STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2019

Tentamen i Regressions- och tidsserieanalys, (4,5 hp)

Kurs: Regressionsanalys och undersökningsmetodik

2019-12-02

Skrivtid:	kl. 9.00 - 14.00 (5 timmar)
Godkända hjälpmedel:	Miniräknare utan lagrade formler och text
Vidhäftade hjälpmedel:	Formelsamling och Statistiska tabeller (endast de tabeller som krävs)

- Tentamen består av 5 uppgifter, i förekommande fall uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.
- Svar med fullständiga redovisningar ska lämnas.
  - 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!
  - Använd minst fem värdesiffror i dina beräkningar (1,2345 och 1234,5 är exempel på tal med fem värdesiffror). I förekommande fall är det inte möjligt pga. avrundning i t.ex. SAS-utskrifter men utgå då ifrån det som är givet. Du kan dock avrunda ditt slutliga svar.
- Tentamen kan maximalt ge 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 Athena kort efter tentamen.

LYCKA TILL!

Uppgift 1. (30p)

En livsmedelskedja i USA genomförde ett försök där man ville studera hur Y = veckoförsäljningen i antal paket av en viss kaffesort påverkades av X = exponeringsytan i kvadratfot. Tolv butiker ingick i experimentet och exponeringsytan valdes slumpmässigt för de olika butikerna till antingen 3, 6 eller 9 kvadratfot. Ingen prisvariation förekom under försöksperioden och pris för och exponering av konkurrerande kaffesorter var konstant och ändrades inte heller. Utgå ifrån att sambandet mellan försäljning och exponering kan beskrivas med en enkel linjär regressionsmodell enligt:

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

Datamaterialet från studien och lite allmän statistik återges i tabellerna nedan.

a) Skatta parametrarna i modellen med minsta-kvadrat-metoden. (6p)

b) Tolka parameterskattningarna i ord. Är båda dessa tolkningar meningsfulla i detta fall? Förklara kortfattat. (6p)

c) Beräkna ett 95 % konfidensintervall för lutningskoefficienten. Förklara kortfattat vilken slutsats du kan dra. (6p)

d) Anta att du ska beräkna 95% konfidensintervall för den genomsnittliga försäljningen för fallen när $X = 3, 6$ respektive 9, dvs. tre olika konfidensintervall. För vilket av dessa värden på X blir intervallet kortast och varför? Motivera ditt svar med hjälp av formler. OBS! Du ska inte behöva göra några beräkningar alls för att besvara frågan. (6p)

e) Två av residualerna saknas nedan. Beräkna dessa först, plotta sedan samtliga residualer mot den förklarande variabeln och kommentera kortfattat det du ser. OBS! Om du inte kan beräkna residualerna som saknas, plotta det du har. (6p)

Data och residualer

Butik nr i	Försäljning i antal y_i	Yta x_i	Residual e_i
1	526	6	8,000
2	421	3	1,875
3	581	6	63,00
4	640	9	23,125
5	412	3	-7,125
6	500	9	-116,875
7	444	6	-74,000
8	443	3	23,875
9	580	9	-36,875
10	570	6	52,000
11	376	3	?
12	723	9	?

Deskriptiv statistik

Statistika	Försäljning i antal y	Yta x
Medelvärde	518	6
Standardfel	30,11644	0,738549
Varians	10884	6,545455
Median	513	6
Minimum	376	3
Maximum	723	9
Summa	6216	72
Antal	12	12
Kovarians x, y	215,7273	

Uppgift 2. (30p)

En grupp sociologer undersökte förekomsten av mord i USA och de var särskilt intresserade av vilka faktorer som kunde tänkas förklara skillnaderna mellan olika städer. I detta syfte utgick man ifrån en multipel regressionsmodell med antal mord per 100 000 invånare som beroende variabel (Y), och stadens befolkningsstorlek i 1000-tal (X_1), andelen familjer i % med en årlig inkomst mindre än \$5000 (X_2) samt andelen arbetslösa i % (X_3) som förklarande variabler. En körning med SAS gav följande resultat:

Number of Observations Read	20
Number of Observations Used	20

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model					<.0001
Error			21.06607		
Corrected Total		1855.20200			

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-36.76493	7.01093		<.0001	0
X1	1	0.00076294	0.00063630		0.2480	1.05997
X2	1	1.19217	0.56165		0.0497	2.99090
X3	1	4.71982	1.53048		0.0071	3.07838

Observera att det saknas en del uppgifter i utskriften!

a) Färdigställ SAS-utskriften ovan genom att fylla i de värden som saknas. (6p)

b) Genomför ett formellt test på 5 % signifikansnivå för att pröva om modellen som helhet är signifikant. Ange hypoteserna som du testar, testvariabel och dess fördelning, beslutsregel med kritiska gräns samt beräkningar och slutsats med förklaring. (8p)

c) Genomför ett formellt test på 5 % signifikansnivå för att pröva om X_2 tillför något till modellen. Redovisa på samma sätt som i b) ovan. (8p)

d) Beräkna förklaringsgraden och den justerade förklaringsgraden. (4p)

e) Utifrån utskriften och dina lösningar, vilka slutsatser kommer du fram till? Finns det anledning att definiera modellen annorlunda? Motivera ditt svar. (4p)

