Stockholms universitet
Statistiska institutionen
Raul Cano

SKRIVNINGSDATUM: 08-01-2018

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).
Tentamensgenomgång och återlämning: tisdagen den 23 januari, 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: (20 poäng)
En villaägareförening i Grönköping vill veta hur mycket olja som förbrukas vid uppvärmning av friliggande småhus inom ett visst område. För ett urval av 7 hus noterade man den uppvärmda bostadsytan i m^2 och oljeförbrukningen i m^3 under ett år. Antag att sambandet mellan oljeförbrukning (Y) och uppvärmd bostadsyta (X) kan beskrivas med regressionsmodellen: $Y = \beta_0 + \beta_1 X + \varepsilon$ där avvikelserna (ε) antas vara oberoende och normalfördelade med väntevärdet 0 och konstant varians σ^2. Följande information har erhållits:

The regression equation is
Y = 1,20 + 0,02 X

Predictor	Coef	SE Coef	T	p
Constant	1,200		1,80	0,132
X	0,020		4,00	0,010

S = R-sq = R-sq(adj) =

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	2,5600	2,5600	16,00	0,010
Residual Error	5				
Total	6				

a). Beräkna residualvariansen. (5 poäng)

b). Hur stor andel av den totala variationen i oljeförbrukning förklaras av den anpassade regressionslinjen? (5 poäng)

c). Testa om $\beta_1 > 0{,}00$. Ställ upp hypoteser och gör sedan en hypotesprövning på signifikansnivå 1% ($\alpha = 0{,}01$). Vilken blir din slutsats? (10 poäng)

Uppgift 2: (20 poäng)
Med hjälp av $n = 10$ observationer vill man skatta parametrarna β_0, β_1 och β_2 i den multipla regressionsmodellen $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Slumpfelen ($\varepsilon$) antas vara oberoende och normalfördelade med väntevärde 0 och konstant varians σ^2. Följande information har erhållits:

```
The regression equation is
Y = 15,32 + 4,27 X1 + 0,94 X2

Predictor      Coef      SE Coef           T          P
Constant      15,32         3,67        4,17      0,003
X1            4,270         1,10        3,88      0,005
X2            0,940         0,17        5,53      0,000

S =           R-sq =           R-sq(adj) =

Analysis of Variance

Source              DF        SS        MS          F
Regression                  4256
Residual Error               224
Total                       4480
```

a). Beräkna en skattning av slumpfelsvariansen σ^2. (5 poäng)
b). Beräkna determinationskoefficienten R^2 och tala om vad detta värde betyder. (5 poäng)
c). Beräkna ett 99 procents konfidensintervall för β_1. (5 poäng)
d). Pröva på 5 % signifikansnivå om modellen som helhet är signifikant. Ange de hypoteser som du testar. (5 poäng)

Uppgift 3: (20 poäng)
Följande tabell visar elförbrukningen (i 1 000 kWh) i Grönköping, åren 2013-2017. Grönköping växer!

2013	2014	2015	2016	2017
4	6	10	14	20

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 elförbrukningen (i 1 000 kWh) i Grönköping, år 2018. (5 poäng)

Uppgift 4: (20 poäng)
Följande tabell visar antal flygpassagerare (i tusentals) på Grönköpingsflygplats "Lilla Arla" under åren 2013-2017. Grönköping fortsätter att växa!

2013	2014	2015	2016	2017
1	3	8	14	22

a) Anpassa en andragradskurva till tidsserien med hjälp av minsta-kvadrat-metoden. (15 poäng)

b) Enligt den anpassade modellen, gör en prognos för antal flygpassagerare (i tusentals) på Grönköpingsflygplats "Lilla Arla", år 2018. (5 poäng)

Uppgift 5: (20 poäng)
I Grönköping har man studerat sambandet mellan dödligheten , blodtryck, kolesterol och motionsvanor. Man har mätt blodtryck, kolesterol och motionsvanor hos ett stort antal personer samt följt dessa fem år framåt i tiden, varefter man har tagit reda på om personen är död eller lever.
Beteckning:
Y= "död/lever" (död = 1, lever = 0),
X_1 = blodtryck (systoliskt blodtryck i mm Hg),
X_2 = kolesterol (plasma-lipid nivåer) och
X_3 = motionsvanor (icke idrottsman = 1, idrottsman = 0).

