SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2024-06-05

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 `kidiq` innehåller 434 observationer. Varje observation innehåller resultatet på ett IQ-test för en person, tillsammans med motsvarande resultat för personens moder och några fler variabler. I den här uppgiften kommer du använda variablerna

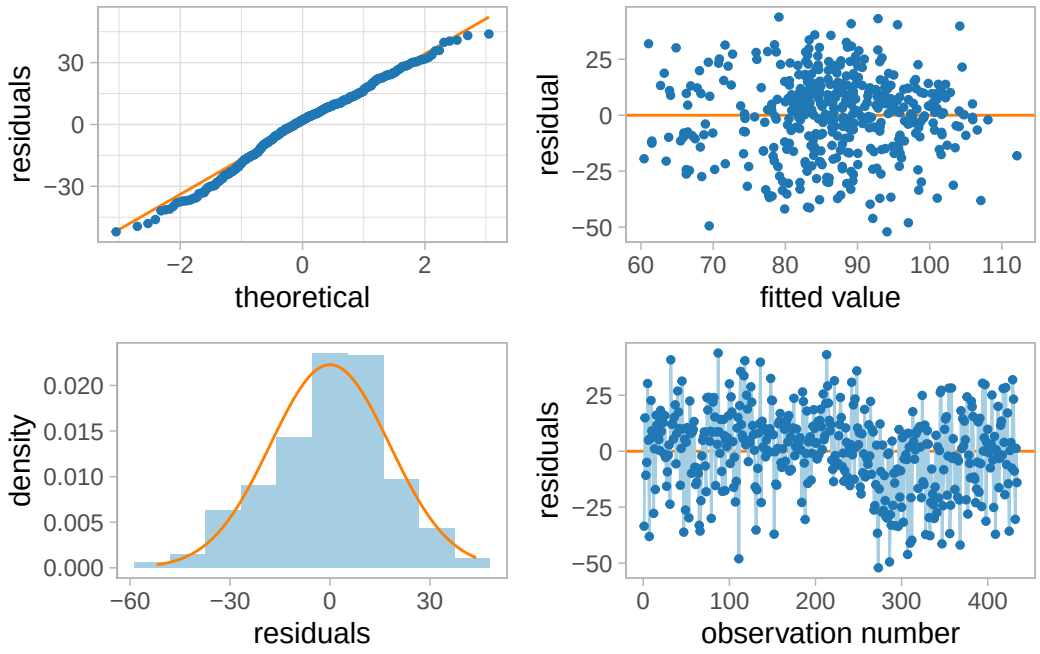
- `kid_score` Resultat på ett IQ testet för personen.
- `mom_iq` Resultat på ett IQ test för personens moder.
- `mom_hs` En dummyvariabel som antar värdet 1 om modern har en högskoleutbildning och 0 annars.

Antalet frihetsgrader i ANOVA-tabellen har dolts. Du kan behöva ta fram dessa värden själv för att lösa vissa deluppgifter.

```
Parameter estimates
-----------------------------------------------
               Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   -11.48202   13.75797 -0.83457 4.0442e-01
mom_hs         51.26822   15.33758  3.34265 9.0239e-04
mom_iq          0.96889    0.14834  6.53138 1.8431e-10
mom_hs:mom_iq  -0.48427    0.16222 -2.98535 2.9942e-03

Analysis of variance - ANOVA

      df        SS         MS        F       Pr(>F)
Regr  NA  41507.51 13835.8364 42.83891 3.066596e-24
Error NA 138878.65   322.9736       NA           NA
Total NA 180386.16         NA       NA           NA
```



a) Tolka `mom_hs`, `mom_iq` och `mom_hs:mom_iq`. (5p)

b) Beräkna $\widehat{\texttt{kid\_iq}}$ för en person vars moder har en högskoleutbildning och vars moder fick 87 poäng på IQ-testet. (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, och vilka figurer i residualplotten ovan används för att undersöka vilket antagande? Vilka antaganden verkar vara (approximativt) uppfyllda? (7p)

d) R^2 är ett mått som beskriver hur stor del av variationen i responsvariabeln som förklaras av modellen. Förklara varför R^2 är ett dåligt verktyg att använda för att jämföra modeller. Ett alternativ till R^2 är den *justerade* förklaringsgraden, R^2_{adj}. Beräkna den justerade förklaringsgraden för modellen. (5p)

