Stockholms universitet
Statistiska institutionen
Raul Cano

SKRIVNINGSDATUM: 27-04-2018

Skriftlig tentamen i **Regressionsanalys och tidsserieanalys** (4,5 hp), ingående som moment 1 i kursen **Regressionsanalys och undersökningsmetodik, 15 hp.**

Skrivtid: 5 timmar
Hjälpmedel: Miniräknare utan lagrade formler eller lagrad text. Vidhäftade formel- och tabellblad (obs! vidhäftas endast de tabellsidor som behövs för den här tentamen).
Återlämning av tentamen: måndagen den 14 maj, kl. 16.00 i B705.
Därefter kan skrivningarna hämtas på studentexpeditionen, plan 7 i B-huset.

Tentamen består av fem uppgifter som kan ge totalt 100 poäng. För betyget A gäller 90-100 p., för betyget B gäller 80-89 p., för betyget C gäller 70-79 p., för betyget D gäller 60-69 p., för betyget E gäller 50-59 p., för betyget Fx gäller 40-49 p. och för betyget F gäller 0-39 p. För detaljerade betygskriterier se kursbeskrivningen på kurshemsidan.
För full poäng på en uppgift krävs fullständiga och väl motiverade lösningar.

Uppgift 1: (30 poäng)
I Grönköping har man studerat sambandet mellan y = sparande och x = inkomst. Data som gäller:

y	8	52	30	31	30	25	8	45	3	16
x	18	44	31	38	36	26	24	51	22	30

Du har tillgång till följande information:

```
Regression Analysis:  y versus x

The regression equation is
y = - 20,9 + 1,43 x

Predictor      Coef   SE Coef
Constant    -20,862     7,711
x            1,4269    0,2304

Analysis of Variance

Source          DF        SS
Regression
Residual Error          407,0
Total                  2357,6
```

a). Testa H_0: $\beta = 0$ mot H_1: $\beta \neq 0$ (dels med t-test (5 poäng), dels med F-test (5 poäng)). Använd signifikansnivå 5% ($\alpha = 0,05$). (totalt 10 poäng)

b). Testa H_0: $\beta = 0$ mot H_1: $\beta > 0$ (med t-test). Använd signifikansnivå 5% ($\alpha = 0,05$). (5 poäng)

c). Beräkna ett 95%-igt konfidensintervall för β. (5 poäng)
d). Beräkna ett 95%-igt konfidensintervall för det förväntat y-värdet för x = 40. (5 poäng)
e). Beräkna ett 95%-igt prediktionsintervall för det enskilda y-värdet för x = 40. (5 poäng)

Uppgift 2: (20 poäng)
I Trumps Amerika har man observerat följande variabler för ett antal amerikanska familjer.
y = nysparande (dollar), x1 = inkomst (dollar), x2 = banktillgångar (dollar) och x3 = antal barn.
En regressionsanalys gav följande information:

```
The regression equation is
y = 66 + 0,120 x1 - 0,0261 x2 - 22,8 x3

Predictor        Coef      SE Coef
Constant         66,2        233,0
x1            0,12002      0,02340
x2          -0,026094     0,008527
x3             -22,76        33,21

Analysis of Variance

Source          DF       SS
Regression
Residual Error          42416
Total            8     640000
```

a). Testa på 5% signifikansnivå om de tre förklarande variablerna tillsammans kan förklara variationen i y. (10 poäng)
b). Testa på 5% signifikansnivå om $\beta_1 > 0$ i modellen. (10 poäng)

Uppgift 3: (20 poäng)
Det framgångsrika företaget AEKI (i Grönköping) har redovisat följande omsättning (i milj. kr.) under åren 2013 till 2017:

År:	2013	2014	2015	2016	2017
Omsättning:	2	4	8	10	12

a). Anpassa en exponentiell modell (exponentiell trendfunktion) till tidsserien. Tolka de skattade koefficienterna i ord och uttryckta i termer av de aktuella variablerna och sorterna. (15 poäng)
b). Enligt den anpassade modellen, gör en prognos för företagets omsättning år 2018. (5 poäng)

Uppgift 4: (20 poäng)
Banken BES (i Grönköping) har redovisat följande avkastning (i procent) på sin aktieportfölj BEST-PLUS under åren 2013 till 2017:

2013	2014	2015	2016	2017
1	2	4	5	4

a). Anpassa en andragradskurva till tidsserien med hjälp av minsta-kvadrat-metoden. (15 poäng)
b). Enligt den anpassade modellen, gör en prognos för aktieportföljens avkastning år 2018. (5 poäng)

