktor A: 2 nivåer, faktor B: 4 nivåer, antal observationer per cell: 15.

(d) ☐ Faktor A: 3 nivåer, faktor B: 5 nivåer, antal observationer per cell: 3.

(e) ☐ Inget av ovanstående.

(f) ☐ vet inte.

7 En enkel linjär regressionsmodell har anpassats till data från 27 mätvärden. Man vill nu testa om β_1 är signifikant d.v.s. skild från noll. Man formulerar hypoteserna:

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

Man har under sina beräkningar fått ut att $\hat{\beta}_1 = 0.64$, $MSE = 0.31$ och $\sum_{i=1}^{27}(X_i - \bar{X})^2 = 4.21$. Vilket av följande interval hamnar p-värdet i?

(a) ☐ p-värdet > 0.20
(b) ☐ $0.10 <$ p-värdet < 0.20
(c) ☐ $0.05 <$ p-värdet < 0.10
(d) ☐ $0.01 <$ p-värdet < 0.05
(e) ☐ p-värdet < 0.01
(f) ☐ Vet ej.

8 För att ta reda på hur mycket avverkningsbar skog en bonde har på sin mark fälldes 30 fullvuxna granar. Diametern på varje träd mättes 1 meter ovanför marken i enheten centimeter och volymen timmer per träd mättes i enheten m^3. Låt X_i vara diametern på träd i och Y_i vara volymen timmer som man får ut från träd i. Det visar sig att man för de data man har samlat in kan anpassa en enkel linjär regressionsmodell. Skattningen av medelresponsen för denna modell blir:

$$\hat{Y} = -0.994. + 0.015 \cdot X$$

Läs nu följande påståenden:

1: Modellen är ej relevant för X nära 0.

2: Om diametern på trädet ökar 1 cm ökar volymen timmer med i genomsnitt 0.015 m^3.

3: Om diametern på trädet ökar 1 cm ökar volymen timmer med i genomsnitt 0.994 m^3.

Vilket eller vilka av dessa påståenden är korrekt/korrekta?

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

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

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

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

(e) ☐ Inget påstående är sant.

(f) ☐ Vet ej.

9 Vid en studie av bensinförbrukning är 4 olika bilar och 5 olika förare involverade. Alla förare kör en och samma runda. Varje förare kör alla bilarna på en ”egen” dag. Under den dagen kör föraren i fråga rundan en gång per bil, i slumpvis ordning och under liknande trafikförhållanden. Bensinförbrukningen under varje runda antecknas sedan. Man observerar för övrigt att trafikförhållandena under de *olika* dagarna är något olika.

Man vill nu ha ut mest möjliga information ur försöket. Om man antar att övriga modellantaganden stämmer, kan/bör man då baserat på informationen ovan:

(a) ☐ Analysera effekten av bil med ett χ^2-test.

(b) ☐ Analysera effekten av bil med en ensidig variansanalys med blockning.

(c) ☐ Analysera effekten av både bil och förare med en tvåsidig variansanalys.

(d) ☐ Analysera effekten av både bil och förare med ett χ^2-test.

(e) ☐ Inget av ovanstående är korrekt, eftersom förarna körde under olika dagar vilka hade olika trafik-förutsättningar.

(f) ☐ Vet ej.

10 Ett distributionsföretag vill beräkna kostnaderna för att frakta ett paket. I en multipel regressionsmodell vill man använda prediktorerna Z_1 = paketets vikt (i kg) och Z_2 = hur långt paketet fraktas (i km). Svarsvariabeln Y är kostnaden for frakten (i SEK).

Man beslutar sig för att använda en modell med följande regressorer:

$X_1 = Z_1$

$X_2 = Z_2$

$X_3 = Z_1 Z_2$

vilket ger modellen

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, där ϵ är $N(0, \sigma)$.

Anpassning av data till modellen ger:

$\hat{\beta}_0 = -1.52$

$\hat{\beta}_1 = 0.41$

$\hat{\beta}_2 = 0.052$

$\hat{\beta}_3 = 0.105$

Vad säger denna modell om den skattade förväntade förändringen av kostnaden Y, då distansen Z_2 ökar 1 km och vikten Z_1 hålls konstant på värdet 2 kg?

(a) ☐ Den skattade förväntade förändringen är +0.052 SEK

(b) ☐ Den skattade förväntade förändringen är +0.41 + 0.052 SEK

(c) ☐ Den skattade förväntade förändringen är +0.052 + 2*0.105 SEK

(d) ☐ Den skattade förväntade förändringen är +0.41 + 0.052 + 0.105 SEK

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej

11 I ett test vill man undersöka hur livslängden på ett batteri påverkas av två faktorer: materialtypen och temperatur. Temperaturfaktorn sätts till nivåerna -10^oC, 25^oC och 50^oC och man testar tre olika materialtyper.

