SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2023-12-06

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 (25 poäng)

Datasetet `uswages` innehåller information om lön, utbildning och erfarenhet för 2000 slumpmässigt utvalda personer från en undersökning som genomfördes 1988. Några av variablerna i datasetet är:

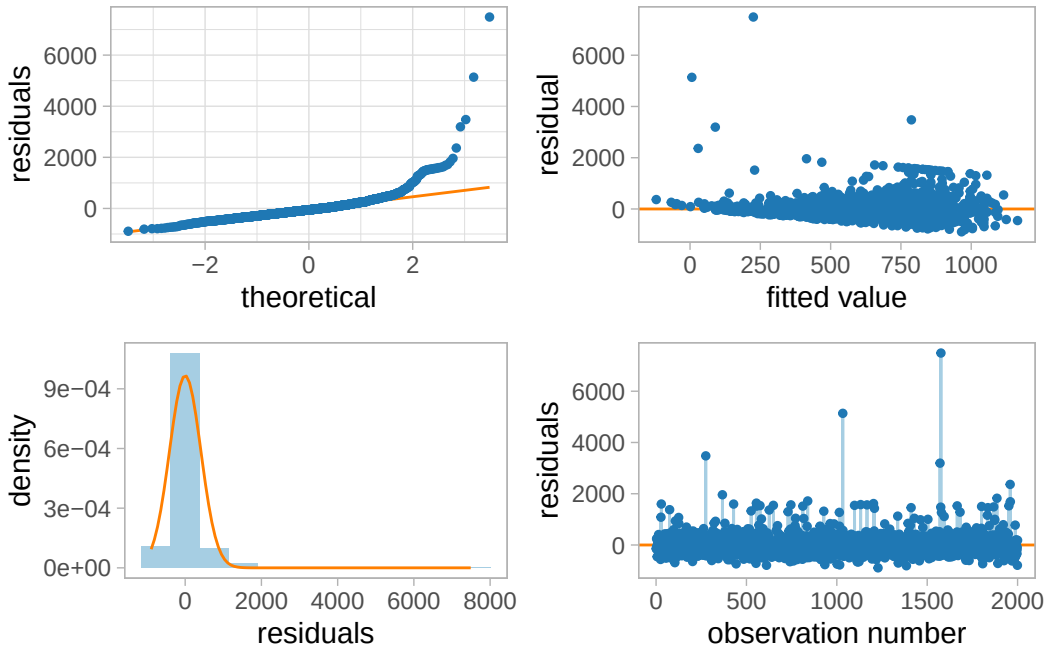
- `wage` Veckolön i dollar.
- `educ` Utbildningsnivå i år. Antar värden mellan 0 och 18 i det här datasetet.
- `exper` Arbetslivserfarenhet i år. Antar värden mellan 0 och 59 i det här datasetet.
- `smsa` Dummyvariabel som antar värdet 1 om personen lever i en "Standard Metropolitan Statistical Area", vilket är ett område med hög befolkningstäthet.
- `pt` Dummyvariabel som antar värdet 1 om personen jobbar deltid.

```
Parameter estimates
-----------------------------------------------
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -217.7637    55.0481  -3.9559 7.8919e-05
educ          50.0398     3.2416  15.4369 7.1827e-51
exper          6.3017     1.4662   4.2979 1.8073e-05
pt          -337.6973    32.0145 -10.5482 2.3753e-25
smsa          45.8232    37.2098   1.2315 2.1829e-01
exper:smsa     3.6071     1.6560   2.1782 2.9506e-02
```

a) Tolka interceptet samt de skattade parametrarna för `exper`, `smsa` och `exper:smsa` i utskriften ovan. Verkar interceptet rimligt att tolka? (6p)

b) Vad är den skattade förväntade veckolönen för en person med 12 års utbildning och 3 års erfarenhet som jobbar heltid och bor i ett tätbefolkat område? (3p)

c) 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 på nästa sida, vilka antaganden verkar vara uppfyllda? Om du tycker att residualerna ser problematiska ut, föreslå en förbättring av modellen som skulle kunna åtgärda detta. (7p)



d) Datasetet innehåller också information om vart i USA dom olika respondenterna bor (södra, västra eller övriga delar av USA). Prediktorn `so` antar värdet 1 om personen är bosatt i södra USA och prediktorn `we` antar värdet 1 om personen är bosatt i västra USA. Din odugliga kusin vill inkludera `we`, `so` och interaktionen `so·we` i modellen. Förklara varför detta inte kommer att fungera. (4p)

e) Finns det någon anledning att oroa sig om multikollinearitet? Motivera ditt svar. (5p)

