STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)
Bo Wallentin

VT2019 A-B

TENTAMEN I GRUNDLÄGGANDE STATISTIK FÖR EKONOMER

2019-04-25

Skrivtid:	kl. 15.00 - 20.00
Godkända hjälpmedel:	Miniräknare utan lagrade formler och text
Bifogade hjälpmedel:	Häftet *Formelsamling och Tabeller över statistiska fördelningar* (återlämnas efter skrivningen)

- Tentamen består av 7 uppgifter, i förekommande fall uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.
- **Uppgift 1 – 5**: Svar lämnas på särskild **SVARSBILAGA**,
  - Totalt 12 flervalsfrågor där ett av fem alternativ är korrekt svar.
  - Har fler än ett svarsalternativ markerats för en deluppgift ges noll poäng.
  - Uträkningar lämnas ej in för dessa, om uträkningar ändå lämnas in kommer de inte att beaktas vid bedömningen.
- **Uppgift 6 – 7**: Svar med **FULLSTÄNDIGA REDOVISNINGAR** ska lämnas in.
  - 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!
- Tentamen kan maximalt ge 60 + 40 = 100 poäng och för godkänt resultat krävs minst 50.
- Betygsgränser:

  A: 90 – 100 p
  B: 80 – 89 p
  C: 70 – 79 p
  D: 60 – 69 p
  E: 50 – 59 p
  Fx: 40 – 49 p
  F: 0 – 40 p

  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 Mondo kort efter tentamen.

LYCKA TILL!

Endast svar ska anges för uppgifterna 1-5 med deluppgifter. Markera rätt svar på Svarsbilagan. OBS! Om du inte hittar rätt svar bland de givna svarsalternativen eller anser att något är fel med uppgiften, skriv ditt svar på Svarsbilagan vid den aktuella frågan och kommentera på baksidan av bladet.

Uppgift 1

RUFS är benämningen på den regionala utvecklingsplanen för Stockholmsregionen och ligger ytterst under Stockholms läns landsting. I en rapport från RUFS presenteras statistik kring hushållens bilinnehav i Stockholms län och i den definieras att antal olika variabler, bl.a.

- Antal bilar i trafik som disponeras av hushållet (ägs, leasas, tjänstebil osv.)
- Hushållets geografiska lokalisering (ort, kodat med s.k. ISP-kod)
- Hustyp (villa/radhus eller flerfamiljshus)
- Ålder (ungdoms-, medelålders- eller pensionärshushåll)
- Hushållsstorlek (antal vuxna i hushållet)
- Hushållets förvärvssituation (fem kategorier: ensam och arbetar respektive arbetar inte, alternativt sammanboende med underalternativen ingen arbetar, en arbetar respektive bägge arbetar).
- Inkomst

a) Utifrån ovanstående beskrivningar av variablerna vilket av följande påståenden anser du vara falsk, dvs. inte korrekt? (5p)

A. Hushållets geografiska lokalisering är en kategorisk variabel på nominal nivå.

B. Förvärvssituationen kan beskrivas med ett s.k. *five-number-summary*.

C. Inkomst kan med fördel åskådliggöras med ett histogram.

D. För antalet bilar kan medelvärde och varians beräknas utan problem.

E. Ålder är enligt definitionen en kategorisk variabel på ordinal skala.

På sista sidan hittar du en tabell som visar fördelningen över antalet bilar per hushåll för hela Stockholms kommun och för olika delar av kommunen. Notera att antalen är givna i 1000-tal.

b) Vad är det genomsnittliga antalet bilar per hushåll i Västra Stockholm? (5p)

A. 0,657

B. 0,194

C. 1,500

D. 0,806

E. 1,000

Uppgift 2

Anta att X och Y är två stokastiska variabler med simultan sannolikhetsfunktion enligt följande tabell:

		y		
		0	**1**	**2**
x	**0**	0,0	0,1	0,1
	1	0,1	0,2	0,5

a) Beräkna korrelationskoefficienten $Corr(X, Y)$ och ange korrekt slutsats om beroendet mellan X och Y. (5p)