Uppgift 3. (20p)

För var och en av följande deluppgifter ska du svara kortfattat. Hela uppgiften bör kunna redovisas på maximalt ca två A4-sidor. Du får gärna komplettera med bilder och skisser.

a) Vilka grundantaganden gäller för en linjär regressionsmodell? (5p)

b) Vad menas med residualanalys och varför är det viktigt? (5p)

c) Variansinflationsfaktor, vad är det och hur ska det användas? (5p)

d) Vad menas med ett exponentiellt samband, hur formulerar man en sådan modell och hur skattar man den enklast? (5p)

Uppgift 4. (10p)

Följande tabell visar driftskostnaden per kvartal för en idrottshall i en kommun räknat i 100 000-tal kronor (lönekostnader ej inräknat):

Kostnad	Kvartal 1	Kvartal 2	Kvartal 3	Kvartal 4
2014	4,8	4,1	6,0	6,5
2015	5,8	5,2	6,8	7,4
2016	6,0	5,6	7,5	7,8
2017	6,3	5,9	8,8	8,4

Innan din statistikerkollega hann sluta skattade hon trenden med glidande medelvärden, trendrensade serien och sammanställde slutligen medelvärdena av de trendrensade värdena för respektive kvartal. Dessa ges nedan:

	Kvartal 1	Kvartal 2	Kvartal 3	Kvartal 4
Medelvärde	0,93220	0,83776	1,09335	1,14331

Du lyckas inte få tag i henne nu och det är bråttom att färdigställa beräkningarna!

a) Är det en additiv eller multiplikativ modell som har använts? Motivera ditt svar. (2p)

b) Gör nödvändiga justeringar och säsongsrensa sedan serien för år 2017, dvs. endast de fyra sista observationerna. (4p)

c) Visa de fyra sista observationerna i ett lämpligt diagram tillsammans med de säsongsrensade värdena. (4p)

Uppgift 5. (10p)

Logistisk regression har här använts för att modellera sannolikheten att man som passagerare överlevde Titanics förlisning. Ålder, kön och klass (1:a, 2:a eller 3:e) används som förklaringsvariabler och modellen kan skrivas som:

$$\text{LogOdds}(Y = 1|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

där

$$Y = \begin{cases} 1, & \text{överlevde} \\ 0, & \text{överlevde ej} \end{cases} \qquad x_1 = \text{ålder i år} \qquad x_2 = \begin{cases} 1, & \text{kvinna} \\ 0, & \text{man} \end{cases} \qquad x_3 = \begin{cases} 1, & \text{1:a klass} \\ 0, & \text{ej 1:a klass} \end{cases}$$

Baserat på riktiga data[1] skattades modellen med SAS och man fick bland annat följande resultat:

Number of Observations Read	891
Number of Observations Used	891

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	549
2	1	342

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1499	0.2291	25.2034	<.0001
Age	1	-0.0283	0.00723	15.3758	<.0001
Sex	1	2.6052	0.1815	206.0644	<.0001
FirstClass	1	1.8960	0.2197	74.4654	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.972	0.958	0.986
Sex	13.534	9.483	19.316
FirstClass	6.660	4.329	10.244

a) Beräkna sannolikheten för överlevnad ($Y = 1$) för en 20-årig kvinna som reste i första klass samt för en 20-årig man som inte reste i första klass. (6p)

b) På vilket sätt påverkade åldern sannolikheten för överlevnad? Du behöver inte tolka siffrorna ovan utan förklara bara kortfattat vad du baserar din slutsats på. (4p)

[1] Man har skattat att Titanic hade 1 317 passagerare så datamaterialet är inte fullständigt.

FORMELSAMLING

DESKRIPTIV STATISTIK

Varians: $s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}$

Kovarians: $s_{xy} = cov(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n-1}$

$= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n(n-1)}$

Korrelation: $r_{xy} = corr(x,y) = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$ Inferens: $t_{n-2} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

ENKEL LINJÄR REGRESSION

Modell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ Betingat medelvärde för $Y|X = x$: $\mu_{Y|X=x} = \beta_0 + \beta_1 x$

Parameterskattningar och dessas varianser:

$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$ $s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$b_0 = \bar{y} - b_1\bar{x}$ $s_{b_0}^2 = s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)$

Prediktion och skattat betingat medelvärde: $\hat{y}_i = \hat{\mu}_{Y|X=x_i} = b_0 + b_1 x_i$

Prediktionsintervall för prediktionen $\hat{y}_i$ givet $X = x$: $\hat{y}_i \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$

Konfidensintervall för betingade medelvärdet $\mu_{Y|X=x}$ givet $X = x$: $\hat{\mu}_{Y|X=x} \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$

ICKE-LINJÄR REGRESSION, exempel

Andragradspolynom: $\hat{y}_i = a + b_1 x_i + b_2 x_i^2$

Exponentiell: $\widehat{\ln(y_i)} = a + bx_i$ $\hat{y}_i = \exp(a + bx_i) = (e^a)(e^b)^{x_i} = (a') \cdot (b')^{x_i}$

