SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2024-05-03

Skrivtid: kl. 08–13 (5 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text.
Bifogade hjälpmedel: Formel- och tabellsamling för statistik och dataanalys II, 15 hp.

Tentamen består av 5 uppgifter uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas för fulla poäng.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
- Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan ge poängavdrag.
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningar i de efterföljande uppgifterna.
- I beräkningar från R-utskrifter får du utgå från det som är givet.

Tentamen kan maximalt ge 100 poäng. För godkänt resultat krävs minst 50 poäng.

Betygsgränser

A	90–100
B	80–89
C	70–79
D	60–69
E	50–59
Fx	40–49
F	0–40

Obs! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Lösningsförslag läggs ut på athena efter att tentamenstiden är över.

Lycka till!

Uppgift 1 - Interaktionseffekter (22 poäng)

I ett forskningspapper från 2005 undersökte Daniel Hamermesh och Amy Parker om studenter tenderar att ge lärare som ser bättre ut mer positiva omdömen i kursutvärderingar. Analysen i pappret är baserad på svar från 463 studenter, där studenterna fått svara på frågor om hur duktiga dom tyckte att olika lärare var på att undervisa (`eval`), tillsammans med ett antal frågor om lärarnas utseende (sammanfattade i variabeln `beauty`). Datasetet innehåller även andra variabler som beskriver olika egenskaper hos lärarna, som kön (`female`) och ålder.

- `eval` Hur duktig studenten tycker att läraren är på att undervisa. Svar på en skala mellan 1.0 till 5.0. (Med svarsalternativ 1.0, 1.1, 1.2, …, 5.0)
- `beauty` Studentens uppfattning om lärarens utseende. Denna variabel är en kontinuerlig numerisk variabel som är konstruerad baserat på ett antal enkätfrågor och tar värden mellan -2 och 2.
- `female` Lärarens kön. Antar i detta datasetet två värden: 1 för kvinnor, 0 för män.

Hamermesh och Parker visar att studenter ofta tenderar att ge mer positiva omdömen av hur duktiga lärarna är på att undervisa när dom tycker att dom ser bra ut. I den här uppgiften har en lite mer komplicerad modell anpassats till datasetet som inkludera kön samt en interaktion mellan kön och utseende.

```
Parameter estimates
-----------------------------------------------
              Estimate Std. Error  t value   Pr(>|t|)
(Intercept)    4.10364   0.033593 122.1579 0.0000e+00
beauty         0.20027   0.043333   4.6217 4.9520e-06
female        -0.20505   0.051030  -4.0182 6.8513e-05
beauty:female -0.11266   0.063975  -1.7610 7.8910e-02


Analysis of variance


      df        SS
Regr  NA  10.32099
Error NA 131.91763
Total NA 142.23862
```

a) Modellen ovan kan beskrivas som två separata regressionslinjer för sambandet mellan `eval` och `beauty`, en för män och en för kvinnor. Ta fram matematiska uttryck för dessa (skattade) regressionslinjer. (5p)

b) Tolka dom skattade regressionskoefficienterna för `female`, `beauty` och `beauty:female`. (5p)

c) Genomför ett F-test av modellen som helhet. Använd $\alpha = 0.05$. (Hint: titta på uttrycken för medel-ANOVA. Notera att antal frihetsgrader har dolts i utskriften, du behöver alltså ta fram dessa själv.) (7p)

d) För vilket värde på `beauty` kommer det skattade förväntade värdet på `eval` vara samma för män som för kvinnor? (5p)

Uppgift 2 - Multikollinearitet (15 poäng)

En multipel regressionsmodell har skattats som utgår från populationsmodellen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

```
Parameter estimates
-----------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  0.13139   0.096696  1.3588 1.7739e-01
x1           0.90508   0.152664  5.9285 4.7947e-08
x2           1.03661   0.099211 10.4485 1.6232e-17
x3           1.00542   0.093698 10.7305 4.0465e-18
```

För att undersöka om multikollinearitet är ett problem för en av prediktorerna har ytterliggare en modell skattats, med följande resultat.

```
Analysis of variance - ANOVA
-----------------------------------------------
      df     SS       MS     F     Pr(>F)
Regr   2 62.026 31.01279 80.12 2.8642e-21
Error 97 37.547  0.38708
Total 99 99.572

Parameter estimates
-----------------------------------------------
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 0.032913   0.064224  0.51247 6.0949e-01
x2          0.509387   0.040974 12.43197 8.6762e-22
x3          0.017284   0.062292  0.27746 7.8202e-01
```

a) Beskriv vad multikollinearitet är och varför det kan vara ett problem. (5p)

b) Ange två typiska tecken på multikollinearitet. (5p)

c) Använd utskrifterna ovan för att beräkna VIF. Vilken prediktor har du beräknat VIF för? Verkar multikollinearitet vara ett problem? (5p)

Uppgift 3 - Logistisk regression (19 poäng)

År 1960 mättes kolesterolhalten hos 3154 friska personer. 9 år senare undersöktes det vilka av dessa personer som utvecklat någon hjärt- och kärlsjukdom. En logistisk regression har skattats för att undersöka hur sambandet mellan kolesterol och hjärt- och kärlsjukdomar ser ut.