Lösningsförslag - Uppgift 1

1a (5p)

Det skattade förväntade skillnaden på IQ-testet mellan en person vars moder har högskoleutbildning jämfört med om personens moder inte har högskoleutbildning är 51 poäng, givet att modern fick 0 poäng på IQ-testet.

Den skattade förväntade ökningen av antal poäng på IQ-testet när moderns poäng på IQ-testet ökar med ett poäng är 0.97 poäng, givet att modern *inte* har högskoleutbildning.

Den sista tolkar vi enklast genom att lägga ihop `mom_iq` och `mom_hs:mom_iq`. Vi får då att den skattade förväntade ökningen av antal poäng på IQ-testet när moderns poäng på IQ-testet ökar med ett poäng är $0.97 - 0.48 = 0.49$ poäng, givet att modern *har* högskoleutbildning.

1b (3p)

$$-11.48 + 51.27 \cdot 1 + 0.97 \cdot 87 - 0.48 \cdot 87 \cdot 1 = 82.42$$

1c (6p)

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

Utifrån dom två figureren till vänster ser det ut som att normalitet är uppfyllt. Möjligtvis är histogrammet något skevt och lite för "platt".

Utifrån figuren uppe till höger, dvs fitted value vs residual, ser vi inga tecken på heteroscedasticitet.

Det är svårt att se om antagandet om oberoende 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.

1d (5p)

Förklaringsgraden är ett dåligt verktyg för att jämföra modeller eftersom att den aldrig kan minska när vi lägger till fler prediktorer. En modelljämförelse baserat på R^2 kommer alltså alltid att föredra den modell som har flest prediktorer, alltså den mer komplicerade modellen.

$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{138879/(434-3-1)}{180386/(434-1)} = 0.2247296.$

Uppgift 2 - Multipel regression (17 poäng)

En alternativ modell som innehåller två till prediktorer, moderns ålder när personen föddes (`mom_age`) och hur snabbt modern gick tillbaks till att jobba efter födseln (`mom_work`) har skattats baserat på samma dataset som i uppgift 1.

```
Analysis of variance - ANOVA
------------------------------------------------
       df     SS      MS     F     Pr(>F)
Regr    5  41876 8375.27 25.88 7.6518e-23
Error 428 138510  323.62
Total 433 180386

Parameter estimates
------------------------------------------------
                Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   -20.618915   16.21460 -1.271626 2.0420e-01
mom_hs         52.775108   15.46206  3.413201 7.0318e-04
mom_iq          0.984556    0.14948  6.586680 1.3220e-10
mom_age         0.350821    0.33177  1.057415 2.9092e-01
mom_work        0.039796    0.76071  0.052315 9.5830e-01
mom_hs:mom_iq  -0.505725    0.16379 -3.087610 2.1490e-03
```

a) Jämför dom två modellerna med ett F-test. Använd $\alpha = 0.05$. Ställ upp hypoteser, beräkna teststatistikan, ta fram det kritiska värdet och dra korrekta slutsatser. (10p)

b) För att identifiera om multikollinearitet är ett problem så finns det ett antal tecken vi kan titta efter. Ange minst två av dessa tecken. (4p)

c) Utifrån modellen ovan (alltså den som även inkluderar `mom_work` och `mom_age`), verkar vi ha problem med multikollinearitet? (3p)

Lösningsförslag - Uppgift 2

2a (10p)