Via en logistik regression modell har man fått följande skattningar:

$\hat{Y} = -14{,}26 + 0{,}03\ X_1 + 0{,}06\ X_2 + 3{,}47\ X_3$

a). Beräkna den skattade sannolikheten att en person i Grönköping kommer att dö om fem år, om personen är en icke idrottsman och har följande mättningar: $X_1 = 150$ och $X_2 = 75$. (10 poäng)

b). Beräkna den skattade sannolikheten att en person i Grönköping kommer att dö om fem år, om personen är en idrottsman och har följande mättningar: $X_1 = 150$ och $X_2 = 75$. (10 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

$$\underbrace{b_0+b_1x_{n+1}}_{\hat{y}_{n+1}} \pm t_{n-2,\frac{\alpha}{2}}\sqrt{s_\epsilon^2\left(1+\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}.$$

Konfidensintervall förväntat y-värde för ett nytt x-värde

$$\underbrace{b_0+b_1x_{n+1}}_{\hat{y}_{n+1}} \pm t_{n-2,\frac{\alpha}{2}}\sqrt{s_\epsilon^2\left(\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}.$$

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	$n-1$		

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

Normalekvationerna för fallet $\hat{y}=a+b_1t+b_2t^2$

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

Säsongsrensning med regression

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

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

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

Logistisk regression

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

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

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

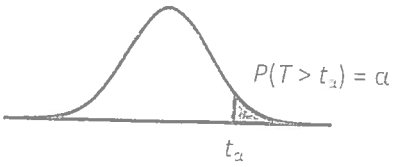
$$P(D) = \frac{\text{odds}(D)}{1 + \text{odds}(D)}$$

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

TABELL 3. *t*-fördelningens kvantiler

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

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

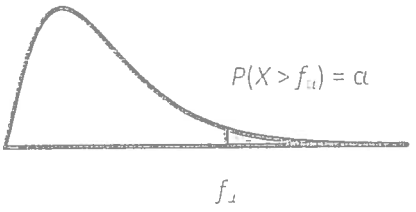


ν	$\alpha = 0{,}1$	0,05	0,025	0,010	0,005	0,0025	0,0010	0,0005
1	3,078	6,314	12,706	31,821	63,657	127,321	318,309	636,619
2	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	1,476	2,015	2,571	3,365	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_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	24 [illegible]	241	[illegible]	[illegible]	244 7	245 4	245,9
2	13 51	19,00	19 16	19,25	19,30	19 33	19,35	19,37	19 38	19 4	1 [illegible]	1 [illegible]	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 7	[illegible]	[illegible]	8,73	8,71	8,70
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6 [illegible]	5 9	[illegible]	[illegible]	5,89	5,87	5,86
5	6,61	5 79	5,41	5 19	5,05	4,95	4,83	4,82	4 77	4 71	4 [illegible]	4 [illegible]	4 66	4,64	4 62
6	5,99	5 14	4,76	4,53	4,39	4,28	4,21	4 15	4 1	4 [illegible]	4 [illegible]	[illegible]	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 61	3 [illegible]	[illegible]	3 55	3,53	3,51
8	5 32	4 46	4,07	3,84	3,69	3 58	3,50	3,44	3 3 [illegible]	3 [illegible]	[illegible]	[illegible]	3 26	3,24	3,22
9	5,12	4,26	3 86	3,63	3,48	3 37	3 29	3 23	3 1 [illegible]	14	1 [illegible]	[illegible]	3 05	3 03	3,01
10	4,96	4,10	3,71	3,48	3,33	3 22	3 14	3,07	3, 2	2,9	2, 1	, 1	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 [illegible]	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 [illegible]	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 [illegible]	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

