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:

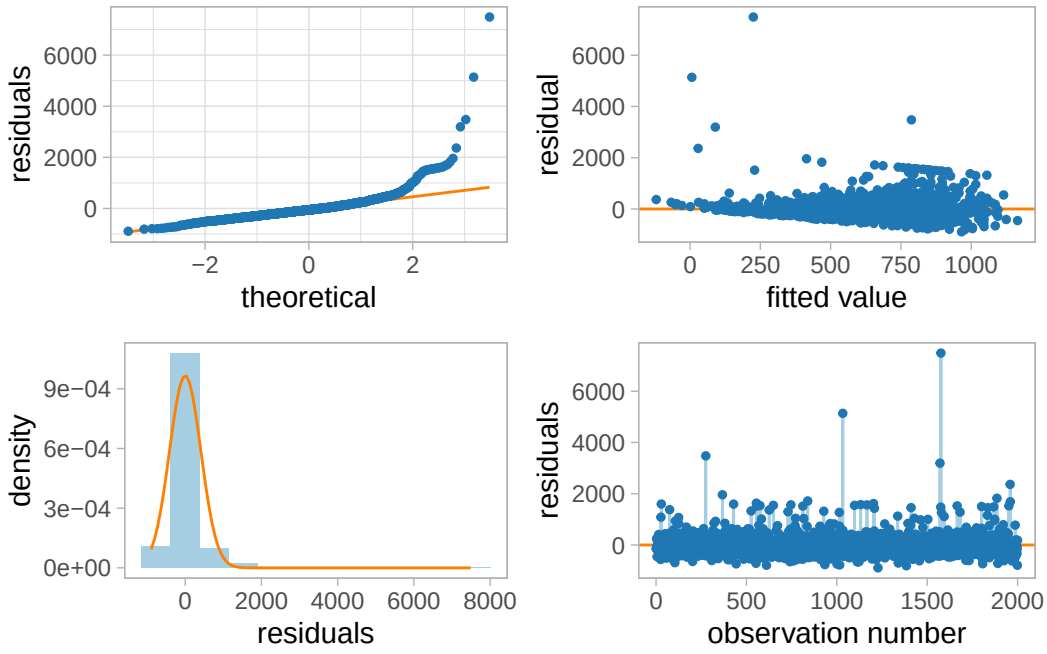
- `wage` Veckolön i dollar.
- `educ` Utbildningsnivå i år. Antar värden mellan 0 och 18 i det här datasetet.
- `exper` Arbetslivserfarenhet i år. Antar värden mellan 0 och 59 i det här datasetet.
- `smsa` Dummyvariabel som antar värdet 1 om personen lever i en "Standard Metropolitan Statistical Area", 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)

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)

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)

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)

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 for 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