Resultatet visas i tabellen nedan. Man vet att där är fyra observationer i varje cell, d.v.s. i varje kombination av temperaturer och material.

Variation	SS
Materialtyp	10684
Temperatur	39119
Samspel	9614
fel	18231
total	77647

F-statistikan för test av hypotesen att det inte finns något samspel mellan materialtyp och temperatur är

(a) ☐ (9614/9)/(77647/35)

(b) ☐ (9614/4)/(18231/27)

(c) ☐ ((10684+39119)/6)/((18231)/27)

(d) ☐ ((10684+39119)/6)/((77647)/35)

(e) ☐ Inget av ovanstående

(f) ☐ Vet ej

12 Bensinförbrukning för 4 typer av bilar undersöks. Ett slumpvis stickprov tas med 3 bilar av varje typ och 12 olika förare. Förarna tilldelas sedan en bil på måfå och ordningen i vilken de kör en och samma bestämda tur väljs på måfå. Varje bil blir alltså körd rundan en gång, varefter bensinförbrukningen antecknas. Hur ska man på bästa sätt analysera de insamlade mätvärdena?

(a) ☐ Med ^{2}-test för oberoende i tvåsidig tabell för kategoriska data.

(b) ☐ Med enkel linjär regression.

(c) ☐ Med ensidig variansanalys utan blockning.

(d) ☐ Med ensidag variansanalys med blockning.

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej

13 Efter injektion av ett antibiotikum i blodet binds en viss del av den injicerade mängden till serumproteiner. Detta fenomen har stor farmakologisk betydelse, eftersom det påverkar hur effektiv antibiotikan ifråga blir mot infektioner. I en studie ville man undersöka hur stor del av fem olika antibiotikatyper som bands. Varje medel injicerades på fyra olika individer. De tjugo frivilliga försökspersonerna tilldelades genom lottning en av de fem antibiotikatyperna.

Antibiotikum	Mängd bundet i serum (okänd enhet)			
Penicillin G	29.6	24.3	28.5	32.0
Tetracycline	27.3	32.6	30.8	34.8
Streptomycin	5.8	6.2	11.0	8.3
Erythromycin	21.6	17.4	18.3	19.0
Chlomphenicol	29.2	32.8	25.0	24.2

Detta analyserades först med en ensidig variansanalys, varvid det visades att skillnader mellan antibiotikumen finns. Man vill nu förutsättningslöst utföra test för att se vilka antibiotikum som skiljer sig åt sinsemellan.

Vilken av följande metoder är dels korrekt och dels mest effektiv om man vill utföra de testen med en total signifikansnivå på 5%?

(a) ☐ Parvisa t-test på 5%-nivån.

(b) ☐ Bonferronis metod på 5%-nivån.

(c) ☐ Scheffés metod på 5%-nivån.

(d) ☐ Tukeys metod på 5%-nivån.

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej

14 För att kontrollera att de modellantaganden som man gör i regressionsanalys är gilitiga konstrueras grafer där man plottar residualerna mot olika variabler som de antas vara oberoende av.

Vilken av följande variabler ska man INTE plotta residualerna mot?

(a) ☐ Responsen Y.

(b) ☐ Prediktorn X.

(c) ☐ Försöksordningen.

(d) ☐ De anpassade värdena $\hat{Y}$.

(e) ☐ Det går bra att plotta residualerna mot alla dessa alternativ.

(f) ☐ Vet ej.

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

För att besvara frågan om 17 till 19-åriga ungdomars användande av cigaretter påverkas av om föräldrarna röker gjordes insamling av data som presenteras nedan.

	B: någon av föräldrarna röker	
	j = 1: Ja	j = 2: Nej
A: Ungdomen röker		
i = 1: Ja	410	373
i = 2: Nej	120	295

X^2 beräknas till 60.45.

Vad kan man säga om nollhypotesen "17-19 åriga ungdomars användande av cigaretter är oberoende av om någon av föräldrarna röker"?

(a) ☐ Vi kan förkasta nollhypotesen om oberoende mellan A och B på signifikansnivå 10%, men inte på 5%.

(b) ☐ Vi kan förkasta nollhypotesen om oberoende mellan A och B på signifikansnivå 5%, men inte på 2,5%.

(c) ☐ Vi kan förkasta nollhypotesen om oberoende mellan A och B på signifikansnivå 2,5%, men inte på 1%.

(d) ☐ Vi kan förkasta nollhypotesen om oberoende mellan A och B på signifikansnivå 1%

(e) ☐ Inget av ovanstående.

(f) ☐ Vet ej