Lösningsförslag - Uppgift 1

1a (6p)

Interceptet: det skattade förväntade veckolönen för en person som jobbar heltid, har noll års utbildning, noll års erfarenhet och inte bor i ett tätbefolkat område är -217.67. Det är orimligt att tolka interceptet här, främst eftersom att det är negativt. Det är möjligt att vara nyanställd på sitt första jobb (erfarenhet noll) och vi kan tänka oss en person utan utbildning, men det är en väldigt ovanlig person som antagligen inte förekommer i datasetet.

`exper`: den skattade förväntade löneökningen (per vecka) för ett extra år arbetslivserfarenhet för en person som inte lever i ett tätbefolkat område är 6 USD, givet att vi håller alla andra prediktorer konstanta.

`smsa`: den skattade förväntade skillnaden i lön mellan en person med noll års erfarenhet som lever i ett tätbefolkat område och en person med noll års erfarenhet som inte bor i ett tätbefolkat område är 46 USD, givet att vi håller övriga prediktorer konstanta.

Interaktionseffekt: den skattade förväntade löneökningen (per vecka) för ett extra år arbetslivserfarenhet för en person som lever i ett tätbefolkat område är 10 USD, givet att vi håller alla andra prediktorer konstanta. (Indirekt tolkning, enklare.)

Det är även OK att tolka interaktionseffekten som en skillnad i löneöknin när erfarenhet ändras ett år mellan personer i tätbefolkat respektive icke-tätbefolkat område.

1b (3p)

$$-217.7637 + 50.0398 \cdot 12 + 6.3017 \cdot 3 + 45.8238 \cdot 1 + 3.6071 \cdot 3 \cdot 1 = 458.2641$$

Det finns många alternativa sätt att göra beräkningarna, som att slå ihop 6.3017 och 3.6071 för att få "lutningen" för `experience`. Svaret kommer bli samma oavsett dock.

1c (7p)

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

Utifrån dom två figureren till vänster är det uppenbart att normalitet inte är uppfyllt. Residualerna har en skev fördelning, vilket inte är så konstigt när lön är responsvariabeln.

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

Det är svårt att se om antagandet är uppfyllt när vi har så mycket data, men eftersom att data kommer från ett slumpmässigt urval finns ingen anledning att tro att det skulle finnas ett problem med beroende.

Residualerna ser mycket problematiska ut. En rimlig lösning är att transformera responsvariabeln genom att logaritmera. (Andra rimliga transformationer är rot eller kubrot.) Detta kommer att hjälpa till med skevheten och heteroskedasticiteten.

1d (4p)

Både `so` och `we` kommer ifrån samma kategoriska variabel (vart personen bor), och det funkar inte att interagera en variabel med sig själv. Ett annat sätt att förklara detta på är att påpeka att interaktionen alltid kommer vara noll eftersom att max en av `so` och `we` kan vara 1. Man kan bara bo på en plats.

Ett resonemang om att det faktiskt är rimligt kan vara OK om den lösningen explicit innehåller att det endast är rimligt om vi antar att en person kan bo på två platser samtidigt. Interaktionen blir då effekten för en person som har boenden i både södra och västra USA.

1e (5p)

I det här fallet behöver vi inte oroa oss för multikollinearitet. Multikollinearitet är ett problem som uppstår när vi har korrelerade prediktorer, och det finns ingen direkt anledning att tro att några av prediktorerna i vår modell skulle vara starkt korrelerade. Vi ser inte heller några av dom typiska tecknen på multikollinearitet, så som konstiga skattningar eller insignifikant. Förvisso är smsa inte signifikant, men det problemet skulle försvinna om vi centraliserade utbildningsnivå.

För full poäng ska det framgå att du har förstått vad multikollinearitet är, när vi behöver oroa oss om det, och vad vi kan titta efter för varningstecken.

