SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2023-05-04

Skrivtid: kl. 8–13 (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 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 också 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

Datasetet CPS85 innehåller information om löner för ett slumpmässigt urval av personer från 1985. För att studera sambandet mellan lön, utbildningsnivå och kön har en regressionsmodell med följande variabler anpassats

- `wage` - lön i USD
- `sex` - i detta dataset kodad som `M` (male) eller `F` (female)
- `educ` - utbildning, mätt i år

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -3.2658785  1.6191895 -2.016984 4.420094e-02
educ         0.8556754  0.1222198  7.001122 7.718271e-12
sexM         4.3704499  2.0850574  2.096081 3.654859e-02
educ:sexM   -0.1725303  0.1571232        NA           NA
```

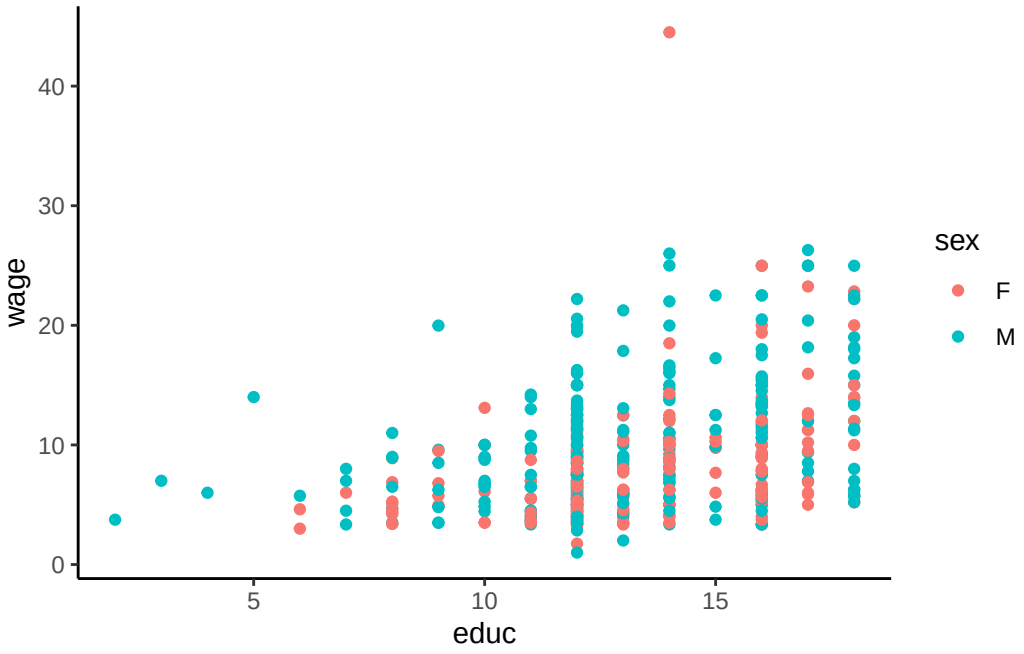
```
       df        SS        MS        F       Pr(>F)
Regr    3  2677.432 892.47742 41.49504 4.240326e-24
Error 530 11399.266  21.50805       NA           NA
Total 533 14076.699        NA       NA           NA
```

a) Ställ upp *populationsmodellen* som korresponderar till den skattade modellen ovan. Ställ även upp den skattade modellen.

b) Den skattade modellen kan visualiseras som två separata regressionslinjer. Rita ut dom två regressionslinjerna på spridningsdiagrammet på nästa sida, och skriv ned respektive regressionslinjes ekvation.

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

d) Genomför ett formellt test huruvida sambandet mellan `wage` och `education` skiljer sig åt mellan män (`M`) och kvinnor (`F`). Använd $\alpha = 0.05$.