$\widehat{\log_{10}(y_i)} = a + bx_i$ $\hat{y}_i = (10^a)(10^b)^{x_i} = (a') \cdot (b')^{x_i}$

ENKEL OCH MULTIPEL LINJÄR REGRESSION (sätt $k = 1$ om enkel regression)

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

Kvadratsummor: $SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s_y^2 = SSR + SSE$

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

$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ = [om enkel regression] $= b^2 \sum_{i=1}^n (x_i - \bar{x})^2$

Förklaringsgrad: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ $\quad R^2_{\text{adj}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$

Inferens för β_j: KI: $b_j \pm t_{n-k-1,\alpha/2} \cdot s_{b_j}$ Test: $t_{n-k-1} = \frac{b_j - \beta_j^*}{s_{b_j}}$

Test för hela modellen: $F_{k;n-k-1} = \frac{SSR/K}{SSE/(n-K-1)} = \frac{MSR}{MSE}$

Beräkningsformler för KORRELATION och REGRESSIONSKOEFFICIENT

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

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sum (x_i - \bar{x})^2/(n-1)} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x s_y}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy} \cdot \frac{s_y}{s_x}$$

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sqrt{\sum (x_i - \bar{x})^2/(n-1)} \cdot \sqrt{\sum (y_i - \bar{y})^2/(n-1)}}$$

$$= \frac{s_{xy}}{\sqrt{s_x^2} \cdot \sqrt{s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b_1 \cdot \frac{s_x}{s_y}$$

TIDSSERIEANALYS

– Komponenter

Additiv modell: $Y_t = T_t + S_t + C_t + E_t$ Multiplikativ modell: $Y_t = T_t \cdot S_t \cdot C_t \cdot E_t$

där T = trend, S = säsong, C = cyklisk/konjunktur samt E = slumpkomponent

- Skattning av trendkomponenten:

– med glidande medelvärden utan säsongvariation, exempel:

3-punkter centrerat: $\hat{T}_t = \frac{1}{3} \cdot y_{t-1} + \frac{1}{3} \cdot y_t + \frac{1}{3} \cdot y_{t+1}$

5-punkter centrerat: $\hat{T}_t = \frac{1}{5} \cdot y_{t-2} + \frac{1}{5} \cdot y_{t-1} + \frac{1}{5} \cdot y_t + \frac{1}{5} \cdot y_{t+1} + \frac{1}{5} \cdot y_{t+2}$

– med centrerade glidande medelvärden med säsongvariation, exempel:

halvårsdata: $\hat{T}_t = \frac{1}{4} \cdot y_{t-1} + \frac{1}{2} \cdot y_t + \frac{1}{4} \cdot y_{t+1}$

kvartalsdata: $\hat{T}_t = \frac{1}{8} \cdot y_{t-2} + \frac{1}{4} \cdot y_{t-1} + \frac{1}{4} \cdot y_t + \frac{1}{4} y_{t+1} + \frac{1}{8} \cdot y_{t+2}$

månadsdata: $\hat{T}_t = \frac{1}{24} \cdot y_{t-6} + \frac{1}{12} \cdot y_{t-5} + \cdots + \frac{1}{12} \cdot y_{t+5} + \frac{1}{24} \cdot y_{t+6}$

– med regressionsanalys, linjär trend och exponentiell trend:

Modell: $Y_t = \beta_0 + \beta_1 t + \varepsilon_1$ Skattad modell: $\hat{y}_t = b_0 + b_1 t = \hat{T}_t$

$\ln Y_t = \beta_0 + \beta_1 t + \varepsilon_1$ $\hat{y}_t = \exp(b_0 + b_1 t) = \hat{T}_t$

- Justering av säsongsindex $\bar{S}_j$ med p säsonger (halvår, kvartal el. månader osv.):

Additiv modell: $S_j^+ = \bar{S}_j - \left(\frac{\sum \bar{S}_i}{p}\right)$ Multiplikativ modell: $S_j^+ = \frac{\bar{S}_j}{(\sum \bar{S}_i / p)}$

- Trend- och säsongsrensning:

Additiv modell: $y_t - \hat{T}_t$ resp. $y_t - S_t^+$ Multiplikativ modell: $y_t / \hat{T}_t$ resp. y_t / S_t^+

LOGISTISK REGRESSION och ODDS

Odds för en händelse A:	$\text{Odds}(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1-P(A)} \Leftrightarrow P(A) = \frac{\text{Odds}(A)}{1+\text{Odds}(A)}$
Oddskvot för händelsen A mot B:	$\text{OR} = \frac{\text{Odds}(A)}{\text{Odds}(B)}$

- Logistisk regression:

