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 183.74 1.1735e-88
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?

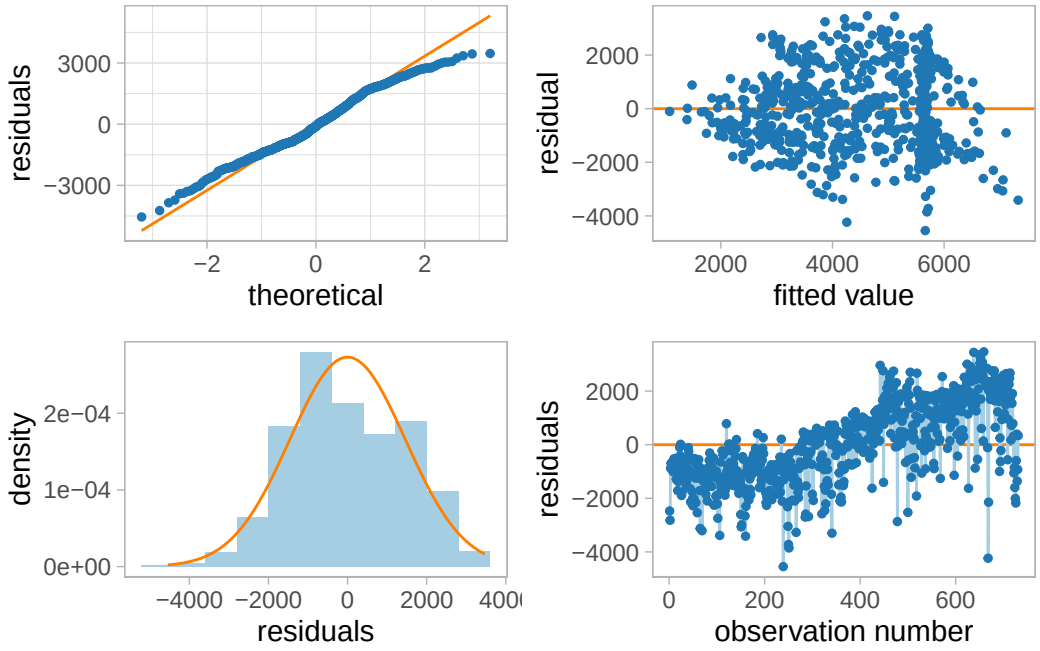
b) Testa om modellen som helhet är signifikant. Använd $\alpha = 0.01$.

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 $\varepsilon \overset{iid}{\sim} \text{N}(0, \sigma_\varepsilon^2)$. Vilka är dessa tre antaganden? Utifrån residualplottarna nedan, vilka antaganden verkar vara uppfyllda? (6p)



Lösningsförslag - Uppgift 1

1a (6p)

Interceptet: det skattade förväntade antalet uthyrningar när det inte är sommar (`summer = 0`) och temperaturen är 0 Celsium (`temp = 0`) är 1081.85. Det är helt rimligt att tolka intercepetet. Datasetet innehåller flera datapunkter när det inte är sommar (eftersom att vi har 731 observationer för 2011-2012). Det skulle kunna vara problematiskt att tolka om 0 grader celsius var en exceptionellt låg temperatur som aldrig händer i Washington, men 0 grader kan absolut hända på vintern.

`temp`: den skattade förväntade ökningen av antalet uthyrningar när temperaturen ökar med en grad C, och det *inte* är sommar, är 191.

`summer`: den skattade förväntade skillnaden i antalet uthyrningar mellan en sommardag och en dag som inte är sommar, givet att temperaturen är 0 grader celsius, är 5000.

Interaktionseffekt: Den skattade förväntade skillnaden i förändringen av antalet uthyrningar när temperaturen ökar med en grad mellan sommar och inte sommar är -200. Det är även OK att tolka denna som att den skattade förväntade minskningen av antalet uthyrningar när temperaturen ökar med en grad C och det är sommar är $191.14 - 206.64 = -15.5$.

