SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2024-01-17

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 (20 poäng)

Datasetet `earnings` innehåller information om lön, utbildning, erfarenhet, längd och kön för 1605 slumpmässigt utvalda personer.

- `earnk` Årslön i tusental dollar.
- `educ` Utbildningsnivå i år. Antar värden mellan 2 och 18 i datasetet. Den genomsnittliga utbildningsnivån i datasetet är 13 år.
- `height` Längd i tum (inches). Antar värden mellan 57 och 82 i det här datasetet. Den genomsnittliga längden i datasetet är 70 tum för män och 64 tum för kvinnor.
- `male` Variabel som i detta datasetet antar två värden: 1 för män, 0 för kvinnor.

```
Analysis of variance - ANOVA
------------------------------------------------
        df     SS       MS      F     Pr(>F)
Regr     4 132030 33007.60 76.273 2.9775e-59
Error 1600 692405   432.75
Total 1604 824436
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.17	16.84	-0.54	0.59
height	-0.10	0.26	-0.38	0.70
male	-43.57	25.68	-1.70	0.09
education	2.58	0.20	12.68	0.00
height:male	0.80	0.38	NA	NA

a) Skiljer sig *sambandet* mellan längd (height) och inkomst åt mellan män och kvinnor? Genomför ett t-test. Använd $\alpha = 0.05$. (6p)

b) Tolka den skattade regressionskoefficienten för `male` (-43.57). (3p)

c) Använd den genomsnittliga utbildningsnivån och dom två genomsnittliga längderna i variabelbeskrivningen för att beräkna hur stor den skattade genomsnittliga löneskillnaden är mellan en genomsnittlig man och en genomsnittlig kvinna. (5p)

d) Ett av dom tre antagandena vi gör om feltermernas fördelning kan testas med White's test. Vilket antagande? För att genomföra White's test behöver en regressionsmodell skattas. Vad används som *responsvariabel* i den regressionsmodellen? (6p)

Lösningsförslag - Uppgift 1

1a (6p)

Det vi ska göra här är ett t-test av β_4, alltså själva *interaktionseffekten*, eftersom att det är lutningen som beskriver hur sambandet mellan `earnk` och `height` ser ut och interaktionseffekten beskriver hur lutningen skiljer sig mellan män och kvinnor.

Vi börjar med att ställa upp noll- och alternativhypotes

$$H_0 : \beta_4 = 0, \quad H_a : \beta_4 \neq 0$$

(Det är OK att kalla parametern något annat så länge det är tydligt vad du testar, tex $\beta_{interaction}$ eller $\beta_{height \times male}$).

t-värdet ges av

$$t_{obs} = \frac{b_4 - 0}{s_{b_4}} = \frac{0.80}{0.38} \approx 2.11$$

$$t_{crit} = t_{0.025,1605-4-1} \approx 1.962$$

Här har jag använt t-fördelningen med 1000 frihetsgrader. Det är helt OK att normalapproximera, då värdet blir 1.96. Vi kan se att det inte kommer spela någon roll (vi kommer förkasta oavsett).

Eftersom att $|t_{obs}| > t_{crit}$ förkastar vi nollhypotesen. Sambandet mellan lön och längd skiljer sig åt mellan män och kvinnor.

1b (3p)

Den skattade förväntade skillnaden i årslön mellan män och kvinnor som är 0 tum långa, givet att övriga prediktorer hålls konstanta.

1c (5p)

Den skattade förväntade årslönen för en "genomsnittlig man" är

$$-9.17 + (-0.10 + 0.80) \cdot 70 - 43.57 \cdot 1 + 13 \cdot 2.58 = 29.8$$

Den skattade förväntade årslönen för en "genomsnittlig kvinna" är

$$-9.17 + (-0.10) \cdot 64 - 43.57 \cdot 0 + 13 \cdot 2.58 = 17.97$$

Löneskillnaden mellan dessa är $29.8 - 17.97 = 11.83$

1d (6p)

White's test används för att testa om antagandet om *homoscedasticitet* är uppfyllt. Responsvariabeln i regressionsmodellen vi behöver skatta för att genomföra White's test är dom kvadrerade residualerna.

Uppgift 2 - Multipel regression (15 poäng)

En alternativ modell som innehåller två till prediktorer, vikt (`weight`) och ålder (`age`) har skattats baserat på samma dataset som i uppgift 1.