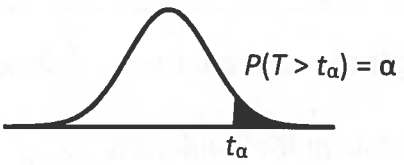
Enkel modell:	$P(Y=1\|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \frac{1}{1+\exp(-\beta_0-\beta_1 x)}$
	$P(Y=0\|x) = \frac{1}{1+\exp(\beta_0+\beta_1 x)}$
	$\text{Odds}(Y=1\|x) = \exp(\beta_0+\beta_1 x)$
	$\text{LogOdds}(Y=1\|x) = \beta_0+\beta_1 x$
Multipel modell:	$\text{LogOdds}(Y=1\|x_1,\dots x_k) = \beta_0+\beta_1 x_1+\cdots+\beta_k x_k$

Intercept β_0:	$P(Y=1\|x_1=\cdots=x_k=0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
Oddskvot för $Y=1$ när $X_j = x_j+1$ mot $X_j = x_j$:	$\text{OR}(X_j) = \frac{\text{Odds}(Y=1\|x_j+1, \text{allt annat lika})}{\text{Odds}(Y=1\|x_j, \text{allt annat lika})} = \exp(\beta_j)$
KI för $\text{OR}(X_j)$:	$\left(\exp(b_j - z_{\alpha/2}\cdot s_{b_j})\,;\ \exp(b_j + z_{\alpha/2}\cdot s_{b_j})\right)$

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	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

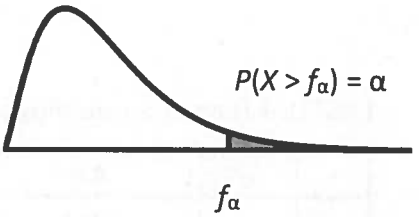
Forts. nästa sida

TABELL 3 forts. *t*-fördelningens kvantiler

ν	$\alpha = 0{,}1$	0,05	0,025	0,010	0,005	0,0025	0,0010	0,0005
80	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
85	1,292	1,663	1,988	2,371	2,635	2,882	3,189	3,409
90	1,291	1,662	1,987	2,368	2,632	2,878	3,183	3,402
95	1,291	1,661	1,985	2,366	2,629	2,874	3,178	3,396
100	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
125	1,288	1,657	1,979	2,357	2,616	2,858	3,157	3,370
150	1,287	1,655	1,976	2,351	2,609	2,849	3,145	3,357
175	1,286	1,654	1,974	2,348	2,604	2,843	3,137	3,347
200	1,286	1,653	1,972	2,345	2,601	2,839	3,131	3,340
300	1,284	1,650	1,968	2,339	2,592	2,828	3,118	3,323
400	1,284	1,649	1,966	2,336	2,588	2,823	3,111	3,315
500	1,283	1,648	1,965	2,334	2,586	2,820	3,107	3,310
1000	1,282	1,646	1,962	2,330	2,581	2,813	3,098	3,300
2000	1,282	1,646	1,961	2,328	2,578	2,810	3,094	3,295
3000	1,282	1,645	1,961	2,328	2,577	2,809	3,093	3,294
4000	1,282	1,645	1,961	2,327	2,577	2,809	3,092	3,293
5000	1,282	1,645	1,960	2,327	2,577	2,808	3,092	3,292

TABELL 5. *F*-fördelningens kvantiler

$X \in F(\nu_1,\nu_2)$ där ν_1,ν_2 = antal frihetsgrader i täljaren respektive nämnaren. Vilket värde har f_α om $P(X > f_\alpha) = \alpha$ där α är en given sannolikhet.



α = 0,05

	ν_1 = 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ν_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