wage
40
30
20
10
0
5
10
15
educ
sex
F
M

Lösningsförslag - Uppgift 1

1a

Populationsmodell

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{sex} + \beta_3 \text{educ} \cdot \text{sex} + \varepsilon$$

Skattad modell

$$\widehat{\text{wage}} = -3.27 + 0.86 \cdot \text{educ} + 4.37 \cdot \text{sex} - 0.17 \cdot \text{educ} \cdot \text{sex} + \varepsilon$$

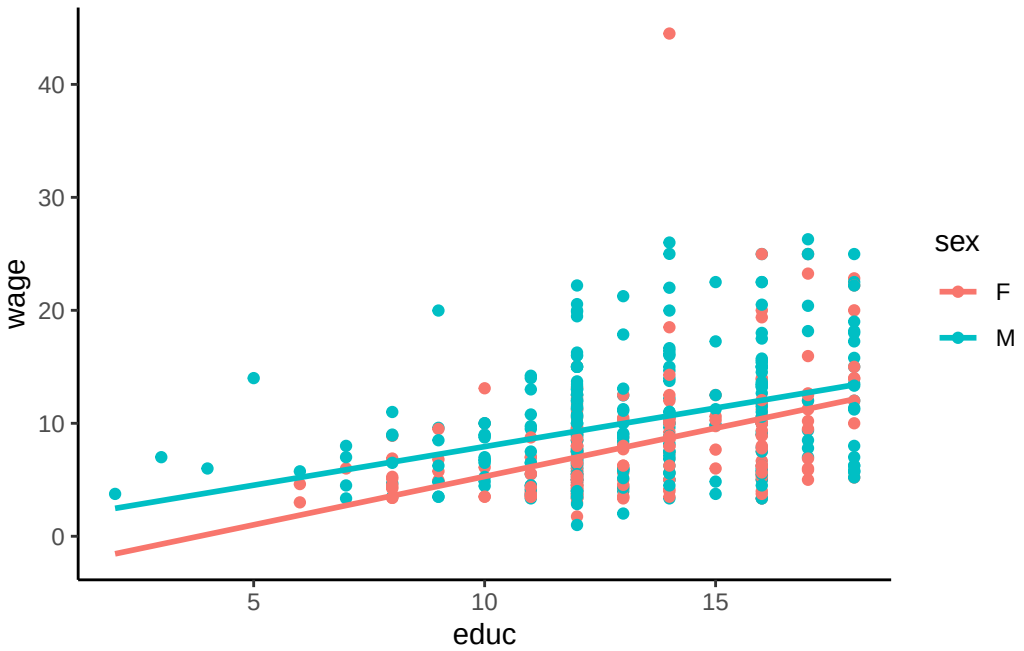
1b

Linje för män har intercept $-3.26588 + 4.37045 = 1.10457$ och lutning $0.85568 - 0.17253 = 0.68315$

$$\widehat{\text{wage}} = 1.10457 + 0.68315 \cdot \text{educ}$$

Linje för kvinnor har intercept -3.26588 och lutning 0.85568

$$\widehat{\text{wage}} = -3.26588 + 0.85568 \cdot \text{educ}$$



1c

Intercept: den skattade förväntade lönen för en kvinna med 0 års utbildning. Inte helt otänkbart att kunna sakna utbildning (det beror lite på vad vi menar, räknas grundskola?). Att interceptet är negativ är dock orimligt, och tyder på att alla observationer har värden på utbildning som är skiljda från noll.

`educ`: Den skattade förväntade ökningen i lön för en kvinna när vi ökar utbildningsnivå med ett år är 0.86 dollar. Rimlig att tolka.

`sex`: Den skattade förväntade skillnaden i lön mellan män och kvinnor och män med 0 års utbildning är 4.37 USD (till fördel för män). Rimlig eller orimlig beroende på vad du skrev om interceptet. Är interceptet orimligt att tolka blir denna också orimlig (eftersom att den också rör utbildning 0 år). (Ni behöver inte resonera om huruvida denna är rimlig att tolka, det är endast för interceptet som detta efterfrågas explicit, men denna parameter representerar också ett slags "intercept" så jag inkluderar den här.)

Interaktionseffekten: Den skattade förvänta skillnade i ökningen av lön vid ett år längre utbildning mellan män och kvinnor är 0.17 dollar (till kvinnors fördel). Det är även OK att skatta interaktionseffekten genom att slå ihop denna med educ-skattningen och säga att den skattade förväntade ökningen i lön för en man när vi ökar utbildningsnivån med ett år är 0.68 dollar ($0.85568 - 0.17253$).

1d

Det vi ska göra här är ett t-test av β_3, alltså själva *interaktionseffekten*, eftersom att det är lutningen som beskriver hur sambandet mellan `educ` och `wage` ser ut.

Vi börjar med att ställa upp noll- och alternativhypotes

$$H_0 : \beta_3 = 0, \quad H_a : \beta_3 \neq 0$$

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

t-värdet ges av

$$t_{obs} = \frac{b_3 - 0}{s_{b_3}} = \frac{-0.17253}{0.15712} = -1.098078$$

$$t_{crit} = t_{0.025,530} \approx 1.966$$

Här har jag använt t-fördelningen med 400 frihetsgrader. Det är helt OK att normalapproximera, då värdet blir 1.96. Vi kan se att det inte kommer spela någon roll (vi kommer inte förkasta oavsett).

