Tentamen
Tillämpad statistik A5 (15hp)
2016-02-13

Statistiska institutionen, Uppsala universitet

Upplysningar

1. Tillåtna hjälpmedel: Miniräknare, A4/A8 Tabell- och formelsamling (alternativ Statistik för samhällsplanerare Tabell- och formelsamling) samt nuvarande formelsamling för A5. Formelsamlingar för A4/A8 samt A5 som användes HT2014-VT2015 är också tillåtna. **Inga anteckningar är tillåtna i formelsamlingarna.**
2. Skrivtid: **9.00-14.00**. Skrivningen omfattar 5 uppgifter, om sammanlagt 100 poäng.
3. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.
4. Om du känner dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren (besök, alternativt telefon).
5. Efter skrivningens slut får du behålla sidorna med frågeställningarna. Preliminära lösningar anslås på Studentportalen.

Uppmaningar

1. Följ noga de anvisningar som finns på skrivningsförsättsbladet.
2. Alla lösningar ska redovisas i en form som gör det lätt att följa din tankegång! Motivera alla väsentliga steg i lösningen. Ange alla antaganden du gör och alla förutsättningar du utnyttjar. Alla uppgifter kräver en verbal slutsats.
3. Vid konfidensintervall måste du ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts.
4. Vid alla hypotestest måste du ange H_0, H_1, signifikansnivå, testfunktion (inklusive antal frihetsgrader), förkastelseområde och resultat.
5. Vid variansanalys måste du ange modell.

Uppgift 1 (22 poäng)

I artikeln *New observations on maternal age effect on germline de novo mutations*[1] studeras associationen mellan antal genförändringar av en speciell typ (DNM) hos nyfödda och föräldrarnas ålder vid befruktningstillfället. Artikeln redovisar följande:[2]

> In the final dataset of 693 trios[3], the ages of the fathers at conception ranged from 17 to 63 years, with a mean of 33.4 years. The ages of the mothers ranged from 17 to 43 years with a mean of 31.2 years. The gestational age[4] of the newborns ranged from 23 to 42 weeks with a mean of 37.2 weeks. We use a linear model to test the parental age effects. The model used is:
>
> $$y_i = \beta_0 + \beta_1 \times fathersAge_i + \beta_2 \times mothersAge_i + \varepsilon_i.$$
>
> where y_i is the number of DNMs observed, β_0 denotes the intercept, $fathersAge_i$ is the age of the father, $mothersAge_i$ is the age of the mother.

Använd resultatet från artikeln i Table 1 på sida 3 för att besvara följande:

A) (3p) Ge en verbal tolkning av parameterskattningen för β_0.

B) (2p) Ge en verbal tolkning av parameterskattningen för β_2.

C) (9p) Beräkna ett konfidensintervall för β_2. Ge en verbal slutsats! (Utgå från att en granskning av residualerna har gjorts och att denna granskning tyder på att alla nödvändiga förutsättningar är uppfyllda.)

D) (5p) Anta att vi skattar en ny modell

$$\begin{aligned} y_i =& \beta_0 + \beta_1 \times fathersAge_i + \beta_2 \times mothersAge_i \\ &+ \beta_3 \times IVF_i + \beta_4 \times pregWeeks_i + \varepsilon_i. \end{aligned}$$

där IVF är en dummyvariabel som indikerar om befruktning skett i provrör och $pregWeeks$ är antal gravididetsveckor. Utskriften i ert statistikprogram ger $R^2 = 0.38$. Motivera på basis av förklaringsgrad om den ursprungliga modellen eller den nya modellen är att föredra.

E) (3p) Motivera varför författarna väljer att redovisa VIF-kolumnen. Vilken slutsats kommer författarna att göra baserat på VIF-kolumnen?

[1] Wong, W. S., et al. (2016). New observations on maternal age effect on germline de novo mutations. Nature communications, 7.

[2] Vissa förändringar har gjorts för att passa tentamen.

[3] En trio definieras som ett barn, en pappa och en mamma, dvs antalet observationer (rader) i datamaterialet är $n = 693$.

[4] Antal graviditetsveckor vid födsel.

Table 1 | Regressions on the effect of parental ages on the number of DNMs.