TABELL 5 forts. *F*-fördelningens kvantiler

$\alpha = 0{,}05$

	$\nu_1 =$ 16	17	18	19	20	25	30	35	40	50	60	70	80	100	∞
$\nu_2 = 1$	246,5	246,9	247,3	247,7	248,0	249,3	250,1	250,7	251,1	251,8	252,2	252,5	252,7	253,0	254,3
2	19,43	19,44	19,44	19,44	19,45	19,46	19,46	19,47	19,47	19,48	19,48	19,48	19,48	19,49	19,50
3	8,69	8,68	8,67	8,67	8,66	8,63	8,62	8,60	8,59	8,58	8,57	8,57	8,56	8,55	8,53
4	5,84	5,83	5,82	5,81	5,80	5,77	5,75	5,73	5,72	5,70	5,69	5,68	5,67	5,66	5,63
5	4,60	4,59	4,58	4,57	4,56	4,52	4,50	4,48	4,46	4,44	4,43	4,42	4,41	4,41	4,37
6	3,92	3,91	3,90	3,88	3,87	3,83	3,81	3,79	3,77	3,75	3,74	3,73	3,72	3,71	3,67
7	3,49	3,48	3,47	3,46	3,44	3,40	3,38	3,36	3,34	3,32	3,30	3,29	3,29	3,27	3,23
8	3,20	3,19	3,17	3,16	3,15	3,11	3,08	3,06	3,04	3,02	3,01	2,99	2,99	2,97	2,93
9	2,99	2,97	2,96	2,95	2,94	2,89	2,86	2,84	2,83	2,80	2,79	2,78	2,77	2,76	2,71
10	2,83	2,81	2,80	2,79	2,77	2,73	2,70	2,68	2,66	2,64	2,62	2,61	2,60	2,59	2,54
11	2,70	2,69	2,67	2,66	2,65	2,60	2,57	2,55	2,53	2,51	2,49	2,48	2,47	2,46	2,40
12	2,60	2,58	2,57	2,56	2,54	2,50	2,47	2,44	2,43	2,40	2,38	2,37	2,36	2,35	2,30
13	2,51	2,50	2,48	2,47	2,46	2,41	2,38	2,36	2,34	2,31	2,30	2,28	2,27	2,26	2,21
14	2,44	2,43	2,41	2,40	2,39	2,34	2,31	2,28	2,27	2,24	2,22	2,21	2,20	2,19	2,13
15	2,38	2,37	2,35	2,34	2,33	2,28	2,25	2,22	2,20	2,18	2,16	2,15	2,14	2,12	2,07
16	2,33	2,32	2,30	2,29	2,28	2,23	2,19	2,17	2,15	2,12	2,11	2,09	2,08	2,07	2,01
17	2,29	2,27	2,26	2,24	2,23	2,18	2,15	2,12	2,10	2,08	2,06	2,05	2,03	2,02	1,96
18	2,25	2,23	2,22	2,20	2,19	2,14	2,11	2,08	2,06	2,04	2,02	2,00	1,99	1,98	1,92
19	2,21	2,20	2,18	2,17	2,16	2,11	2,07	2,05	2,03	2,00	1,98	1,97	1,96	1,94	1,88
20	2,18	2,17	2,15	2,14	2,12	2,07	2,04	2,01	1,99	1,97	1,95	1,93	1,92	1,91	1,84
25	2,07	2,05	2,04	2,02	2,01	1,96	1,92	1,89	1,87	1,84	1,82	1,81	1,80	1,78	1,71
30	1,99	1,98	1,96	1,95	1,93	1,88	1,84	1,81	1,79	1,76	1,74	1,72	1,71	1,70	1,62
35	1,94	1,92	1,91	1,89	1,88	1,82	1,79	1,76	1,74	1,70	1,68	1,66	1,65	1,63	1,56
40	1,90	1,89	1,87	1,85	1,84	1,78	1,74	1,72	1,69	1,66	1,64	1,62	1,61	1,59	1,51
45	1,87	1,86	1,84	1,82	1,81	1,75	1,71	1,68	1,66	1,63	1,60	1,59	1,57	1,55	1,47
50	1,85	1,83	1,81	1,80	1,78	1,73	1,69	1,66	1,63	1,60	1,58	1,56	1,54	1,52	1,44
60	1,82	1,80	1,78	1,76	1,75	1,69	1,65	1,62	1,59	1,56	1,53	1,52	1,50	1,48	1,39
70	1,79	1,77	1,75	1,74	1,72	1,66	1,62	1,59	1,57	1,53	1,50	1,49	1,47	1,45	1,35
80	1,77	1,75	1,73	1,72	1,70	1,64	1,60	1,57	1,54	1,51	1,48	1,46	1,45	1,43	1,32
100	1,75	1,73	1,71	1,69	1,68	1,62	1,57	1,54	1,52	1,48	1,45	1,43	1,41	1,39	1,28
∞	1,64	1,62	1,60	1,59	1,57	1,51	1,46	1,42	1,39	1,35	1,32	1,29	1,27	1,24	1,00

Statistiska institutionen

Stockholms universitet

Rättningsblad

Datum: 2/12-2019

Sal: Laduvikssalen

Tenta: Regressions- och tidsserieanalys

Kurs: Regressionsanalys och undersökningsmetodik

ANONYMKOD: 0084-DTC

[X] 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					5
Lär.ant. 28	30	20	10	10					

POÄNG	BETYG	Lärarens sign.
98	A	ML

SU, STATISTIK

Skrivsal: LA Anonymkod: 0084-DTC Blad nr: 1

(1) givet: $n = 12$ $\bar{x} = 6$ $\bar{y} = 518$

X är en st. variabel $s_x^2 = 6.545455$ $s_y^2 = 10884$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $S_{xy} = 215.7273$

1-a) $\hat{\beta}_1 = b_1 = \frac{S_{xy}}{s_x^2} = \frac{215.7273}{6.545455} = 32.958$

$b_0 = \bar{y} - b\bar{x} = 518 - \frac{215.7273}{6.545455} \cdot 6 = 320.25$

$\Rightarrow \hat{y} = 320.25 + 32.958\,x$ R /6

1-b) $b_0 = 320.25$ betyder att om exponeringsytan i kvadratfoten är 0, 320.25 packet av kaffesorten säljs i genomsnitt i en vecka. Är detta en meningsfull tolkning?

