SDAII (ST1201), Tentamen 1, 7.5 hp

Stockholms universitet, statistiska institutionen

Kurs: Statistik och dataanalys II, 15 hp

Tentamensdatum: 2024-01-17

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 `earnings` innehåller information om lön, utbildning, erfarenhet, längd och kön för 1605 slumpmässigt utvalda personer.

- `earnk` Årslön i tusental dollar.
- `educ` Utbildningsnivå i år. Antar värden mellan 2 och 18 i datasetet. Den genomsnittliga utbildningsnivån i datasetet är 13 år.
- `height` Längd i tum (inches). Antar värden mellan 57 och 82 i det här datasetet. Den genomsnittliga längden i datasetet är 70 tum för män och 64 tum för kvinnor.
- `male` Variabel som i detta datasetet antar två värden: 1 för män, 0 för kvinnor.

```
Analysis of variance - ANOVA
------------------------------------------------
        df     SS       MS      F     Pr(>F)
Regr     4 132030 33007.60 76.273 2.9775e-59
Error 1600 692405   432.75
Total 1604 824436
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.17	16.84	-0.54	0.59
height	-0.10	0.26	-0.38	0.70
male	-43.57	25.68	-1.70	0.09
education	2.58	0.20	12.68	0.00
height:male	0.80	0.38	NA	NA

a) Skiljer sig *sambandet* mellan längd (height) och inkomst åt mellan män och kvinnor? Genomför ett t-test. Använd $\alpha = 0.05$. (6p)

b) Tolka den skattade regressionskoefficienten för `male` (-42.37). (3p)

c) Använd den genomsnittliga utbildningsnivån och dom två genomsnittliga längderna i variabelbeskrivningen för att beräkna hur stor den skattade genomsnittliga löneskillnaden är mellan en genomsnittlig man och en genomsnittlig kvinna. (5p)

c) Ett av dom tre antagandena vi gör om feltermernas fördelning kan testas med White's test. Vilket antagande? För att genomföra White's test behöver en regressionsmodell skattas. Vad används som responsvariabel i den regressionsmodellen? (6p)

Uppgift 2 - Multipel regression (15 poäng)

En alternativ modell som innehåller två till prediktorer, vikt (`weight`) och ålder (`age`) har skattats baserat på samma dataset som i uppgift 2.

```
Analysis of variance - ANOVA
------------------------------------------------
        df     SS       MS      F     Pr(>F)
Regr     6 148216 24702.68 58.376 1.7823e-65
Error 1598 676220   423.17
Total 1604 824436

Parameter estimates
------------------------------------------------
               Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -31.9792573  17.131118 -1.86673 6.2122e-02
height        0.1012198   0.269008  0.37627 7.0677e-01
male        -38.4296935  25.461930 -1.50930 1.3142e-01
education     2.7548779   0.203623 13.52928 1.4492e-39
age           0.1886192   0.030787  6.12664 1.1280e-09
weight       -0.0054024   0.018763 -0.28793 7.7344e-01
height:male   0.7166586   0.376889  1.90151 5.7415e-02
```

a) Jämför dom två modellerna med ett F-test. Använd $\alpha = 0.01$. (10p)

b) Vad är skillnaden mellan en outlier (uteliggare) och en inflytelserik observation? (5p)

Uppgift 3 - Logistisk regression (22 poäng)

Det amerikanske presidentvalet 1992 stod mellan Bill Clinton (demokrat) och George Bush (republikan). En modell har skattats för att undersöka sambandet mellan inkomst, ålder och vilken av dom två kandidaterna en person röstade på. Variablerna i modellen beskrivs nedan.

- `rvote`: dummyvariabel som antar värdet 1 om personen röstade på Bush och 0 om personen röstade på Clinton. Personer som röstade på någon annan kandidat har sorterats bort.
- `income`: månadsinkomst i **tusentals** dollar. Om personen tjänar 2000 dollar ska alltså `income` vara 2, om personen tjänar 1500 dollar ska `income` vara 1.5 och så vidare.
- `age`: ålder.

`Parameter estimates`

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.80	0.27	-6.57	0.00
income	0.29	0.06	5.08	0.00
age	0.01	0.00	1.76	0.08

a) Dina kusiner Orvar och Chrissy röstade i valet. Orvar hade då en inkomst på 2000 dollar och var 23 år gammal, Chrissy var 51 och hade en inkomst på 4000 dollar. Beräkna dom skattade sannolikheterna att Orvar respektive Chrissy röstade för Bush. (4p)

b) Tolka interceptet i termer av odds *och* i termer av sannolikheter. Är det rimligt att tolka interceptet? Tolka parameterskattnigarna för `income` och `age` i termer av oddskvoter. (6p)

c) Det visar sig att Orvar röstade för Clinton och att Chrissy röstade för Bush. Beräkna summan av dom logaritmerade prediktionsvärdena för dom två skattade sannolikheterna i del a). I termer av logaritmerade prediktionsvärden, är modellen du använt bättre eller sämre än en modell som chansar vilt och alltid ger den skattade sannolikheten 0.5? (6p)

d) Din kompis har beräknat att sannolikheten för en viss person med en inkomst på 4000 dollar att rösta på Bush är 0.50, men vägrar berätta hur gammal personen är. Beräkna åldern på personen. (6p)

Uppgift 4 - Ickelinjär regression (18 poäng)

Populationsmodellen för utvecklingen av antal bakterier (`bact`) över tid ges av

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

där tiden `t` anges i timmar.

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
```