```
Analysis of variance - ANOVA
------------------------------------------------
        df     SS       MS      F     Pr(>F)
Regr     6 148216 24702.68 58.376 1.7823e-65
Error 1598 676220   423.17
Total 1604 824436

Parameter estimates
------------------------------------------------
               Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -31.9792573  17.131118 -1.86673 6.2122e-02
height        0.1012198   0.269008  0.37627 7.0677e-01
male        -38.4296935  25.461930 -1.50930 1.3142e-01
education     2.7548779   0.203623 13.52928 1.4492e-39
age           0.1886192   0.030787  6.12664 1.1280e-09
weight       -0.0054024   0.018763 -0.28793 7.7344e-01
height:male   0.7166586   0.376889  1.90151 5.7415e-02
```

a) Jämför dom två modellerna med ett F-test. Använd $\alpha = 0.01$. (10p)

b) Vad är skillnaden mellan en outlier (uteliggare) och en inflytelserik observation? (5p)

Lösningsförslag - Uppgift 2

2a (10p)

Den fulla modellen ges av

$$\texttt{earnk} = \beta_0 + \beta_1\texttt{height} + \beta_2\texttt{male} + \beta_3\texttt{education} + \beta_4\texttt{height:male} + \beta_5\texttt{age} + \beta_6\texttt{weight} + \varepsilon$$

Vår reducerade modell inkluderar varken `age` eller `weight`, vilket ger oss följande noll- och alternativhypotes

$$\begin{aligned} H_0 &: \beta_5 = \beta_6 = 0 \\ H_A &: \beta_5 \neq 0 \text{ och/eller } \beta_6 = 0 \end{aligned}$$

Vår teststatistika är

$$F = \frac{(R^2_{FM} - R^2_{RM})/(p-q)}{(1-R^2_{FM})/(n-p-1)}$$

För att kunna beräkna det observerade värdet behöver vi ta fram förklaringsgraden för dom två modellerna. $R^2_{FM} = 148216/824436 = 0.1797787$, $R^2_{RM} = 132030/824436 = 0.1601458$. Vi har 1605 observationer, så $n = 1605$, $p = 6$ och $q = 4$.

$$F_{obs} = \frac{(0.1797787 - 0.1601458)/(6-4)}{(1-0.1797787)/(1605-6-1)} = 19.38824.$$

Slutligen behöver vi ta fram F_{crit}, där vi har 2 respektive 1598 frihetsgrader. Tabellen går bara till 1000, så vi får nöja oss med att $F_{crit} = F_{\alpha=0.01}(2, 1598) \approx 4.63$.

Eftersom att $F_{obs} > F_{crit}$ så förkastar vi nollhypotesen.

2b (5p)

En outlier är en observation som är extrem på något sätt. Det kan vara ett extremt värde på responsvariabeln eller någon av prediktorerna.

En inflytelserik observation är en observation som påverkar skattningarn i modellen mycket. Exempelvis så kan det hända att en skattning går från signifikant till icke-signifikant när vi tar bort en specifik observation. Den observationen kallas i så fall en inflytelserik observation.

Uteliggare kan ofta vara inflytelserika, men så måste inte vara fallet. En observation kan vara inflytelserik utan att vara en uppenbar outlier, men ofta sticker inflytelserika observationer ut på något sätt.

Uppgift 3 - Logistisk regression (22 poäng)

Det amerikanske presidentvalet 1992 stod mellan Bill Clinton (demokrat) och George Bush (republikan). En modell har skattats för att undersöka sambandet mellan inkomst, ålder och vilken av dom två kandidaterna en person röstade på. Variablerna i modellen beskrivs nedan.

- `rvote`: dummyvariabel som antar värdet 1 om personen röstade på Bush och 0 om personen röstade på Clinton. Personer som röstade på någon annan kandidat har sorterats bort.
- `income`: månadsinkomst i **tusentals** dollar. Om personen tjänar 2000 dollar ska alltså `income` vara 2, om personen tjänar 1500 dollar ska `income` vara 1.5 och så vidare.
- `age`: ålder.

`Parameter estimates`

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.80	0.27	-6.57	0.00
income	0.29	0.06	5.08	0.00
age	0.01	0.00	1.76	0.08

a) Dina kusiner Orvar och Chrissy röstade i valet. Orvar hade då en inkomst på 2000 dollar och var 23 år gammal, Chrissy var 51 och hade en inkomst på 4000 dollar. Beräkna dom skattade sannolikheterna att Orvar respektive Chrissy röstade för Bush. (4p)

b) Tolka interceptet i termer av odds *och* i termer av sannolikheter. Är det rimligt att tolka interceptet? Tolka parameterskattnigarna för `income` och `age` i termer av oddskvoter. (6p)

c) Det visar sig att Orvar röstade för Clinton och att Chrissy röstade för Bush. Beräkna summan av dom logaritmerade prediktionsvärdena för dom två skattade sannolikheterna i del a). I termer av logaritmerade prediktionsvärden, är modellen du använt bättre eller sämre än en modell som chansar vilt och alltid ger den skattade sannolikheten 0.5? (6p)

d) Din kompis har beräknat att sannolikheten för en viss person med en inkomst på 4000 dollar att rösta på Bush är 0.50, men vägrar berätta hur gammal personen är. Beräkna åldern på personen. (6p)

Lösningsförslag - Uppgift 3

3a (4p)

Den skattade sannolikheten för Orvar blir

$$\mathrm{P}(\text{rvote} = 1 \mid \text{income} = 2, \texttt{age} = 23) = \frac{\exp(-1.80 + 0.29 \cdot 2 + 0.01 \cdot 23)}{1 + \exp(-1.80 + 0.29 \cdot 2 + 0.01 \cdot 23)} = 0.2709121 \approx 0.27$$

Den skattade sannolikheten för Chrissy blir

$$\mathrm{P}(\text{rvote} = 1 \mid \text{income} = 4, \texttt{age} = 51) = \frac{\exp(-1.80 + 0.29 \cdot 4 + 0.01 \cdot 51)}{1 + \exp(-1.80 + 0.29 \cdot 4 + 0.01 \cdot 51)} = 0.4675457 \approx= 0.47$$

3b (6p)

Vi tolkar interceptet genom att sätta `income = 0` och `age = 0`.

Det skattade oddset att en nyfödd person utan inkomst röstar på Bush är $\exp(-1.80) \approx 0.17$.

Den skattade sannolikheten att en nyfödd person utan inkomst röstar på Bush är $\frac{\exp(-1.80)}{1+\exp(-1.80)} \approx 0.14$.

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.)

I detta fall verkar det inte rimligt att tolka interceptet, då nyfödda inte får rösta. Om dom nu skulle släppas lös i ett röstbås så skulle jag tro att det är 50/50 vilken kandidat dom röstar på.

Den skattade oddsration för `income` är $\exp(0.29) = 1.336427$. Vår skattning är att varje ytterligare 1000 dollar i inkomst ökar oddsration med en faktor 1.34, vilket motsvarar en ökning med 34 procent.

Den skattade oddsration för `age` är $\exp(0.01) = 1.01005$. Vår skattning är att oddsration med en faktor 1.01 för varje ytterligare år, vilket motsvarar en ökning med 1 procent.

3c (6p)

För Orvar blir prediktionsvärdet $1 - 0.27 = 0.73$ och för Chrissy 0.47. Summan av dom två logaritmerade prediktionsvärdena blir alltså $\log(0.73) + \log(0.47) = -1.069733$.

En modell som bara gissar får alltid ett prediktionsvärde på 0.5 (eftersom att $P(y = 1) = P(y = 0) = 0.5$ så spelar inte det faktiska utfallet någon roll), så summan av dom två chansningarnas logaritmerade prediktionsvärden blir -1.386294. Vår modell har presterat bättre än en modell som helt chansar!

3d (6p)

Det är enklast att räkna om sannolikheten till odds först. Oddset för personen är $05/05 = 1$. Vi kan nu ställa upp

$$\exp(-1.80 + 0.29 \cdot 4 + 0.01 \cdot x) = 1$$

Vi logaritmerar sen båda sidorna och löser ut x

$$-1.80 + 0.29 \cdot 4 + 0.01 \cdot x = \log(1)$$

$$x = \frac{\log(1) + 1.80 - 0.29 \cdot 4}{0.01} = 64$$

Svar: personen som din kompis tagit fram sannolikheten för är 64 år.

Uppgift 4 - Ickelinjär regression (18 poäng)

Populationsmodellen för utvecklingen av antal bakterier (`bact`) över tid ges av

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

där tiden `t` anges i timmar.

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
```