Uppgift 5: (10 poäng)
I Grönköping har man studerat sambandet mellan lungcancer, ålder och rökvanor.
Beteckning: Y= (lungcancer = 1, ej lungcancer = 0), X_1 = ålder och
X_2 = rökvanor (tung rökare = 1, ej tung rökare = 0).
Via en logistik regression modell har man fått följande skattningar:

$\hat{Y} = -5{,}9 + 0{,}04\, X_1 + 2{,}3\, X_2$
a). Enligt modellen, beräkna den skattade sannolikheten att en person har lungcancer, om personen är en tung rökare och är 50 år gammal. (5 poäng)
b). Enligt modellen, beräkna den skattade sannolikheten att en person har lungcancer, om personen är en ej tung rökare och är 50 år gammal. (5 poäng)

Stockholms universitet
Statistiska institutionen
Regressionsanalys och undersökningsmetodik
Vårterminen 2017
Jörgen Säve-Söderbergh

Formelsamling – regressionsanalys

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

Enkel lineär regression

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$b_0 = \hat{\beta}_0 = \overline{y} - b_1\overline{x}$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SSE}}.$$

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 - b_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

$$s_e^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Konfidensintervall för β_1 ges av

$$b_1 \pm t_{n-2,\frac{\alpha}{2}} s_{b_1}.$$

där

$$s_{b_1} = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Prediktionsintervall

$$\underbrace{b_0 + b_1 x_{n+1}}_{\hat{y}_{n+1}} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}.$$

Konfidensintervall förväntat y-värde för ett nytt x-värde

$$\underbrace{b_0 + b_1 x_{n+1}}_{\hat{y}_{n+1}} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{s_e^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}.$$

Multipel regression

n st observationer och p förklarande variabler.

Variationsorsak	SS	df	MS	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	MSR/MSE
Residual	SSE	$n-p-1$	$MSE = \frac{SSE}{(n-p-1)}$	
Totalt	SST	$n-1$		

$$R_{adj}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Normalekvationerna för fallet $\hat{y} = a + b_1 t + b_2 t^2$

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= a \cdot n + b_1 \sum_{i=1}^{n} t_i + b_2 \sum_{i=1}^{n} t_i^2 \\
\sum_{i=1}^{n} y_i t_i &= a \sum_{i=1}^{n} t_i + b_1 \sum_{i=1}^{n} t_i^2 + b_2 \sum_{i=1}^{n} t_i^3 \\
\sum_{i=1}^{n} y_i t_i^2 &= a \sum_{i=1}^{n} t_i^2 + b_1 \sum_{i=1}^{n} t_i^3 + b_2 \sum_{i=1}^{n} t_i^4
\end{aligned}$$

Säsongsrensning med regression

$$a_0 = \bar{y} - b \cdot \bar{t}$$

$$T_t = a_0 + b \cdot t$$

$$\begin{aligned} S_1 &= a - a_0 + c_1 \\ S_2 &= a - a_0 + c_2 \\ S_3 &= a - a_0 + c_3 \\ S_4 &= a - a_0 \end{aligned}$$

Logistisk regression

$$P(Y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

$$\log(\text{odds}) = \beta_0 + \beta_1 x_i.$$

$$\text{odds}(D) = \frac{P(D)}{P(D \text{ inträffar inte})} = \frac{P(D)}{1 - P(D)}$$

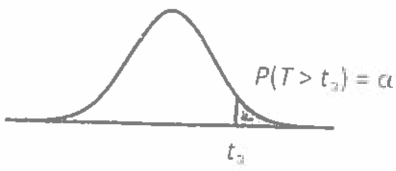
$$P(D) = \frac{\text{odds}(D)}{1 + \text{odds}(D)}$$

Konfidensintervall för oddskvoten e^{β_1}: $e^{b_1 \pm z \times s_{b_1}}$

TABELL 3. t-fördelningens kvantiler

$T \in t(v)$ där v = antal frihetsgrader.