Eftersom att $|t_{obs}| < t_{crit}$ kan vi *inte* förkasta nollhypotesen. Vi hittar alltså inte något stöd för att sambandet mellan lön och utbildningsnivå skiljer sig åt mellan män och kvinnor. (Observera att detta inte betyder att det inte finns några systematiska skillnader i lön mellan män och kvinnor.)

Uppgift 2 - F-test av restriktioner

En mer komplicerad modell än den i uppgift 1 som även innehåller följande förklarande variabler har skattats

- `age` - ålder, i år
- `union` - fackligt medlemskap

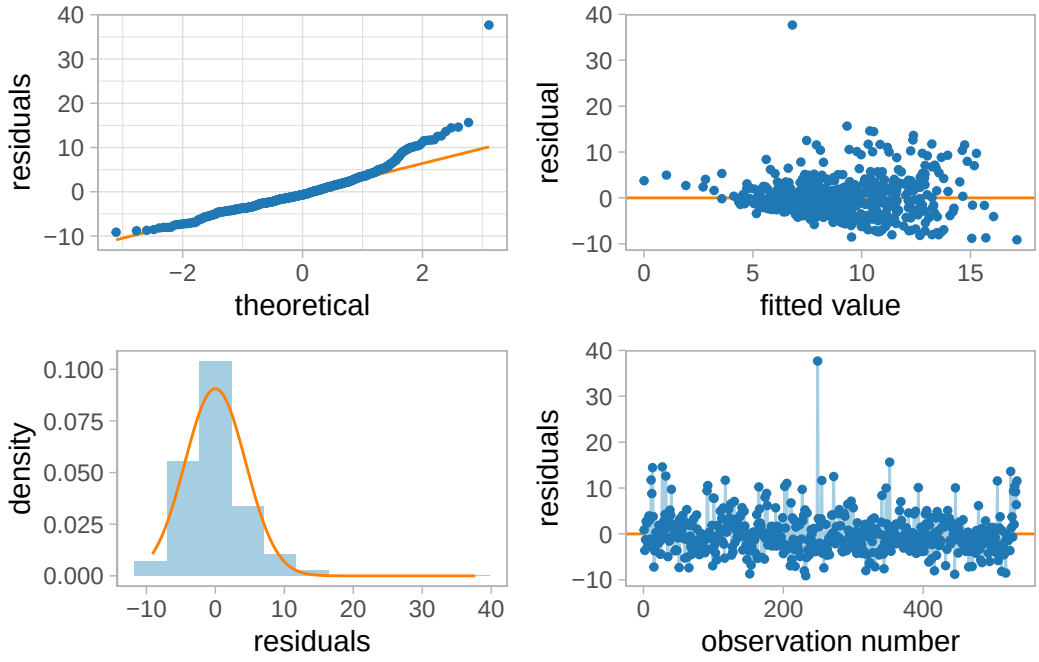
```
Analysis of variance - ANOVA
------------------------------------------------
       df      SS      MS      F     Pr(>F)
Regr    5  3751.8 750.362 38.372 1.3144e-33
Error 528 10324.9  19.555
Total 533 14076.7

Parameter estimates
------------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) -8.52072   1.745657 -4.8811 1.3996e-06
educ         0.93431   0.117507  7.9511 1.1298e-14
sexM         4.42617   1.999957  2.2131 2.7315e-02
age          0.10748   0.016733  6.4232 2.9800e-10
unionUnion   1.43120   0.510359  2.8043 5.2283e-03
educ:sexM   -0.17468   0.150229 -1.1628 2.4545e-01
```

a) Använd ett F-test för att jämföra modellen i uppgift 1 och den nya, utökade modellen. Använd 1 % signifikansnivå.

b) 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?

c) Ett av dom tre antaganden kan testas med White's test. Vilket? Beskriv i grova drag hur White's test fungerar. Utgå ifrån en modell som bara inehåller `educ` och `age` som förklarande variabler. (Du behöver inte ta fram någon teststatisika eller kritisk gräns.)


residuals
theoretical
residual
fitted value
density
residuals
residuals
observation number

Lösningsförslag - Uppgift 2

2a

Första steget är att ställa upp noll- och alternativhypotes.