(a) Multiple linear regression of the total number of DNMs on the father's and mother's ages ($R^2 = 0.35$)

Model	β	s.e.	t	Pr($>t$)	VIF
(constant)	6.61	1.79	3.69	2.39×10^{-4}	
Father's age	0.64	0.06	10.03	$< 2.00 \times 10^{-16}$	1.99
Mother's age	0.35	0.08	4.51	7.61×10^{-6}	1.99

Uppgift 2 (14 poäng)

I en undersökning studerades sambandet mellan bostadskostnad och disponibel inkomst. Både inkomstklass och boendekostnad redovisas i en tregradig skala. För 510 slumpmässigt valda enpersonshushåll blev resultatet:

	Låg kostnad	Mellan kostnad	Hög kostnad	Summa
Låg inkomst	26	77	34	137
Mellan inkomst	70	88	62	220
Hög inkomst	26	42	85	153
Summa	122	207	181	510

Finns det ett samband mellan disponibel inkomst och bostadskostnad i den population som stickprovet är utvalt från? Undersök med 5% signifikansnivå.

Uppgift 3 (20 poäng)

Du har fått i uppdrag av ett franschiseföretag att undersöka vad som förklarar franschisetagarnas omsättning. Eftersom det handlar om företagshemligheter är variablerna i datasetet anonymiserade. Du vet dock att $X1 - X10$ är oberoende variabler och att beroende variabel Y är logaritmerad försäljning (där e^Y är tusentals kronor). Variablerna $X1, X2, X4, X5, X7, X8, X10$ är kvantitativa och $X3, X6, X9$ är kategorivariabler. Du skattar 2 stycken modeller enligt Minitabutskrifterna på sidorna 6-7. En analys av residulerna tyder på att alla förutsättningar är uppfyllda.

A) (7p) Studera skattningarna av Modell 1. Beskriv, så utförligt som möjligt, sambandet mellan $X1$ och Y.

B) (13p) Genomför en fullständig hypotesprövning och testa på 10% signifikansnivå om minst en av kategorivariablerna bör vara med i regressionsmodellen.

Modell 1:

Regression Analysis: Y versus X1; X2; X4; X5; X7; X8; X10; X3; X6; X9

Method

Categorical predictor coding (1; 0)

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	10	6,16747	92,29%	6,16747	0,61675	106,54	0,000
X1	1	4,13643	61,90%	1,43512	1,43512	247,91	0,000
X2	1	0,32522	4,87%	0,41581	0,41581	71,83	0,000
X4	1	0,45185	6,76%	0,63133	0,63133	109,06	0,000
X5	1	0,17309	2,59%	0,08810	0,08810	15,22	0,000
X7	1	0,01789	0,27%	0,00073	0,00073	0,13	0,724
X8	1	0,00001	0,00%	0,00153	0,00153	0,26	0,609
X10	1	0,03975	0,59%	0,00063	0,00063	0,11	0,742
X3	1	1,00861	15,09%	0,99872	0,99872	172,53	0,000
X6	1	0,00476	0,07%	0,00482	0,00482	0,83	0,364
X9	1	0,00984	0,15%	0,00984	0,00984	1,70	0,196
Error	89	0,51520	7,71%	0,51520	0,00579		
Total	99	6,68267	100,00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0,0760840	92,29%	91,42%	0,654218	90,21%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	10,022	0,148	(9,728; 10,316)	67,69	0,000	
X1	0,02792	0,00177	(0,02440; 0,03144)	15,75	0,000	2,90
X2	0,02903	0,00343	(0,02223; 0,03584)	8,48	0,000	1,07
X4	0,000514	0,000049	(0,000416; 0,000612)	10,44	0,000	1,16
X5	0,002048	0,000525	(0,001005; 0,003091)	3,90	0,000	1,12
X7	-0,00051	0,00144	(-0,00337; 0,00235)	-0,35	0,724	2,91
X8	-0,00263	0,00513	(-0,01282; 0,00756)	-0,51	0,609	1,09
X10	-0,00098	0,00296	(-0,00686; 0,00490)	-0,33	0,742	1,13
X3						
1	0,2243	0,0171	(0,1904; 0,2583)	13,13	0,000	1,13
X6						
1	-0,0154	0,0169	(-0,0489; 0,0181)	-0,91	0,364	1,23
X9						
1	-0,0266	0,0204	(-0,0670; 0,0139)	-1,30	0,196	1,06