$b_1 = 32.958$ betyder att om exponeringsytan ökar med en enhet då antal packet som säljs per vecka ökar med ca 33 packet i genomsnitt. men exponeringsytan kan ta värdet av 3,6,9 och det finns ingen 3+1=4 eller 7 och 10 ytan, då b_1 kan inte vara meningsfull. Spelar ingen roll /4

1-c) β_1 s lutningskoefficient

95% KI för β_1 s $b_1 \pm t_{10;0.025}\, s_b$

$s_b^2 = \frac{s_e^2}{(n-1)s_x^2}$

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

$\hat{y}_{11} = 320.25 + 32.958 \cdot 3 = 419.124 \Rightarrow e_{11} = y_{11} - \hat{y}_{11} = 376 - 419.124 = -43.124$

$\hat{y}_{12} = 320.25 + 32.958 \cdot 9 = 616.872 \Rightarrow e_{12} = y_{12} - \hat{y}_{12} = 106.128$

$\Rightarrow \sum_{i=1}^{12} e_i = 8.00 + 1.875 + \dots - 43.124 + 106.128 = 0.004$

Σe_i måste vara 0 men för avrundningsfel det har blivit 0.004. OK

1-c forts:

$\sum_{i=1}^{12} e_i^2 = 41514.42551$ /

$\Rightarrow s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{41514.42551}{10} = 4151.4426$

$\Rightarrow s_b^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{4151.4426}{(11)\cdot 6.545455} = 57.659$

$\rightarrow$ 95% KI $\beta_1 = 32.958 \pm t_{10;0.025}\cdot\sqrt{57.659}$ $\Big\}$ 95% KI $\in (16.0400;\ 49.8759)$

$t_{10;0.025} = 2.228$

dvs med 95% säkerhet kan vi säga att β_1 ligger i intervallet ovan. Intervallet innehåller inte 0 vilket betyder att β_1 skiljer från 0 och exponeringsytan påverkar veckoförsäljningen i antal packet på vissa Kaffesorten. BRA!

/6

1-d) Ett konfidensintervall för genomsnittlig försäljningen givet ett värde på x skapas med hjälp av förmeln:

$$\hat{\mu}_{Y|X=x} \pm t_{n-2;\alpha/2}\sqrt{s_e^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$$

Intervallet blir kortare om $t_{n-2;\alpha/2}\sqrt{s_e^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$ blir lägre

$t_{n-2;\alpha/2}$, s_e^2, s_x^2 och n är konstant. Därför att talen blir mindre behöver att $(x-\bar{x})^2$ blir mindre $\Rightarrow$ hur närmare till $\bar{x}$, är x, då kortare intervall har vi.

I detta fal $\bar{x}=6$ $\Rightarrow$ när $x=6$ blir intervallet kortast.

Om $x=3$ eller $x=9$ $\Rightarrow (x-6)^2 = 9$ $\Rightarrow$ samma intervall när $x=3$ eller $x=9$

Om $x=6$ $\Rightarrow (x-\bar{x})^2 = (6-6)^2 = 0$ $\rightarrow$ vi får kortaste intervall

/6

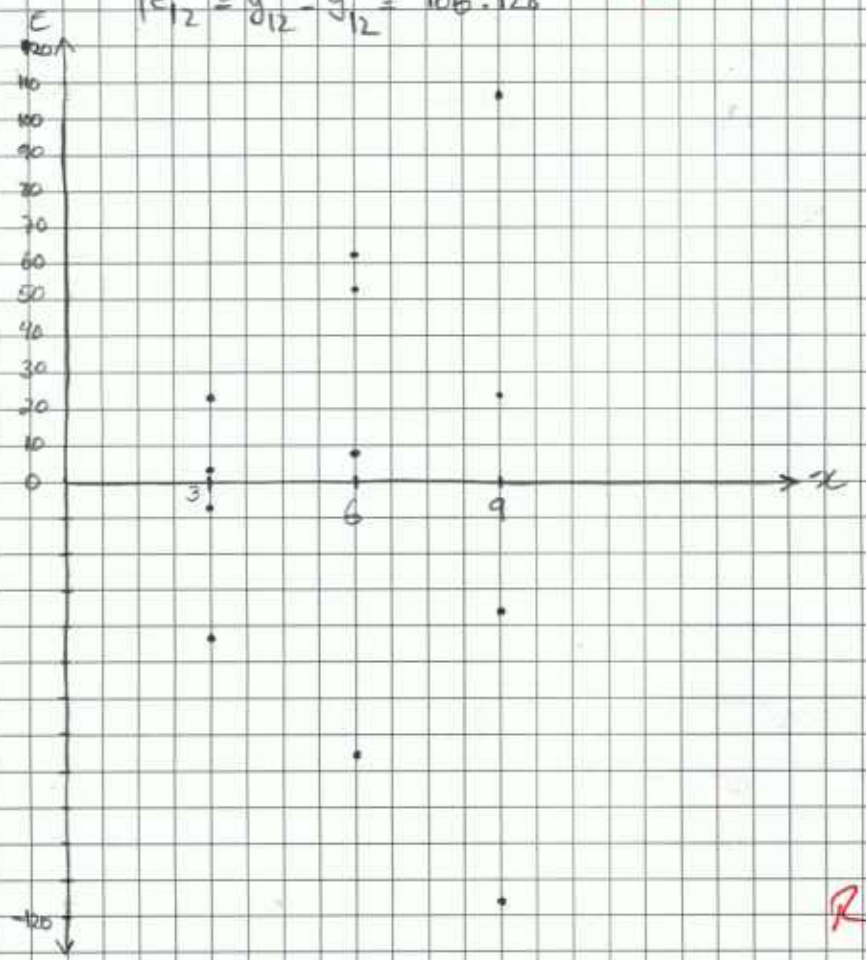
SU, STATISTIK

Skrivsal: LA Anonymkod: 0084-DJC Blad nr: 2

1-e) Residualerna beräknas enligt: $y_i - \hat{y}_i$

$\hat{y}_i = 320.25 + 32.958 \cdot x_i \Rightarrow \hat{y}_{11} = 419.124 \quad \hat{y}_{12} = 616.872$