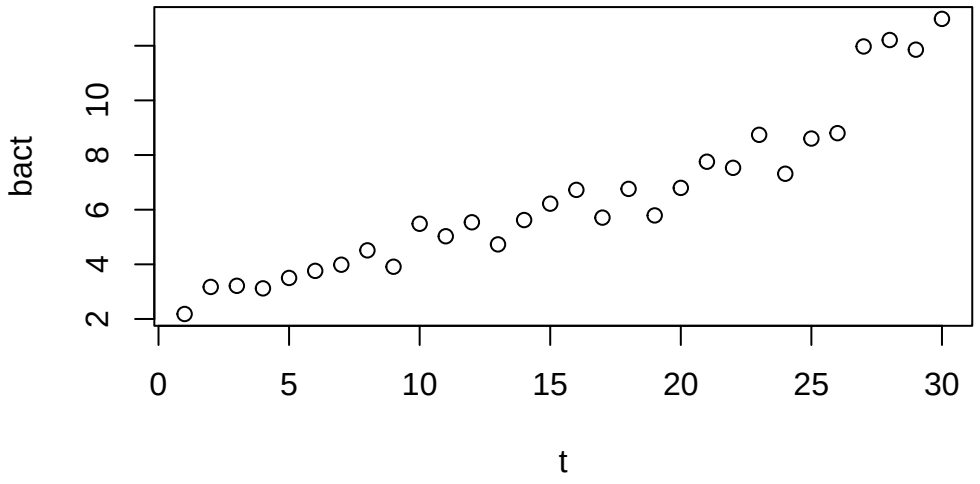
```
aa
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.994      0.042  23.807        0
t              0.050      0.002  21.205        0
```