1b (8p)

Vi börjar med att ställa upp noll- och alternativhypotes

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \beta_3 = 0 \\ H_A: &\quad \text{Åtminstone en } \beta_j \neq 0, \text{ där } j = 1, 2, 3 \end{aligned}$$

$k = 3$, $n = 731$

$F_{crit} = F_{0.01}(3, 731 - 3 - 1) = F_{0.01}(3, 727) \approx F_{0.01}(3, 1000) = 3.80$. (Det är OK att använda approximationen $F_{0.01}(3, 400)$ istället.)

Vi förkastar nollhypotesen om $F_{obs} > F_{crit}$, där

$$F_{obs} = \frac{MSR}{MSE} = \frac{393794251}{2143264} = 183.7358$$

Då $183.74 > 3.80$ så förkastar vi nollhypotesen. Alltså lyckas dom förklarande variablerna förklara någon del av variationen av utfallsvariabeln. (Alternativt, vår modell är bättre än en modell som endast innehåller ett intercept.)

1c (5p)

- Lutningen för vår är β_1
- Lutningen för sommar är $\beta_1 + \beta_5$
- Lutningen för höst (fall) är $\beta_1 + \beta_6$
- Lutningen för vinter (winter) är $\beta_1 + \beta_7$

1d (6p)

Antagandena vi gör är oberoende, normalitet och homoskedasticitet (samma varians).

Utifrån dom två figureren till vänster ser det ut som att normalitetsantagandet inte är uppfyllt. Resdiualerna har för tjocka svansar.

Utifrån figuren uppe till höger, dvs fitted value vs residual, ser vi tydliga tecken på heteroscedasticitet. Variansen har ett tydligt trattformat utseende.

Utifrån figuren nere till höger, observation number vs residuals, ser vi tydliga tecken på att residualerna *inte* är oberoende över tid. Nästan alla residualer i början är negativa och alla i slutet positiva.

Uppgift 2 - Multikollinearitet (10 poäng)

Du funderar på att utveckla 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?

Lösningsförslag - Uppgift 2

2a (3p)

Multikollinearitet är linjärt beroende mellan dom förklarande variablerna. Vi kan även förklara/förstå multikollinearitet som att dom olika förklarande variablerna (delvis) förklarar samma sak.

Problem som uppstår vid multikollinearitet är att det är svårt att separerar dom olika förklarande variablernas effekt på y, stora standardfel för dom förklarande varaiblerna, samt problem med insignifikans för dom förklarande variablerna.

Perfekt multikollinearitet, vilket betyder att en av dom förklarande variablerna kan beskrivas som en linjär funktion av dom andra förklarande variablerna, är mycket problematiskt och innebär att det inte går att skatta någon modell.

(Svaret ovan är väldigt utförligt för en 3-poängs-fråga, och svar som utelämnar multikollinearitet, samt endast anger några utav problemen i andra stycket kan också få full poäng.)

2b (5p)

Modellen vi behöver skattat kan beskrivas med populationsmodellen

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

Det är såklart även ok att ange svaret som - utfallsvariabel: `windspeed` - förklarande variabler: `temp`, `summer`, och interaktionen mellan `temp` och `summer`.

Det är helt OK att misstolka frågan och utgå ifrån den mer komplicerade modellen i 1a, men då ska det vara korrekt utifrån den modellen.

2c (2p)

$$\text{VIF} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.1} = \frac{1}{0.9} \approx 1.11$$

Ett VIF-värde på 1.11 är inte högt (tumregel 10) och vi behöver inte oroa oss för 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.

Lösningsförslag - Uppgift 3

3a (4p)

$$\mathrm{P}(\text{union} = 0 \mid \text{wage} = 10) = 1 - \mathrm{P}(\text{union} = 1 \mid \text{wage} = 10)$$

Enklast att använda

$$\mathrm{P}(\text{union} = 0 \mid \text{wage} = 10) = \frac{1}{1 + \exp(-2.206957 + 0.071736 \cdot 10)} = 0.8160178$$

3b (5p)

Vi tolkar interceptet genom att sätta `wage = 0`.

Den skattade sannolikheten att en person med en lön på 0 USD i timmen är med i facket är 0.0991274.