Uppgift 2 - Multipel regression (15 poäng)

Två modeller har skattats baserat på datasetet i uppgift 1. En innehåller samma prediktorer som modellen i uppgift 1, och den andra har lagt till `so` och `we` (se uppgift 1d).

```
Analysis of variance - ANOVA
------------------------------------------------
        df        SS       MS     F     Pr(>F)
Regr     7  83388381 11912626 69.94 1.4504e-90
Error 1992 339292268   170327
Total 1999 422680648

Parameter estimates
------------------------------------------------
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -231.8930    56.3506  -4.11518 4.0264e-05
educ          50.0441     3.2435  15.42887 8.0739e-51
exper          6.4144     1.4660   4.37537 1.2749e-05
pt          -337.2480    31.9783 -10.54615 2.4295e-25
smsa          47.3049    37.1663   1.27279 2.0324e-01
so            -2.8911    21.3405  -0.13547 8.9225e-01
we            59.8631    24.1935   2.47434 1.3431e-02
exper:smsa     3.5477     1.6560   2.14237 3.2285e-02


Analysis of variance - ANOVA
------------------------------------------------
        df        SS       MS     F     Pr(>F)
Regr     5  82152183 16430437 96.21 5.4134e-91
Error 1994 340528465   170777
Total 1999 422680648

Parameter estimates
------------------------------------------------
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -217.7637    55.0481  -3.9559 7.8919e-05
educ          50.0398     3.2416  15.4369 7.1827e-51
exper          6.3017     1.4662   4.2979 1.8073e-05
pt          -337.6973    32.0145 -10.5482 2.3753e-25
smsa          45.8232    37.2098   1.2315 2.1829e-01
exper:smsa     3.6071     1.6560   2.1782 2.9506e-02
```

a) Jämför dom två modellerna med ett F-test. Använd $\alpha = 0.05$. (10p)

b) Förklaringsgraden R^2 beskriver hur stor del av variationen i responsvariabeln som förklaras av prediktorerna, men är inte lämpligt att använda för att jämföra modeller. Varför? Ett alternativ till R^2 är den *justerade* förklaringsgraden, R^2_{adj}. R^2 ligger alltid mellan 0 och 1, men R^2_{adj} kan i vissa fall bli negativ. Använd det matematiska uttrycket för R^2_{adj} till att förklara varför. (5p)

Lösningsförslag - Uppgift 2

2a (10p)

Den fulla modellen ges av

$$\texttt{wage} = \beta_0 + \beta_1\texttt{educ} + \beta_2\texttt{exper} + \beta_3\texttt{pt} + \beta_4\texttt{smsa} + \beta_5\texttt{so} + \beta_6\texttt{we} + \beta_7\texttt{exper:smsa} + \varepsilon$$

Vår reducerade modell inkluderar varken `so` eller `we`, 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. $R^2_{FM} = 83388381/422680648 = 0.1972846$, $R^2_{RM} = 82152183/422680648 = 0.1943599$. Vi har 2000 observationer, så $n = 2000$, $p = 7$ och $q = 5$.

$$F_{obs} = \frac{(0.1972846 - 0.1943599)/(7-5)}{(1-0.1972846)/(2000-7-1)} = 3.628934.$$

Slutligen behöver vi ta fram F_{crit}, där vi har 2 respektive 1992 frihetsgrader. Tabellen går bara till 1000, så vi får nöja oss med att $F_{crit} = F_{\alpha=0.05}(2, 1992) \approx 3.00$.

Eftersom att $F_{obs} > F_{crit}$ så förkastar vi nollhypotesen. Om vi skulle använt $\alpha = 0.01$ så skulle vi inte fått ett signifikant resultat!

2b (5p)

Förklaringsgraden är problematisk eftersom att den aldrig kan miska när vi lägger till ytterliga prediktorer. Om vi jämför två modeller, exempelvis dom i uppgift 2, som kommer R^2 alltid att vara "bättre" för den fulla modellen.

För att förstå varför R^2_{adj} kan bli negativ räcker det att vi tittar på det matematiska uttrycket:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{SSE}{SST} \cdot \frac{n-1}{n-k-1}$$

$\frac{SSE}{SST}$ kan aldrig bli större än 1 (det är därför R^2 aldrig kan bli negativt), men $\frac{n-1}{n-k-1}$ kan bli (mycket) större än 1, vilket leder till att $\frac{SSE}{SST} \cdot \frac{n-1}{n-k-1}$ kan bli större än 1, vilket i sin tur leder till att R^2_{adj} blir negativt.