Den begränsade (reducerade) modellen ges av

$$\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{sexM} + \beta_3 \texttt{educ:sexM} + \varepsilon$$

den obegränsade (fulla) modellen ges av

$$\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{sexM} + \beta_3 \texttt{educ:sexM} + \beta_4 \texttt{age} + \beta_5 \texttt{union} + \varepsilon$$

$$\begin{aligned} H_0 &: \beta_4 = \beta_5 = 0 \\ H_A &: \text{minst en av } \beta_4, \beta_5 \neq 0 \end{aligned}$$

Nästa steg är att ta fram vår teststatistika, vilket för F-test av restriktioner ges av

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

Vi hittar dom värden vi behöver för att beräkna teststatistikan

- $n = 533 + 1 = 534$, antalet observationer
- $p - q = 2$, antalet restriktioner
- $p = 5$, antalet förklarande variabler i den obegränsade modellen

Vi behöver också beräkan R^2_{FM} och R^2_{RM}, dvs förklaringsgraden i dom två modellerna.

Eftersom att vi har tillgång till SSR och SST för båda modellerna är den enklaste formeln

$$R^2 = SSR/SST$$

För vår UR-modell (den utan restriktioner) så får vi

$$R^2_{FM} = 3751.8/14076.6 = 0.2665274$$

På motsvarande sätt beräknar vi

$$R^2_{RM} = 2677.4/14076.6 = 0.1902022$$

Vi kan nu beräkna värdet på vår teststatistika

$$F_{obs} = \frac{(0.2665274 - 0.1902022)/2}{(1 - 0.2665274)/(534 - 5 - 1)} = 27.47185$$

Nästa steg är att hitta det kritiska värdet. Vi ska använda $\alpha = 0.01$. Frihetsgraderna i täljaren och nämnaren är 2 respektive 528, vilket ger

$$F_{crit} = F_{0.01}(2, 528) \approx 4.66$$

Det sanna kritiska värdet ligger någonstans mellan 4.66 och 4.63. Oavsett, så ser vi att

$$F_{obs} > F_{crit}$$

så vi förkastar nollhypotesen och föredrar modellen med ålder och fackligt medlemsskap utöver modellen som endast innehåller utbildning, kön, och interaktionen mellan utbildning och kön.

2b

iid-antagandet säger att feltermerna är *oberoende*.

Vi kan kontrollera detta antagande genom figuren nere i det högra hörnet som visar residualerna mot observationsnummer. Om vi ser tydliga mönster i denna figur, till exempel om det ofta kommer flera positiva (eller negativa) residualer i följd så är det ett tecken på att detta antagande inte är uppfyllt. Med så här mycket data är det svårt att se något beroende även när det finns. Vårt dataset är dock ett slumpmässigt urval (och inte en tidsserie) så antagligen har vi inga problem med beroende.

Ett annat antagande som ingår i $\varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2)$ är *konstant varians*, eller homoskedasticitet. Detta antagande säger att variansen inte varierar beroende på dom förklarande variablernas värden (eller utfallsvariabeln). Figuren uppe till höger visar residualerna plottade mot *anpassade värden*. Om vi ser ett mönster i denna figur, som text en trattform, så innebär det att det finns ett samband mellan det anpassade värdet och residualerna, vilket indikerar ett samband mellan feltermernas varians och värdet på dom förklarande varaiblerna.

Det tredje antagandet är normalitet. Detta antagandet kan vi kontrollera med dom två figurerna till vänter. Histogrammet visar fördelningen för residualerna med ett histogram, tillsammans med en linje som visar densiteten för en normalfördelning.Om antagandet om normalitet för feltermerna är uppfyllt förväntar vi oss att histogrammet ska följa linjen hyffsat väl (men för små stickprov kan det se ut lite hur som helst). Den andra figuren som vi kan använda för att undersöka normalitet är QQ-plotten. QQ-plotten visar sample-kvantilerna plottade mot dom teroetiska kvantilerna för en normalfördelning. Om dessa “matchar” så ska samtliga punkter

ligga på en rak linje. Dest mer punkterna avviker från linjen, desto starkare indikationer har vi på att feltermerna inte följer en normalfördelning.

2c

White's test testar homoskedasticitetsantagandet, alltså konstant varians. Den gör det genom att skapa en ny utfallsvariabel genom att *kvadrera* residualerna. Sen skapar vi en regressionsmodell där vi använder dom kvadrerade residualerna som utfall, och som förklarande variabel använder vi dom förklarande variablerna från orginalmodellen, tillsammans med deras kvadrater och interkationer. För modellen med educ och age skulle vi alltså skatta modellen

$$e^2 = \tilde{\beta}_0 + \tilde{\beta}_1 \text{age} + \tilde{\beta}_2 \text{age}^2 + \tilde{\beta}_3 \text{educ} + \tilde{\beta}_4 \text{educ}2^2 + \tilde{\beta}_5 \text{age} \cdot \text{educ} + \tilde{\varepsilon}$$

Om feltermerna verkligen är homoskedastiska så ska dom förklarande variablerna här inte bidra med någonting alls i modellen ovan, så vi kan testa om heteroskedasticitet föreligger genom att göra ett F-test på modellen ovan. Om modellen ovan har hög förklaraningsgrad så är det ett starkt tecken på att det finns problem med heteroskedasticitet.