Det skattade oddset att en person med en lön på 0 USD i timmen är med i facket är 0.1100349.

Notera att om du skrivit "skattade förväntade" så kommer du få poängavdrag. Det finns ingen felterm i en logistisk regressionsmodell, och vi kan därför inte tala om en förväntad sannolikhet. Däremot är den fortfarande skattad! (Och inte den "faktiska" sannolikheten, vad det nu betyder.)

3c (5p)

Den skattade oddsration för `wage` är $\exp(0.071736) = 1.074372$. För en ökning av `wage` med 1 USD/timme är skattar vi att oddsration ökar med en faktor 1.074372, vilket motsvarar en ökning med 7 procent.

3d (4p)

15 USD ger en sannolikhet på 0.2439917.

25 USD ger en sannolikhet på 0.3980592.

35 USD ger en sannolikhet på 0.5753714.

3e (7p)

Prediktionsvärdena ges av

$0.2439917 \cdot 0 + (1 - 0.2439917) \cdot (1 - 0) = 0.7560083$

$0.3980592 \cdot 1 + (1 - 0.3980592) \cdot (1 - 1) = 0.3980592$

$0.5753714 \cdot 0 + (1 - 0.5753714) \cdot (1 - 0) = 0.4246286$

Det totala logaritmerade prediktionsvärdet ges alltså av

$$\log 0.7560083 + \log 0.3980592 + \log 0.4246286 = -2.057398$$

En modell som alltid ger sannolikheten 0.5 kommer alltid få prediktionsvärdet 0.5 för varje observation, och alltså totalt logaritmerat prediktionsvärde i det här fallet på

$$3 \cdot \log 0.5 = -2.079442$$

Vår modell har alltså gjort bättre prediktioner är en modell som bara gissar!

Kommentar: inga avdrag för att använda bas 10 i logaritmeringen. (Men du kommer ha fått något annorlunda svar än ovan såklart.)

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 GDP.

Lösningsförslag - Uppgift 4

4a (2p)

Exponentiellt samband.

4b (3p)

$$\log \mathrm{GDP} = \log \alpha + \log \beta \cdot \mathrm{t} + \log \varepsilon$$

4c (5p)

$b = \exp(0.16643) = 1.181081$ och $a = \exp(6.56879) = 712.5072$

Detta ger det skattade värdet $712.5072 \cdot 1.181081^{1}5 = 8649.371$

Alternativ

$$6.56879 + 0.16643 \cdot 15 = 9.06524$$

vilket ger skattningen $\exp(9.06524) = 8649.355$

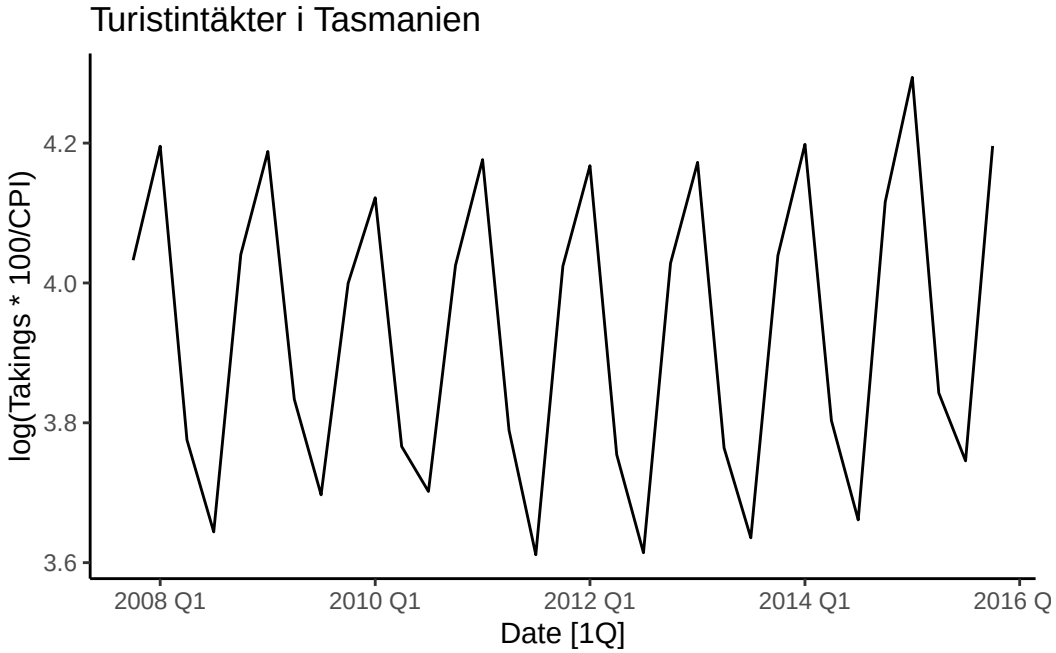
4d (4p)