3

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 8/01/18

Sal: Värtasalen

Tenta: Regressionsanalys och undersökningsmetodik

Kurs: Regressions- och tidsserieanalys

ANONYMKOD: REG-ANY-LLO

☑ 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					3 17-
Lär.ant. 20p	14p	20p	20p	20p					

POÄNG	BETYG	Lärarens sign.
94p	A	RC

SU, DEPARTMENT OF STATISTICS

Room: Väitasalen Anonymous code: REG-ANY-LLO Sheet number: 1

1) 20p

1.) Y = Oljeförbrukning X = Uppvärmd bostadsyta

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$ $\hat{Y}_i = 1,2 + 0,02 X_i$ $n = 7$

	SE Coef	
Constant	$\frac{2}{3} \approx 0,667$	$\frac{1,2}{S_a} = 1,8$ $S_a = \frac{1,2}{1,8} = \frac{2}{3}$
X	$\frac{1}{200} = 0,005$	$\frac{0,02}{S_b} = 4$ $S_{b_1} = \frac{0,02}{4} = \frac{1}{200}$

Source	DF	SS	MS	F
Regression	1	SSR = 2,56	MSR = 2,56	16
Residual Error	5	SSE = 0,8	MSE = SSE/DF = 0,16	
Total	6	SST = 3,36		

$F = \frac{MSR}{MSE}$ $16 = \frac{2,56}{MSE}$ $MSE = \frac{2,56}{16} = 0,16$

$\frac{SSE}{5} = 0,16$ $SSE = 0,8$ $SST = SSR + SSE = 2,56 + 0,8 = 3,36$

a.) Residualvariansen $S_e^2 = MSE = 0,16$

b.) R^2 - determinationskoefficienten, anger hur stor andel av den totala variationen i oljeförbrukningen (Y) förklaras av den anpassade linjen.

$R^2 = \frac{SSR}{SST}$ $R^2 = \frac{2,56}{3,36} = 0,761905$

c.) $H_0: \beta_1 = 0 \quad H_1: \beta_1 > 0$ R R $t_{kvot} = t_{0,01}(7-2) = 3,365$

Beslutsregel: Förkasta H_0 om $t_{obs} > t_{kvot} = 3,365$

Testvariabel $t = \frac{b_1 - \beta_1^0}{s_{b_1}} = \frac{0,02-0}{0,005} = 4 > 3,365$ R R

Vi kan därmed på 1% signifikansnivå säga att β_1 är positiv och skiljd från 0

2 a.) $n=10 \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ 2) 14 p

Source	DF	SS	MS	F
Regression	$k=2$	4256	$SSR/DF = \frac{4256}{2} = 2128$	$\frac{MSR}{MSE} = \frac{2128}{32} = 66,5$
Residual Error	$n-k-1=7$	224	$SSE/DF = \frac{224}{7} = 32$	
Total	$n-1=9$	4480		

s_ε^2 är en väntevärdesriktig skattning av σ^2 F

$s_\varepsilon^2 = SSE = 224$ F

b.) $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{224}{4480} = 0,95$ R

- Detta värde säger oss att 95% av variationen av y kan förklaras av vår modell, vilket också är ett väldigt högt värde. Eftersom vi har en multipel regressionsmodell så kan det vara värt att också titta på R^2-adj. Detta kommer ge oss en indikation på ifall tillagda X-variabler är relevanta/kör hemma i modellen. Detta eftersom R^2 ökar automatiskt när vi lägger till fler X-variabler.

$R^2\text{-adj} = 1 - \left(\frac{SSE}{df} / \frac{SST}{df}\right) = 1 - \left(\frac{224}{7} / \frac{4480}{9}\right) \approx 0,93571$

Värdet indikerar att det inte har uppstått något knasigt av att inkludera X-variablerna i modellen.

