IDA arkiv

Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2011-01-18
Sal (1) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	TER1
Tid	8-13
Kurskod	732G21
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Sambandsmodeller Tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	4
Jour/Kursansvarig Ange vem som besöker salen	Oleg Sysoev
Telefon under skrivtiden	0735673482
Besöker salen ca kl.	10 båda lärare
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Carita Lilja, 1463, carita.lilja@liu.se
Tillåtna hjälpmedel	Valfri räknedosa, kursbok (Kutner m fl) utan anteckningar. Markering av texten i boken med färg är tillåtet. Flärpar tillåtna
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	Rutigt
Antal exemplar i påsen	27

Tentamen

Linköpings Universitet, Institutionen för datavetenskap, Statistik

---

Kurskod och namn:	732G21 Sambandsmodeller
Datum och tid:	2011-01-18, 8-13
Jourhavande lärare:	Oleg Sysoev
Tillåtna hjälpmedel:	Valfri räknedosa, kursbok (Kutner m fl) utan anteckningar. Markering av texten i boken med färg är tillåtet.
Betygsgränser:	Tentamen omfattar totalt 30p. Godkänt från och med 18p, väl godkänt från och med 24p.

Redovisa och motivera tydligt alla dina lösningar!

Uppgift 1 (7p)

Datamängden "Bilar78" innehåller information om egenskaper av bilar som tillverkats på olika ställen i världen under 1978-1979. Datamängden innehåller 34 observationer följande variabler:

- Country: Ursprunglandet (US, Japan, eller Tyskland). Ges som 2 binära variabler:
  - Cn1=1 om landet är US, 0 annars
  - Cn2=1 om landet är Japan, 0 annars
- MPG: Antalet mil per gallon
- Drive_Ratio: Antalet gånger motorn roterar för att vrida ett hjul 360 på grader
- Horsepower: Hästkraft
- Displacement: Större värden motsvarar kraftigare motorer

En forskare vill bygga upp en linjär regressionsmodell där Drive_Ratio är responsvariabel och de andra variabler är förklarande. För att välja den bästa modellen, genomför man modellurvalsprocessen:

Best Subsets Regression: Drive_Ratio versus MPG; Weight; ...

Response is Drive_Ratio

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	MPG	Weight	Horsepower	Displacement
1	67,1	66,0	6,8	0,30407				X
1	58,9	57,6	15,8	0,33958		X		
2	73,0	71,3	2,1	0,27947	X			X
2	70,9	69,0	4,5	0,29052	X	X		
3	74,0	71,4	3,0	0,27895	X	X		X
3	73,0	70,3	4,1	0,28408	X		X	X
4	74,0	70,4	5,0	0,28372	X	X	X	X

1. Vilken modell skulle man välja enligt C_p kriteriet? Motivera. Forskaren bestämde dock att använda modellen med följande förklaringsvariabler : MPG, Weight och Displacement. Motivera detta val. (1p)

Dessutom har forskaren lagt till variabler Cn1 och Cn2 och sedan anpassat följande regressionsmodell:

Regression Analysis: Drive_Ratio versus MPG; Weight; ...

The regression equation is
Drive_Ratio = 4,95 - 0,0246 MPG + 0,004 Weight - 0,00502 Displacement
- 0,499 Cn1 - 0,288 Cn2

Predictor	Coef	SE Coef	T	P
Constant	4,947	1,301	3,80	0,001
MPG	-0,02455	0,02272	-1,08	0,289
Weight	0,0037	0,3993	0,01	0,993
Displacement	-0,005020	0,002180	-2,30	0,029
Cn1	-0,4992	0,1592	-3,14	0,004
Cn2	-0,2877	0,1510	-1,91	0,067

S = 0,248378 R-Sq = 80,8% R-Sq(adj) = 77,3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	7,2545	1,4509	23,52	0,000
Residual Error	28	1,7274	0,0617		
Total	33	8,9819			

Source	DF	Seq SS
MPG	1	2,9958
Weight	1	3,3697
Displacement	1	0,2820
Cn1	1	0,3831
Cn2	1	0,2240

2. Använd ovanstående utskrifter och testa om det finns regressionsrelation genom att utnyttja signifikansnivå α=0.01. Ange hypotes, mothypotes, beslutsregel och slutsats. (1p)
3. Redovisa anpassade regressionsekvationer för USA, Japan och Tyskland och tolka regressionskoefficienterna som motsvarar Cn1 och Cn2. (2p)
4. Testa om variablerna Cn1 och Cn2 kan tas bort från modellen genom att använda signifikansnivå α=0.1. Ange hypotes, mothypotes, beslutsregel och slutsats. Vilket är P-värde av testet? (3p)

Uppgift 2 (8p)

Målet av denna uppgift är att utforska ett samband mellan variablerna HP (hästkraft) and WT (bilens vikt, givet i 100 skålpund). Datamängder har 82 observationer.