Den skattade förväntade ökningen av GDP är cirka 16.6 procent per år. (Också OK: den skattade förväntade ökningen av GDP när t ökar med en enhet är 16.6 procent.)

Notera: När vi tolkar parameterskattningar för ett exponentiellt samband så är tolkningen baserad på approximationen $\exp(a) - 1 \approx a$. Denna approximation är endast bra när a ligger nära noll. Som vi ser ovan så är den faktiska förväntade ökningen per år i själva verket drygt 18 procent. Båda den exakta tolkningen och approximationen är OK!

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.

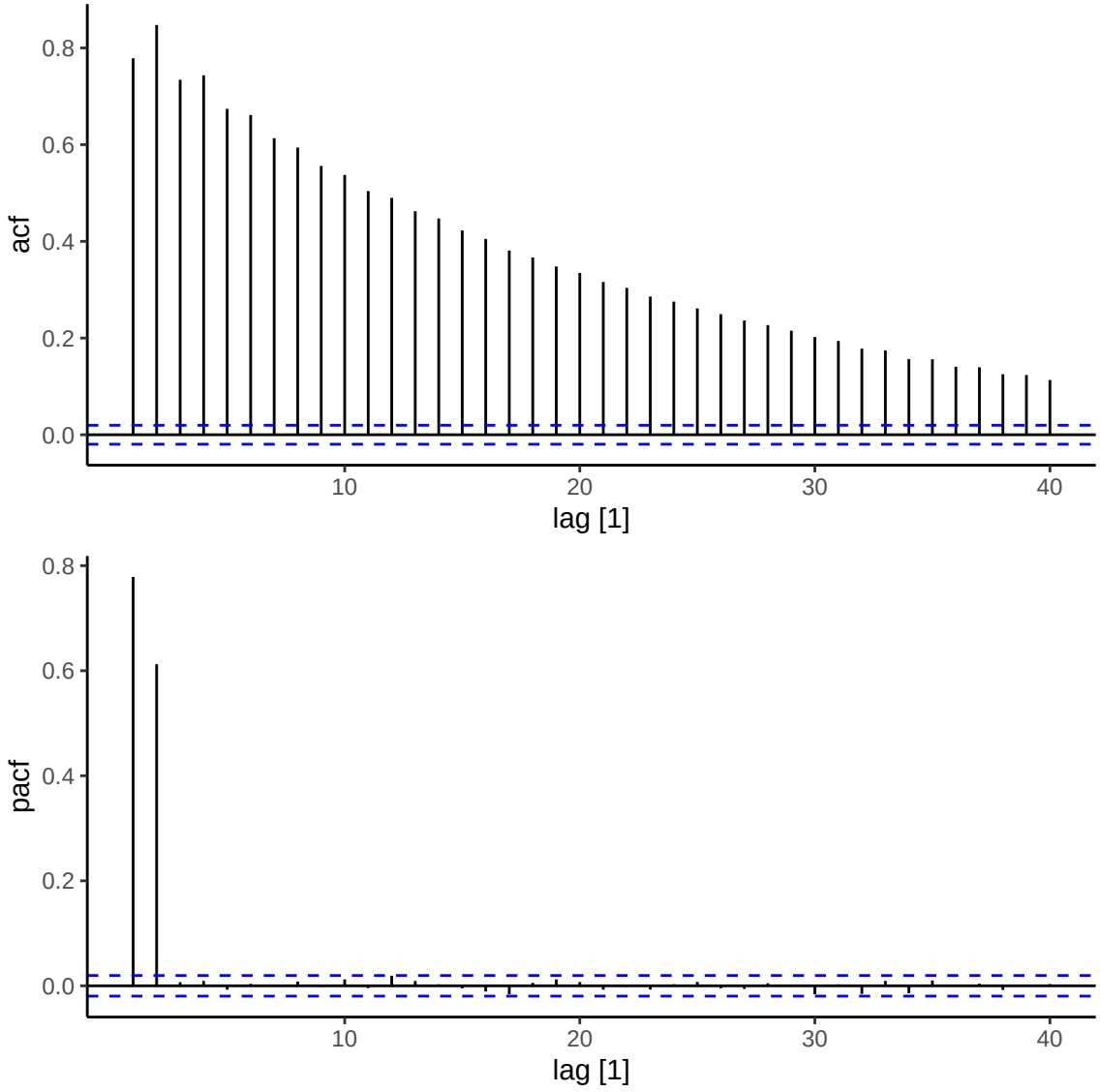



```
# A tsibble: 6 x 3 [1Q]
     Date Takings   CPI
    <qtr>   <dbl> <dbl>
1 2014 Q4    65.4  107.
2 2015 Q1    78.2  107.
3 2015 Q2    50.1  108.
4 2015 Q3    45.7  108
5 2015 Q4    72.0  108.
6 2016 Q1    84.5  108.
```