A. $Corr(X, Y) = 0$ X och Y är beroende

B. $Corr(X, Y) = 1{,}2$ X och Y är oberoende

C. $Corr(X, Y) = 0{,}522$ X och Y är beroende

D. $Corr(X, Y) = 0$ X och Y är oberoende

E. $Corr(X, Y) = 0{,}522$ X och Y är oberoende

b) Beräkna det betingande väntevärdet och den betingade variansen för Y givet att $X = 1$. (5p)

A. $E(Y|X = 1) = 1{,}5$ $Var(Y|X = 1) = 0{,}25$

B. $E(Y|X = 1) = 1{,}0$ $Var(Y|X = 1) = 0{,}667$

C. $E(Y|X = 1) = 1{,}5$ $Var(Y|X = 1) = 0{,}5$

D. $E(Y|X = 1) = 1{,}0$ $Var(Y|X = 1) = 0{,}667$

E. $E(Y|X = 1) = 1{,}5$ $Var(Y|X = 1) = 0{,}5$

Uppgift 3

Enligt Ekonomifakta är andelen kvinnor i Sverige som är i sysselsättning 80 % (jämfört med 84 % av männen). Anta att du frågar $n = 16$ slumpmässigt valda kvinnor om de är i sysselsättning. Utgå ifrån att deras svar är korrekta och att de svarar oberoende av varandra.

a) Vad är sannolikheten att minst 4 svarar "nej" på frågan? (5p)

A. 0,402

B. 0,200

C. 0,120

D. 0,798

E. 0,598

Carl och David arbetar på samma kontor men de är ofta sena. En statistikintresserad medarbetare har analyserat deras ankomsttider och kommit fram till att det antal minuter de kommer sent till jobbet kan beskrivas med en slumpmodell. Låt X = antalet minuter Carl är sen, och Y = antalet minuter David är sen en viss dag. För enkelhet antas det att $X \sim N(30;\ 30^2)$ och $Y \sim N(50;\ 40^2)$. Det kan dessutom antas att X och Y är oberoende eftersom de bor på olika håll.

b) Vad är sannolikheten att <u>båda</u> kommer för sent en viss dag? (5p)

A. 0,263

B. 0,752

C. 0,983

D. 0,822

E. 0,513

Medarbetaren klockar i smyg Carls ankomsttider under en arbetsvecka och antecknar $n = 5$ mätvärden för Carl. Anta att ankomsttiderna för de fem dagarna är oberoende av varandra.

c) Vad är sannolikheten att den genomsnittliga förseningen under fem dagar är mindre än 20 minuter? (5p)

A. 0,227

B. 0,773

C. 0,047

D. 0,492

E. 0,953

OBS! Svarsalternativen i a) – c) har avrundats till 3 decimaler. Välj det som är närmast.

Uppgift 4

Som en del av ett personalprojekt på ett stort företag samlade man in stickprovsdata där man bland annat mätte X = antalet arbetade timmar per anställd under olika slumpvist valda arbetsveckor. Endast heltidsanställda ingick i undersökningen och anställda som var sjukskrivna eller på semester uteslöts från analysen. Varje deltagande anställd deltog endast en gång under en slumpvist vald helgfri vecka. Baserat på ett stickprov av storlek $n = 25$, skattades medelvärdet till 42,3 och variansen till 10.

a) Vilket av följande alternativ är ett 95 % konfidensintervall för μ, det förväntade värdet? (5p)

A. (41,1; 43,5)

B. (41,2; 43,4)

C. (41,0; 43,6)

D. (38,4; 46,2)

E. (38,2; 46,4)

OBS! Svarsalternativen har avrundats till en decimal, välj det som är närmast.