- Variabeln `chol` anger mängden kolesterol i mmol per liter.
- Responsvariabeln `chd` antar värdet 1 om personen utvecklat en hjärt- och kärlsjukdom, annars 0.

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.36       0.36  -14.91        0
chol            0.01       0.00    8.65        0
```

a) Vad är den skattade sannolikheten att en person med ett kolesterolvärde på 210 mmol per liter inte utvecklade någon hjärt- och kärlsjukdom? (4p)

b) Tolka interceptet i termer av odds *och* i termer av sannolikheter. Tolka parameterskattnigen för `chol` i termer av oddskvot. (5p)

c) För vilket värdet på `chol` är den skattade sannolikheten exakt 0.5? Detta värde har ett speciellt namn, vilket? (6p)

d) Oavsett vilken modell vi använder så kommer alltid dom skattade sannolikheterna från en logistisk regression ligga mellan 0 och 1. Vilka värden kan dom skattade oddsen ligga mellan? (4p)

Uppgift 4 - Ickelinjär regression (20 poäng)

En populationsmodellen som beskriver hur antalet gulhårig bergrävar (`fox`) utvecklas över tid ges av

$$\texttt{fox} = \alpha\beta^{\texttt{t}}\varepsilon, \qquad \log\varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

där tiden `t` anges i år.

a) Vad kallas sambandet som ges av populationsmodellen? (2p)

b) Populationsmodellen som skrivits ut ovan behöver logaritmeras för att det ska gå att estimera den med minsta kvadrat-metoden. Skriv ut den logaritmerade modellen. (3p)

Följande modell skattas med hjälp av minsta kvadrat-metoden

