Stockholms universitet

OBS! Läs noga igenom anvisningarna i tentamen, t.ex. hur du ska skriva svaren. Det är ditt ansvar som student att följa de anvisningar som ges.

NOTE! Read the examination instructions carefully, e.g. how to write the answers. It is your responsibility as a student to follow the given instructions.

Skriv din anonymiseringskod och dagens datum på allt material du lämnar in.
(*Enter your anonymization code and todays date on all submitted materials*)

Anonymiseringskod (*Anonymization code*)	3	1	1	–	0	0	4	0	–	Z	H	G

Datum (*Date* YYYY-MM-DD)	2021-11-29	Plats nr. (*Seat No.*)	12

Kurs/Kurskod (*Course/Course code*)	ST123G
Kursmoment (*Course component*)	Regressions och tidsserieanalys

Fylls i av tentamensvärd (*To be filled in by invigilator*)

Direkt i skrivning: (kryss)		Svarsblankett: (kryss)		Lösa svarsblad: (antal)	11.

Lämnat in blankt: (kryss)		Dator: (kryss)	

Inlämningstid: 17:52 Signatur tentamensvärd: HS2

Fylls i av lärare/examinator (*To be filled in by teacher/examinator*)

Betyg:	A	Poäng:	90

29 28 7 15 11

Signatur rättande lärare/examinator: MW

1. a) Vi behöver följande summor:

$\Sigma x_i = 8$ $\Sigma y_i = 52$ $\Sigma x_i y_i = 60$

$\Sigma x_i^2 = 12$ $\Sigma y_i^2 = 358$ $n = 8$

Med minsta kvadrat metoden kan vi få fram b_1

$$b_1 = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x)^2} = \frac{64}{32} = 2$$ (Skattning av β_1)

Med b_1 kan vi få fram a

$a = \bar{y} - b_1\bar{x} = 6{,}5 - 2 \cdot 1 = 4{,}5$ (Skattning av β_0)

1.b) b_1:s koefficient säger att en ökning med en timmas fysisk aktivitet ökar oavbruten sömn med två timmar.

Det måste dock tas med försiktighet, vi har visat en korrelation, men inte kausalitet.

Det kan vara så att de som sover längre och tränar mer

a ger oss ~~värdet~~ mängden sömn en individ som inte utför någon fysisk aktivitet uppskattas ha.

I många fall ger interceptet ingen särskilt användbar information, om datat inte innehåller x-värden med noll t.ex. (extrapolering)

I detta fall gör datat det, så man kan säga att att de som inte tränar sover 4,5 timmar.

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

Lärarens kommentar: (Teacher's note)

1. b) fort. Vanligtvis vill man också testa koefficienterna innan man hävdar att det råder något samband.

1.c) σ_ε^2 är variansen för feltermerna. Den är ett mått på det teoretiska "bruset" som gör att inte ens vår teoretiska modell kan ge perfekta prediktioner. Våra modeller ställer särskilda krav på σ_ε^2:s egenskaper, t.ex att den är normalfördelad med väntevärde noll. Att feltermerna är oberoende av varandra, och att variansen är konstant. (Homoskedastisk)

Tyvärr är dessa feltermer omöjliga att komma åt, så vårt enda sätt att undersöka dom är att använda våra residualer för att bilda en skattning.

Vi skattar σ_ε^2 med s_e^2.

$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-k-1} = \frac{4}{6} = \frac{2}{3}$$

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

Lärarens kommentar: (Teacher's note)

1.d) $s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{2/3}{4} = \frac{1}{6}$

$s_{b_1} = \sqrt{\frac{1}{6}} \approx 0,40825$

Det visar hur mycket b_1 kan variera.
Om s_{b_1} är väldigt hög gör det skattningen mer osäker och konfidens intervallet blir större.

1.e) Jag väljer att testa b_1. (T-test)

Hypoteser

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$\alpha = 0,05$, tvåsidigt test så vi använder $\frac{\alpha}{2} = 0,025$

$\frac{b_1 - \beta_0}{s_{b_1}} = \frac{2}{\sqrt{1/6}} \approx 4,89898$

$T_{obs} = 4,89898 > T_{krit} = T_{8-1-1, 0,975} = T_{6, 0,975} = 2,447$

Förkasta H_0, b_1 är signifikant under en 5% signifikansnivå.
Det betyder att ~~[illegible]~~ timmar av fysisk aktivitet har en effekt på timmar sömn, med en 5% risk att sambandet ej gäller och bara visades av en slump.
Vi kan inte bevara kausalitet.