Notera: ni behöver inte ställa upp den specifika teststatistikan eller veta vilken fördelning den följer.

Uppgift 3 - Logistisk regression

Iris-datasetet innehåller information om sepallängd och sepalvidd för två olika iris-blommor: *iris versicolor* och *iris virginica*.

Vi är nu intresserade om vi kan klassificera irisar som tillhörande *iris versicolor* respektive *iris virginica* baserat på endast deras sepallängd (`Sepal.Length`) och sepalbredd (`Sepal.Width`). För att göra detta skattar vi en logistisk regressionsmodell.

Utfallsvariabeln antar värdet 1 om irisen är en *versicolor* och 0 om den är en *virginica*. `Sepal.Width` och `Sepal.Length` är angivna i *centimeter*.

```
Parameter estimates
------------------------------------------------
              Estimate Std. Error  z value   Pr(>|z|)
(Intercept)  -13.04603    3.09736 -4.21198 2.5314e-05
Sepal.Width    0.40466    0.86283  0.46899 6.3908e-01
Sepal.Length   1.90238    0.51691  3.68027 2.3299e-04
```

a) Ställ upp *populationsmodellen* som korresponderar till den skattade modellen ovan, antingen i termer av sannolikheter eller odds.

b) Vad är sannolikheten att en iris med sepalvidd 6.3 och sepallängd 2.8 är en `versicolor`?

c) Vad är sannolikheten att en iris med sepalvidd 5.3 och sepallängd 2.3 är en `virginica`?

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

e) Tolka parameterskattnigen för `Sepal.Width` och `Sepal.Length` i termer av oddsration.

Lösningsförslag - Uppgift 3

3a

Populationsmodell i odds

$$\text{Odds}(y = 1 \mid \text{Sepal.Width}, \text{Sepal.Length}) = \exp(\beta_0 + \beta_1 \cdot \text{Sepal.Width} + \beta_2 \cdot \text{Sepal.Length})$$

- Även ok att använda svenska namn (sepalvidd, sepalbredd), samt OK att ha versicolor som utfallsvariabel. Att skriva ut en batt är inkorrekt. Att skriva ut dom skattade värdena är inkorrekt.

3b

$$\widehat{\text{P}}(y = 1 \mid \text{Sepal.Width} = 6.3, \text{Sepal.Length} = 2.8) = \frac{\exp(-13.04603 + 0.40466 \cdot 6.3 + 1.90238 \cdot 2.8)}{1 + \exp(-13.04603 + 0.40466 \cdot 6.3 + 1.90238 \cdot 2.8)}$$

$$\widehat{\text{P}}(y = 1 \mid \text{Sepal.Width} = 6.3, \text{Sepal.Length} = 2.8) = 0.005652392$$

3c

Den skattade sannolikheten att det är en `virginica` är 1 minus sannolikheten att det är en `versicolor`. Vi kan antingen räkna ut sannolikheten att den är en `versicolor` och sen ta 1 minus den sannolikheten, eller så kan vi använda att

$$\widehat{\text{P}}(y = 0 \mid \text{Sepal.Width} = 5.3, \text{Sepal.Length} = 2.3) = \frac{1}{1 + \exp(-13.04603 + 0.40466 \cdot 5.3 + 1.90238 \cdot 2.3)}$$

$$\widehat{\text{P}}(y = 0 \mid \text{Sepal.Width} = 5.3, \text{Sepal.Length} = 2.3) = 0.9985371$$

3d

Här hade jag nog tänkt fråga efter att tolka det *skattade* interceptet, men det framgår inte i frågan! Om vi tolkar själva interceptet (inte det skattade) så får vi att oddset att en iris med sepallängd och sepalbredd 0 är en versicolor är $\exp(\beta_0)$. Detta är lite av en miss från min sida, och jag kommer ge båda tolkningarna korrekt eftersom att jag inte minns om nån frågade om detta på tentan.