```
Parameter estimates
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.994      0.042 119.605        0
t              0.050      0.002  21.205        0
```

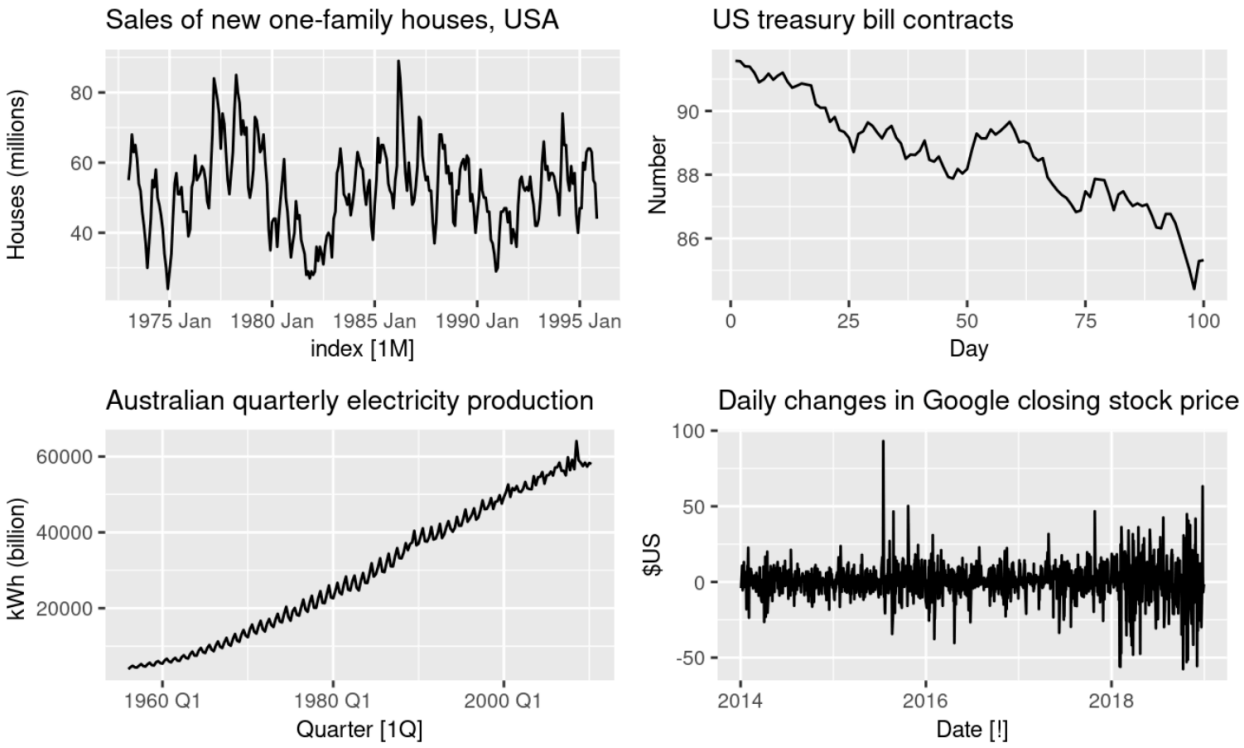
c) Enligt den skattade modellen, hur många rävar förväntas det finnas efter 23 år? Under skattandet av modellen har den naturliga logaritmen, e, använts. (5p)

d) Tolka den skattade regressionskoefficienten för `t` i termer av orginalmodellen. (Alltså, det är inte OK att tolka estimatet i termer av log-rävar. Vad är ens en log-räv?) (5p)

e) Som ett alternativ till modellen ovan har en polynomregression av ordning 5 anpassats. Datasetet består endast av 30 observationer, vilket innebär att ett polynom av grad 5 kan leda till överanpassning. För att hantera detta har Ridge och LASSO använts. Förklara på ett övergripande sätt hur Ridge och LASSO fungerar. Vilken av raderna i tabellen nedan korresponderar till modellen som använder Ridge? Vilken korresponderar till modellen som använder LASSO? Vilken korresponderar till modellen som varken använder Ridge eller LASSO? (5p)

Modell	Intercept	t	t^2	t^3	t^4	t^5
1	352.22	756.78	134.04	83.45	0.00	0.00
2	352.22	806.25	183.51	132.92	32.11	-22.66
3	352.22	701.15	159.59	115.60	27.93	-19.71

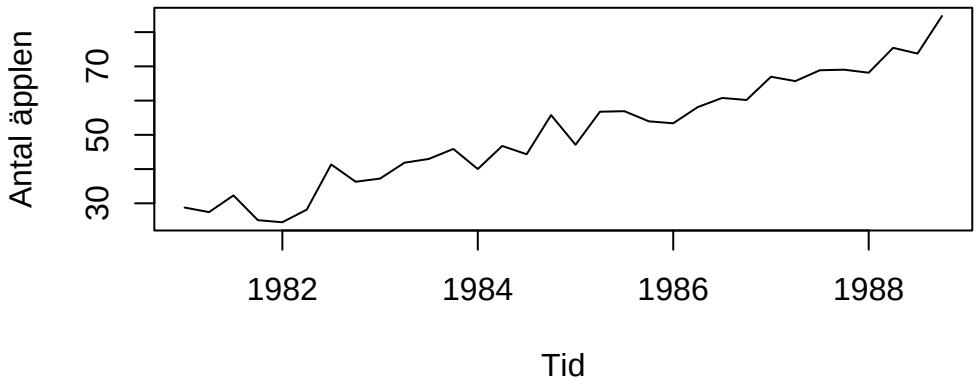
Uppgift 5 - Tidsserieanalys (24 poäng)



a) För var och en av tidsserierna ovan, svara på följande (6p)

- har tidsserien trend?
- har tidsserien en cykel?
- har tidsserien säsong?
- om tidsserien har säsgon, är säsongen additiv eller multiplikativ?

b) Figuren nedan visar den genomsnittliga äppelförbrukningen i Tasmanien per kvartal från kvartal 1 (`Qtr1`) 1981 till kvartal 4 (`Qtr4`) 1988. På nästa sida finns värdena från kvartal 4 1981 till kvartal 1 1983 utskrivna. Ta fram trendskattningar för samtliga kvartal 1982 med $3 - MA$. Ta sedan fram restkomponenten för samtliga kvartal 1982. (Tidsserien har ingen säsongskomponent.) (8p)




```
      Qtr1  Qtr2  Qtr3  Qtr4
1981                    25.08
1982 24.47 28.17 41.35 36.31
1983 37.23
```

c) Beräkna första ordningens autokorrelationen baserat på restkomponenterna du tog fram i uppgift c. (6p)

d) Vilket test kan du använda för att avgöra om första ordningens autokorrelation är signifikan skiljd från noll? Beskriv kortfattat hur testet fungerar. (4p)