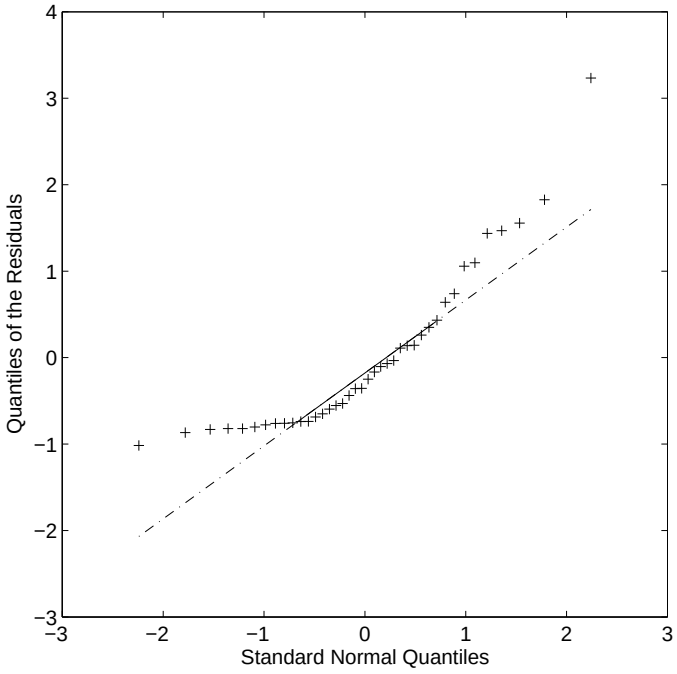
-2.5cm

TENTAMEN: Statistisk modellering för I3, TMS160, lördagen den 11 december 2004 kl 8:30 - 11:30 på M. **Jour:** John Gustavsson, mob 0705-330375
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/ 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 i studieportalen.

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

1. En odlare är intresserad av att hitta en enkel metod för att snabbt kunna uppskatta en pumpas vikt, och väljer att undersöka sambandet mellan vikt och diameter hos 30 slumpmässigt uttagna pumpor. Vid en regressionsanalys fås följande qq-plott av residualerna.



Vilken av följande slutsatser kan man dra från qq-plotten?

a ☐ Modellantagandet om oberoende feltermer är felaktigt, och därmed är det inte relevant att gå vidare och resonera om antagandet om normalitet.

b ☐ Antagandet om normalfördelade feltermer i den använda regressionsmodellen stämmer inte överens med data.

c ☐ Normalfördelningsantagandet verkar vara OK, men det verkar finnas ett par avvikande värden (*outliers*) i datamaterialet.

d ☐ Man kan se en variansheterogeneitet – variansen är högre för stora diametrar och lägre för mindre diametrar.

e ☐ Inget av ovanstående.

f ☐ Vet ej.

2. Man är intresserad av om det i en viss befolkningsgrupp finns någon skillnad mellan manliga och kvinnliga rökares debutålder. Man har tillgång till enkätsvar från 30 slumpvis valda rökare av vardera kön från befolkningsgruppen. I enkäten fick rökarna ange sin debutålder med något av alternativen $<$13, 13-15, 16-20 och $>$20 år. Om man ska utnyttja resultatet av enkätundersökningen till att undersöka frågeställningen är det bäst att använda:

   a ☐ Regressionsanalys.

   b ☐ Ensidig variansanalys utan blockindelning (*one-way model*).

   c ☐ Ensidig variansanalys med blockindelning (*RCBD*).

   d ☐ Flersidig variansanalys med den generella modellen (alltså med en samspelsterm).

   e ☐ Inget av ovanstående.

   f ☐ Vet ej.

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

Analysis of variance				
Source	DF	Sum of squares	Mean square	F Stat
A	*	20.48	*	?
B	1	1.70	1.70	*
A×B	*	157.25	78.62	*
Error	18	112.05	6.23	
Total	23	291.48		

Värdena i några av fälten saknas och är markerade med (*). Från de givna siffrorna kan man ändå beräkna värdet av F-statistikan (markerad med ett frågetecken i tabellen). Den blir:

a ☐ 10.24/6.23

b ☐ 10.24/112.05

c ☐ 20.48/6.23

c ☐ 20.48/112.05

e ☐ 20.48/291.48

f ☐ Vet ej.