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

2.a) T-test $\frac{\alpha}{2} = 0,025$

Hypoteser

$\begin{cases} H_0: \beta_3 = 0 \\ H_A: \beta_3 \neq 0 \end{cases}$ $T_{obs} = \frac{0,64817 - 0}{0,48619} \approx 1,3332$

(approx. till 2 kan)

$T_{obs} \approx 1,3332 < T_{krit} = T_{n-k-1, \alpha/2} = T_{100-3-1, 0,975} = T_{96, 0,975} \approx 1,96$

Vi kan ej förkasta H_0, med vår signifikansnivå är b_3 insignifikant.

Tolkning av koefficienten: Balkong är en dummyvariabel, när den är lika med ett ökar $\hat{y}$ med 0,64817. Man kan se det som att interceptet ökar med summan av koefficienten.

2.b) KI: $b_1 \pm T_{96, 0,975} \cdot s_{b_1}$ Använder T för mer precisa siffror

(A) $0,05978 - 1,985 \cdot s_{b_1} = 0,03994$ $s_{b_{1A}} = 0,009994622$

(B) $0,05978 + 1,985 \cdot s_{b_1} = 0,07961$ $s_{b_{1B}} = 0,009989244$

approx. s_{b_1} $\frac{s_{b_{1A}} + s_{b_{1B}}}{2} = 0,009992433$

Hypoteser T-test, $\frac{\alpha}{2} = 0,025$

$\begin{cases} H_0: \beta_1 = 0 \\ H_A: \beta_1 \neq 0 \end{cases}$ $T_{obs} = \frac{0,05978 - 0}{0,009992433} \approx 5,9825$

$T_{obs} \approx 5,9825 > 1,985 \approx T_{krit}$

Förkasta H_0, b_1 signifikant.

Samma slutsats kan dras direkt från "95% conf. int.". Eftersom den utnyttjar en 5% signifikans för att skapa →

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

2.h) föreställning.

→ intervallet. Eftersom intervallet inte inte innehåller 0 vet vi med 95% konfidens att b_1 är mellan 0,039944 och 0,07961.

2C) vi ser att rum och krm har ett höst VIF, över 10.

Dom har alltså ett starkt samband, vilket är intuitivt. Ett problem som kan uppstå i sådana fall är att koefficienter blir överskattade. Här ser det ut som att antalrum (b_2) är negativ pga att b_1 är överskattad, och b_2 "balanserar".

Man ska också komma ihåg att b_2 är insignifikant, så koefficienten bör förkastas.

Regressionen förbättras nog av att ta bort antingen krm eller antalrum, eftersom dom är korrelerade kommer en beskriva en stor del av bådas påverkan.

Förmodligen är det ändå bättre att välja bort antalrum.

Lärarens kommentar: (Teacher's note)

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

Lärarens kommentar: (Teacher's note)

2.d) $R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = \frac{582,2127}{732,29457} = 0,79505$

Detta modellens förklaringsvärde, men R^2 kan ökas av att lägga till onödiga koefficienter. Därför används ofta R^2_{adj} för multipel regression.

$R^2_{adj} = 1 - \frac{MSE}{SST/(n-1)} = \frac{1,56335}{7,396915} \approx 0,211352$

Alltså beskriver modellen ca 21,1352% av variationen av y. Vi ser här en väldig skillnad i R^2 och R^2_{adj}.

2.e) F-test med 5% sign. $\alpha = 0,05$

Hypoteser

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_A: \beta_1$ eller β_2 eller $\beta_3 \neq 0$

$F_{obs} = \frac{MSR}{MSE} \approx 124,138$

$F_{obs} = 124,138 > F_{krit} = F_{k, n-k-1} = F_{3,96} \approx 2,7$

$\alpha = 0,05$

Förkasta H_0, åtminstone en av b_1, b_2, b_3 är signifikant under en signif.nivå på 5%.