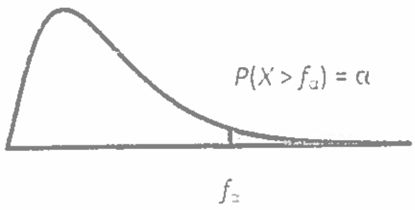
Vilket värde har t_α om $P(T > t_\alpha) = \alpha$ där α är en given sannolikhet. Utnyttja även $P(T \leq -t_\alpha) = P(T > t_\alpha)$.



v	α = 0,1	0,05	0,025	0,010	0,005	0,0025	0,0010	0,0005
1	3,078	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
55	1,297	1,673	2,004	2,396	2,668	2,925	3,245	3,476
60	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
65	1,295	1,669	1,997	2,385	2,654	2,906	3,220	3,447
70	1,294	1,667	1,994	2,381	2,648	2,899	3,211	3,435
75	1,293	1,665	1,992	2,377	2,643	2,892	3,202	3,425

Forts. nästa sida

TABELL 5. *F*-fördelningens kvantiler



$X \in F(v_1, v_2)$ där v_1, v_2 = antal frihetsgrader i täljaren respektive nämnaren. Vilket värde har f_α om $P(X > f_\alpha) = \alpha$ där α är en given sannolikhet.

$\alpha = 0{,}05$

v_2 \ v_1 =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
v_2 = 1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	244,7	245,4	245,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	2,07	2,04	2,01	1,99	1,96
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05	2,01	1,97	1,94	1,92	1,89
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,89	1,86	1,84
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,93	1,89	1,86	1,84	1,81
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,91	1,88	1,84	1,82	1,79
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67

Forts. nästa sida

(37)

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 27/04/18

Sal: Laduvikssalen

Tenta: Regressions- och tidsserieanalys

Kurs: Regressions- och undersökningsmetodik

ANONYMKOD: 0003-GCZ

☑ 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					12 RC
Lär.ant. 30p	20p	20p	20p	10p					

POÄNG	BETYG	Lärarens sign.
100p	A	RC

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 GCZ Blad nr: 1

1. a) $H_0: \beta = 0$ $H_1: \beta \neq 0$

1) 30p

t-test:

$t = \frac{b_1}{s_{b_1}}$ följer en t-fördelning med $n-2$ frihetsgrader om H_0 sann

Vi har $n = 10$ i vårt fall

Vi ska testa på signifikansnivån $\alpha = 0{,}05$

enligt tabell $t_{8;0{,}025} = 2{,}306$ R

Vi förkastar H_0 om $|t_{obs}| > 2{,}306$

$t_{obs} = \frac{b_1}{s_{b_1}} = \frac{1{,}4269}{0{,}2304} = 6{,}193 > 2{,}306$ R R R

Alltså förkastas H_0, vi har visat att $\beta \neq 0$ | R

F-test: Anova-tabellen fylls i

	DF	SS	MS	F = MSR/MSE
Regression	1	1950,6	1950,6/1 = 1950,6	$\frac{1950{,}6}{50{,}875} = 38{,}34$
Residual Error	8	407,0	407/8 = 50,875	
Total	9	2357,6		

p; n-p-1; n-1

(SST = SSR + SSE)

p = 1 (en förklarande variabel)

n = 10

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 6CZ Blad nr: 2

1. forts.

$F = \frac{MSR}{MSE}$ följer en F-fördelning med 1 frihetsgrad i täljaren och 8 i nämnaren om H_0 sann

Enligt tabell $F_{1,8,0,05} = 5,32$ R

Om $F_{obs} > 5,32$ så förkastas H_0

$F_{obs} = \frac{MSR}{MSE} = \frac{1750,6}{50,875} \approx 38,34 > 5,32$ R

Alltså förkastas H_0, vi har visat att $\beta \neq 0$ R

b) $H_0: \beta = 0$ $H_1: \beta > 0$

Samma som i a)-uppgiften fast ett ensidigt test

t_{obs} är fortfarande 6,193 R

enligt tabell $t_{8,0,05} = 1,86$ R

$t_{obs} > 1,86$ R

Alltså förkastas H_0 återigen R

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 GCZ Blad nr: 3

1. forts.

c) Konfidensintervall för β_1

$$b_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot s_{b_1}$$

$t_{8,0{,}025} = 2{,}306$

$s_{b_1} = 0{,}2304$

$b_1 = 1{,}4269$

$1{,}4269 \pm 2{,}306 \cdot 0{,}2304$

Konf intervall 95%: $[0{,}896 ; 1{,}958]$

R

R

d) se nästa blad

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 6CZ Blad nr: 4

d) konfidensintervall för förväntat y-värde för $x = 40$:

$$b_0 + b_1 X_{n+1} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{S_e^2\left(\frac{1}{n} + \frac{(X_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