Från registerdata på företaget samlade du data på närvaron bland ett urval anställda under en viss helgfri vecka. Du har beräknat ett 95 % konfidensintervall för andelen anställda som av olika skäl var frånvarande från jobbet minst en dag under arbetsveckan till (0,06; 0,14) men du har glömt bort hur stort ditt stickprov var!

b) Hur stort stickprov hade du? (5p)

A. $n = 95$

B. $n = 720$

C. $n = 110$

D. $n = 20$

E. $n = 220$

OBS! Konfidensintervallets gränser har avrundats, välj det alternativ för n som är närmast.

Uppgift 5

En ny och dyr tillverkningsprocess som vi kan kalla Y hoppas man ska ge kortare tillverkningstider än den tidigare processen som vi kan kalla X. I efterhand har man insett att om process Y lyckas tillverka en enhet i snitt 20 sekunder snabbare än X så var det värt investeringen. Följande uppgifter om tillverkningstid per enhet från respektive process har samlats in:

$$\bar{x} = 122 \qquad s_x^2 = 180 \qquad n_x = 240 \qquad \bar{y} = 100 \qquad s_y^2 = 110 \qquad n_y = 180$$

Du ska nu med hjälp av uppgifterna avgöra om differensen $\mu_X - \mu_Y$ kan anses vara större än 20 eller inte.

a) Ange korrekt noll- och mothypotes för testet. (5p)

A. $H_0: \mu_X - \mu_Y = 20$ mot $H_1: \mu_X - \mu_Y > 20$

B. $H_0: \mu_X - \mu_Y = 20$ mot $H_1: \mu_X - \mu_Y < 20$

C. $H_0: \mu_X - \mu_Y > 20$ mot $H_1: \mu_X - \mu_Y = 20$

D. $H_0: \mu_X - \mu_Y = -20$ mot $H_1: \mu_X - \mu_Y > -20$

E. $H_0: \mu_X - \mu_Y = 20$ mot $H_1: \mu_X - \mu_Y \neq 20$

b) Ange det observerade värdet på testvariabeln och den korrekta slutsatsen för $\alpha = 10\ \%$. (5p)

A. $z_{\text{obs}} = 1{,}469$; H_0 förkastas, Y är inte i genomsnitt mer än 20 sekunder snabbare

B. $z_{\text{obs}} = 1{,}714$; H_0 förkastas inte, Y är i genomsnitt mer än 20 sekunder snabbare

C. $z_{\text{obs}} = 1{,}469$; H_0 förkastas inte, Y är i genomsnitt mer än 20 sekunder snabbare

D. $z_{\text{obs}} = 1{,}714$; H_0 förkastas, Y är i genomsnitt mer än 20 sekunder snabbare

E. $z_{\text{obs}} = 1{,}714$; H_0 förkastas, Y är inte i genomsnitt mer än 20 sekunder snabbare

c) Statistiker använder olika begrepp såsom Typ I och Typ II fel, p-värde, signifikans och styrka. Ett av alternativen nedan är <u>inte</u> korrekt. Ange vilket. (5p)

A. Styrkan i ett test är sannolikheten att förkasta en falsk nollhypotes

B. Sannolikheten att förkasta en sann nollhypotes är lika med signifikansnivån

C. p-värdet är sannolikheten att man förkastar H_0 för givna data

D. Typ I fel innebär att man förkastar H_0 när H_0 är sann

E. Typ II fel innebär att man accepterar H_0 när H_0 är falsk

Fullständig redovisning krävs för följande uppgifter. Använd separata pappersark för uppgift 6 resp. uppgift 7.