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

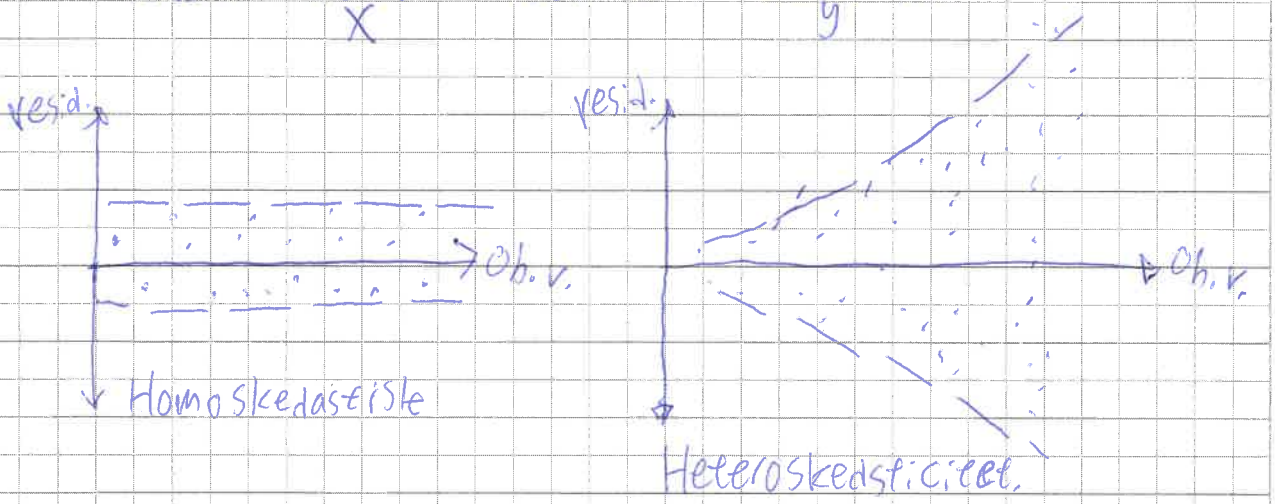
Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

3.a) Variansen är konstant när den oberoende variabeln och förändras

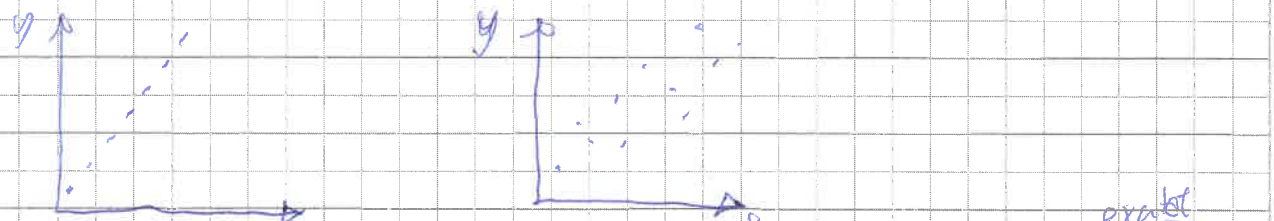
Oberoende variabel och residualer

x y



3.b) Mäter linjärt samband, men inte lutning.

En beräkning på styrkan av det linjära sambandet.



Korrelation mellan och 1 (inte 0 och inte 1) exakt

3.c) test för tidsserier

Auto Korrelation?

Ska vara mellan $\frac{2}{\text{två}}$ och $\frac{\text{fyra}}{4}$

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

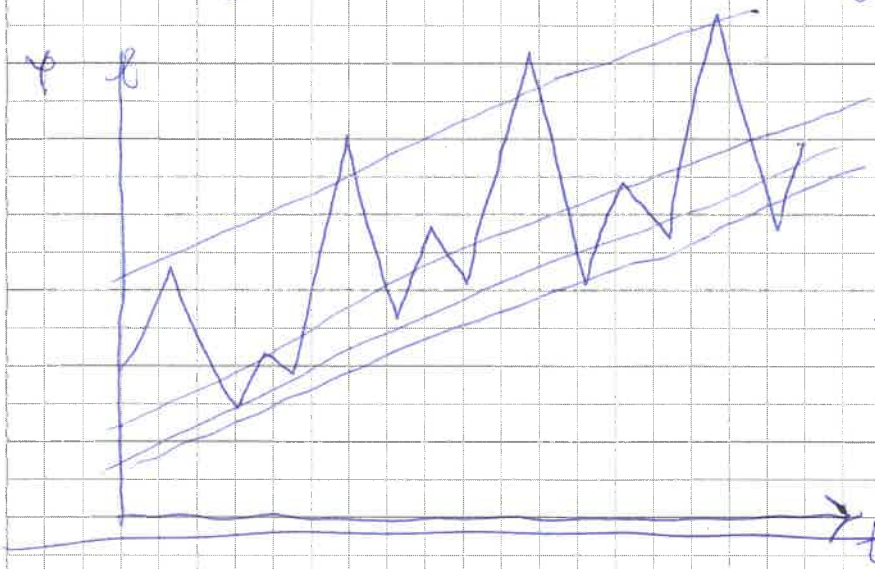
Lärarens kommentar: (Teacher's note)

3,d) En uppskattning av trenden i tidsserier.

