TENTAMEN I SAMBANDSMODELLER, 2018-01-19

Skrivtid: kl: 8-13
Hjälpmedel: Räknedosa. Läroboken: *Applied linear statistical models* av Kutner, Nachtsheim m fl som inte får innehålla anteckningar men får ha markeringar och flärpar. Flärpar får ha en liten anteckning.
Jourhavande lärare: Lotta Hallberg
Betygsgränser: För godkänt krävs minst 12 av 20 poäng och för väl godkänt krävs minst 16 av 20 poäng.

Redovisa och motivera kort alla dina lösningar

Tolka (om möjligt) alla dina resultat!

1

En myndighet ville undersöka åttonde-klassares utbildnings-resultat i matematik med deras hemförhållanden. Resultat från 40 delstater i USA observerades. Enheten för Y är poäng.
Följande förklarande variabler användes:
x_1= andelen elever som hade båda föräldrarna boende hemma.
x_2= andelen elever som hade mer än tre typer av läsmedel hemma, såsom böcker, tidningar encyklopedi, osv.
x_3= andelen elever som läste mer än 10 sidor per dag.
x_4= andelen elever som spelade datorspel el dyl, mer än 6 timmar per dag.
x_5= andelen elever som var borta från skolan mer än tre dagar förra månaden.
Ett litet utdrag ur datamaterialet:

Alabama	252	75	78	34	18	18
Arizona	259	75	73	41	12	26
Arkansas	256	77	77	28	20	23

Best Subsets Regression: Y versus x1; x2; x3; x4; x5

Response is Y

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	x1	x2	x3	x4	x5
1	76,3	75,7	1883,6	72,3	22,0	6,5079					X
1	55,5	54,3	3367,2	50,4	72,8	8,9157		X			
1	55,0	53,8	4778,9	29,6	74,2	8,9700	X				
2	84,2	83,4	1392,6	79,5	4,6	5,3810		X			X
2	79,2	78,1	2344,4	65,5	16,8	6,1743	X	X			
2	77,6	76,4	1975,0	70,9	20,7	6,4047		X	X		
3	85,1	83,9	1412,8	79,2	4,4	5,2939		X	X	X	
3	85,1	83,8	1669,2	75,4	4,5	5,3062	X	X		X	
3	84,3	82,9	1545,7	77,2	6,5	5,4496		X		X	X
4	85,9	84,3	1629,3	76,0	4,5	5,2327	X	X	X	X	
4	85,4	83,7	1863,6	72,6	5,8	5,3285	X	X		X	X
4	85,2	83,5	1571,0	76,9	6,3	5,3661		X	X	X	X
5	86,1	84,1	1832,5	73,0	6,0	5,2680	X	X	X	X	X

a) Vilka antaganden måste gälla på Y för att man ska få anpassa en multipel linjär regressionsmodell? 1p

b) Vilken modell är bäst enligt C_p kriteriet? 1p

c) Förklara vad PRESS är. Hur ska det tolkas? Vad ska PRESS jämföras med? 1p

Följande modell har anpassats:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	186,5	29,3	6,36	0,000	
x1	0,346	0,241	1,43	0,161	3,40
x2	0,787	0,172	4,58	0,000	1,62
x4	-1,120	0,298	-3,75	0,001	4,43

d) Om andelen elever som spelar datorspel el dyl, mer än 6 timmar per dag ökar med 10% (de andra variablerna hålls fixa), vad händer då med utbildningsresultatet i matematik prediktivt? Svara numeriskt (med en siffra). 1p

e) Beräkna ett prediktionsintervall för Y då $x_1 = 80, x_2 = 80, x_4 = 20$. Nedan ges $(X'X)^{-1}$ där X är modellens designmatris. $\sqrt{MSE}$ för modellen hittar du i utskriften för Best subsets ovan 3p

30,5865	-0,215718	-0,122753	-0,280774
-0,2157	0,002068	0,000325	0,002064
-0,1228	0,000325	0,001048	0,000947
-0,2808	0,002064	0,000947	0,003162

2

a) Vilken typ av regression är lämplig att använda om responsvariabeln har värdena 'Ja' eller 'Nej'? Hur hanteras denna nominella variabel? 1p

b) Vilken typ av regression är lämplig att använda om responsvariabeln är normalfördelad? 1p

c) Ge exempel på två olika länkfunktioner inom generaliserade linjära modeller. 1p

3

Låt X vara antalet olyckstillfällen under en månad på en viss väglänk. X kan antas vara Poissonfördelad med parameter λ. Man observerade antalet olyckstillfällen under 10 månader. De observerade värdena blev 0, 0, 2, 1, 1, 0, 3, 0, 1, 2 olyckstillfällen.
Det gäller att $E[X] = Var]X\} = \lambda$. Härled maximumlikelihood-skattningen av variansen för X.
Sannolikhetsfunktionen för en Poissonfördelad slumpvariabel är $f(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0,1, \ldots$ 3p

4

Frukost-m$\ddot{u}$sli måste torkas innan den förpackas. En laboratorie-tekniker utförde ett experiment där han mätte fuktigheten i m$\ddot{u}$slin då den utsatts för olika temperaturer och torkningstid i en ugn.
Responsvariabelnn är alltså fuktigheten i % i m$\ddot{u}$slin.
Faktor ***Tid*** har tre nivåer vilka är 30, 60 och 90 minuter i ugnen.
Faktor ***Temperatur*** har tre nivåer vilka är 125, 130 och 135 grader i ugnen.
Varje cell har 4 observationer.
Resultatet presenteras i följande tabell:

	Temperatur °C		
	125	130	135
30min	$\bar{Y}_{11\cdot} = 0{,}138$	$\bar{Y}_{12\cdot} = 0{,}11$	$\bar{Y}_{13\cdot} = 0{,}09575$
Tid 60min	$\bar{Y}_{21\cdot} = 0{,}1125$	$\bar{Y}_{22\cdot} = 0{,}1005$	$\bar{Y}_{23\cdot} = 0{,}098$
90min	$\bar{Y}_{31\cdot} = 0{,}09775$	$\bar{Y}_{32\cdot} = 0{,}045$	$\bar{Y}_{33\cdot} = 0{,}037$

Använd en modell för balanserad två-vägs ANOVA med interaktion och utför följande analyser.

Du får använda direkt att SSTO= 0,0396912 och SSE= 0,0067115

a) Testa om det finns interaktion mellan faktornivåerna på 5% signifikansnivå. 2p
b) Både Temperatur och Tid har effekter som är signifikant skilda från 0. Skatta dessa sex effekter. 1p
c) Beräkna konfidensintervall för alla differenser mellan tiderna i ugn med Tukeys metod. 95% familjekonfidensgrad. 2p
d) Beräkna ett 95% konfidensintervall för $\mu_{\cdot 1} - \frac{\mu_{\cdot 2} + \mu_{\cdot 3}}{2}$. Tolka 2p