Uppgift 6

En viss bank analyserade inkomna ansökningar för tre olika kreditkortstyper (Silver, Gold och Platinum) i tre olika länder (Sverige, Danmark och Norge). Man drog stickprov av storleken $n = 200$ från respektive land och fick följande siffror:

Frekvens	Kreditkortstyp		
	Silver	Gold	Platinum
Sverige	120	55	25
Danmark	110	60	30
Norge	100	65	35

a) Ställ upp lämpliga hypoteser för att testa om fördelningarna över kreditkortstyp är lika för de tre olika länderna. Ange vilka antaganden som gäller för testet, testvariabel och dess fördelning samt beslutsregel och kritiskt värde. Använd 5 % signifikansnivå. (8p)

b) Genomför testet dvs. kontrollera för eventuella tumregler, beräkna det observerade värdet på testvariabeln och fatta ditt beslut, tolka även slutsatsen i ord. (8p)

c) Förklara kortfattat skillnaden mellan ett anpassningstest (*goodness-of-fit test*) och ett oberoende-/homogenitetstest. Någon lång utläggning krävs inte, ett par meningar ska räcka för att besvara frågan. (4p)

Uppgift 7.

Efter att din kollega har slutat på det elektronikföretag där du arbetar blir du ombedd att göra klart en regressionsanalys som ska användas för en kostnad- och prissättningsmodell. Datamaterialet omfattar $n = 10$ observationer på två variabler: X_i = antal tillverkade enheter av en viss typ under en slumpmässigt vald dag i och Y_i = genomsnittlig kostnad per enhet samma dag. Man antar att Y beror på X enligt en linjär modell:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

På nästa sida hittar du den tabell som din kollega efterlämnade; tabellen innehåller data på de två variablerna samt några påbörjade beräkningar.

a) Skatta modellens parametrar och tolka de skattade parametervärdena i ord. (5p)

b) Beräkna residualvariansen och förklaringsgraden R^2 och tolka den senare, dvs. förklara kortfattat vad värdet på R^2 betyder så att din chef förstår. (6p)

c) Beräkna ett 95 % konfidensintervall för lutningskoefficienten. Ange om man kan anse att koefficienten är signifikant skild från noll eller inte. Motivera ditt svar. (5p)

d) Åskådliggör X och Y i ett lämpligt diagram tillsammans med din skattade regressionslinje och kommentara kortfattat vad du ser. Är modellen lämplig för dessa data? (4p)

Bilaga till Uppgift 1

Antalet hushåll i 1000-tal i Stockholms kommun efter antalet bilar som disponeras av hushållet; totalt hela kommunen samt uppdelat efter olika delar av kommunen (Källa: PM 1:2002, www.rufs.se/globalassets/h.-publikationer/2002_1_hushallens_bilinnehav_en_kartlaggning.pdf)

Antal hushåll	0	1	2	3	Antal hushåll totalt
Innerstaden norr	51,4	33,2	3,6	0,9	89,1
Innerstaden Södermalm	33,0	20,0	0,6	0,0	53,6
Järvafältet	8,9	8,1	0,5	0,1	17,6
Sydvästra Stockholm	24,2	26,6	3,6	0,4	54,8
Sydöstra Stockholm	37,5	32,3	3,2	0,3	73,3
Västra Stockholm	21,2	32,6	6,8	1,2	61,8
Stockholm totalt	176,2	152,8	18,3	2,9	350,2

Not: Kolumnen för "3 bilar" anger egentligen antalet hushåll med "3 eller fler" bilar men för att förenkla uppgiften ska ni utgå ifrån att det är exakt 3 bilar.

Bilaga till Uppgift 7

Obs. i	1	2	3	4	5	6	7	8	9	10	Summa
x_i	3	5	5	7	9	9	11	13	13	15	90
y_i	26	18	20	14	10	15	12	10	11	14	150
x_i^2	9	25	25	49	81	81	121	169	169	225	954
y_i^2	676	324	400	196	100	225	144	100	121	196	2482
$x_i y_i$	78	90	100	98	90	135	132	130	143	210	1206

