STOCKHOLMS UNIVERSITET
Statistiska institutionen
Hans Nyquist

TENTAMEN I STATISTISK REGRESSIONSANALYS OCH UNDERSÖKNINGSMETODIK

DELKURS 1, REGRESSIONSANALYS

2021-01-11

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 25 januari.
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)

En butik har noterat pris, x, och såld kvantitet, y, av jordgubbar under 8 dagar i juli. Följande data erhölls

Vecka	Pris (kr) x	Kvantitet (liter) y
1	40	110
2	35	145
3	35	140
4	32	160
5	30	160
6	30	150
7	25	170
8	30	155

a) Gör en lämplig figur över datat och beräkna urvalskovariansen och urvalskorrelationen

b) Sätt upp en regressionsmodell med såld kvantitet som beroende variabel och pris som förklaringsvariabel. Ange fullständiga antaganden. Ge en tolkning av modellens parametrar. Är sambandet kausalt? Motivera!

c) Bestäm minstakvadratskattningarna av modellens parametrar och R^2. Tolka skattningarna och det erhållna värdet på R^2.

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

e) Har priset en signifikant påverkan på såld kvantitet?

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) I ett laboratorium undersöktes ett samband mellan elektrisk strömstyrka, x, och utvecklad effekt, y i en krets. Följande observationer gjordes:

x_i	y_i	$\widehat{\mu}_i$	$y_i - \widehat{\mu}_i$
2,30	11,41	9,42	1,99
4,90	4,85	10,13	-5,28
6,50	8,07	10,58	-2,51
3,60	6,87	9,78	-2,91
2,40	11,60	9,45	2,16
7,30	10,12	10,80	-0,68
4,10	5,23	9,91	-4,68
5,00	5,71	10,16	-4,45
1,50	16,81	9,20	7,62
8,90	19,99	11,24	8,75

Minstakvadratuppskattningarna av en linjär regressionsmodell blev

$$\begin{aligned} \widehat{\mu}_i &= 8,785 + 0,275x_i \\ s_e^2 &= 28,503 \end{aligned}$$

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)

För att uppskatta hur elkostnader för bergvärmepumpar för uppvärmning av villor beror på villornas storlek observerades elkostnaden, Y, och bostadsytan, X, hos sju friliggande villor i ett visst område. En regressionsanalys av observationerna gav $SS_R = 15,36$ och $F = 16$. Uppskatta observationernas varians runt regressionslinjen, s^2, och regressionslinjens förklaringsgrad, R^2. Pröva med signifikansnivån 5 procent om villornas storlek signifikant påverkar uppvärmningskostnaden.

Uppgift 4. (25 poäng)

a) En tidsserie kan tänkas bestå av ett antal komponenter. Beskriv komponenterna och hur de kan komponeras till en additiv respektive en multiplikativ modell. Ange eventuella skillnader och likheter mellan modellerna.

b) Tabellen nedan visar kvartalsvisa observationer på elförbrukningen (i MWh) för ett samhälle från första kvartalet 2016 till fjärde kvartalet 2019. Välj en lämplig modell och utjämna serien med hjälp av löpande medelvärden.

2,9	1,2	0,8	2,4	4,1	1,7	1,2	4,0	6,6	2,7	1,8	5,1	10,7	3,9	2,7	7,5

c) Uppskatta säsongkomponenten.

Uppgift 5. (10 poäng)

I en undersökning användes logistisk regression för att studera om variablerna $x_1 = \textit{ålder}$ och $x_2 = \textit{inkomst}$ (i tusentals kr, tkr) påverkar sannolikheten för bortfall. I analysen hade den beroende variabeln y värdet 1 om respondenten är ett bortfall, och värdet 0 om respondenten deltar i undersökningen. Data från $n = 20$ respondenter analyserades.
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	-3,388	3,279	1,068	0,301
ålder	1	-0,102	0,117	0,745	0,388
inkomst	1	0,119	0,155	0,589	0,443

a) Beräkna sannolikheten att en respondent som är 25 år och med en inkomst om 28000 kt (28 tkr) blir ett bortfall.

b) Testa med signifikansnivån 5 procent om variablerna ålder respektive inkomst påverkar sannolikheten för bortfall.