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

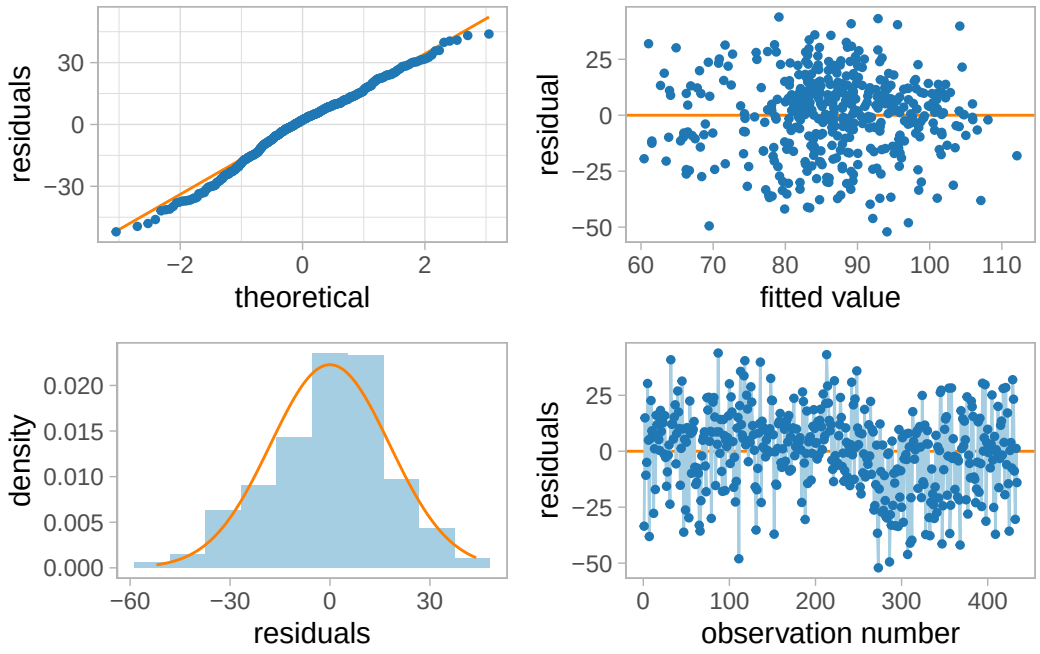
- `kid_score` Resultat på ett IQ testet för personen.
- `mom_iq` Resultat på ett IQ test för personens moder.
- `mom_hs` En dummyvariabel som antar värdet 1 om modern har en högskoleutbildning och 0 annars.

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)

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)

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)

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
```

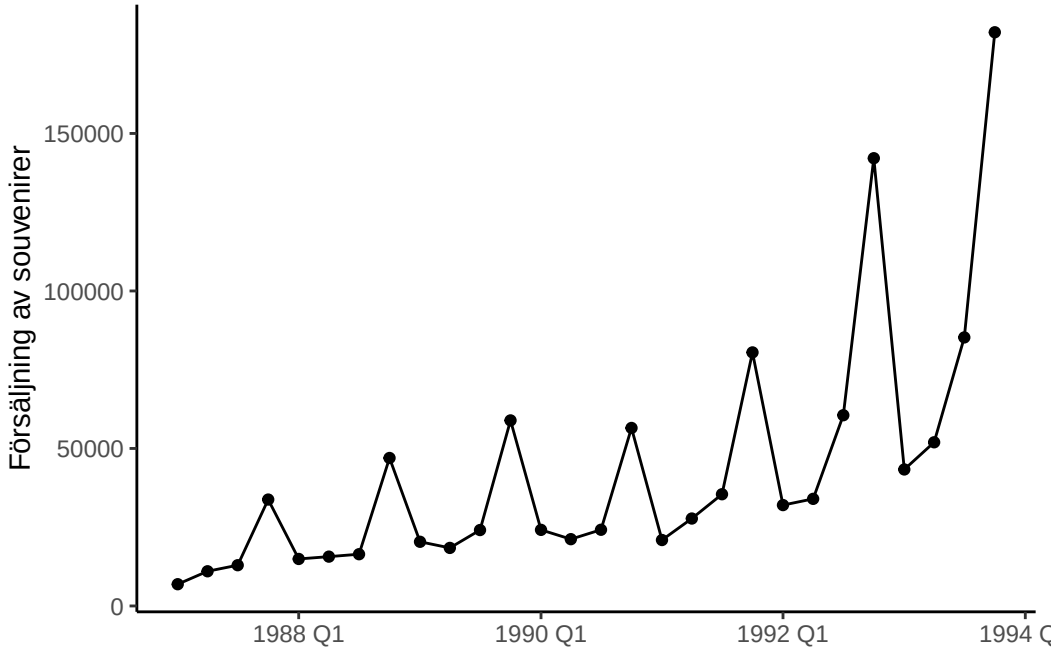
c) Enligt den skattade modellen, hur många förväntas blivit smittade efter $\texttt{t} = 12$ veckor? Under skattandet av modellen har den naturliga logaritmen, e, använts. (5p)

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)

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
```