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 25/4-2019

Sal: Värtasalen

Tenta: Statistik för ekonomer

Kurs: Grundläggande statistik för ekonomer

ANONYMKOD: 0089-XXP

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN

Markera besvarade uppgifter med kryss

1	2	3	4	5	6	7	8	9	Antal inl. blad
X	X	X	X	X	X	X	X	X	3
Lär.ant. 10	5	10	10	10	18	13			

POÄNG	BETYG	Lärarens sign.
76	C	ML

STOCKHOLMS UNIVERSITET
Statistiska institutionen

VT2019 A-B

SVARSBILAGA till Tentamen i Grundläggande statistik för ekonomer
2019-04-25

Skrivsal: VÅ

Anonymkod: 0089-XXP *(skriv tydligt!)*

Markera ditt svar med ett tydligt kryss (X) i rutorna nedan.
OBS! Endast ett kryss per uppgift. Har fler än ett svarsalternativ markerats ges noll poäng.

OBS! Om du efter att ha kontrollerat dina beräkningar ordentligt kommer fram till att svaret inte finns bland de givna svarsalternativen eller anser att något är fel med uppgiften, skriv ditt svar på vid den aktuella frågan och kommentera på baksidan av detta blad.

		A	B	C	D	E	
Uppgift 1	a)		X				R
	b)				X		R
Uppgift 2	a)	X					R
	b)		X				–
Uppgift 3	a)	X					R
	b)					X	–
	c)	X					R
Uppgift 4	a)			X			R
	b)					X	R
Uppgift 5	a)	X					R
	b)				X		R
	c)	X					–

45/60

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: 0089-XX Sheet number: 1

6. a) Hypotes: H_0: Lika fördelning av kreditkortstyp mellan länderna.

mot H_1: Ej lika fördelning av kreditkortstyp mellan länderna, minst en skiljer sig. R

Antaganden: Ansökningarna i stickprovet är oberoende av varandra & lika fördelning. (iid) R + ober. storor (mellan länderna)

Då $n > 30$ kan vi med hjälp av CGS approximera stickprovet. ✓

Testvariabel: $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ där $E_{ij} = \frac{R_i C_j}{n}$.

OK

Testvariabeln är χ^2-fördelad med $(r-1)\cdot(c-1) = (3-1)\cdot(3-1) = 4$ frihetsgrader.

Kritiskt värde: $\chi^2_{4;0,05} = 9,488$ R

7

Beslutsregel: om $X^2_{obs} > X^2_{krit} = 9,488$ så förkastar vi H_0.

b)

O_{ij}	Silver	Gold	Platinum	tot
Sverige	120	55	25	200
Danmark	110	60	30	200
Norge	100	65	35	200
	330	180	90	600

E_{ij}	Silver	Gold	Platinum
Sverige	110	60	30
Danmark	110	60	30
Norge	110	60	30

$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	Silver	Gold	Platinum	tot
Sverige	0,9091	0,4167	0,8333	2,1591
Danmark	0	0	0	0
Norge	0,9091	0,4167	0,8333	2,1591
				4,3182 R

Då $X^2_{obs} = 4,318 < X^2_{krit} = 9,488$ så kommer vi alltså inte att förkasta H_0. Dvs. att vi kan inte på 5% signifikansnivå säga att ~~testvariabeln är skild från noll.~~ Den är skild från noll. Detta betyder alltså att vi inte kan utesluta att fördelningen av kreditkortstyp kan vara lika mellan länderna. R

8

c) Skillnaden mellan ett goodness-of-fit test (anpassningstest) $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ där $E_i = np_i$ och ett homogenitetstest/oberoende-test $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ där $E_{ij} = \frac{R_i C_j}{n}$ är att i ett anpassningstest vill man undersöka om fördelningen av en ~~kategori~~ variabel (ex glassmåker, dock uppdelad i olika kolumner med smaker) kan vara lika. Alltså att alla undersökta smaker är lika populära.

I homogenitetstestet/oberoendetestet jämför man sambandet mellan ~~flera~~ två olika ~~kategorier~~ variabler.