Regression Equation

Y = 10,022 + 0,02792 X1 + 0,02903 X2 + 0,000514 X4 + 0,002048 X5 - 0,00051 X7 - 0,00263 X8
- 0,00098 X10 + 0,0 X3_0 + 0,2243 X3_1 + 0,0 X6_0 - 0,0154 X6_1 + 0,0 X9_0 - 0,0266 X9_1

Modell 2:

Regression Analysis: Y versus X1; X2; X4; X5; X7; X8; X10

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	7	5,14426	76,98%	5,14426	0,73489	43,95	0,000
X1	1	4,13643	61,90%	1,43076	1,43076	85,56	0,000
X2	1	0,32522	4,87%	0,43902	0,43902	26,25	0,000
X4	1	0,45185	6,76%	0,36450	0,36450	21,80	0,000
X5	1	0,17309	2,59%	0,17869	0,17869	10,69	0,002
X7	1	0,01789	0,27%	0,00842	0,00842	0,50	0,480
X8	1	0,00001	0,00%	0,00002	0,00002	0,00	0,973
X10	1	0,03975	0,59%	0,03975	0,03975	2,38	0,127
Error	92	1,53841	23,02%	1,53841	0,01672		
Total	99	6,68267	100,00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0,129313	76,98%	75,23%	1,84602	72,38%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	10,127	0,233	(9,665; 10,590)	43,46	0,000	
X1	0,02680	0,00290	(0,02105; 0,03256)	9,25	0,000	2,68
X2	0,02958	0,00577	(0,01812; 0,04105)	5,12	0,000	1,05
X4	0,000373	0,000080	(0,000214; 0,000531)	4,67	0,000	1,05
X5	0,002834	0,000867	(0,001112; 0,004556)	3,27	0,002	1,06
X7	0,00165	0,00232	(-0,00296; 0,00625)	0,71	0,480	2,62
X8	-0,00029	0,00853	(-0,01723; 0,01666)	-0,03	0,973	1,04
X10	-0,00758	0,00492	(-0,01734; 0,00218)	-1,54	0,127	1,08

Regression Equation

Y = 10,127 + 0,02680 X1 + 0,02958 X2 + 0,000373 X4 + 0,002834 X5 + 0,00165 X7 - 0,00029 X8
- 0,00758 X10

Uppgift 4 (24 poäng)

På uppdrag av Vårdförbundet genomförde UserAwards och Innova undersökningen "Vård-IT-rapporten 2010, Enkätundersökningar, flödesstudier och uppföljning av Vård-IT-kartan 2004". Enkätundersökningen innehöll flera frågor angående vårdgivares nöjdhet beträffande IT och journalsystem. I rapporten står:[5]

> Urvalsramar i denna undersökning har skapats utifrån medlemsregistren hos fackförbunden Kommunal, SKTF, Läkarförbundet och Vårdförbundet. Fackförbunden har själva dragit urvalen och levererat dem till UsersAward. Urvalet ska vara draget som ett proportionellt stratifierat obundet slumpmässigt urval.

Totalt besvarade 1019 individer som arbetar i vården enkäten, varav 373 var sjuksköterskor och barnmorskor (SSK/BSK), 252 var undersköterskor (USK), 306 var läkare och 88 var läkarsekreterare (LÄKSEKR). Vi antar att det inte finns något bortfall. Skattade andelar i respektive stratum som håller med ett antal påståenden presenteras i Tabell 6.

	Enkelt att lära	Lita på att det fungerar	Patient integritet skyddas	Lätt att korrigera	Tillgängligt när det behövs	Tillräckligt snabbt
SSK/BM	76 %	57 %	73 %	63 %	81 %	57 %
USK	72 %	63 %	67 %	60 %	74 %	53 %
LÄKARE	58 %	56 %	61 %	55 %	79 %	48 %
LÄKSEKR	79 %	68 %	77 %	80 %	95 %	70 %

Tabell 6: Journalsystemanvändare. Teknisk utformning

A) (6p) Ge en punktskattning för andelen individer i vården som håller med om att journalsystemet är enkelt att lära.