Det skattade oddset att en iris med sepallängd och sepalbredd 0 är en `versicolor` är $\exp(-13.04603) = 0.000002158645$, och den skattade sannolikheten är ugefär 0.00000216. En

blomma med sepalbredd/vidd 0 existerar inte och det är därför inte meningsfullt att tolka inteceptet. (Obs, om du skrivit "det är fullt möjligt att det finns irisar av typen versicolor som inte har sepalblad, och alltså är det rimligt att ha värden på noll" så är det helt OK, detta är inte en biologitenta.)

3e

Sepalbredd: exp(0.40466) = 1.498793. Den skattade oddskvoten är ca 1.5, alltså är vår skattning att oddset att en iris är en `versicolor` ökar med en faktor 1.5 för varje centimeter vi ökar sepalbredd, givet att sepallängden hålls konstant. Vi kan även uttrycka det som att den skattade ökningen av oddset för varje ytterligare centimeter sepalbredd är 50 procent, givet att sepallängden hålls konstant.

Sepallängd: exp(1.90238) = 6.701826. Den skattade oddskvoten är ca 6.7, alltså är vår skattning att oddset att en iris är en `versicolor` ökar med en faktor 6.7 för varje centimeter vi ökar sepallängden, givet att sepalbredden hålls konstant. Alternativt, vår skattning är att oddset att en iris är en `versicolor` ökar med 570 procent för varje (ytterligare) centimeter sepallängd, givet att sepalbredd hålls konstant.

Uppgift 4 - Cykeluthyrning

En modell har anpassats för att undersöka sambandet mellan antalet uthyrningar av cyclar (`nRides`) en given dag, och antalet uthyrningar dagen innan (`nRides_lagged`) och den standardiserade luftfuktigheten (`humidity`). Såväl utfallsvariabeln som dom två förklarande variablerna har *logaritmeras* (med basen e) innan modellen skattades med hjälp av R.

Modellen utgår ifrån att sambandet kan beskrivas med följande populationsmodell:

$$\texttt{nRides} = \alpha \cdot \texttt{nRides\_lagged}^{\beta_1} \cdot \texttt{humidity}^{\beta_2} \cdot \varepsilon$$

```
Parameter estimates
-----------------------------------------------
                  Estimate Std. Error  t value    Pr(>|t|)
(Intercept)        2.121009   0.207200 10.23653  4.5739e-23
log_nRides_lagged  0.742770   0.024686 30.08889 8.1192e-130
log_humidity      -0.022564   0.042990 -0.52487  5.9983e-01
```

a) Vad kallas sambandet som ges av populationsmodellen?

b) Vad är $\widehat{\texttt{nRides}}$ när $\texttt{nRides\_lagged} = 1339.431$ och $\texttt{humidity} = 0.59$?

c) Tolka dom skattade parametrarna ovan, exklusive interceptet.

d) Modellen ovan har skattats om med två olika typer av regularisering: Ridge och LASSO. Parameterestimatet för dom två regulariserade modellerna hittar du i tabellen nedan. Vilken modell (av 1 och 2) korresponderar till LASSO och vilken till Ridge? Utifrån vad du vet om regularisering, verkar det rimligt att använda det här? Glöm inte att motivera ditt svar.

Modell	Intercept	`log_nRides_lagged`	`log_humidity`
1	5.06	0.39	0.00
2	5.23	0.37	0.01

Lösningsförslag - Uppgift 4

4a

Elasticitetssamband.

4b

Detta kan vi beräkna på två olika sätt. Vi kan antingen "antilogaritmera" α och beräkna vår skattning av `nRides` på orginalskala, eller så kan vi logaritmera dom två förklarande variablerna, beräkna vår skattning av `log_nRides` och sen antilogaritmera log-skattningen. Eftersom att vi har fått värden på förklarande varaiblerna på orginalskalan är den första approachen enklare.

$$\hat{\alpha} = \exp(2.121009) = 8.339548$$

$$\widehat{nRides} = 8.339548 \cdot 1339.431^{0.742770} \cdot 0.59^{-0.022564} = 1773.763$$

Alternativt kan vi logaritmera dom två förklarande variablerna och ta fram en skattning av logaritmen av `nRides`

$$2.121009 + 0.742770 \cdot \log(1339.431) - 0.022564 \cdot \log(0.59) = 7.480859$$

vilket vi sen behöver översätta till orginalskalan

$$\widehat{nRides} = \exp(7.480859) = 1773.763$$

4c

Den skattade förväntade ökningen av antalet uthyra cyklar idag vid en procents ökning av antalet uthyrda cyklar igår är 0.74 procent. (Behöver inte vara idag / igår, utan det går bra att skriva "vid en viss tidpunkt för en procents ökning tidpunkten innan" eller liknande.) Det som ska finnas med är att det är skattningar, och förväntad ökning/minskning (vi har en felterm), samt att det är procent förändring i förklarande variabel till procent förändring i utfallsvariabel.