4. För att undersöka överlevnadstiden hos möss vid exponering av tre gifter och med fyra typer av behandling genomfördes ett fullständigt faktoriellt försök med tre observationer per cell. Vid en tvåsidig variansanalys av de observerade överlevnadstiderna fick man följande ANOVA-tabell.

Analysis of variance				
Source	DF	Sum of squares	Mean square	F Stat
Poison	2	73.26	36.63	25.82
Treatment	3	66.65	22.21	15.66
Poison×Treatment	6	24.04	4.00	2.82
Error	24	34.04	1.41	
C Total	35	197.99		

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

a ☐ Både typen av gift och valet av behandling har effekt på överlevnadstiden, och någon eller några av behandlingarna har olika effekt på olika typer av förgiftning.

b ☐ Både typen av förgiftning och valet av behandling har effekt på överlevnadstiden, men data visar inte på någon signifikant skillnad mellan behandlingseffekterna för olika typer av förgiftning.

c ☐ De olika gifterna har olika effekt på överlevnadstiden, men behandlingseffekten skiljer sig inte signifikant åt mellan behandlingarna.

d ☐ Vare sig typ av förgiftning eller behandling har någon signifikant effekt på överlevnadstiden.

e ☐ Inget av ovanstående.

f ☐ Vet ej.

5. Vilket av följande uttalanden a - e är falskt?

a ☐ Bonferroni-konfidensintervall med den simultana konfidensnivån L för k stycken parametrar består av k stycken "vanliga" konfidensintervall på nivå $1-(1-L)/k$.

b ☐ Idén bakom Tukey-konfidensintervall för differenser mellan alla par av medelvärden är att först hantera skillnaden mellan det största och det minsta av de observerade medelvärdena. Man använder sedan att alla andra skillnader i observerade medelvärden är mindre.

c ☐ De individuella konfidensintervallen är alltid smalare vid Bonferroni-jämförelse än vid Tukey-jämförelse.

d ☐ Idén bakom Bonferroni-jämförelsen är att bygga konfidensintervallen på "värsta-fall analys".

e ☐ Om man vill ha simultana konfidensintervall för många kontraster är det vanligen bäst att använda Scheffe's metod.

f ☐ Vet ej.

6. Ett företag som målar aluminiumytor vill undersöka hur vidhäftningen beror av vilken typ av primer som används, och av hur man behandlar med primern.

   Man gör därför ett försök med tre olika slags primer och med två olika behandlingsmetoder, sprejning och neddoppning. För varje kombination av primer och behandlingsmetod målade man tre stycken ytor. Tabellen nedan visar de uppmätta vidhäftningarna.

Vidhäftning						
	Behandlingsmetod					
Primer	Sprejning			Doppning		
1	4.0	4.5	4.3	5.4	4.9	5.6
2	5.6	4.9	5.4	5.8	6.1	6.3
3	3.8	3.7	4.0	5.5	5.0	5.0

   Nästa tabell visar lämpliga medelvärden av observationerna.

Genomsnittlig vidhäftning			
	Behandlingsmetod		
Primer	Sprejning	Doppning	medelvärde
1	4.27	5.30	4.78
2	5.30	6.07	5.68
3	3.83	5.17	4.50
medelvärde	4.47	5.51	4.99

   Vid en tvåsidig variansanalys fås följande resultat.

Variation	Kvadrat-summa
Mellan primertyper	4.58
Mellan behandlingsmetoder	4.91
Växelverkan	0.24
Inom celler	0.99
Totalt	10.72

Växelverkanseffekten mellan primertyp 3 och behandling med doppning skattas med:

a ☐ $4.50 - 5.51 = -1.01$

b ☐ $4.50 - 4.99 = -0.49$

c ☐ $5.51 - 4.99 = 0.52$

d ☐ $5.17 - 4.99 - (4.50 - 4.99) - (5.51 - 4.99) = 0.15$

e ☐ $\frac{1}{2}\left[4.27 + 5.17 - 3.83 - 5.30\right] = 0.16$

f ☐ Vet ej.