SU, DEPARTMENT OF STATISTICS

Room: Växtasalen Anonymous code: REG-ANY-LLO Sheet number: 2

forts. 2 c.) 99% igt KI för β_1 $\Rightarrow$ $b_1 \pm t_{\alpha/2}(n-2) \times s_{b_1}$ (n-3) (-1p)

$b_1 = 4,27$ $t_{0,005}(8) = 3,355$ $s_{b_1} = 1,1$

$4,27 \pm 3,355 \cdot 1,1 = 4,27 \pm 3,6905 \Rightarrow [0,5795 ; 7,9605]$ R

d.) $H_0: \beta_1 = \beta_2 = 0$ H_1: Minst en β är $\neq 0$

Testvariabel: F, $F_{0,05}\left(\frac{2}{7}\right) = 4,74$

Beslutsregel: Förkasta H_0 om $F_{obs} > F_{krit} = 4,74$

- Från tabellen såg vi att $F = \frac{MSR}{MSE} = \frac{2128}{32}$

Detta gav oss ett värde på: 66,5, $66,5 > 4,74$ med stor marginal så därför kan vi dra slutsatsen att förkasta H_0 på 5% ig signifikansnivå. R

3a.) Exponentiell modell: $y = a \cdot b^x$ i detta fall kan vi istället skriva $y = a \cdot b^t$

För att få fram de skattade värdena kan vi logaritmera:

$\log(y) = \underbrace{\log(a)}_{a'} + t \cdot \underbrace{\log(b)}_{b'}$ och har nu en linjär funktion.

$a = 10^{a'}$ $b = 10^{b'}$

Årtal − 2015 = t_i	t_i^2	Y_i	$\log(Y_i)$	$t \cdot \log(Y_i)$
−2	4	4	≈ 0,602206	≈ −1,20412
−1	1	6	≈ 0,77815	≈ −0,77815
0	0	10	1	0
1	1	14	≈ 1,14613	≈ 1,14613
2	4	20	≈ 1,30103	≈ 2,60206
Σ:	10		4,82737	1,76592

3) 2p

$$b' = \frac{\Sigma \log(y_i) \cdot t_i - \frac{\Sigma t_i \cdot \Sigma \log(y_i)}{n}}{\Sigma t_i^2 - \frac{(\Sigma t_i)^2}{n}}$$

$$a' = \frac{\Sigma \log(y_i)}{n} - \frac{\Sigma t_i}{n} \cdot b'$$

Eftersom vi har transformerat årtalen till t och att $\Sigma t_i = 0$ så kommer många termer att försvinna och kvar blir två enklare uttryck för b' och a':

R

$$b' = \frac{\Sigma \log(y_i) \cdot t_i}{\Sigma t_i^2} = \frac{1,76592}{10} = 0,176592 \quad R$$

$$a' = \frac{\Sigma \log(y_i)}{n} = \frac{4,82737}{5} = 0,965474 \quad R$$

$$a = 10^{0,965474} \approx 9,23579 \quad R$$

$$b = 10^{0,176592} \approx 1,50173 \quad R$$

$$\Rightarrow Y = 9,23579 \cdot 1,50173^t \quad R$$

Y mäter den skattade elförbrukningen.

Om $t = 0$ så är förbrukningen endast "a" dvs 9,23579 (1000 kWh). R

Det skattade värdet på R_1 dvs $b_1 = 1,50173$, detta anger den procentuella ökningen av elförbrukningen som växer exponentiellt. Dvs ökningen är på ca 50,2 % för varje år R

forts. 3 b.) Prognos för elförbrukningen år 2018:

År 2018 är $t = 3$

$\hat{y}_{2018} = 9{,}23579 \cdot 1{,}50173^3 \approx 31{,}27877$ (1000 kWh).