$b_0 = -20{,}862$

$b_1 = 1{,}4269$

$t_{8,\,0{,}025} = 2{,}306$

$S_e^2 = MSE = 50{,}875$ (enligt tidigare uppgift)

$n = 10$

$X_{n+1} = 40$

$$\bar{x} = (18+44+31+38+36+26+24+51+22+30)/10 = 32$$

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 11198 - \frac{1}{10} \cdot 320^2 =$$

$$11198 - 10240 = \underline{\underline{958}}$$

$$-20{,}862 + 1{,}4269 \cdot 40 \pm 2{,}306\sqrt{50{,}875\left(\frac{1}{10} + \frac{(40-32)^2}{958}\right)}$$

$$-20{,}862 + 1{,}4269 \cdot 40 \pm 2{,}306\sqrt{8{,}4862}$$

$$-20{,}862 + 57{,}076 \pm 6{,}71764$$

R R R

$[29{,}496 \; ; \; 42{,}932]$ konfidensintervall

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 GCZ Blad nr: 5

e) Prediktionsintervall för det enskilda y-värdet för $x = 40$:

$$b_0 + b_1 x_{n+1} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

Samma värden på variablerna som i föregående uppgift

$$-20,862 + 1,4269 \cdot 40 \pm 2,306 \sqrt{50,875\left(1 + \frac{1}{10} + \frac{(40-32)^2}{958}\right)}$$

$$-20,862 + 1,4269 \cdot 40 \pm 2,306 \sqrt{59,361247}$$

$$-20,862 + 57,076 \pm 17,766865$$

$$[18,447 \,;\, 53,981]$$ prediktionsintervall

R

R

R

R

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 GCZ Blad nr: 6

2) 20p

2.

a) 3 förklarande variabler ⇒ DF Regression = 3

DF Total = 8 vilket innebär att vi har 9 observationer

DF Residual Error blir då 9 − 3 − 1 = 5

	DF	SS	MS	F = MSR/MSE
Regression	3	597584	199194,6667	23,48
Residual error	5	42416	8483,2	
Total	8	640000		

(SST = SSR + SSE)

H_0 $\beta_1 = \beta_2 = \beta_3 = 0$ H_1: någon av $\beta_1, \beta_2, \beta_3 \neq 0$

R

$F = \frac{MSR}{MSE}$ följer en F-fördelning med 3 frihetsgrader i täljaren och 5 i nämnaren då H_0 sann

Enligt tabell $F_{3,5,0,05} = 5,41$ R

Vi förkastar H_0 om $F_{obs} > 5,41$ R R

R

$F_{obs} = 23,48 > 5,41$ Alltså förkastar vi H_0, dvs. vår modell är inte värdelös, den kan förklara variationen y.

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 GCZ Blad nr: 7

(2) forts.

b) $H_0: \beta_1 = 0$ $H_1: \beta_1 > 0$

t-test:

$t = \frac{b_1}{S_{b_1}}$ följer en t-fördelning med $n-4$ frihetsgrader om H_0 sann.

↑ $n=9$ enligt a)-uppgiften

Vi gör ett ensidigt test. Enligt tabell

$t_{5,\,0{,}05} = 2{,}015$

$t_{obs} = \frac{0{,}12002}{0{,}02340} = 5{,}129 > 2{,}015$

H_0 kan förkastas, dvs vi har visat att $\beta > 0$

SU, STATISTIK

Skrivsal: ________ Anonymkod: 0003 GCZ Blad nr: 8

(3.)

$$\hat{y} = a \cdot b^x$$

$$\log \hat{y} = \log a + x \log b$$

$$\log \hat{y} = a' + b' x$$

eller $\log \hat{y} = a' + b' t$

3) 20p

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} \left(\sum x_i\right)^2}$$

ÅR	t	log y	t^2	t · log y
2013	−2	0,30103	4	−0,60206
2014	−1	0,60206	1	−0,60206
2015	0	0,90309	0	0
2016	1	1	1	1
2017	2	1,0792	4	2,1584
SUMMA	0	3,88536	10	1,95428

$$b_1 = \frac{\sum t \log y - \frac{1}{n} \overbrace{\sum t}^{=0} \sum \log y}{\sum t^2 - \frac{1}{n} \underbrace{\left(\sum t\right)^2}_{=0}} = \frac{\sum t \log y}{\sum t^2} = \frac{1,95428}{10} =$$

$$= 0,195428$$

SU, STATISTIK

Skrivsal: ____________ Anonymkod: 0003 GCZ Blad nr: 9

(3) forts.