7. Betrakta igen försöket från föregående uppgift, uppgift 6. Vad betyder en eventuell växelverkan?

   a ☐ Det är bara typ av primer som har betydelse för vidhäftningen.

   b ☐ Det är bara behandlingsmetoden som har betydelse för vidhäftningen.

   c ☐ Skillnaden mellan vidhäftningen för de två olika behandlingsmetoderna beror av vilken primer man använder.

   d ☐ Det är samma skillnad i vidhäftning mellan de två olika behandlingsmetoderna för alla primertyper.

   e ☐ Vare sig primertyp eller behandlingsmetod har betydelse för vidhäftningen.

   f ☐ Vet ej

8. Vilket av följande uttalanden a - e är falskt?

a ☐ Faktoreffektsmodellen kan formuleras som en multipel linjär regressionsmodell, fast med diskreta regressorer.

b ☐ χ^2 test i 2×2 tabeller kan inte hantera kontinuerliga variabler.

c ☐ Den statistiska analysen av både den additiva faktoreffektmodellen och den linjära regressionsmodellen bygger, i den här kursen, på antagandet att avvikelsetermen ε är normalfördelad med konstant varians.

d ☐ Om man vid en tvåsidig variansanalys finner samspelstermer som inte kan transformeras bort kan det vara lämpligt att använda konfidensintervall för väl utvalda kontraster som sammanfattning av resultaten av analysen.

e ☐ Man behöver aldrig plotta residualer i en variansanalys mot tidsordningen om man har randomiserat.

f ☐ Vet ej

9. Det är känt att 10% av alla producerade 512 MB ram-minnen är defekta. Med inspektion och kassation av upptäckta defekta minnen, kan man försäkra sig om att bara 1% av alla ram-minnen som säljs lagligt är defekta. Tyvärr stjäls en del ram-minnen innan denna inspektion. Man tror sig veta att ungefär 1% av hela världsmarknaden för 512 MB ram-minnen består av stulna minnen. Vad är då sannolikheten att ett givet ram-minne är stulet givet att det är defekt?

   a ☐ 0.092

   b ☐ 0.132

   c ☐ 0.066

   d ☐ 0.056

   e ☐ Inget av ovanstående.

   f ☐ Vet ej

10. Vid första föreläsningen i höstas gjorde vi ett försök där alla fick välja mellan att antingen dricka Cola eller juice samt antingen äta frukt eller bulle. Samma försök gjordes även hösten 2003. Om man säger att alla som valde både frukt och juice gjorde ett "nyttigt" val medan resten gjorde ett "onyttigt" val man ställa upp följande tabell över resultaten:

	2003	2004	Total
Nyttigt	19	29	48
Onyttigt	24	38	62
Total	42	67	n = 110

Frågan är nu om det är någon skillnad i val av nyttigt eller onyttigt alternativ i år jämfört med förra året. Den kan besvaras genom att testa

H_0 : Val av nyttigt eller onyttigt alternativ är oberoende av år

mot

H_a : Val av nyttigt eller onyttigt alternativ är beroende av år.

Vad blir då p-värdet?

a ☐ p-värde<0.01.

b ☐ $0.01 \leq$p-värde< 0.025.

c ☐ $0.025 \leq$p-värde< 0.05.

d ☐ $0.05 \leq$p-värde<0.1.

e ☐ p-värde≥ 0.1.

f ☐ Vet ej.

11. Betrakta följande två situationer och avgör vad Pearsons korrelationskoefficient är för variablerna sträcka och tid.

    1) En cyklist håller alltid konstant fart 20 km/h. Cyklisten cyklar ett antal olika långa uppmätta sträckor och tar tiden. Låt r_1 vara korrelationskoefficienten.

    2) En annan cyklist som inte håller konstant hastighet, cyklar ett antal olika långa uppmätta sträckor och tar tiden. Låt r_2 vara korrelationskoefficienten.

    a ☐ $r_1 = 1$ och $r_2 = 1$.

    b ☐ $-1 < r_1 < 1$ och $r_2 = 1$.

    c ☐ $r_1 = 1$ och $-1 < r_2 < 1$.

    d ☐ $-1 < r_1 < 1$ och $-1 < r_2 < 1$.

    e ☐ Inget av ovanstående.

    f ☐ Vet ej.

12. En apotekare har hittat på en ny metod att bestämma vikterna x_1 och x_2 av två föremål med hjälp av en balansvåg. I stället för att lägga upp föremålen i en viktskål en åt gången och balansera dem med vikter i den andra skålen gör han som följer: Först lägger han båda föremålen i samma viktskål och balanserar dem med vikter i den andra skålen och mäter på så sätt $x_1 + x_2$. Sedan lägger han ett föremål i var skål och balanserar med vikter, och mäter på så sätt $x_1 - x_2$. Från dessa två mätningar kan man så räkna ut x_1 och x_2.

