STOCKHOLMS UNIVERSITET
Statistiska institutionen
Hans Nyquist

TENTAMEN I STATISTISK REGRESSIONSANALYS OCH UNDERSÖKNINGSMETODIK

DELKURS 1, REGRESSIONSANALYS

2020-12-01

Skrivtid: 09.00-15.00

Godkända hjälpmedel: Miniräknare, dator, kurslitteratur, föreläsningsanteckningar och språklexikon, formelsamling (från Athena) och statistiska tabeller (från Athena)

Obs! Det är inte tillåtet att ta hjälp av andra personer under skrivningen

Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
Resultatet meddelas senast den 9 november.
Kortfattade svar läggs ut strax efter tentamen på Athena

Kontakt med examinator under tentamen: För eventuella frågor om innehållet i tentan kan du kontakta examinator under pågående tentamen på mail: Hans.Nyquist@stat.su.se Inkommande mailfrågor besvaras kontinuerligt under tentans gång. Om examinator behöver informera om någonting under tentan görs detta till din registrerade mailadress. Kontrollera därför din mail under tentans gång.

Observera att praktisk hjälp endast finns tillgänglig under tentans första timme på mailadressen expedition@stat.su.se Läs noggrant bifogad instruktion för inlämning av tentan. Där finns all nödvändig information om inlämning, anonymkod etc. Om du trots instruktionerna skulle få problem att lämna in tentan, maila istället tentan till tenta@stat.su.se Detta görs dock bara i undantagsfall.

Uppgift 1. (30 poäng)

I ett försök ville man undersöka om samma arbete utfört med olika tempo ger olika träningseffekter. Man lät därför en grupp om 15 personer med så lika fysiska förutsättningar som möjligt springa en viss sträcka på ett löpband varje dag under två månader. Personerna sprang samma sträcka men med olika hastighet. Som mått på träningseffekten användes personernas VO2max (maximal syreupptagning per kg kroppsvikt). Följande observationer, med några uträkningar, erhölls

Obs	Tempo, km/h	VO2max			
	x	y	x^2	y^2	xy
1	8,5	37	72,25	1369	314,50
2	9,0	35	81,00	1225	315,00
3	9,0	38	81,00	1444	342,00
4	9,5	40	90,25	1600	380,00
5	10,0	42	100,00	1764	420,00
6	10,0	43	100,00	1849	430,00
7	10,5	45	110,25	2025	472,50
8	11,0	44	121,00	1936	484,00
9	11,0	42	121,00	1764	462,00
10	11,5	45	132,25	2025	517,50
11	12,0	49	144,00	2401	588,00
12	12,0	48	144,00	2304	576,00
13	12,5	52	156,25	2704	650,00
14	13,0	52	169,00	2704	676,00
15	13,0	53	169,00	2809	689,00
Summa	162,5	665	1791,25	29923	7316,50

a) Sätt upp en enkel linjär regressionsmodell där löptempo förklarar förväntad VO2max och uppskatta modellens parametrar med minstakvadratmetoden.

b) En av experimentets medarbetare räknade ut residualkvadratsumman till $SS_E = 32,076$. Bestäm R^2 och ge en tolkning.

c) Bilda ett 95%-igt konfidensintervall för lutningsparametern.

d) Bilda ett 95%-igt konfidensintervall för förväntat CO2max då tempot 12 km/tim har använts i träningen.

e) Bilda ett 95%-igt prediktionsintervall för en ny observation på CO2max då tempot 12 km/tim har använts i träningen. Jämför och kommentera resultatet med resultatet i deluppgift d

Uppgift 2. (10 poäng)

a) Ett av de antaganden som behöver göras i en regressionsmodell för att inferens ska vara giltig är att beroendevariabeln är normalfördelad. Ange vilka andra antaganden som måste gälla.

b) Man vill undersöka om strömmen mellan två elektroder nedstoppade i en vätska kan relateras till alkoholhalten i vätskan. Man gjorde därför exakta bestämningar av alkoholhalten, x, och strömstyrkan mellan elektroderna, y, i tio vätskeblandningar. Minstakvadratuppskattningarna av en linjär regressionsmodell blev

$$\begin{aligned} \widehat{\mu}_i &= 8,617 + 1,334x_i \\ s_e^2 &= 9,584 \\ R^2 &= 0,58 \end{aligned}$$

x_i	y_i	$\widehat{\mu}_i$	$y_i - \widehat{\mu}_i$
8	14,56	19,29	-4,73
1	11,45	9,95	1,50
6	15,86	16,62	-0,76
6	17,95	16,62	1,32
4	16,15	13,96	2,20
5	15,92	15,29	0,63
7	14,43	17,96	-3,52
5	13,99	15,29	-1,30
10	27,42	21,96	5,46
3	11,83	12,62	-0,79

Gör en utvärdering av modellantagandena. Eftersom antalet observationer är litet är det svårt att utvärdera normalfördelningsantagandet. Du behöver därför inte (i den här uppgiften) utvärdera det antagandet.

Uppgift 3. (25 poäng)

Resultaten från en regressionsanalys av 22 observationer gav minstakvadratuppskattningarna $a = 5,2$ och $b = -4,4$ samt kvadratsummorna $SS_E = 140$ respektive $SS_R = 260$. Bestäm förklaringsgraden R^2 och pröva hypotesen $H_0 : \beta = 0$ mot $H_A : \beta \neq 0$. Använd signifikansnivån 1 procent.

Uppgift 4. (25 poäng)

Tabellen nedan visar kvartalsvisa observationer på elförbrukningen (i MWh) för ett samhälle från första kvartalet 2016 till fjärde kvartalet 2019.

7,5 1,5 2,5 5 6,5 2 3 6,5 8,5 4 4 8 9 4,5 5,5 8,5

a) Vilka komponenter kan serien tänkas bestå av?

b) Utjämna serien med hjälp av löpande medelvärden

c) Uppskatta säsongkomponenten

Uppgift 5. (10 poäng)

I en analys användes logistisk regression för att studera om en behandling mot oro påverkar sannolikheten att en patient som har haft en hjärtinfarkt får en ny hjärtinfarkt inom ett år. Data från 20 patienter analyserades där den beroende variabeln y definieras som $y = 1$ om patienten får en ny hjärtinfarkt inom ett år och 0 annars. Som förklaringsvariabler användes $x_1 = \textit{behandling}$, med värdet 1 om patienten har genomgått en behandling baserad på samtalsterapi för att minska sin oro, och 0 annars och $x_2 = \textit{oro index}$, som är ett index där högre värden anger en större oro.
Modellen definierades genom

$$Logodds = \alpha + \beta_1 x_1 + \beta_2 x_2$$

En utskrift från SAS gav följande tabell

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-7,388	3,279	5,076	0,024
behandling	1	1,024	1,171	0,765	0,382
oroindex	1	0,119	0,055	4,688	0,030

a) Beräkna sannolikheten att en patient får en ny hjärtinfarkt inom ett år om patienten har genomgått behandlingsprogrammet och har värdet 60 på oroindex.

b) Avgör om behandlingen mot oro är effektiv för att minska sannolikheten att drabbas av en ny hjärtinfarkt.