Den fulla modellen ges av

$$\texttt{kid\_score} = \beta_0 + \beta_1 \texttt{mom\_hs} + \beta_2 \texttt{mom\_iq} + \beta_3 \texttt{mom\_age} + \beta_4 \texttt{mom\_work} + \beta_5 \texttt{mom\_hs:mom\_iq} + \varepsilon$$

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

$$\begin{aligned} H_0 &: \beta_4 = \beta_5 = 0 \\ H_A &: \beta_4 \neq 0 \text{ och/eller } \beta_5 = 0 \end{aligned}$$

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. $R^2_{FM} = 41876/180386 = 0.2321466$, $R^2_{RM} = 41508/180386 = 0.2301065$. Vi har 434 observationer, så $n = 434$, $p = 5$ och $q = 3$.

$$F_{obs} = \frac{(0.2321466 - 0.2301065)/(5-3)}{(1 - 0.2321466)/(434-5-1)} = 0.5685739.$$

Slutligen behöver vi ta fram F_{crit}, där vi har 2 respektive 428 frihetsgrader. Vi har inte tillgång till exakta värden, så vi får nöja oss med $F_{crit} = F_{\alpha=0.05}(2, 428) \approx F_{\alpha=0.05}(2, 400) = 3.02$.

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

2b (4p)

Typiska tecken på multikollinearitet är - Prediktorer vi vet är viktiga har skattade regressionskoefficienter som inte är signifikanta, samtidig som ett F-test indikerar att modellen som helhet kan förklara delar av variationen i responsvariabeln. - Skattade regressionskoefficienter har konstiga värden, exempelvis kan dom vara alldeles för små eller ha fel tecken. - Skattade regressionskoefficienter ändras radikalt n är prediktorer tas bort eller läggs till i modellen.

2c (3p)

Utifrån utskriften ser vi att dom skattade regressionskoefficienterna inte ändras radiktalt när vi lägger till nya variabler, vilket är ett tecken på att vi *inte* har problem med multikollinearitet. Det verkar inte heller rimligt att det skulle finnas några mycket starka samband mellan prediktorerna. Obs att andra svar mycket väl kan ge full poäng, så länge dom är väl motiverade.

Uppgift 3 - Logistisk regression (20 poäng)

Datasetet `wbca` innehåller information om 681 potentiella cancertumörer. För varje tumör har information samlats in om dess tjocklek (`Thick`) och antal cellkärnor (`BNucl`). Både `Thick` och `BNulc` antar värden på en skala mellan 1 och 10.

En logistisk regressionsmodell har anpassats. Responsvariabeln `Class` antar värdet 1 om tumören är godartad och värdet 0 om den är elakartad.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.08	0.73	11.05	0
BNucl	-0.84	0.09	-9.82	0
Thick	-0.95	0.11	-8.28	0

a) Beräkna den skattade sannolikheten att en tumör med `BNucl` = 7 och `Thick` = 4 är elakartad. (4p)

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

c) Du överväger en mer komplicerad modell som inkluderar tre ytterligare variabler: `Chrom`, `Epith` och `Mitos`. Vilket test kan du använda för att jämföra om den enklare modellen är korrekt eller om du bör använda den mer komplicerade? Vilken *specifika* fördelning kommer test-statistikan följa, och vilken information behöver du för att genomföra testet? (Du behöver inte ställa upp testet, utan endast svara på dom tre specifika frågorna!) (5p)

d) Din kusin har hört om maskininlärning och undrar om det är möjligt att använda en maskininlärningsmodell för att förutsäga om en tumör är godartad eller elakartad. Du föreslår kNN som ett alternativ. Förklara hur du skulle kunna använda kNN för att predicera om en tumör är elak- eller godartad för en person med `BNucl` 3 och `Thick` 7. Du får själv välja vilket värde på k du vill använda. (5p)

Lösningsförslag - Uppgift 3

3a (4p)

$$\mathrm{P}(\mathrm{Class} = 0 \mid \mathrm{Thick} = 4, \texttt{BNulc = 7}) = \frac{1}{1 + \exp 8.08 - 0.95 \cdot 4 - 0.84 \cdot 7} = 0.83$$

3b (6p)

Vi tolkar interceptet genom att sätta `Thick = 0` och `BNulc = 0`.

