SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2023-06-08

Skrivtid: kl. 14–19 (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 uppgift.

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 (25 poäng)

Datasetet `bike` innehåller information om användning av hyrcyklar i Washington, D.C. åren 2011 och 2012. En modell har skattats för att studera sambandet mellan antalet dagliga uthyrningar (`nRides`) och dom två förklarande variablerna temperatur i Celsius (`temp`) och huruvida det är sommar eller inte (`summer`, en dummy som antar värdet 1 när det är sommar, och 0 annars).

Eftersom att det är mycket möjligt att sambandet mellan temperatur och antalet uthyrningar skiljer sig åt beroende på om det är sommar eller inte används en modell med en *interaktionseffekt*:

$$\text{nRides} = \beta_0 + \beta_1 \text{temp} + \beta_2 \text{summer} + \beta_3 \text{temp} \cdot \text{summer} + \varepsilon$$

Modellen skattas med hjälp av R, vilket ger följande resultat

```
Analysis of variance - ANOVA
------------------------------------------------
       df         SS        MS      F     Pr(>F)
Regr    3 1181382753 393794251 [illegible]
Error 727 1558152639   2143264
Total 730 2739535392

Parameter estimates
------------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  1081.85   164.3390  6.5830 8.8289e-11
temp          191.14     9.5863 19.9388 6.7935e-71
summer       5000.09   997.3933  5.0132 6.7326e-07
temp:summer  -206.64    35.9505 -5.7480 1.3288e-08
```

a) Tolka de skattade parametrarna i utskriften. Är interceptet rimligt att tolka?

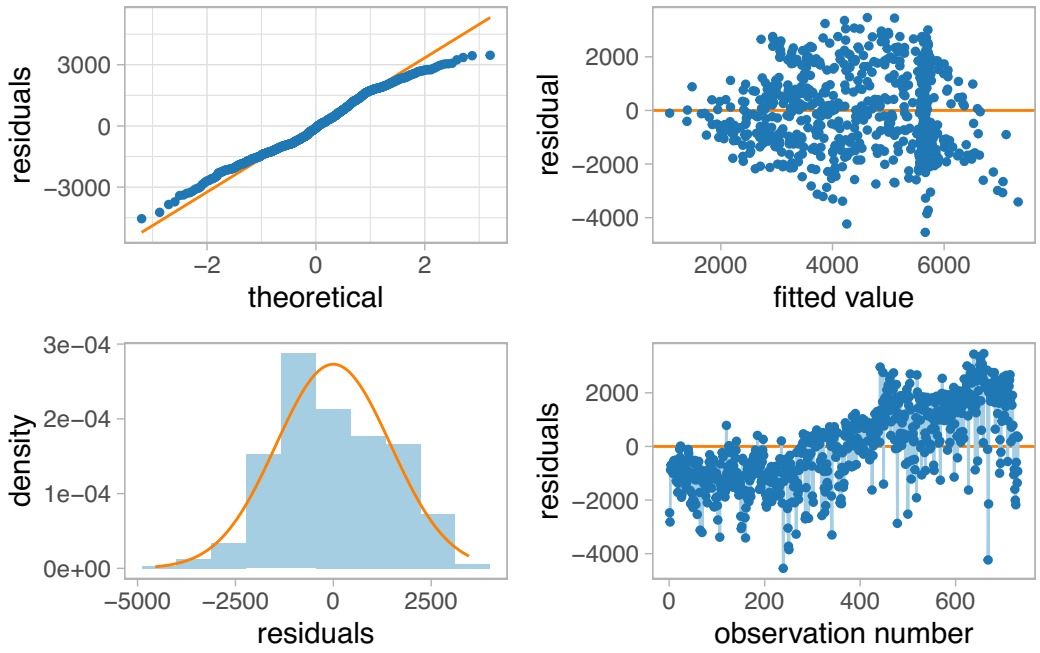
b) Testa om modellen som helhet är signifikant. Använd $\alpha = 0.01$.

c) Modellen ovan använder en dummy för sommar, vilket innebär att varje observation antingen är sommar eller ”inte sommar”. En mer realistisk indelning är att använda tre olika dummy-variabler, samt interaktioner, som i modellen nedan. Modellen med endast sommar kan representeras som en regressionslinje för sommar och en för ”inte sommar”. Modellen nedan kan beskrivas som fyra olika regressionslinjer. Ange *lutningen* för dom fyra olika regressionslinjerna (en för varje årstid) givet den alternativa modellen nedan.