Den skattade förväntade minskningen av antalet uthyrda cyklar idag vid en procents ökning av luftfuktigheten är 0.02 procent.

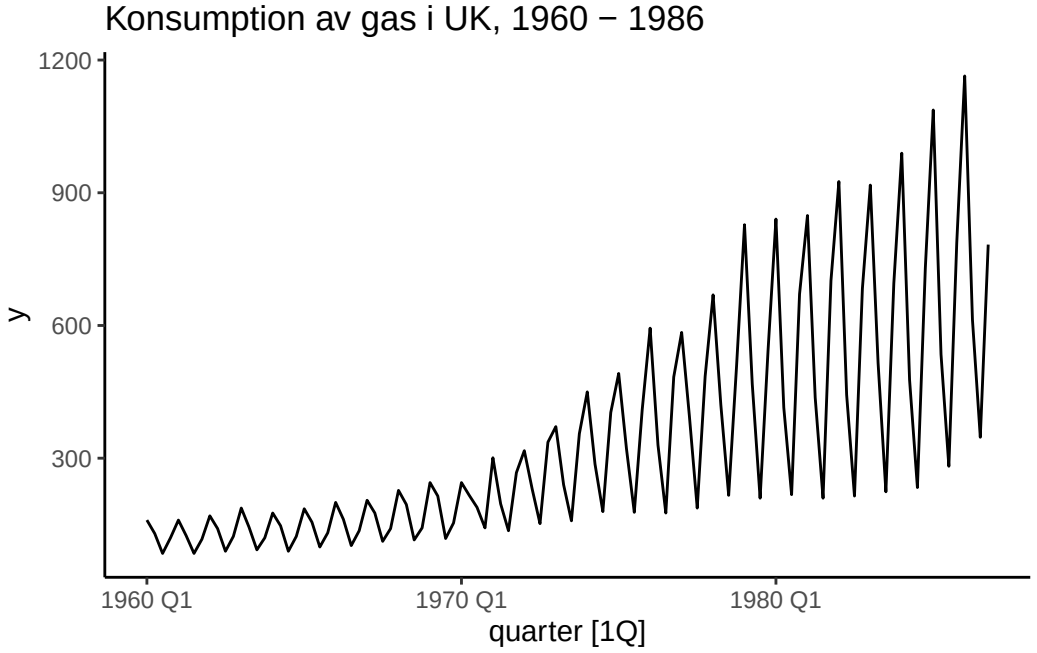
4d

Rad 1 motsvarar LASSO och rad 2 Ridge. Vi vet att LASSO straffar små avvikelser från noll hårdare, vilket får följden att lasso ofta sätter parameterskattningar till exakt noll. (Att bara gissa ger noll poäng.)

Det verkar inte vara rimligt med regularisering här. Modellen är mycket enkel (endast två förklarande variabler).

Uppgift 5 - Klassisk dekomponering

Figuren nedan visar konsumptionen av gas i UK mellan första kvartalet 1960 och fjärde kvartalen 1986. Tabellen visar dom faktiska värdena för dom tre sista åren.

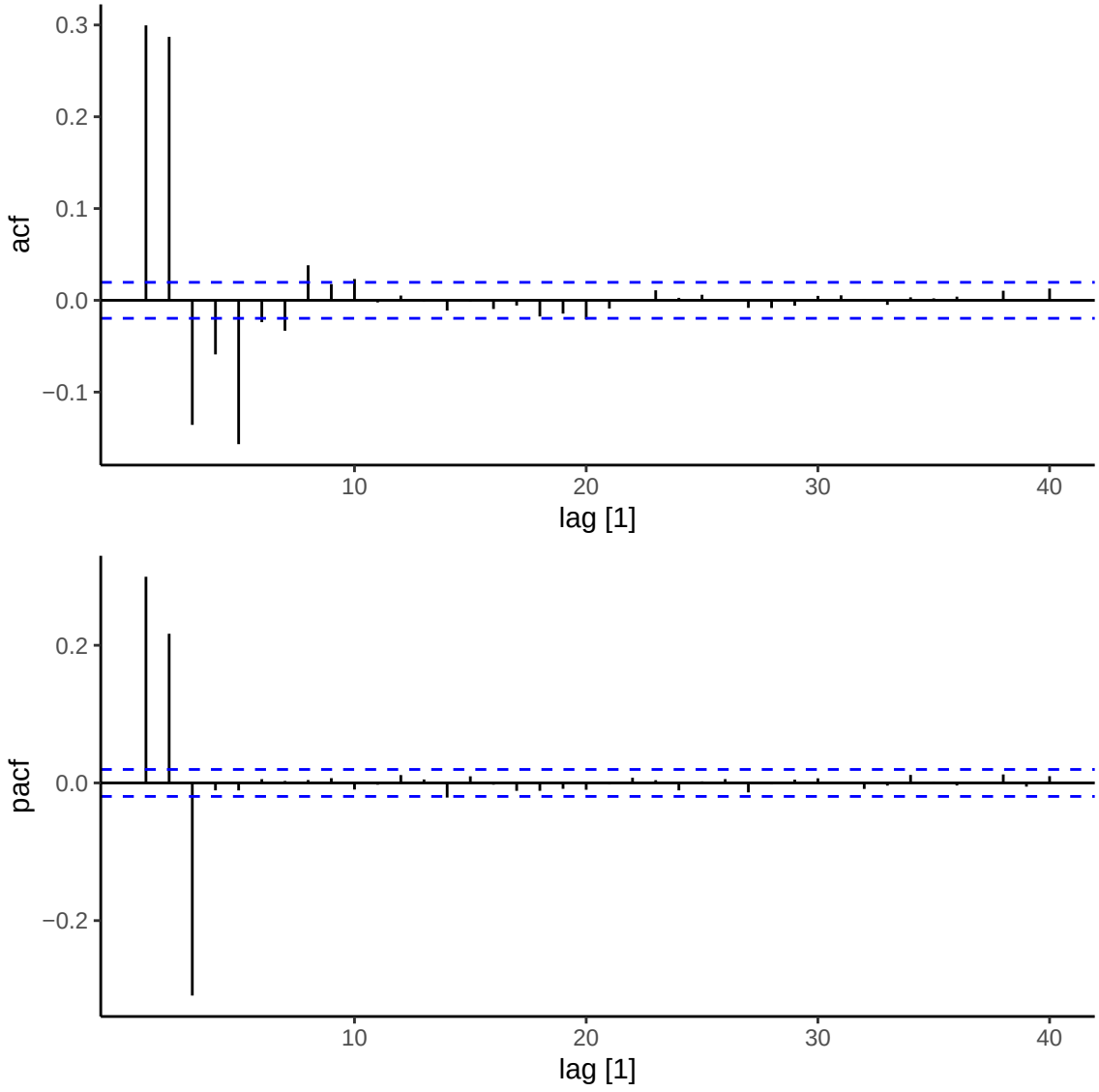



```
# A tibble: 12 x 2
   kvartal gasanvändning
     <qtr>         <dbl>
 1 1984 Q1          989.
 2 1984 Q2          477.
 3 1984 Q3          234.
 4 1984 Q4          730
 5 1985 Q1         1087
 6 1985 Q2          535.
 7 1985 Q3          282.
 8 1985 Q4          788.
 9 1986 Q1         1164.
10 1986 Q2          613.
11 1986 Q3          347.
12 1986 Q4          783.
```

a) Givet att du skall använda klassisk dekomponering för att analysera tidsserien ovan, bör du använda en additiv eller multiplikativ modell?