Uppgift 3 - Logistisk regression (20 poäng)

Datasetet `hsb` innehåller information om 200 amerikanska elevers val av gymnasieprogram (high school). En modell har skattats för att undersöka sambandet mellan elevernas poäng på ett matteprov (`math` i utskriften nedan) och deras val av gymnasieprogram.

Responsvariabeln antar värdet 1 om eleven valde ett akademiskt gymnasieprogram (som natur- eller sammhällsvetarprogrammet) och 0 annars.

```
Parameter estimates
-----------------------------------------------
            Estimate Std. Error z value   Pr(>|z|)
(Intercept) -6.19910    1.06564 -5.8173 5.9813e-09
math         0.12061    0.02034  5.9298 3.0339e-09
```

a) Beräkna den skattade sannolikheten att en person med 35 poäng på matteprovet *inte* kommer välja ett akademiskt program. (4p)

b) Tolka interceptet i termer av odds *och* i termer av sannolikheter. Är det rimligt att tolka interceptet? Tolka parameterskattnigen för `math` i termer av oddskvot. (6p)

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

d) Du överväger en mer komplicerad modell som också inkluderar två ytterligare variabler: skoltyp (en dummy där 1 innebär friskola och 0 kommunal skola) och föräldrarnas genomsnittliga utbildningsnivå (numerisk variabel, mätt i år). Vilket test kan du använda för att jämföra om den enklare modellen är korrekt eller om du bör använda den mer komplicerade? Vilken fördelning kommer test-statistikan följa? (4p)

Lösningsförslag - Uppgift 3

3a (4p)

$$\mathrm{P}(\text{akademiskt} = 0 \mid \text{math} = 35) = 1 - \mathrm{P}(\text{akademiskt} = 1 \mid \text{math} = 35)$$

Enklast att använda

$$\mathrm{P}(\text{akademiskt} = 0 \mid \text{math} = 35) = \frac{1}{1 + \exp(-6.19910 + 0.12061 \cdot 35)} = 0.8784411$$

3b (6p)

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

Det skattade oddset att en person med noll poäng på matteprovet väljer ett akademiskt program är 0.002031258.

Den skattade sannolikheten att en person med noll poäng på matteprovet väljer ett akademiskt program är 0.00202714.

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 rimligt att tolka interceptet, då det helt klart kan hända att man får 0 poäng på ett prov.

Den skattade oddsration för `math` är $\exp(0.12061) = 1.128185$. Vår skattning är att varje ytterligare poäng på matteprovet ökar oddsration med en faktor 1.128185, vilket motsvarar en ökning med 13 procent.

3c (6p)

Detta värde kallas för en *beslutsgräns*.

Vi får fram värdet genom att sätta $P(academic = 1 \mid math) = P(academic = 0 \mid math)$, vilket ger

$$-6.19910 + 0.12061 \cdot math = 0$$

alltså

$$math = \frac{6.19910}{0.12061} = 51.39789$$

3d (4p)

För att jämföra två nästlade modeller så kan vi använda ett likelihoodkvottest. Fördelningen på teststatistikan är χ^2_2, där antalet frihetsgrader beror på skillnaden i antal prediktorer mellan dom två modellerna. Eftersom att vi lägger till två prediktorer blir $df = 2$.

Uppgift 4 - Ickelinjär regression (15 poäng)

En lektor på statistiska institutionen har hittat på följande polynomregression.

$$y = \beta_0 + \beta_1 x + \beta_2 x^3 + \varepsilon$$

a) Vilken grad har polynomregressionen ovan? (2p)

b) Hur stor är den förväntade förändringen av y när x ökar "lite grann" (från x till $x + \Delta x$)? (4p)

c) Antag att $\beta_0 = 1.2, \beta_1 = 1.8$ och $\beta_2 = -1.4$. Vad är det förväntade värdet på y när $x = 3.2$? (4p)

d) Polynomregression låter oss skatta modeller med i princip hur många prediktorer som helst. Om vi exempelvis har tillgång till en enda prediktor x så kan vi skatta en polynomregression av grad 100 och på så vis inkludera hela 100 prediktorer. Att inkludera många prediktorer kan leda till överanpassning. Förklara vad överanpassning är, och välj sedan en metod som kan användas för att hantera överanpassning. Beskriv hur denna metod fungerar. (5p)

Lösningsförslag - Uppgift 4

4a (2p)

Polynomregression är av grad tre eftersom att den största potensen är 3.