3

18

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: 0089-XXP Sheet number: 2

7.

a) $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$b_1 = \frac{Cov(x,y)}{S_x^2}$$

$\bar{x} = 90/10 = 9 \quad \bar{y} = 150/10 = 15$

$$Cov(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{1206 - 10 \cdot 9 \cdot 15}{10-1} = \frac{1206-1350}{9} = -16$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} = \frac{36+16+16+4+0+0+4+16+16+36}{10-1} = \frac{144}{9} = 16$$

$$b_1 = \frac{Cov(x,y)}{S_x^2} = \frac{-16}{16} = -1$$

R

$b_0 = \bar{y} - b_1\bar{x} = 15 + 1 \cdot 9 = 24$

Detta betyder att grund-tillverkningskostnaden per enhet är 24 kr. Men då b_1 är negativ betyder det att för varje enhet tillverkad så kommer den genomsnittliga kostnaden per enhet minska.

Ju fler enheter desto lägre genomsnittlig kostnad per enhet. /4

b) Residualvarians: $S_e^2 = \frac{SSE}{n-k-1}$

$SSE = \sum e_i^2 \quad e_i = y - \hat{y} \quad \hat{y} = b_0 + b_1 x_i$

e_i		e_i^2
$26-(15-1\cdot3)=14$	$\longrightarrow$	196
$18-(15-1\cdot5)=8$	$\longrightarrow$	64
$20-(15-1\cdot5)=10$	$\longrightarrow$	100
$14-(15-1\cdot7)=6$	$\longrightarrow$	36
$10-(15-1\cdot9)=4$	$\longrightarrow$	16
$15-(15-1\cdot9)=9$	$\longrightarrow$	81
$12-(15-1\cdot11)=8$	$\longrightarrow$	64
$10-(15-1\cdot13)=8$	$\longrightarrow$	64
$11-(15-1\cdot13)=9$	$\longrightarrow$	81
$14-(15-1\cdot15)=14$	$\longrightarrow$	196
		898 = SSE

$= \sum e_i = 0$; kolll!

mitt fel

15? $b_0 = 24$ enl a)

Residualvariansen $S_e^2 = \frac{SSE}{n-k-1} = \frac{898}{10-1-1} = \underline{112,25}$

Förklaringsgrad $R^2 = 1 - \frac{SSE}{SST}$

$SST = \sum (y_i - \bar{y})^2$

$(y_i - \bar{y})^2 =$
$(26-15)^2 = 121$
$(18-15)^2 = 9$
$(20-15)^2 = 25$
$(14-15)^2 = 1$
$(10-15)^2 = 25$
$(15-15)^2 = 0$
$(12-15)^2 = 9$
$(10-15)^2 = 25$
$(11-15)^2 = 16$
$(14-15)^2 = 1$
$232 = SST$

$0 \leq R^2 \leq 1$

Förklaringsgrad $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{898}{232} \approx -2,871$ negativt

Förklaringsgraden beskriver de variationer i Y som kan förklaras av modellen ju högre förklaringsgrad desto bättre modell. Fler förklaringsvariabler leder oftast till en högre förklaringsgrad. Hur mycket är det i detta fall?

2

c) $b_1 \pm t_{n-k-1;\alpha/2} \cdot s_{b_1}$

$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{112,25}{(10-1)\cdot 16} = 0,7795138889$ följdfel från c)

$s_{b_1} = \sqrt{0,7795138889} = 0,88290008375$

$-1 \pm t_{8;0,025} \cdot 0,88290008375 = -1 \pm \underbrace{2,306}_{R} \cdot 0,88290008375 \approx -1 \pm 2,036$