$\Rightarrow \begin{cases} e_{11} = y_{11} - \hat{y}_{11} = 376 - 419.124 = -43.124 \\ e_{12} = y_{12} - \hat{y}_{12} = 106.128 \end{cases}$ R



R

Som ses från tabellen, residualerna är beroende av x-värdena, när x-värden ökar då residualernas (genomsnittliga) avvikelse från noll ökar också => när x ökar då residualerna ökar, vilket betyder att felltermena är inte oberoende av x, Det finns heterogedacitet mellan felltermena. BRA!

Det kan testas om y är en exponentiell funktion av x. /6

28

(2)

Det finns 3 oberoende variabler ⇒ K=3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

n=20

2-a

Source	DF	Sum of Sq	Mean S	F.value	Pr>F
Model	K=3	1518.14488	506.048293	24.02196012	<.0001
Error	n-k-1=16	337.05712	21.06607		
Total	n-1=19	1855.202			

$SSE = MSE \cdot (n-k-1) = 21.06607 \cdot 16 = 337.05712$

$SSR = SST - SSE = 1518.14488$

$MSR = \frac{SSR}{K} = 506.0482933$

$F = \frac{MSR}{MSE} = 24.02196012$

Parameter Estimates

Var	df	$\hat{Par}$	St.Error	t.value	Pr>|t|	VIF
Intercept	1	-36.76493	7.01093	-5.243945	<.0001	0
X_1	1	0.00076294	0.0006363	1.1990	0.2480	1.05997
X_2	1	1.19217	0.56165	2.1226	0.0497	2.99090
X_3	1	4.71982	1.53048	3.0839	0.0071	3.07838

$t_i = \frac{b_i}{s_{b_i}} \rightarrow$

$t_0 = \frac{b_0}{s_{b_0}} = \frac{-36.76493}{7.01093} = -5.2439$

$t_1 = \frac{b_1}{s_{b_1}} = \frac{0.00076294}{0.0006363} = 1.1990$

$t_2 = \frac{b_2}{s_{b_2}} = 2.1226$

$t_3 = \frac{b_3}{s_{b_3}} = 3.0839$

/6

2-b) $\alpha = 5\% = 0.05$

Vi kommer att göra ett F-test om modellen som helhet är signifikant

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: \text{minst en av } \beta_j \neq 0 \end{cases}$$

Test-variabel: $F = \frac{MSR}{MSE} \sim F(3,16)$

Beslutsregel: Förkasta H_0 om $F_{obs} > F_k = F_\alpha(k; n-k-1) = F_{0.05}(3,16) = 3.24$

Beräkningar: $F_{obs} = \frac{MSR}{MSE} = \frac{506.05}{21.066} = 24.022$

Slutsats: $F_{obs} = 24.022 > 3.24 \rightarrow H_0$ förkastas

Det betyder att modellen som helhet är signifikant med 95% säkerhet och minst en av parametrarna finns i modellen. /8

2-C) $\alpha = 0.05$

Vi kör på t-test

$$\begin{cases} H_0: \beta_2 = 0 \mid \beta_1 \text{ och } \beta_3 \text{ finns i modellen} \\ H_1: \beta_2 \neq 0 \mid \beta_1 \text{ och } \beta_3 \text{ finns i modellen} \end{cases}$$

Test-variabel: $t = \frac{b_2}{s_{b_2}} \sim t$

Beslutregel: Förkasta H_0 om $|t_{obs}| > t_k = t_{n-k-1;\alpha/2} = t_{16;0.025} = 2.120$

Beräkningar: Enligt tabellen har vi $\begin{cases} b_2 = 1.19217 \\ s_{b_2} = 0.56165 \end{cases}$

$\rightarrow |t_{obs}| = \left|\frac{1.19217}{0.56165}\right| = 2.1226 > 2.120 \rightarrow H_0$ förkastas (Litet skillnad)

Med 95% säkerhet β_2 finns i modellen när de andra parametrarna finns i modellen också

/8

2-d) Förklaringsgraden $R^2 = \frac{SSR}{SST} = \frac{1518.14488}{1855.26200} = 0.8183$

Vilket betyder att ca 82% av y förklaras av modellen.

justerade förklaringsgrad: $R^2_{adj} = 1 - \frac{SSE/n-k-1}{SST/n-1} = 1 - \frac{MSE}{MST} = 0.7842$ /4

2-e) Enligt tabellen och lösningar modellen som helhet är signifikant med 95% säkerhet, dvs minst en av de tre oberoende variabler samvarierar med beroende variabeln y.