En enkel linjär regressionsmodell var anpassad till datamängden med utfall HP och förklaringsvariabel WT

Regression Analysis: HP versus WT

```
The regression equation is
HP = - 62,5 + 5,81 WT
```

Predictor	Coef	SE Coef	T	P
Constant	-62,49	13,83	-4,52	0,000
WT	5,8103	0,4328	13,43	0,000

```
S = 31,7115   R-Sq = 69,3%   R-Sq(adj) = 68,9%
```

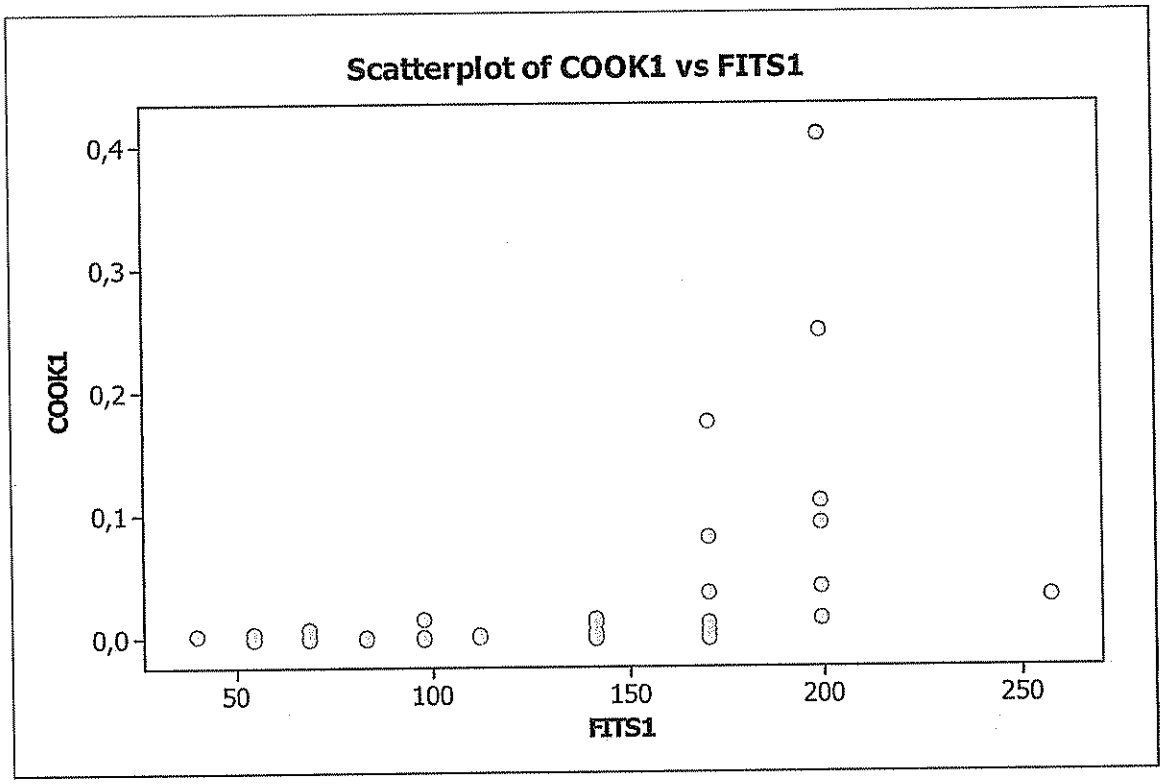
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	181252	181252	180,24	0,000
Residual Error	80	80450	1006		
Total	81	261702			

Descriptive Statistics: WT

Variable	N	N*	Mean	Variance	Minimum	Q1	Median	Q3	Maximum
WT	82	0	30,915	66,283	17,500	25,000	30,000	36,250	55,000