4b (4p)

Vi behöver först derivera.

$$\frac{\partial dy}{\partial x} = \beta_1 + 3\beta_2 x^2$$

Den förväntade ändringen när x ändras lite grann, från x till $x + \Delta x$, är $(\beta_1 + 3\beta_2 x^2)\Delta x$.

4c (4p)

$$\mathrm{E}(y \mid x = 3.2) = 1.2 + 1.8 \cdot 3.2 - 1.4 \cdot 3.2^3 = -38.9152$$

4d (5p)

Överanpassning innebär att modellen anpassat sig mycket väl (för väl) till det dataset vi har använt för att estimera (träna) modellen, men att modellen presterar mycket dåligt på nya datapunkter, exempelvis när vi vill göra prediktioner.

Ridge regression och LASSO är två alternativ. Båda fungerar genom att introducera en straffterm i minsta kvadrat-metoden. Detta leder till att dom olika regressionskoefficienterna väljs på ett sånt sätt att en balans uppnås mellan att välja dom värden som bäst passar det observerade datasetet utan att överanpassa genom att föredra värden som ligger närmare noll.

Andra alternativ som man kan diskutera är korsvalidering, och säkert fler jag inte kommer på på rak hand.

Uppgift 5 - Tidsserieanalys (25 poäng)

Tabellen nedan visar det genomsnittliga antalet uthyrningar av cyklar *per kvartal* från första kvartalet 2011 till sista kvartalet 2012 i Washington D.C.

```
# A tsibble: 8 x 2 [1Q]
  avg_rentals quarter
        <dbl>   <qtr>
1        1678 2011 Q1
2        4147 2011 Q2
3        4397 2011 Q3
4        3402 2011 Q4
5        4008 2012 Q1
6        6296 2012 Q2
7        6920 2012 Q3
8        5165 2012 Q4
```

a) Rita en tidsserieplot som visar hur det genomsnittliga antalet uthyrningar har utvecklats över tid. Verkar det finnas en trend och säsong? (5p)

b) Välj antingen en additiv eller multplikativ modell, och motivera ditt svar. Beräkna sedan $2 \times S$ glidande medelvärden, alltså viktade glidande medelvärden, utifrån den modell du valt för tidsserien ovan. (6p)

c) stl-dekomponering är ett alternativ till klassisk dekomponering. Under laboration 5 så använde du `STL`-funktionen i R för att genomföra en stl-dekomponering. Delarna `trend()` och `season()` i kodbiten nedan kontrollerar inställningar för trend och säsong för stl-dekomponeringen. Vad blir skillnaden när vi använder `trend(window = 1)` jämfört med `trend(widow = 50)`? Och vad händer med säsongskomponenten när vi använder `season(window = "periodic")`? (5p)

```
model(STL(min_variabel ~ trend() + season(), robust = TRUE))
```

d) Nedan finner du skattade värden från en AR(1)-modell. Givet att $y_T = 0.5$, ta fram prediktioner för y_{T+1}, y_{T+2} och y_{T+3}. (6p)