Alternativ modell:

$$\text{nRides} = \beta_0 + \beta_1 \text{temp} + \beta_2 \text{summer} + \beta_3 \text{ fall} + \beta_4 \text{winter} + \beta_5 \text{temp} : \text{summer} + \beta_6 \text{temp} : \text{fall} + \beta_7 \text{temp} : \text{winter} + \varepsilon$$

d) Vi gör vanligtvis tre antaganden om feltermerna, sammanfattat som $\varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2)$. Vilka är dessa tre antaganden? Utifrån residualplottarna nedan, vilka antaganden verkar vara uppfyllda? (6p)



Uppgift 2 - Multikollinearitet (10 poäng)

Du funderar på att utvecka den skattade modellen i uppgift 1 genom att lägga till variabeln `windspeed`, men oroar dig för *multikollinearitet*.

a) Vad innebär multikollinearitet och varför kan det vara ett problem?

b) Ett mått som kan användas för att avgöra om en variabel har allvarliga problem med multikollinearitet är VIF. Beskriv den modell du skulle behöva skatta (med tex R) för att beräkna VIF för `windspeed`. Alltså, ange vilken variabel du skulle använda som utfallsvariabel och vilken/vilka variabler du skulle använda som förklarande variabler.

c) Vi hittar på att du kört regressionen i b-frågan och fått ett R^2 på 0.1. Beräkna VIF. Bör du oroa dig över multikollinearitet?

Uppgift 3 - Logistisk regression (25 poäng)

Datasetet CPS85 innehåller information om löner för ett slumpmässigt urval av personer från 1985. För att undersöka om det finns något samband mellan timlön och fackligt medlemsskap har en logistisk regressionsmodell anpassats med variablerna

- `wage` - lön, i USD per timme (förklarande variabel)
- `union` - fackligt medlemskap, där 1 indikerar fackligt medlemsskap och 0 avsaknad av fackligt medlemskap (utfallsvariabel)

Populationsmodellen ges av

$$\mathrm{P}(\text{union} = 1 \mid \text{wage}) = \frac{\exp(\beta_0 + \beta_1 \text{wage})}{1 + \exp(\beta_0 + \beta_1 \text{wage})}$$

```
Parameter estimates
------------------------------------------------
             Estimate Std. Error z value   Pr(>|z|)
(Intercept) -2.206957   0.232985 -9.4725 2.7314e-21
wage         0.071736   0.020055  3.5770 3.4750e-04
```

a) Beräkna den skattade sannolikheten att en person med en timlön på 10 USD *inte* är facklig medlem.

b) Tolka interceptet i termer av odds *och* i termer av sannolikheter. Är det rimligt att tolka interceptet?

c) Tolka parameterskattnigen för `wage` i termer av oddskvot.

d) Beräkna den skattade sannolikheten för fackligt medlemsskap för personer med timlöner på 15, 25 och 35 USD.

e) Av dom tre personerna i d visar det sig att endast personen med en timlön på 25 USD har ett fackligt medlemsskap. Beräkna det totala logaritmerade prediktionsvärdet (summan) för dom tre personerna. Jämför detta värde med det totala logaritmerade prediktionsvärdet för en modell som helt och hållet gissar, och alltid ger en skattad sannolikhet på 0.5.

Uppgift 4 - Ickelinjär regression (14 poäng)

Du vill använda följande modell för att undersöka hur Kinas GDP (bruttonationalprodukt, ett mått på ekonomins storlek) förändras över tid.

$$\texttt{GDP} = \alpha \cdot \beta^{\texttt{t}} \cdot \varepsilon$$

Till din hjälp har du följande variabler

- `GDP`, din utfallsvariabel, som anger Kinas GDP (i miljarder USD)
- `t`, år. Denna variabel är omskalad så att $t = 1$ motsvarar år 2000.

```
Parameter estimates
-----------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  6.56879  0.0413436 158.883 2.5961e-21
t            0.16643  0.0048556  34.276 2.4210e-13
```

a) Vad kallas sambandet som ges av populationsmodellen?