$(-3,06 ; 1,036)$

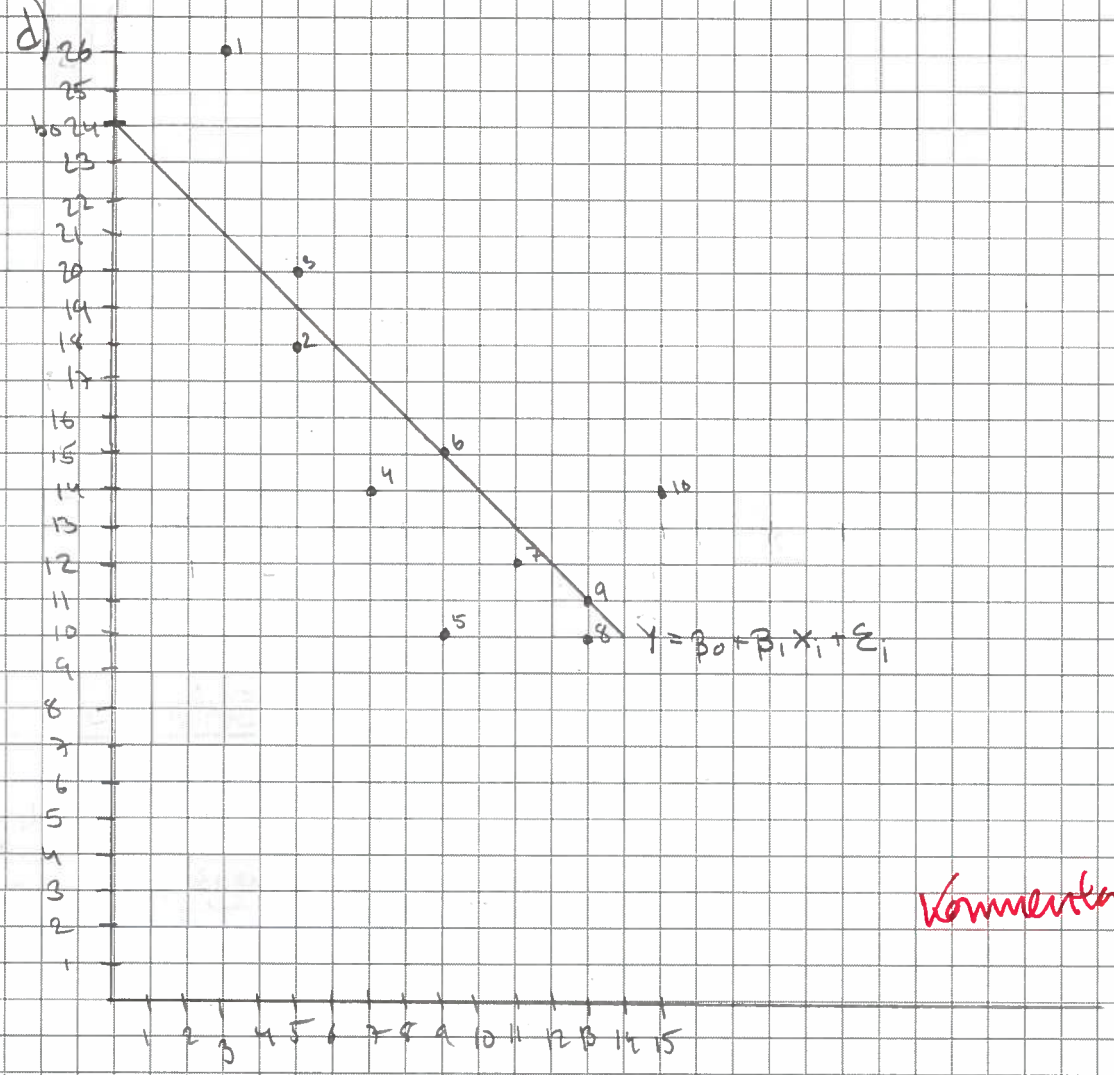
Då intervallet täcker värdet på noll så är koeficienten alltså inte signifikant skilt från noll. Rätt slutsats givet siffrorna

5-

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: 0089-XXP Sheet number: 3

7 d)



Kommentarer saknas

2