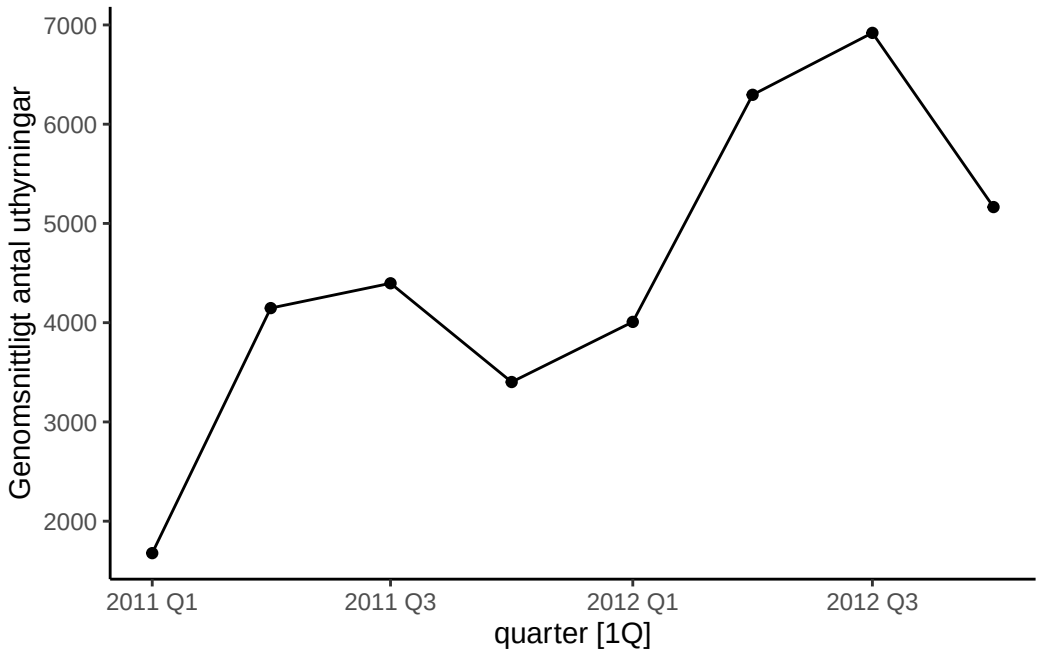
e) Skissa upp stickprovsautokorrelationsfunktionen för den skattade modellen nedan, samt den partiella stickprovsautokorrelationsfunktionen. (3p)

```
     Estimate Std. Error z-ratio Pr(>|z|) 2.5 % 97.5 %
ar1     -0.70       0.01  -97.04        0 -0.71  -0.68
mean     2.54       0.01  429.41        0  2.52   2.55
```

Lösningsförslag - Uppgift 5

5a (5p)

```
Plot variable not specified, automatically selected `.vars = avg_rentals`
```



Det ser ut att finnas både trend och säsong, även om det inte är så lätt att avgöra från en så pass kort tidsserie. Det verkar, utifrån vad vi vet om människors cykelvanor, rimligt att anta säsong på teoretiska grunder.

5b (6p)

Antalet uthyrningar har en tydlig säsongseffekt, och denna är rimligtvis multiplikativ. Vi kan tänka oss att en viss andel av dom som vanligtvis hyr en cykel på sommaren inte gör det på vintern.

Om du antar multiplikativ säsong (vilket är rimligt) så behöver du först logaritmera. Vi kan sen beräkna våra viktade glidande medelvärden.

2011, Q3

$$\frac{\log(1678) + 2\log(4147 + 2\log(4397) + 2\log(3402) + \log(4008))}{8} = 8.17791$$

2011, Q4

$$\frac{\log(4147 + 2\log(4397) + 2\log(3402) + 2\log(4008) + \log(6296))}{8} = 8.338937$$

2012, Q1

$$\frac{\log(4397) + 2\log(3402) + 2\log(4008) + 2\log(6296) + \log(6920))}{8} = 8.435883$$

2012, Q2

$$\frac{\log(3402) + 2\log(4008) + 2\log(6296) + 2\log(6920) + \log(5165))}{8} = 8.532831$$

Stilpoäng om du översätter tillbaks till orginalskala, men det är inget krav.

5c (5p)

Window-argumentet anger hur mycket data som varje glidande medelvärde är baserat på, och motsvarar alltså ungefär m i ett glidande medelvärde av ordning m. Väljer vi `window = 1` så kommer vi få ett glidande medelvärde som endast baseras på få observationer och ändras snabbt, medans `window = 50` kommer ge ett stabilare medelvärde.

Vanligtvis så kan säsongskomponenterna variera över tid för stl-dekomponering, men om vi sätter `window = "periodic"` så får vi en säsongseffekt som är konstant över tid, precis som för klassisk dekomponering.

