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

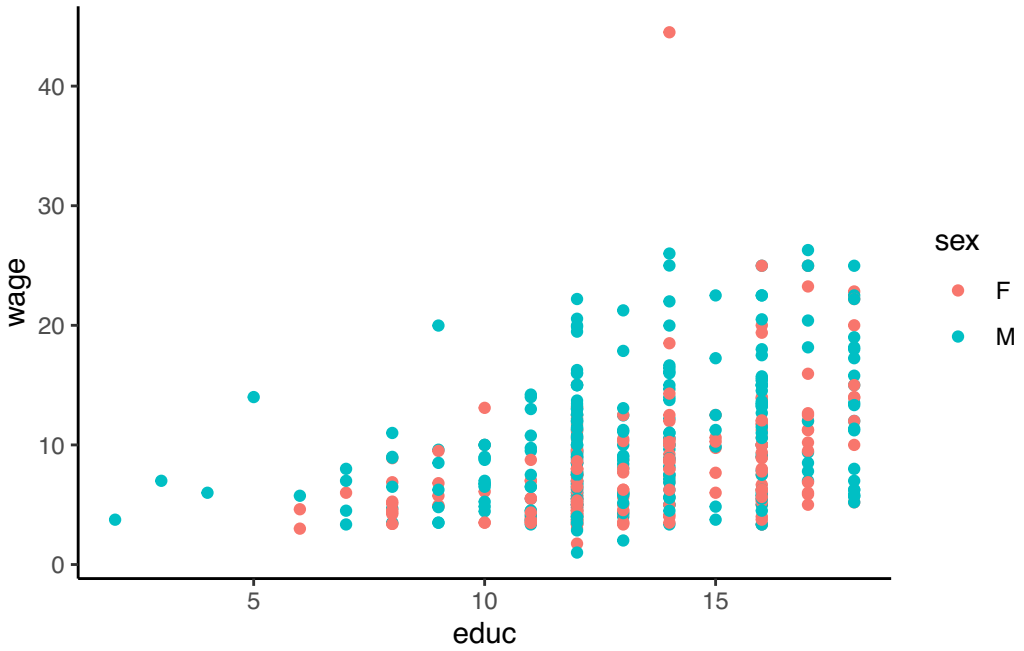
```
Analysis of variance - ANOVA
------------------------------------------------
       df      SS      MS      F     Pr(>F)
Regr    3  2677.4 892.477 41.495 4.2403e-24
Error 530 11399.3  21.508
Total 533 14076.7

Parameter estimates
------------------------------------------------
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) -3.26588    1.61919 -2.0170 4.4201e-02
educ         0.85568    0.12222  7.0011 7.7183e-12
sexM         4.37045    2.08506  2.0961 3.6549e-02
educ:sexM   -0.17253    0.15712 [redacted]
```