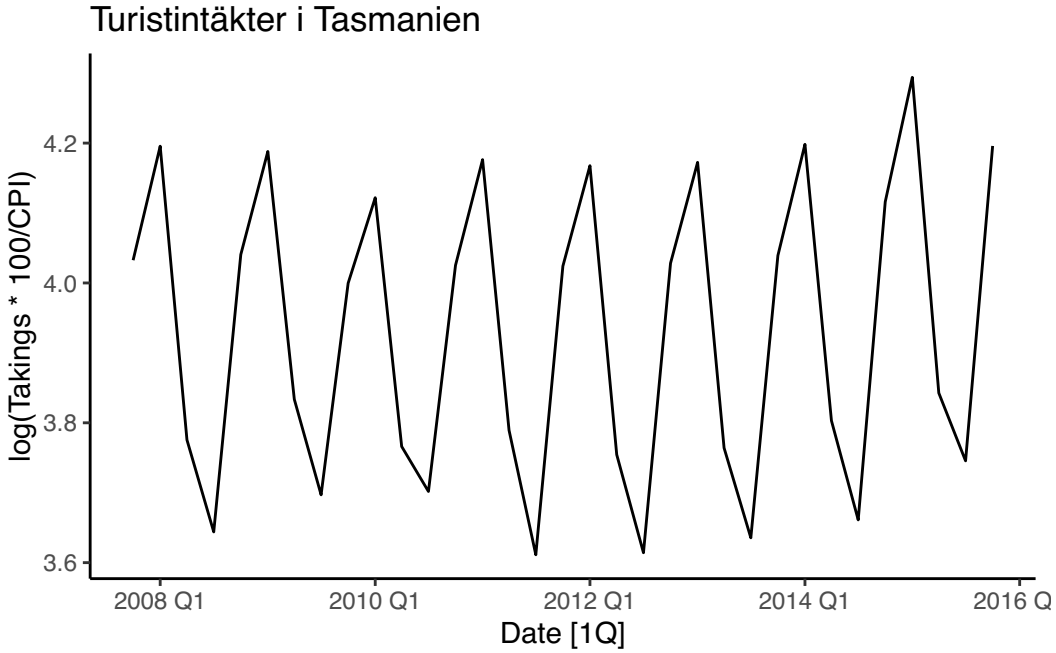
b) Skriv ut den *logaritmerade* populationsmodellen.

c) Vad är $\widehat{\texttt{GDP}}$ när $\texttt{t} = 15$? Skattningarna i R-utskriften bygger på logaritmering med basen e (den naturliga logaritmen).

d) Tolka den skattade parametern t i utskriften ovan. Tolkningen ska vara i termer av förändring av GDP, inte $\log \text{GDP}$.

Uppgift 5 - Tidsserieanalys (26 poäng)

Figuren nedan visar intäkter från turistindustrin i Tasmanien (en del av Australien) mellan 2008 och 2016. Värdena i figuren är inflationsjusterade och logartimerade.