5d (6p)

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

$\hat{c} = \text{mean} \cdot (1 - \text{ar1}) = 2.54 \cdot (1 - (-0.7)) = 4.318.$

Vi kan nu göra våra prognoser

$$\hat{y}_{T+1|T} = \hat{c} + \hat{\phi}_1 y_T = 4.318 - 0.7 \cdot 0.5 = 3.968$$

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

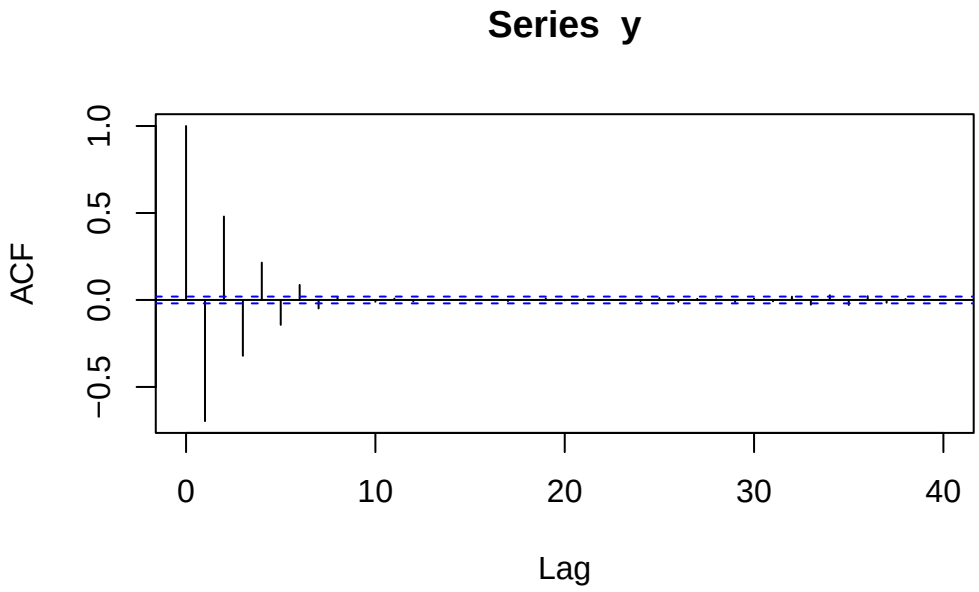
$$\hat{y}_{T+2|T} = \hat{c} + \hat{\phi}_1 \hat{y}_{T+1|T} = 4.318 - 0.7 \cdot 3.968 = 1.5404$$

Slutligen, tre steg framåt

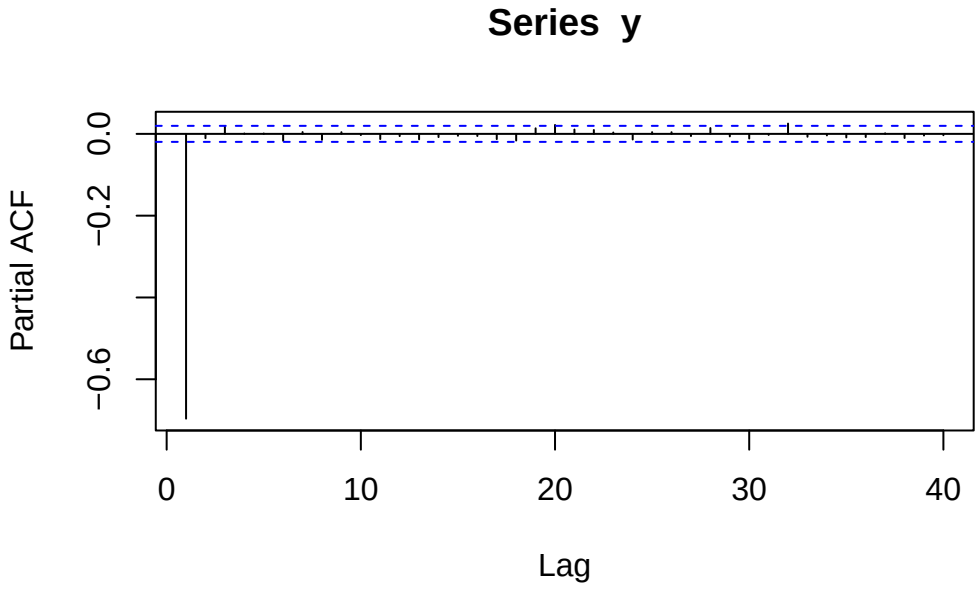
$$\hat{y}_{T+3|T} = \hat{c} + \hat{\phi}_1 \hat{y}_{T+2|T} = 4.318 - 0.7 \cdot 1.5404 = 3.23972$$

5e (3p)

```
acf(y)
```




```
pacf(y)
```


Series y
Partial ACF
0.0
−0.2
−0.6
0
10
20
30
40
Lag