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

Uppgift 2 - F-test av restriktioner

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

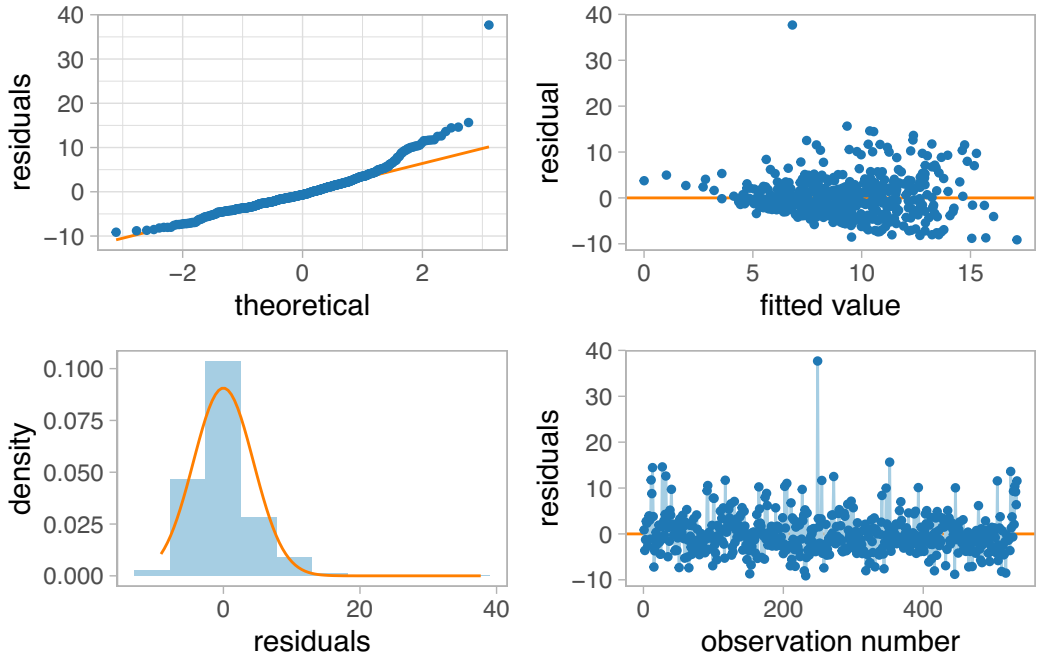
- `age` - ålder, i år
- `union` - fackligt medlemskap

```
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

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 sepalvidd 2.8 är en `versicolor`?

c) Vad är sannolikheten att en iris med sepalvidd 5.3 och sepalvidd 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.

Uppgift 4 - Cykeluthyrning

En modell har anpassats för att undersöka sambandet mellan antaletuthyrningar 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 `nRides_lagged` $= 1339.431$ och `humidity` $= 0.59$?

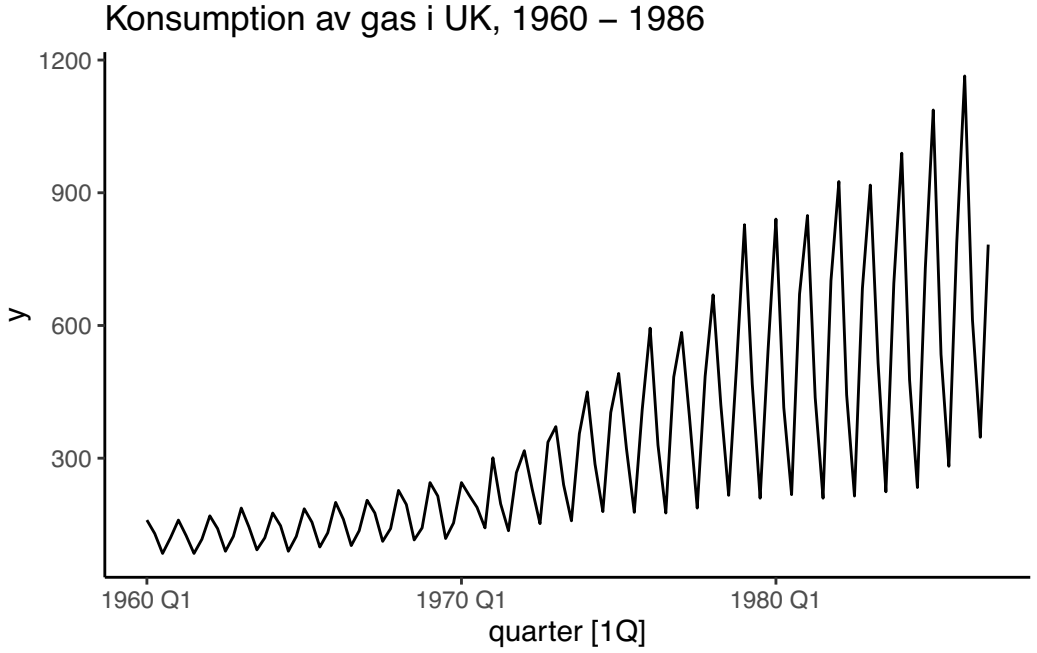
c) Tolka dom skattade parametrarna ovan, exklusive interceptet.

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

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!