1. Skatta simultana prediktionsintervall för nya observationer WT=30 och WT=40 genom att använda Scheffé och Bonferroni procedurer med 95% konfidensnivå. Vilken av procedurer fungerar bäst för denna mängd? Motivera. (TIPS: använd "descriptive statistics") (3p)
2. Bestäm hur många observationer påverkar alla anpassade värdena i den enkla linjära modellen genom att utnyttja det nedanstående diagrammet: (2p)



3. En andragradsregressionsmodell med utfall HP och centrerade förklaringsvariablerna X1=WTc och X2=WTc^2 (där WTc är den centrerade WT) anpassades till datamängden. Använd nedanstående utskrifter för att:
   a. Uttrycka den anpassade modellen i gamla beteckningar (HP, WT) (2p)
   b. Skatta partiell förklaringsgrad (coefficient of partial determination) $R^2_{Y2|1}$ (Tips: kom ihåg SSTO = SSE +SSR) (1p)

Regression Analysis: HP versus X1; X2

The regression equation is
HP = 112 + 5,44 X1 + 0,0835 X2

Predictor	Coef	SE Coef	T	P
Constant	111,670	4,652	24,00	0,000
X1	5,4442	0,4755	11,45	0,000
X2	0,08346	0,04755	1,76	0,083

S = 31,3070 R-Sq = 70,4% R-Sq(adj) = 69,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	184271	92136	94,00	0,000
Residual Error	79	77430	980		
Total	81	261702			

Source	DF	Seq SS
X1	1	181252
X2	1	3019

Descriptive Statistics: HP; WT

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
HP	82	0	117,13	6,28	56,84	49,00	83,25	99,00	140,00	322,00
WT	82	0	30,915	0,899	8,141	17,500	25,000	30,000	36,250	55,000

Uppgift 3 (11p)

Tre olika rengöringsmedel ska jämföras för att studera deras inbromsnings-effekt av bakterietillväxt i 5-liters mjölkförpackningar. Analysen sker i ett laboratorium, och bara tre försök kan köras per dag. Eftersom det kan vara skillnad mellan dagar så beslutar försöksledaren att använda ett randomiserat block försök. Observationer tas i fyra dagar, och resultatet visas nedan.

	Dagar			
Rengörings-medel	1	2	3	4
1	13	22	18	39
2	16	24	17	44
3	5	4	1	22

a) Plotta data på ett lämpligt sätt. Ange modellen. Beräkna de fyra kvadratsummorna. (sums of squares). (4p)

b) Pröva om det finns några behandlingseffekter (treatment effects). Ställ upp hypoteserna. Använd 5% signifikansnivå. (1p)

c) Utför ett test för att se om det är nödvändigt att kontrollera för Dagar. Ställ upp hypoteserna. Använd 5% signifikansnivå. (1p)

d) Skatta med ett 95% konfidensintervall skillnaden mellan medelvärdet för väntevärdena för rengöringsmedel 1 och 2 och väntevärdet för rengöringsmedel 3. (3p)

e) Anta att observationen för rengöringsmedel 1 och Dag 2 saknas. Skatta det saknade värdet (missing value). (1p)

f) Variansanalystabellen då endast rengöringsmedel är med i modellen visas nedan. Ren=rengöringsmedel