b) Genomför en klassisk dekomponering baserat på ditt svar på a. Det vill säga, beräkna $\hat{T}_t$ och $\hat{R}_t$ för Q3 1984 till Q2 1986, samt skattningar av samtliga säsongskomponenter.

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!



Lösningsförslag - Uppgift 5

5a

Multiplikativ modell.

5b

Gasanvändningen här är *inte* logaritmerad, så det första vi behöver göra är att logaritmera alla värden.

Eftersom att vi har kvartalsdata med tydlig säsongseffekt ska vi använda 2xS-MA. (Baserat på dom logaritmerade värdena.)

Vi får då trendvärden för alla tidpunkter utom dom två första och sista.

Säsongsskattning görs genom att först beräkna den grova skattningen (medelvärde för varje säsong), och sen korrigera detta genom att subtrahera medelvärdet för samtliga säsongskomponenter.

Slutligen kan vi beräkna $\hat{R}$ för alla utom dom två första och sista tidpunkterna.

Utifrån frågan så ser det ut som att svaret skall ges på orginalskalan ($\hat{T}_t$ och inte $\log \hat{T}_t$), men det görs inga poängavdrag om svar endast ges på logskala.

Inga poängavdrag görs för enklare räknefel, så länge det är tydligt att rätt metod använts, och att alla steg har gåtts igenom.

5c

Detta är en AR(3), tre spikes på pacf och fler (avtagande) spikes på acf.