STOCKHOLMS UNIVERSITET
Statistiska institutionen
Hans Nyquist

TENTAMEN I REGRESSIONSANALYS OCH UNDERSÖKNINGSMETODIK

DELKURS 1, REGRESSIONSANALYS OCH TIDSSERIEANALYS

2021-04-30

Skrivtid: 09.00-14.00. **Inlämningstid:** 14.00-15.00

Godkända hjälpmedel: Miniräknare, dator, kurslitteratur, föreläsningsanteckningar och språklexikon, formelsamling och statistiska tabeller (bifogas).

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 14 maj.
Lösningsförslag till tentamensuppgifterna läggs ut på Athena strax efter tentamen

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 (25 poäng)

För att studera effekten av reklam på försäljning hos e-handel observerades månadskostnaden för annonsering på internet och intäkter från försäljning under en månad för sju e-handelsbutiker. Observationerna framgår av nedanstående tabell

Butik	Annonskostnad (tkr)	Försäljning (tkr)
	x	y
1	1,7	368
2	1,5	340
3	2,8	665
4	5,0	954
5	1,3	331
6	2,2	556
7	1,3	376

a) Gör en lämplig figur över observationerna och beräkna urvalskovariansen och urvalskorrelationen mellan variablerna

b) Sätt upp en regressionsmodell med intäkter från försäljning som beroende variabel och kostnad för annonsering 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-procentigt konfidensintervall för lutningsparametern.

e) Bestäm ett 95-procentigt prediktionsintervall för försäljning hos en ny butik som annonserar för 2 tkr.

Uppgift 2 (30 poäng)

Finns det ett positivt samband mellan försäljning av blyhaltig bensin och bly i blodet hos nyfödda barn? Den frågan studerades i en forskningsstudie där data analyserades med en enkel linjär regressionsmodell. I modellen är försäljning av blyhaltig bensin (i ton) per månad förklaringsvariabel och koncentration av bly i blodet (i mikroliter bly per deciliter blod) hos nyfödda barn beroendevariabel. $n = 14$ observationer på variablerna gjordes under åren 1980 och 1981, i Massachusetts, USA.
Minstakvadratuppskattningen av lutningskoefficienten blev $b = 0,014885$, med t-värdet 3,15. Vidare vet man att residualkvadratsumman blev $SS_E = 4,5560$.

a) Bestäm residualvariansen, s_e^2.

b) Bestäm förklaringsgraden R^2

c) Bilda ett 95-procentigt konfidensintervall för lutningskoefficienten β. Kan man dra slutsatsen att försäljningen av blyhaltig bensin påverkar koncentrationen av bly i blodet hos nyfödda barn?

Uppgift 3 (15 poäng)

Tabellen nedan visar hushållens konsumtion av livsmedel och alkoholfria drycker i fasta priser (2019 års penningvärde, miljarder kr) från första kvartalet 1981 till andra kvartalet 1987.

81, I	81, II	81, III	81, IV	82, I	82, II	82, III	82, IV	83, I	83, II	83, III	83, IV
36,1	38,7	35,9	37,0	36,3	38,4	36,5	38,0	37,2	37,1	36,5	36,7
84, I	84, II	84, III	84, IV	85, I	85, II	85, III	85, IV	86, I	86, II	86, III	86, IV
37,1	38,3	36,4	37,6	36,9	38,7	36,0	37,8	37,2	39,0	36,9	38,5
87, I	87, II										
37,6	40,5										

a) Rita en lämplig figur som illustrerar tidsseriens utveckling.

b) För att analysera observationerna definierades fyra variabler: tiden t ($t = 1$ för första kvartalet 1981, $t = 2$ för andra kvartalet 1981, osv.) och tre dummyvariabler för kvartalen, $Kv1 = 1$ för kvartal 1 och 0 för övrigt, $Kv2 = 1$ för kvartal 2 och 0 för övrigt samt $Kv3 = 1$ för kvartal 3 och 0 för övrigt. Tre modeller estimerades, en linjär modell, en kvadratisk modell och en exponentiell modell. μ_t betecknar förväntat värde av hushållens konsumtion vid tidpunkten t. Följande resultat erhölls:

Linjär modell

$$\mu_t = \alpha + \beta t + \gamma_1 Kv1_t + \gamma_2 Kv2_t + \gamma_3 Kv3_t$$

ANOVA

Källa	SS	df	MS	F	sig
Regression	23,912	4	5,978	19,562	0.000
Residual	6,388	21	0,304		
Total	30,300	25			

Coefficients

	b	std error	t	sig.
Constant	36,838	0,303	121,686	0,000
t	0,054	0,014	3,763	0,001
$Kv1$	-0.631	0,307	-2,055	0,053
$Kv2$	1,071	0,307	3,492	0,002
$Kv3$	-1,179	0,319	-3,600	0,001

Kvadratisk modell

$$\mu_t = \alpha + \beta_1 t + \beta_2 t^2 + \gamma_1 Kv1_t + \gamma_2 Kv2_t + \gamma_3 Kv3_t$$

ANOVA

Källa	SS	df	MS	F	sig
Regression	25,108	5	5,022	19,343	0.000
Residual	5,192	20	0,260		
Total	30,300	25			