Det skattade oddset att en tumör med `Thick = 0` och `BNulc = 0` är godartad är 3229.233.

Den skattade sannolikheten att en tumör med `Thick = 0` och `BNulc = 0` är godartad är $3229.233 = 0.9996904$.

Det är inte rimligt att tolka interceptet eftersom att det framgår i uppgiften att dom två prediktorerna antar värden mellan 1 och 10.

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.)

Det är inte rimligt att tolka interceptet eftersom att det framgår i uppgiften att dom två prediktorerna antar värden mellan 1 och 10.

Den skattade oddsration för `Thick` är $\exp(-0.95) = 0.386741$. Vår skattning är att varje ytterligare enhet på tjockhetsskalan minskar oddsration med en faktor 0.39, vilket motsvarar en minskning med 61 procent.

3c (5p)

Vi kan jämföra dom två modellerna med ett likelihoodkvottest. Likelihoodkvottestet följer en χ^2_3 fördelning. Fördelningen har tre frihetsgrader eftersom att den reducerade modelle har tre restriktioner (tre färre prediktorer). För att genomföra testet behöver vi tillgång till likelihooden (eller log likelihooden) för dom två modellerna.

3d (5p)

kNN fungerar genom att jämföra en ny observation med dom närmsta observationerna i datamängden. Vi kan använda kNN för att förutsäga om en tumör är godartad eller elakartad genom att jämföra den nya observationen med dom närmsta observationerna i datamängden. Om majoriteten av dom närmsta observationerna är godartade så förutsäger vi att den nya observationen är godartad, och vice versa. För den specifika observationen i uppgiften innebär detta att vi tar fram dom k observationerna vars värde på `BNucl` och `Thick` är närmast den nya observationen. Om majoriteten av dessa observationer är godartade så förutsäger vi att den nya observationen är godartad, och om majoriteten är elakartade så förutsäger vi att den nya observationen är elakartad.

Uppgift 4 - Ickelinjär regression (21 poäng)

Populationsmodellen för utvecklingen av antal coviddrabbade personer i Sverige (`smittade`) under 2020 ges av

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

där tiden `t` räknas i antal veckor.

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)	3.003	0.040	74.661	0
t	0.299	0.002	132.064	0

c) Enligt den skattade modellen, hur många förväntas blivit smittade efter $\texttt{t} = 12$ veckor? 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) Hur många veckor kommer det ta tills det skattade förväntade antalet smittade personer är över 1 miljon? En lösning där du gissat dig fram till rätt svar kommer ej ge full poäng. (6p)

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.

$$\widehat{\texttt{log\_bact}} = 3.003 + 0.299 \cdot 12 = 6.591$$

Det skattade förväntade antalet smittade är alltså $\exp(6.591) = 728.509 \approx 729$.

4d (5p)

Vi börjar med att ta fram $\widehat{\beta}$: $\exp(0.299) = 1.34851$. Alltså, när `t` ökar en enhet så är den skattade förväntade ökningen av antalet smittade trettiofem procent.