c) Enligt den skattade modellen, hur många bakterier förväntas det finnas efter `t` = 12 timmar? 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-bakterier.) (5p)

e) Baserat på plotten nedan, nämn en annan lämplig modell du skulle kunna använda för att beskriva sambandet mellan `t` och `bact`. (Hint: inte vanlig linjär regression!) (3p)



Lösningsförslag - Uppgift 4

4a (2p)

Ett exponentiellt samband.

4b (3p)

$$\log \texttt{bact} = \log \alpha + \texttt{t} \log \beta + \log \varepsilon$$

4c (5p)

Vi kan först räkna ut värdet på logskala.

$$\widehat{\texttt{log\_bact}} = 0.994 + 0.05 \cdot 12 = 1.594$$

Det skattade förväntade antalet bakterier är alltså $\exp(1.594) = 4.923403$.

4d (5p)

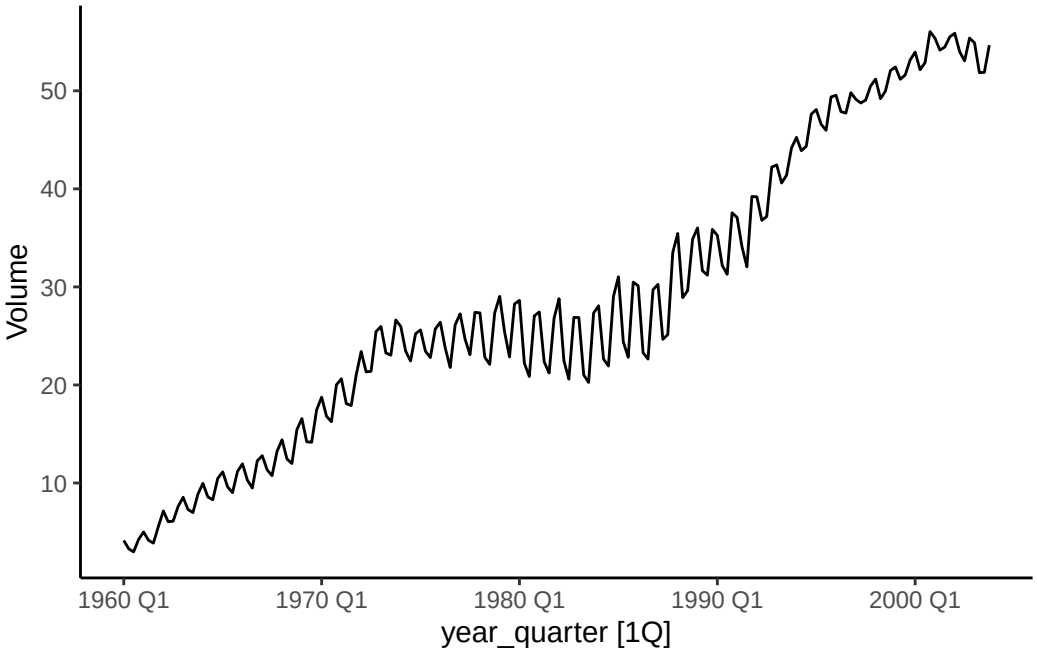
Vi börjar med att ta fram $\widehat{\beta}$: $\exp(0.05) = 1.051271$. Alltså, när `t` ökar en enhet så är den skattade förväntade ökningen av antalet bakterier fem procent.