Coefficients

	b	std error	t	sig.
Constant	37,425	0,391	95,679	0,000
t	-0,063	0,056	-1,118	0,278
t^2	0,004	0,002	2,146	0,044
$Kv1$	-0.706	0,286	-2,471	0,023
$Kv2$	0,996	0,286	3,488	0,002
$Kv3$	-1,179	0,294	-4,004	0,001

Exponentiell modell

$$\mu_t = \alpha \cdot \beta^t \cdot \gamma_1^{Kv1_t} \cdot \gamma_2^{Kv2_t} \cdot \gamma_3^{Kv3_t}$$

Modellen transformerades till

$$\ln \mu_t = \ln \alpha + t \ln \beta + Kv1_t \ln \gamma_1 + Kv2_t \ln \gamma_2 + Kv3_t \ln \gamma_3$$

och estimerades med resultaten

ANOVA

Källa	SS	df	MS	F	sig
Regression	0,017	4	0,004	20,111	0.000
Residual	0,004	21	0,000		
Total	0,021	25			

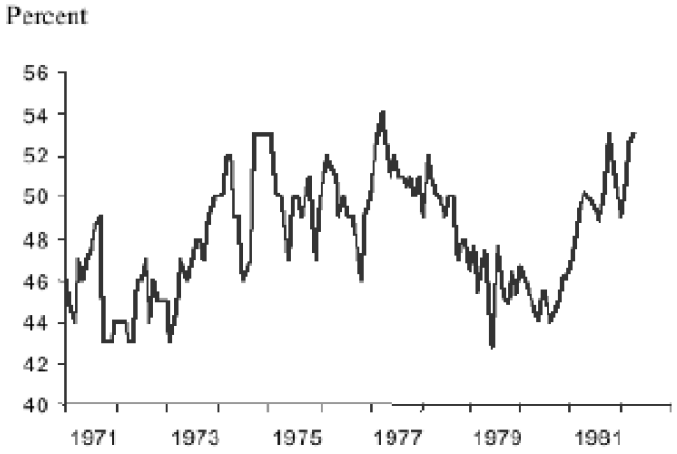
Coefficients

	b	std error	t	sig.
Constant	3,607	0,008	454,832	0,000
t	0,001	0,000	3,794	0,001
$Kv1$	-0.017	0,008	-2,104	0,048
$Kv2$	0,028	0,008	3,475	0,002
$Kv3$	-0,032	0,008	-3,812	0,001

Välj en lämplig modell. Argumentera för ditt val och ge tolkningar av modellens parametrar.

c) Beräkna prognoser för hushållens konsumtion av livsmedel och alkoholfria drycker för de två sista kvartalen 1987. Använd den modell du valde i uppgift b. (De observerade värdena blev 37,1 respektive 38,6).

Uppgift 4 (15 poäng)
Sympatierna för de tyska politiska partierna CDU/CSU har uppmätts varje månad. Nedanstående figur illustrerar observationer på andelen röstberättigade i Tyskland som skulle rösta på CSU/CDU om det hade varit val vid undersökningstillfället, för tiden januari 1971 till april 1982 ($T = 136$ observationer).



För att analysera observationerna ansattes en AR(1) modell,

$$\mu_t = \alpha + \beta x_{t-1},$$

där μ_t är det förväntade värdet vid tidpunkt t och x_{t-1} är det observerade värdet vid tidpunkten $t-1$. Estimation av modellens parametrar gav $a = 8,053$ med den uppskattade variansen $s_a^2 = 5,5122$, och $b = 0,834$ med den uppskattade variansen $s_b^2 = 0,0023787$. Residualvariansen uppskattades till $s_e^2 = 1,586$. En utvärdering av modellen antydde att gjorda antaganden var rimliga.

a) Pröva med signifikansnivån 5 procent om sympatierna för CSU/CDU en viss månad beror av sympatierna föregående månad.

b) Gör prognoser för sympatierna för CDU/CSU en respektive två månader framåt i tiden (x_{T+1} och x_{T+2}) om sympatierna vid x_T är 40 procent. Gör också prognoser två månader framåt i tiden om sympatierna vid x_T är 55 procent.

c) Tidsserien varierar kring ett visst tal och prognoserna närmar sig det talet. Detta tal är tidsseriens ekvilibrium. Bestäm en uppskattning av ekvilibrium för sympatierna för CDU/CSU.

Uppgift 5 (15 poäng)

För att undersöka effekten av ett toxin på en viss sorts insekter utsattes sex grupper om 250 insekter i vardera gruppen för en viss mängd av toxinet. Olika grupper utsattes för olika stor dos av toxinet. Efter en tid noterades antal döda insekter i varje grupp. En logistisk regression användes för att analysera observationerna varvid följande resultat erhölls:

Analysis of Maximum Likelihood Estimates					
Coefficient	DF	Estimate	SE Coef	z-value	P-Value
Constant	1	-2,664	0,156	-16,94	0,000
Dos	1	0,6740	0,0391	17,23	0,000

a) Rita en figur över hur logodds för att en insekt dör beror av dosen toxin. Beräkna speciellt logoddset om dosen är 3 (mg).

b) Avgör om dosen toxin signifikant påverkar dödligheten hos insekterna. Använd signifikansnivån 5 procent.

c) Bestäm sannolikheten att en insekt som utsätts för dosen 3 (mg) överlever.