~~Man väljer per säsong~~

Ander den trenden för att räkna ut $y_t - \hat{T}_t$.

Säsongs indelar $y_t - \hat{T}_t$ och tar medelvärdet för säsongen för att få en skattning av säsongseffekten.



$\hat{y}_1 = \hat{T}_t + \hat{S}_1 + C_0$

$\hat{y}_2 = \hat{T}_t + \hat{S}_2 + C_0$

$\hat{y}_3 = \hat{T}_t + \hat{S}_3 + C_0$

$\hat{y}_4 = \hat{T}_t + \hat{S}_4 + C_0$

3,e) Till för att välja antal Ob.variabler.

Lägger en restriktion på antal ob.v (λ).

Bra för ~~data set~~ enormt dataset med massa variabler. Också bra för att välja polynomgrad. Alltså $x_1^2, x_1^3 \ldots x_1^{27}$ t.ex..

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Stockholms universitet

Datum: (Date YYYY-MM-DD) 2021-11-29
Kurs/Kurskod: (Course/Course code) ST123G
Sidnr.: (Page no.) 9
Anonymiseringskod (Anonymization code) 311-0040-2HC

Uppg.nr.: (Task no.)
Lärarens kommentar: (Teacher's note)

4. a) En autokorrelationsfunktion visar hur mycket y_t korrelerar med y_{t-i} ~~Tu~~ man kan kolla korrelationen så långt tillbaka som datat tillåter.

~~$y_t (y_{t-i})$~~ ~~Man kan ha fler laggade y värden samtidigt~~

~~Från figuren ser vi att konf. dens intervallet skapas med vår första laggning, y_{t-1}. Och att det ökar efter den fjärde, y_{t-4}. Jag tolkar detta som att en ökning i längden för KI, innebär ett mindre höjden~~

~~KI, alltså osäkrare prognos~~

y mot y har en korrelation på 1, den minskar sedan men ökar vid y mot y_{t-4}, det är ett tecken på att det finns en säsongseffekt ~~av~~ var fjärde kvartal.

Osäkerheten i korrelationen ser ut att öka längre bak i tiden, pga KI.

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

Uppg.nr.: (Task no.)

Lärarens kommentar: (Teacher's note)

4.b) Eftersom det ser ut att finnas en säsongseffekt är det nog bäst att ta modell AR(4).
Här kan vi också se att y_{t-4} har en signifikant koefficient. Vi ser också i figuren att vi, inte har något tydligt samband hos residualerna. Det är ett tecken på att vi har lyckats säsongsrensa/anpassa modellen

4.c) $\hat{y}_{2021K1} = 62,2752 + 0,3321 \cdot y_{2020K4} + 0,0137 \cdot y_{2020K3}$

$\hat{y}_{2021K1} = 96,852102$

$\hat{y}_{2021K2} = 62,2752 + 0,3321 \cdot \hat{y}_{2021K1} + 0,0137 \cdot y_{2020K4}$

$\hat{y}_{2021K2} = 95,80772807$

Poäng: (Points)

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)

5.a) $OR(X_1) = \frac{0,008739521}{0,008433866} = 1,036241267$

Ålder ökar sannolikheten att ~~få~~ ha diabetes med ungefär 3,6% per år.

5.b) Logaritmen av en logistisk regression blir linjär. Är icke linjär då man använder den i sin exponentiella form. Beräknas ofta med maximum likelihood

5.c)

$e^{(-5,0009 - 0,0101 \cdot 100 + 0,1062 \cdot 40 + 0,0356 \cdot 60 + 0,0937 \cdot 4)} =$

$= 2,112558981$

$P(\text{Diabetes}) = \frac{2,112558981}{3,112558981} = 0,678720947$

Uppg.nr.:
(*Task no.*)

Lärarens kommentar:
(*Teacher's note*)

Poäng:
(*Points*)