4e (3p)

Polynomregression skulle kunna funka.

Uppgift 5 - Tidsserieanalys (25 poäng)

Figuren nedan visar den kvartalsvisa produktionen av naturgas i Kanada (i miljarder kubikmeter).



	year_quarter	Volume
1	1960 Q3	2.97
2	1960 Q4	4.26
3	1961 Q1	5.01
4	1961 Q2	4.17
5	1961 Q3	3.87
6	1961 Q4	5.57
7	1962 Q1	7.15
8	1962 Q2	6.05

a) Om du skulle göra en klassisk dekomponering av tidsserien ovan, skulle du föredra en additiv eller multiplikativ modell? Skulle den innehålla trend och/eller säsong? Motivera dina svar. (3p)

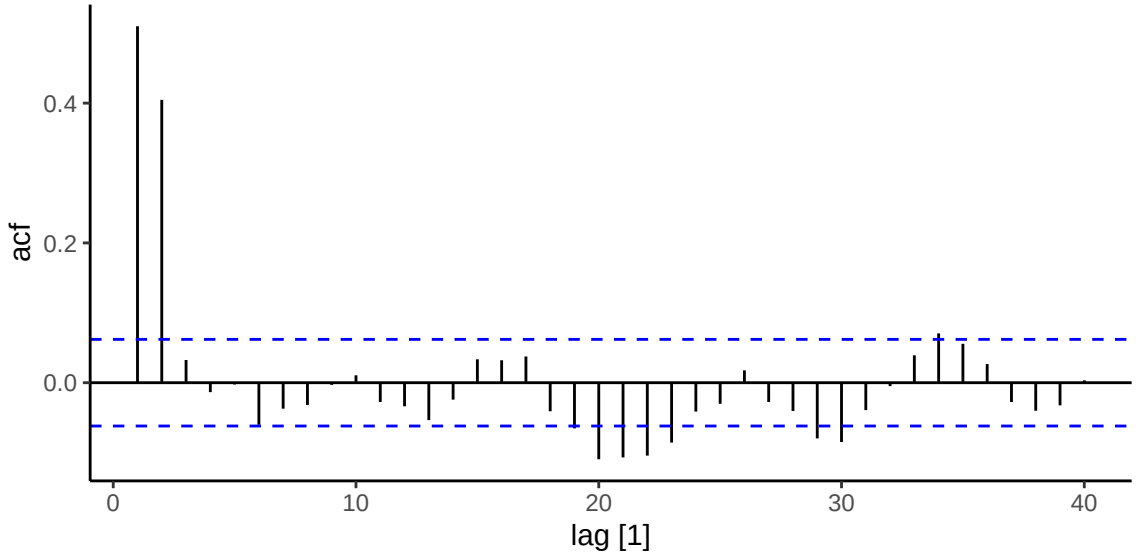
b) Beräkna trendskattningar för hela år 1961 baserat på ditt svar på a). Givet att trendskattningen för första kvartalet 1962 är 5.941762, ge en grov approximering av säsongskomponenten för det första kvartalet. (8p)

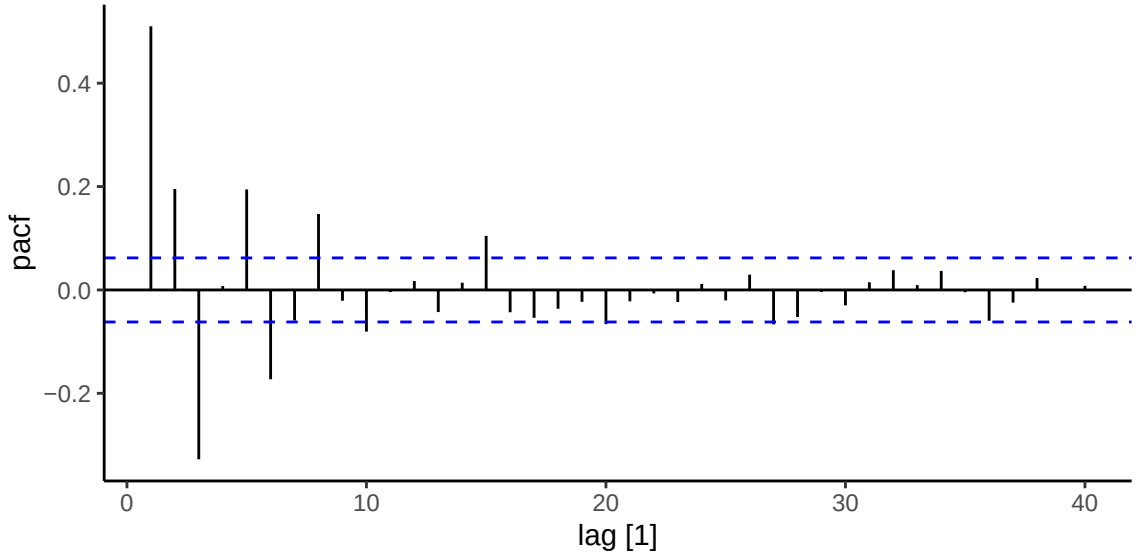
c) Utifrån figuren ovan, finns det något problem med att använda klassisk dekomponering? (4p)