4e (5p)

Vi kan enklast lösa denna fråga genom att räkna ut värdet på t för den logaritmerade modellen. Logaritmen (med bas e) av 1 miljon är ungefär 13.8155. Vi kan nu lösa för t.

$$\log(1000000) = 3.003 + 0.299 \cdot t,$$

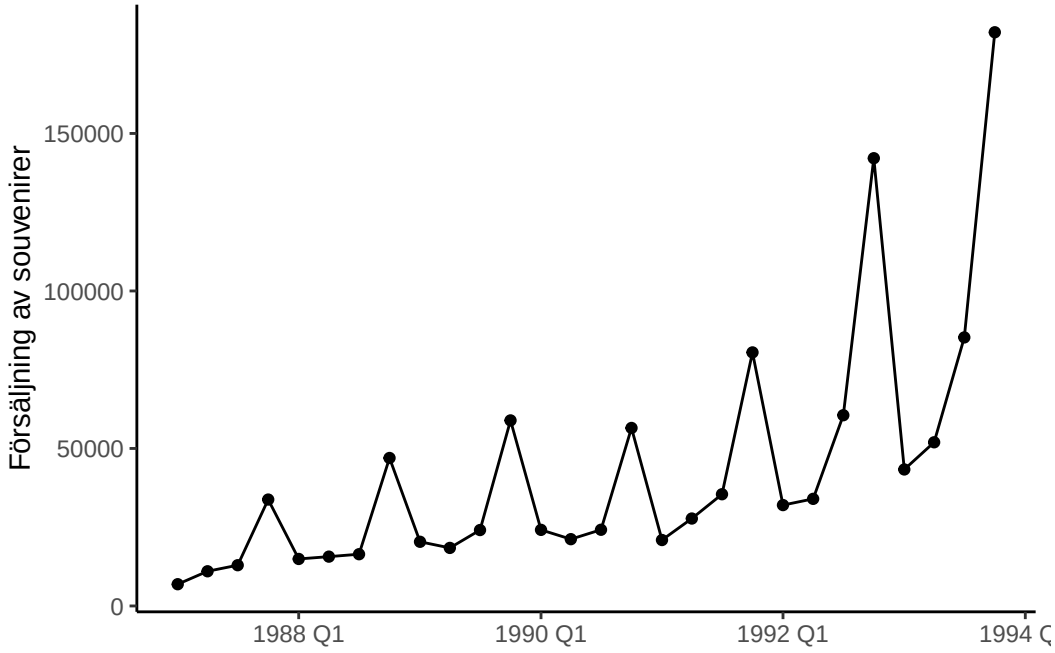
vilket ger

$$\frac{\log(1000000) - 3.03}{0.299} = t = 36.16224.$$

Alltså kommer det att ta drygt 36 veckor innan det skattade antalet smittade personer är över 1 miljon.

Uppgift 5 - Tidsserieanalys (22 poäng)

Figuren nedan visar försäljningen av souvenirer per kvartal i en affär i Queensland, Australien.



a) Har tidsserien någon trend och/eller någon säsong? Om du anser att tidsserien har säsong, skulle du säga att den är multiplikativ eller additiv? (3p)

b) Baserat på ditt svar ovan, använd klassisk dekomponering för att ta fram restkomponenterna för alla kvartal från och med Q1 1988 till och med Q4 1989. Dom siffror du behöver hittar du i tabellen på nästa sida. (11p)

c) Klassisk dekomponering fångar upp två typer av systematisk variation: trend och säsong (ibland räknar vi även cyklisk variation). För att undersöka om en dekomponering har fångat *all* systematisk variation kan vi analysera resttermernas autokorrelation. Förklara kortfattat vad autokorrelation är. Vilket värde på autokorrelationen innebär att vi har fångat upp all systematisk variation i tidsserien? (3p)

```
# A tsibble: 12 x 2 [1Q]
   year_quarter  sales
          <qtr>  <dbl>
 1      1987 Q3 12938.
 2      1987 Q4 33780.
 3      1988 Q1 14923.
 4      1988 Q2 15658.
 5      1988 Q3 16428.
 6      1988 Q4 46977.
 7      1989 Q1 20377.
 8      1989 Q2 18428.
 9      1989 Q3 24099.
10      1989 Q4 58903.
11      1990 Q1 24157.
12      1990 Q2 21204.
```

d) Nedan finner du skattade värden från en AR(1)-modell. Givet att $y_T = 0.5$, ta fram prediktioner för y_{T+1} och y_{T+2}. (5p)

```
     Estimate Std. Error z-ratio Pr(>|z|) 2.5 % 97.5 %
ar1      0.70       0.01   97.99        0  0.69   0.71
mean     4.37       0.03  130.81        0  4.31   4.44
```

Lösningsförslag - Uppgift 5

5a (5p)

Tidsserien har tydlig trend och säsong, och säsongen är glasklart multiplikativ.