4 a.) $\hat{y} = a + b_1 t_i + b_2 t^2$ $\quad n = 5$

Årtal − 2015 = t	t^2	t^4	y_i	$y_i t_i$	$y_i t_i^2$
−2	4	16	1	−2	4
−1	1	1	3	−3	3
0	0	0	8	0	0
1	1	1	14	14	14
2	4	16	22	44	88
Σ: 0	10	34	48	53	109

4) 2 p

$\Sigma y_i = a \cdot n + b_1 \Sigma t_i + b_2 \Sigma t_i^2$

$\Sigma y_i t_i = a \cdot \Sigma t_i + b_1 \Sigma t_i^2 + b_2 \Sigma t_i^3$

$\Sigma y_i t_i^2 = a \cdot \Sigma t_i^2 + b_1 \Sigma t_i^3 + b_2 \Sigma t_i^4$

Eftersom $\Sigma t_i = 0$ så kommer vissa termer att bli 0 så därför kan vi ta bort dem. (udda termer).

$\Sigma y_i = a \cdot n + b_2 \Sigma t_i^2$

$\Sigma y_i t_i = b_1 \Sigma t_i^2$ $\quad \Rightarrow 53 = 10 b_1 \quad b_1 = 5{,}3$

$\Sigma y_i t_i^2 = a \cdot \Sigma t_i^2 + b_2 \Sigma t_i^4$

$$\begin{cases} 48 = 5a + 10 b_2 \\ 109 = 10a + 34 b_2 \end{cases} \quad \frac{5a}{5} = \frac{48 - 10 b_2}{5} \quad a = 9{,}6 - 2 b_2$$

$109 = 10(9{,}6 - 2b_2) + 34 b_2 \Rightarrow 109 = 96 - 20 b_2 + 34 b_2$

$14 b_2 = 13 \quad b_2 = \frac{13}{14}$ eller $\approx 0{,}92857$

$a = 9{,}6 - 2 \cdot \frac{13}{14} \approx 7{,}74286$

$Y = 7,74286 + 5,3t + 0,92857t^2$ R

b.) Prognos för antal passagerare (1000-tals) år 2018:

År 2018 är $t = 3$

$\hat{Y} = a + b_1 \cdot 3 + b_2 3^2 = 7,74286 + 5,3 \cdot 3 + 0,92857 \cdot 3^2 =$ R

$\hat{Y}_{2018} = 31,99999$ R så alltså nästan 32000 stycken totalt

5 a.) Logistisk regression: $\hat{Y} = -14,26 + 0,03X_1 + 0,06X_2 + 3,47X_3$

Y = död/lever
X_1 = blodtryck = 150
X_2 = kolesterol = 75
X_3 = motionsvanor = 1

5) 2op

$$P(Y=1 \mid X_1=150, X_2=75, X_3=1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}} =$$

$$\frac{1}{1+e^{-Z}} \quad \text{där} \quad Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$Z = -14,26 + 0,03 \cdot 150 + 0,06 \cdot 75 + 3,47 \cdot 1 = +1,79$ R

$P(Y=1) = \frac{1}{1+e^{-(-1,79)}} \approx 0,14307$ R

=> Sannolikheten att vara död om 5 år givet dessa värden på X_1, X_2, X_3 är 0,14307 R

b.) $X_1 = 150 \quad X_2 = 75 \quad X_3 = 0 \quad Z = -14,26 + 0,03 \cdot 150 + 0,06 \cdot 75 + 3,47 \cdot 0$

$$P(Y=1 \mid X_1=150, X_2=75, X_3=0) = \frac{1}{1+e^{-Z}}$$

$Z = -5,26$ R

$\frac{1}{1+e^{-(-5,26)}} \approx 0,005168$ R $P(Y=1) \approx 0,005168$

Givet att en person har värdena $X_1 = 150$ $X_2 = 75$ $X_3 = 0$ dvs personen är en idrottsman så är sannolikheten ca 0,005168 att han kommer vara död om fem år.
– Sannolikheten är alltså mycket mindre jämfört med deltagare a. R