d) Nedan finner du skattade värden från en AR(1)-modell. Givet att $y_T = 0.8$, ta fram prediktioner för y_{T+1} och y_{T+2}. (5p)

	Estimate	Std. Error	z-ratio	Pr(>|z|)	2.5 %	97.5 %
ar1	0.40	0.01	43.67	0	0.38	0.42
mean	7.19	0.02	428.57	0	7.15	7.22

e) 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! (5p)





Lösningsförslag - Uppgift 5

5a (3p)

```
 [1] -0.48071123  2.69719560  0.49883437 -1.17805096  0.15759009  1.54931665
 [7] -0.09738848 -0.05865022 -2.18358036 -1.03718051
```

Det ser ut att finnas både trend och säsong. Om vi tittar på första halvan av tiddserien ser det ut som att effekten är multiplikativ (svängningarna blir större när nivån ökar), men mot slutet blir säsongseffekten *mindre* igen vilket går imot det multiplikativa antagandet.

För att kunna lösa nästa fråga måste vi dock utgå ifrån något.

5b (8p)

Vi utgår ifrån en additiv modell (multiplikativ är också OK, men då ska du ha logaritmerat först).

1961, Q1

$$\frac{2.97 + 2 \cdot 4.26 + 2 \cdot 5.01 + 2 \cdot 4.17 + 3.87}{8} = 4.215$$

1961, Q2

$$\frac{4.26 + 2 \cdot 5.01 + 2 \cdot 4.17 + 2 \cdot 3.87 + 5.57}{8} = 4.49125$$

1961, Q3

$$\frac{5.01 + 2 \cdot 4.17 + 2 \cdot 3.87 + 2 \cdot 5.57 + 7.15}{8} = 4.9225$$

1961, Q4

$$\frac{4.17 + 2 \cdot 3.87 + 2 \cdot 5.57 + 2 \cdot 7.15 + 6.05}{8} = 5.425$$

För att få en grov uppskattning av säsongkomponenten behöver vi först avtrenda data för dom två Q1 vi har tillgång till (1961 och 1962). Detta ger oss $S_1 + R_{1961Q1}$ och $S_1 + R_{1962Q1}$. Tar vi medelvärdet av dessa värden får vi en grov skattning av S_1.

$5.01 - 4.215 = 0.795$ respektive $7.15 - 5.942 = 1.208$. Medelvärdet av dessa är $(0.795 + 1.208)/2 = 1.0015$.

5c (4p)

Ett problem är att klassisk dekomponering antar att säsongseffekten antingen är konstant över tid (additiv modell) eller att den beror på tidsseriends nivå (trenden). I det här fallet är säsongseffekten som störst i mitten av tidsperioden. Den är alltså inte konstant, men den ökar inte heller med nivån, då säsongseffekterna verkar vara som minst i börjar *och* slutet av tidsserien.

5d (5p)

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

$\hat{c} = \text{mean} \cdot (1 - \text{ar1}) = 7.19 \cdot (1 - 0.40) = 4.314$.

Vi kan nu göra våra prognoser

$$\hat{y}_{T+1|T} = \hat{c} + \hat{\phi}_1 y_T = 4.314 + 0.4 \cdot 0.8 = 4.634$$

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

$$\hat{y}_{T+2|T} = \hat{c} + \hat{\phi}_1 \hat{y}_{T+1|T} = 4.314 + 0.4 \cdot 4.634 = 6.1676$$

5e (5p)

Två toppar på acf och exponentiellt avtagande pacf betyder att det här är en MA(2).