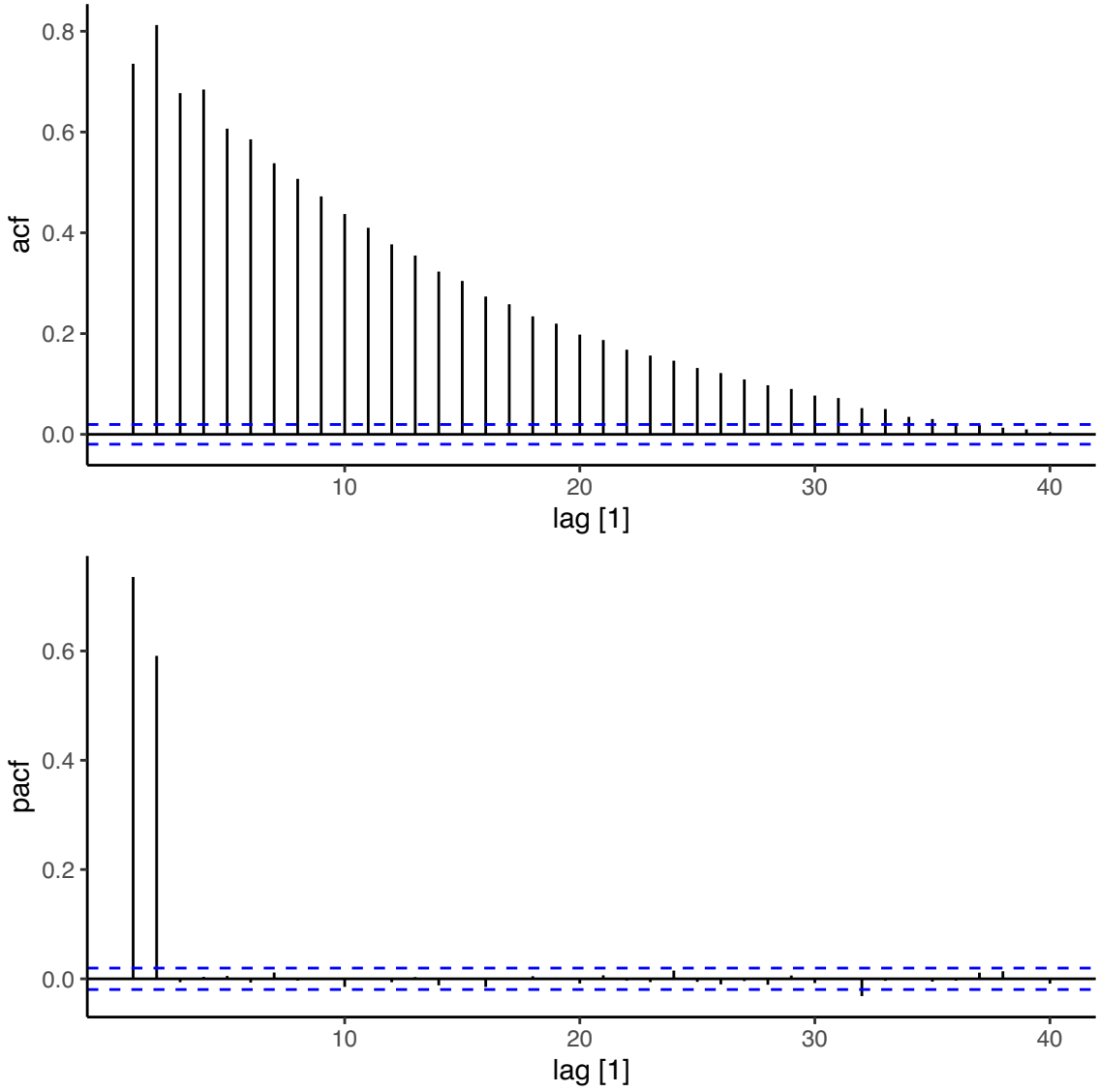



```
# A tsibble: 6 x 3 [1Q]
     Date Takings   CPI
    <qtr>   <dbl> <dbl>
1 2014 Q4    65.4  107.
2 2015 Q1    78.2  107.
3 2015 Q2    50.1  108.
4 2015 Q3    45.7  108
5 2015 Q4    72.0  108.
6 2016 Q1    84.5  108.
```

a) Med hjälp av tabellen ovan, beräkna dom logaritmerade och inflationsjusterade värdena från Q4 2014 till Q1 2016. Använd sedan glidande medelvärde med $k = 1$ för att beräkna trendskattningar för Q1 2015 till Q4 2015.

b) Förklara varför det är en dålig idé att använda glidande värde med $k = 1$ för att skatta trend i detta fallet. Vilken metod skulle passa bättre?

c) Diagrammen nedan visar den (skattade) autokorrelationsfunktionen och den (skattade) partiella autokorrelationsfunktionen för ett simulerat dataset. Verkar det finnas autokorrelation? Om autokorrelation finns, skulle du välja en AR(p) eller MA(q) modell? Glöm inte att specificera värdet på p eller q!



d) Nedan finner du skattade värden från en AR(1)-modell. Givet att $y_T = 0.8$, ta fram prediktioner för y_{T+1}, y_{T+2} och y_{T+3}.

e) När vi jobbar med AR-processer antar vi ofta att dom är *stationära*. I laboration 5 jobbade du med en klassisk *icke*-stationär AR-process: en slumpvandring (random walk på engelska). Hur ser populationsmodellen ut för en slumpvandring?

```
Parameter estimates
------------------------------------------------------------
     Estimate Std. Error z-ratio Pr(>|z|)     2.5 %   97.5 %
ar1  0.298347  0.0095436 31.2613  0.00000  0.279641 0.317052
mean 0.017057  0.0142812  1.1944  0.23232 -0.010934 0.045049
```