Analysis of Variance for Y, using Adjusted SS for Tests

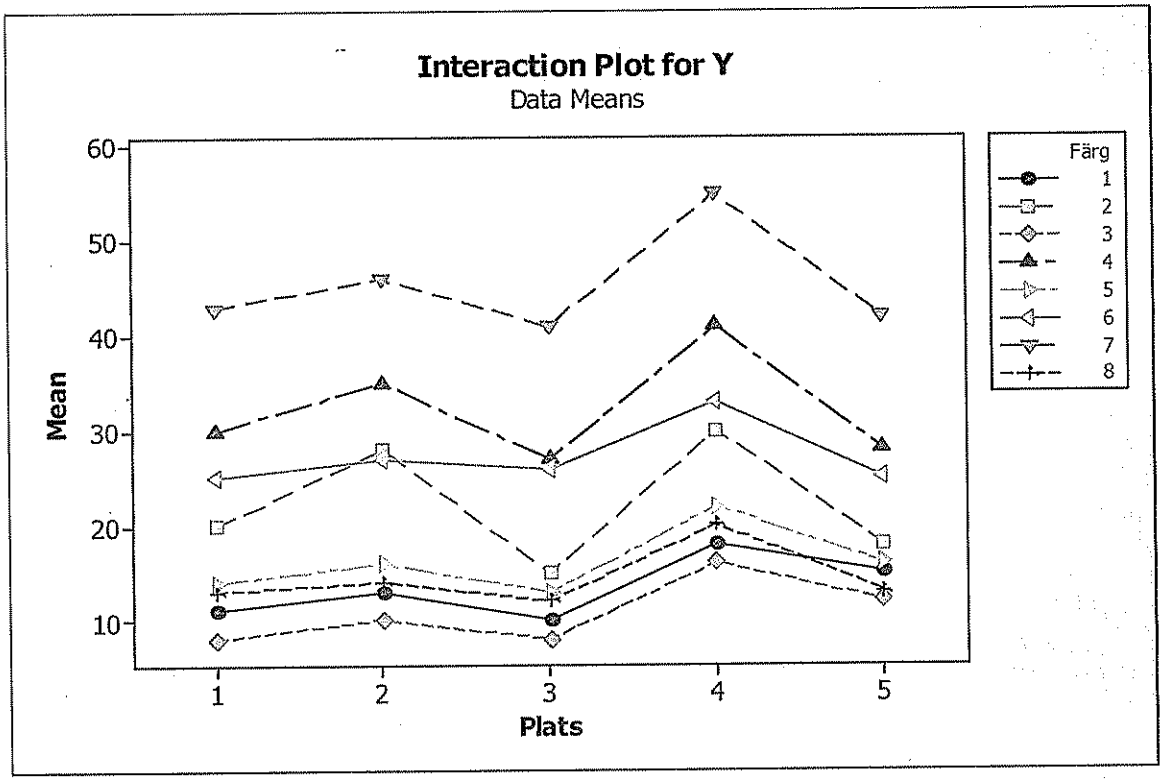
Source	DF	Seq SS	Adj SS	Adj MS
Ren	2	703,5	703,5	351,8
Error	9	1158,8	1158,8	128,8
Total	11	1862,3		

Testa om det finns några behandlingseffekter. Överensstämmer resultatet med det du fick i testen ovan? (1p)

Uppgift 4 (4p)

Kvaliteten hos fem olika vägfärger (den som används för vägmarkeringar) ska utvärderas. Färg nummer ett är standarfärgen och är den som används idag. Åtta platser valdes slumpmässigt ut och på vald plats målades alla fem färgerna i någon slumpmässig ordning. Då färgerna utsatts för väder och trafik efter en tid så mättes synbarhet och varaktighet hos färgen. Y är ett sammansatt mått på

detta. Ju högre värden desto bättre. Nedan visas en linjeplot över data. Upplägget är ett randomiserat blockförsök med slumpmässiga blockeffekter.



General Linear Model: Y versus Färg; Plats

```
Factor  Type    Levels  Values
Färg    fixed        5  1; 2; 3; 4; 5
Plats   random       8  1; 2; 3; 4; 5; 6; 7; 8
```

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS       F      P
Färg     4   531,35   531,35  132,84   30,42  0,000
Plats    7  4826,38  4826,38  689,48  157,92  0,000
Error   28   122,25   122,25    4,37
Total   39  5479,97
```

```
S = 2,08951   R-Sq = 97,77%   R-Sq(adj) = 96,89%
```

```
Bonferroni 95,0% Simultaneous Confidence Intervals
Response Variable Y
Comparisons with Control Level
Färg = 1  subtracted from:

Färg   Lower  Center   Upper  ---------+---------+---------+-------
2      0,336   3,125   5,914            (----*-----)
3     -4,289  -1,500   1,289  (-----*-----)
4      6,086   8,875  11,664                       (-----*----)
5     -2,164   0,625   3,414       (----*-----)
                               ---------+---------+---------+-------
                                      0,0       5,0      10,0
```

Descriptive Statistics: Y

Variable	Färg	Total Count	Mean
Y	1	8	20,50
	2	8	23,63
	3	8	19,00
	4	8	29,37
	5	8	21,13

a) Förklara varför detta är ett randomiserat blockförsök med slumpmässiga effekter och vilka krav vi ställer på modellen. Visa hur modellen ser ut och beskriv behandling och block. (1p)

b) Skatta ett 95% konfidensintervall för σ_ρ^2 med hjälp av Satterthwaites procedur, där σ_ρ^2 är varianskomponenten för blockeffekten. (2p)

c) Beskriv hur konfidensintervallen ovan är beräknade. Det är förstås inte möjligt för dig att slå upp korrekt t-tabellvärde. (1p)