5b (10p)

Eftersom att säsongen är multiplikativ så bör vi först logaritmera tidsserien.

Första steget är att beräkna trenden, vilket vi gör med ett viktat glidande medelvärde eftersom att tidsserien har en tydlig säsong. Exempelvis kommer det glidande medelvärden för den första tidpunkten ges av

1988, Q1

$$\frac{\log(12938) + 2\log(33780) + 2\log(14923) + 2\log(15658) + \log(16428))}{8} = 9.82$$

Efter att vi har tagit fram trenden så kan vi beräkna säsongen genom att ta differensen mellan logaritmen av försäljningen och trenden för dom åtta kvartalet. Genom att sen ta medelvärdet att dessa differenser för varje kvartal så får vi en säsongseffekt för varje kvartal. Detta är den grova skattningen av säsongseffekten, och för att få den slutgiltiga säsongseffekten behöver vi subtrahera medelvärdet av samtliga säsongseffekter från varje enskiljd säsongseffekt.

```
# A tsibble: 12 x 6 [1Q]
# Key:       .model [1]
   .model                       year_quarter `log(sales)` trend seasonal   random
   <chr>                               <qtr>        <dbl> <dbl>    <dbl>    <dbl>
 1 "classical_decomposition(l~      1987 Q3         9.47 NA      -0.209 NA
 2 "classical_decomposition(l~      1987 Q4        10.4  NA       0.714 NA
 3 "classical_decomposition(l~      1988 Q1         9.61  9.82   -0.203 -0.00789
 4 "classical_decomposition(l~      1988 Q2         9.66  9.89   -0.302  0.0690
 5 "classical_decomposition(l~      1988 Q3         9.71  9.97   -0.209 -0.0565
 6 "classical_decomposition(l~      1988 Q4        10.8  10.0     0.714  0.0117
 7 "classical_decomposition(l~      1989 Q1         9.92 10.1    -0.203  0.0248
 8 "classical_decomposition(l~      1989 Q2         9.82 10.2    -0.302 -0.0520
 9 "classical_decomposition(l~      1989 Q3        10.1  10.2    -0.209  0.0734
10 "classical_decomposition(l~      1989 Q4        11.0  10.3     0.714  0.00519
11 "classical_decomposition(l~      1990 Q1        10.1  NA      -0.203 NA
12 "classical_decomposition(l~      1990 Q2         9.96 NA      -0.302 NA
```

5c (3p)

Med autokorrelation menar vi hur starkt korrelerade observationer i en tidsserie är med tidigare värden. Exempelvis, hur starkt är sambandet mellan tidsseriens värde vid tidpunkt t och $t-1$. Om autokorrelationen är noll så innebär det att vi har fångat upp all systematisk variation.

5d (5p)

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

$\hat{c} = \text{mean} \cdot (1 - \text{ar1}) = 4.37 \cdot (1 - 0.7) = 1.311.$

Vi kan nu göra våra prognoser

$$\hat{y}_{T+1|T} = \hat{c} + \hat{\phi}_1 y_T = 1.311 + 0.7 \cdot 0.5 = 1.661$$

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

$$\hat{y}_{T+2|T} = \hat{c} + \hat{\phi}_1 \hat{y}_{T+1|T} = 1.311 + 0.7 \cdot 1.661 = 2.4737$$