    Man vet att variansen av felet vid en vägning är 1 g^2. Resultaten av olika vägningar är oberoende av varandra. Vad blir då standardavvikelsen för felet när man bestämmer x_1 på ovanstående sätt?

    a ☐ 1 g

    b ☐ 2 g

    c ☐ $1/\sqrt{2}$ g

    d ☐ $1/2$ g

    e ☐ $1/4$ g^2.

    f ☐ Vet ej

13. Vad innebär multikollinearitet i den multipla linjära regressionsmodelen:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \ ?$$

a ☐ Feltermens varians beror på Y.

b ☐ Feltermens varians beror på X_1, X_2 och X_3.

c ☐ Det finns ett ickelinjärt samband mellan Y och X_1, X_2 och X_3.

d ☐ Responsvariablen Y är korrelerad med X_1, X_2 och X_3.

e ☐ Det finns (starka) korrelationer mellan två eller flera av X_1, X_2 och X_3.

f ☐ Vet ej.

14. Susanne cyklar till arbetet och försöker varje morgon förutspå hur lång tid det kommer ta så att hon kan stanna hemma och läsa tidningen så länge som möjligt och ändå normalt komma fram i tid. Hon har därför under 50 dagar observerat tiden, vindstyrkan och vindriktningen och har ansatt följande modell

$$\ln(\text{tid}) = \beta_0 + \beta_1 \text{vindstyrka} + \beta_2 \text{vindriktning} + \epsilon.$$

Tiden har mätts i minuter, vindstyrkan i m/s och vindriktningen mellan 0^o och 180^o, där 0^o är rak medvind och 180^o är rak motvind. Minsta kvadratskattningarna av parameterarna är $\hat{\beta}_0 = 2.86$, $\hat{\beta}_1 = 0.0357$ och $\hat{\beta}_2 = 0.000509$. Skattningen av felvariansen är MSE$= 0.0132$. Residualplottarna visar inga tydliga trender och inga stora avvikelser från normalitet.

Vad är ett 95% prediktionsintervall för tiden i minuter det tar att cykla till jobbet en morgon när det blåser 8 m/s rak sidvind (90^o)? Man har räknat ut att standardfelet för den predikterade responsen vid denna vindriktning och vindstyrka är $\hat{\sigma}(\ln(\hat{\text{tid}})) = 0.046$.

a ☐ $(22.2, 26.7)$

b ☐ $(19.8, 29.9)$

c ☐ $(19.3, 30.6)$

d ☐ $(19.0, 31.2)$

e ☐ $(14.9, 39.7)$

f ☐ Vet ej.

15. I ett försök mätte man värmeutvecklingen vid härdning av cement vid användning av olika recept. Man gjorde en multipel regressionsanalys för att beskriva värmeutvecklingen som en linjär funktion av mängden tricalsiumaluminat (x1), tricalsiumsilikat (x2) och dicalciumsilikat (x3), och fick bl a följande resultat:

Summary of Fit

Mean of Response	92.3462	R-Square	0.7092
Root MSE	12.4776	Adj R-Sq	0.6123

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	3	3417.3225	1139.1075	7.32	0.0087
Error	9	1401.2098	155.6900	.	.
C Total	12	4818.5323	.	.	.

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	47.8888	76.4324	0.63	0.5465	.	0.0000
x1	1	0.9511	0.6323	1.50	0.1668	0.9378	1.0663
x2	1	0.8383	1.0031	0.84	0.4250	0.0532	18.7803
x3	1	-0.1002	0.9365	-0.11	0.9172	0.0528	18.9401

Estimated Corr Matrix

	Intercept	x1	x2	x3
Intercept	1.0000	-0.1283	-0.9919	-0.9879
x1	-0.1283	1.0000	0.0457	0.1025
x2	-0.9919	0.0457	1.0000	0.9715
x3	-0.9879	0.1025	0.9715	1.0000

Vilket av följande påståenden a - e är riktigt?

a ☐ Det är inga tecken på multikollinearitet.

b ☐ Analysen ger en klar indikation på multikollinearitet mellan x1 och x2.

c ☐ Analysen ger en klar indikation på multikollinearitet mellan x1 och x3.

d ☐ Analysen ger en klar indikation på multikollinearitet mellanx2 och x3.

e ☐ Inget av ovanstående.

f ☐ Vet inte.