B) (12p) Beräkna ett 95% konfidensintervall för andelen individer i vården som håller med om att journalsystemet är enkelt att lära. Tolka intervallet! (Du behöver inte använda ändlighetskorrektion, men var noga med att ange övriga förutsättningar.)

C) (6p) Anta att du med ett OSU-UÅ vill skatta medelvärdet i en (annan) population bestående av 900 individer. Du känner sedan tidigare till att populationsvariansen är 52. Om felmarginalen får vara vara maximalt 2 enheter, hur stort måste stickprovet vara?

[5]Text och värden är något ändrade för att passa uppgiften.

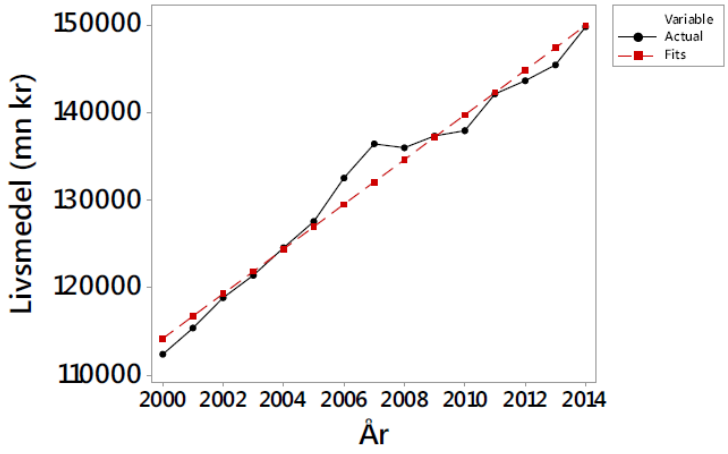
Uppgift 5 (20 poäng)

En linjär trend har anpassats till den årliga livsmedelsförsäljningen i Sverige. Använd Minitabutskrift och figurer på sidorna 10-11 för att besvara följande:

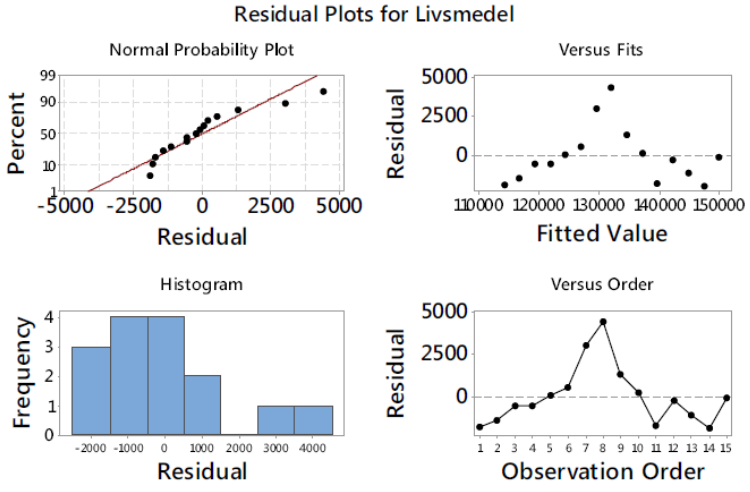
A) (3p) Skriv ut modellen som har skattats och tolka resultatet. Var noga med att definiera alla delar i modellen.

B) (10p) Beräkna ett prediktionsintervall för försäljningen år 2015. Använd konfidensgraden 95%. Tolka resultatet. Utnyttja att $\sum t = 120$ och att $\sum t^2 = 1240$.

C) (7p) Förutom Minitabutskriften och figuren över den anpassade trendlinjen har du även residualfigurer. Anser du att modellen ger en god beskrivning av data? Är nödvändiga förutsättningar uppfyllda? Om inte, vilka är konsekvenserna?



(a) Årlig livsmedelförsäljning (miljoner kronor), 2000-2014



(b) Residualfigurer

Regression Analysis: Livsmedel versus t

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1822276241	1822276241	531,18	0,000
t	1	1822276241	1822276241	531,18	0,000
Error	13	44597711	3430593		
Total	14	1866873952			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1852,19	97,61%	97,43%	96,98%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	111704	1006	110,99	0,000	
t	2551	111	23,05	0,000	1,00