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
```

Lösningsförslag - Uppgift 5

5a (7p)

```
# A tsibble: 6 x 4 [1Q]
     Date Takings   CPI `log(takings*100/CPI)`
    <qtr>   <dbl> <dbl>                  <dbl>
1 2014 Q4    65.4  107.                   4.12
2 2015 Q1    78.2  107.                   4.29
3 2015 Q2    50.1  108.                   3.84
4 2015 Q3    45.7  108                    3.75
5 2015 Q4    72.0  108.                   4.20
6 2016 Q1    84.5  108.                   4.36
```

Dom fyra skattade medelvärden är alltså

2015Q1: 4.083333

2015Q2: 3.96

2015Q3: 3.93

2015Q4: 4.103333

5b (4p)

Det är dåligt eftersom att vi har kvartalsdata med tydlig säsongseffekt. Poängen med att använda ett glidande medelvärde är att ta bort säsongseffekten när vi beräknar trenden. För att göra detta måste vi använda en metod som ger samma vikt till alla kvartal, vilket glidande medelvärde med $k = 1$ inte gör. (Exempelvis, medelvärdet för Q1 2015 är baserat på Q4 2014, Q1 2015 och Q2 2015. Den innehåller alltså inte någon Q3!).

5c (4p)

ACF avatar långsamt, och PACF har två tydliga spikes. Detta är en AR(2).

5d (6p)

Första steget är att beräkna $\hat{\beta}_0$ med hjälp av `mean` och `ar1`.

$\hat{\beta}_0 = \text{mean} \cdot (1 - \text{ar1}) = 0.017057 \cdot (1 - 0.298347) = 0.0119681.$

Vi kan nu göra våra prognoser

$$\hat{y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 y_T = 0.0119681 + 0.298347 \cdot 0.8 = 0.2506457$$

För prognosen två steg frammåt behöver vi använda vår progons ett steg frammåt

$$\hat{y}_{T+2|T} = \hat{\beta}_0 + \hat{\beta}_1 \hat{y}_{T+1|T} = 0.0119681 + 0.29481160 \cdot 0.2506457 = 0.08674749$$

Slutligen, tre steg framåt

$$\hat{y}_{T+3|T} = \hat{\beta}_0 + \hat{\beta}_1 \hat{y}_{T+2|T} = 0.0119681 + 0.29481160 \cdot 0.08674749 = 0.03784895$$

5e (5p)

$$y_t = \beta_0 + 1 \cdot y_{t-1} + \varepsilon_t$$