$a' = \overline{\log y} - b'\,\overline{t}$ $\quad \overline{t} = 0$

$a' = \overline{\log y} = \frac{\sum \log y}{n} = \frac{3,88536}{5} = 0,777072$

R

$a = 10^{a'} = 10^{0,777072} = 5,98511$

$b = 10^{b'} = 10^{0,195428} = 1,56830$

Alltså $\boxed{\hat{y} = 5,98511 \cdot 1,56830^{t} \\ \hat{y} = a \cdot b^{t}}$ R

$t = 0 \implies y = 5,98511$ R dvs omsättningen vid $t = 0$ (år 2015) är i genomsnitt ungefär 6 miljoner kronor.

a

R

$b = 1,56830$ innebär att för varje år så ökar omsättningen i genomsnitt med 56,8 % R R

b) 2018 innebär $t = 3$

$y = 5,98511 \cdot 1,56830^{3} = 23,087$ R

R

Omsättningen kommer att bli ungefär 23 miljoner kronor 2018 om vår modell stämmer

SU, STATISTIK

Skrivsal: ________ Anonymkod: 0003 GCZ Blad nr: 10

4) 20p

(4.)

År	t	y	t^2	t^3	t^4	$y \cdot t$	yt^2
2013	-2	1	4	-8	16	-2	4
2014	-1	2	1	-1	1	-2	2
2015	0	4	0	0	0	0	0
2016	1	5	1	1	1	5	5
2017	2	4	4	8	16	8	16
Summa	0	16	10	0	34	9	27

$\hat{y} = a + b_1 t + b_2 t^2$ med normalekvationerna

$$\sum y = a \cdot n + b_1 \underbrace{\cancel{\sum t}}_{=0} + b_2 \sum t^2$$

$$\sum yt = a \underbrace{\cancel{\sum t}}_{=0} + b_1 \sum t^2 + b_2 \underbrace{\cancel{\sum t^3}}_{=0}$$

$$\sum yt^2 = a \sum t^2 + b_1 \underbrace{\cancel{\sum t^3}}_{=0} + b_2 \sum t^4$$

$$\begin{cases} 16 = 5a + 10b_2 \\ 9 = 10b_1 \\ 27 = 10a + 34b_2 \end{cases}$$

$b_1 = \frac{9}{10} = 0{,}9$

r

SU, STATISTIK

Skrivsal: ____ Anonymkod: 0003 GCZ Blad nr: 11

forts.

(4.)

$$16 = 5a + 10b_2$$ → multiplicera med (−2) och addera till andra ekvationen

$$27 = 10a + 34b_2$$

$$27 - 16 \cdot 2 = 10a - 10a + 34b_2 - 20b_2$$

$$-5 = 14b_2$$

$$b_2 = -\frac{5}{14} \approx -0{,}3571$$

R

$$16 = 5a + 10 \cdot \left(-\frac{5}{14}\right)$$

$$a = 3{,}9143$$

R

R

Alltså får vi $\boxed{\hat{y} = 3{,}9143 + 0{,}9t - 0{,}3571t^2}$

b) År 2018 motsvarar $t = 3$

$$y = 3{,}9143 + 0{,}9 \cdot 3 - 0{,}3571 \cdot 3^2 \approx 3{,}40$$

R

R

Om vår modell stämmer blir avkastningen ungefär 3,4 % år 2018

SU, STATISTIK

Skrivsal: ______ Anonymkod: 0003 6CZ Blad nr: 12

5) 10p

(5)

Y = (lungcancer = 1, ej lungcancer = 0)

X_1 = ålder

X_2 = rökvanor (tung rökare = 1, ej tung rökare = 0)

$$\hat{Y} = -5,9 + 0,04 X_1 + 2,3 X_2$$

a) $X_1 = 50$, $X_2 = 1$

$$\hat{Y} = -5,9 + 0,04 \cdot 50 + 2,3 \cdot 1 = -1,6$$

$$P(Y_i = 1 \mid X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

$$p = \frac{e^{-1,6}}{1 + e^{-1,6}} = 0,1680$$

R

Sannolikheten är knappt 17%

R

b) $X_1 = 50$, $X_2 = 0$

$$\hat{y} = -5,9 + 0,04 \cdot 50 + 2,3 \cdot 0 = -3,9$$

$$p = \frac{e^{-3,9}}{1 + e^{-3,9}} = 0,0198$$

R

Sannolikheten är knappt 2%

R