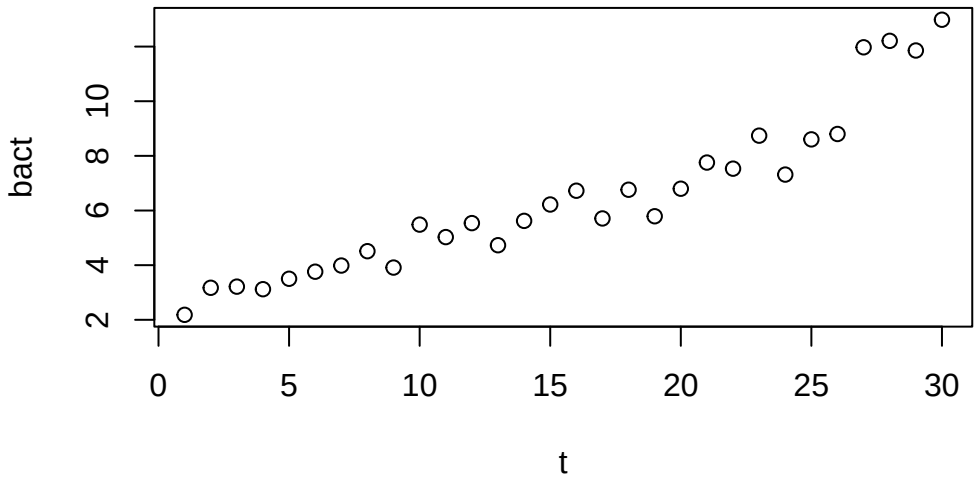
```
aa
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.994      0.042  23.807        0
t              0.050      0.002  21.205        0
```

c) Enligt den skattade modellen, hur många bakterier förväntas det finnas efter `t` = 12 timmar? Under skattandet av modellen har den naturliga logaritmen, e, använts. (5p)

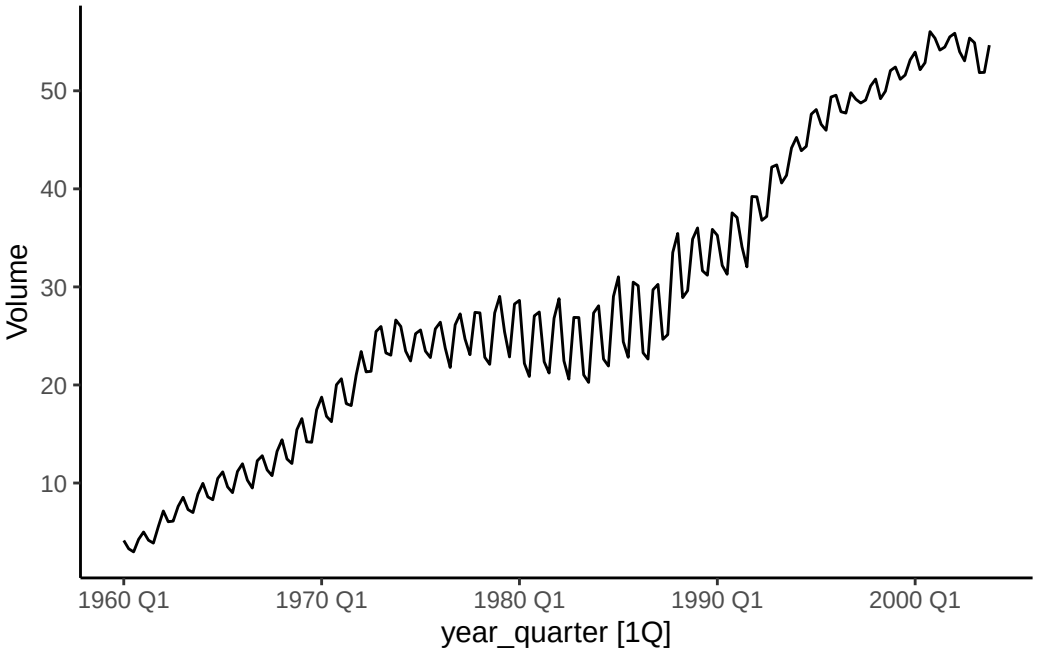
d) Tolka den skattade regressionskoefficienten för `t` i termer av orginalmodellen. (Alltså, det är inte OK att tolka estimatet i termer av log-bakterier.) (5p)

e) Baserat på plotten nedan, nämn en annan lämplig modell du skulle kunna använda för att beskriva sambandet mellan `t` och `bact`. (Hint: inte vanlig linjär regression!) (3p)



Uppgift 5 - Tidsserieanalys (25 poäng)

Figuren nedan visar den kvartalsvisa produktionen av naturgas i Kanada (i miljarder kubikmeter).



	year_quarter	Volume
1	1960 Q3	2.97
2	1960 Q4	4.26
3	1961 Q1	5.01
4	1961 Q2	4.17
5	1961 Q3	3.87
6	1961 Q4	5.57
7	1962 Q1	7.15
8	1962 Q2	6.05

a) Om du skulle göra en klassisk dekomponering av tidsserien ovan, skulle du föredra en additiv eller multiplikativ modell? Skulle den innehålla trend och/eller säsong? Motivera dina svar. (3p)

b) Beräkna trendskattningar för hela år 1961 baserat på ditt svar på a). Givet att trendskattningen för första kvartalet 1962 är 5.941762, ge en grov approximering av säsongskomponenten för det första kvartalet. (8p)

c) Utifrån figuren ovan, finns det något problem med att använda klassisk dekomponering? (4p)

d) Nedan finner du skattade värden från en AR(1)-modell. Givet att $y_T = 0.8$, ta fram prediktioner för y_{T+1} och y_{T+2}. (5p)

	Estimate	Std. Error	z-ratio	Pr(>|z|)	2.5 %	97.5 %
ar1	0.40	0.01	43.67	0	0.38	0.42
mean	7.19	0.02	428.57	0	7.15	7.22

e) 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! (5p)