Om vi tittar på p-värden för skattade parametrar då kommer vi fram att X_1 är inte signifikant i modellen. dvs β_1 är 5% signifikant lika med 0. X_1 kan inte ~~vara med skattning~~ kanske tas bort av y när X_2 och X_3 skattar y.

Då kan definieras en annan modell som:

$$y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Vilket har vi tagit bort X_1 från första modell därför att det var inte signifikant.

Staters befolkningstorlek kan inte påverka antal mord när andel familjer med en årlig inkomst mindre än 5000 $ och andelen arbetslösa skattar antal mord.

Enligt tabellen det finns positiv samband mellan X_2 och X_3 och y. Det betyder att antal mord ökar när arbetslöshet ökar. och mindre inkomst påverkar mord positivt.

/4

övrigt? (R^2, VIF)

SU, STATISTIK

Skrivsal: LA Anonymkod: 0089-DTC Blad nr: 4

(3)

3-a) • Y som en beroende variabel kan vara en linjär funktion av oberoende stokastiska variabler $X_1, \dots, X_k$ plus en slump felterm

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

• Feltermerna ε_i är normalfördelade med samma varians } iid.
$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ (feltermerna kommer från samma fördelning) }

• Feltermerna är oberoende av varandra

• Feltermerna är oberoende av x-värden

• Variationen av Y-värderna är oberoende av varandra med en konstant varians. det finns homogedasticitet mellan feltermerna.

om: $Y = \mu_Y + \varepsilon \rightarrow E(Y) = \mu_Y$

$\text{var}(Y) = \text{var}(\varepsilon) = \sigma_\varepsilon^2$

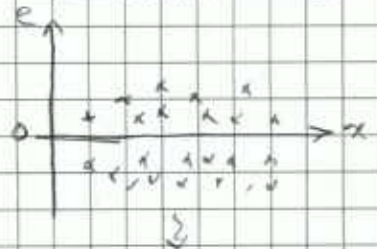
$\Rightarrow Y \sim N(\mu_Y, \sigma_\varepsilon^2)$ BRA!

• X_i-na kan vara korrelerade med varandra men de får inte vara bestämmt av varandra. (Multikullaritet)

/5

3-b) Residualanalys görs därför att det behövs att kontrollera modellen om modelantaganden stämmer eller inte. BRA!

Med residualanalys kollar vi om residulerna är beroende av x-värderna och om de följer normalfördelning.



e är oberoende av x

e är beroende av x.
med ökning på x => residualer ökar

/5

3 forts.

3-c) Variansinflationsfaktor är ett värde som visar om det finns multikullaritet mellan oberoende variabler.

Den formuleras som : $VIF = \frac{1}{1-R^2}$

om $VIF > 10$ $\rightarrow$ Det finns multikullaritet mellan oberoende variabler dvs de är bestämt av varandra

/5

3-d) En exponentiellt samband mellan y och x formuleras s

$$y = \alpha \beta^x \quad R$$

för att skatta parametrar α och β, behövs att först logaritmera ekvationen s

$$\left.\begin{array}{l} \log y = \log \alpha\beta^x = \log\alpha + x \log\beta \\ \log\alpha = \alpha' \\ \log\beta = \beta' \end{array}\right\} \Rightarrow \log y = \alpha' + \beta' x \quad R$$

$\rightarrow$ log y är en linjär funktion av x.

Vi räcknar $\log y_i$ för varje y_i, då s

$$b' = \hat{\beta}' = \frac{\sum_{i=1}^{n} x_i \log y_i - n\bar{x}\overline{\log y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$a' = \hat{\alpha}' = \overline{\log y} - \hat{\beta}'\bar{x}$$

För att kunna använda dessa skattade parametrar för skattning av y, antilogaritmerar vi dem.

$\hat{\alpha} = a = 10^{a'}$ och $\hat{\beta} = b = 10^{b'}$

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

a betyder att om $x=0$ $\rightarrow$ y har värdet av $a = 10^{a'}$

b betyder att med ökning med en enhet av x, beror på tecknet av $(b-1)$, ökar eller minskar y med $(b-1).100$ procent.

/5 (20)

SU, STATISTIK

Skrivsal: LA Anonymkod: 0084 - DTC Blad nr: 5

(4)

4-a) Det är en multiplikativ modell.

För en multiplikativ modell ska summan av trendrensade värden för kvartalerna vara nära 4, som i sluten måste justeras. R

BRA!

/2

$$\sum_{j=1}^{4} \bar{S}_j = 0.93220 + 0.83776 + 1.09335 + 1.14331 = 4.00662$$

I en additiv modell $\rightarrow \sum_{j=1}^{4} \bar{S}_j \approx 0$

4-b) I en multiplikativ modell:

$$S_j^+ = \bar{S}_j \Big/ \sum_{j=1}^{4} \bar{S}_j / 4 \Rightarrow S_1^+ = \frac{0.9322}{4.00662/4} = 0.93066$$

$$S_2^+ = \frac{0.83776}{4.00662/4} = 0.83638$$

$$S_3^+ = \frac{1.09335}{4.00662/4} = 1.09154$$

$$S_4^+ = \frac{1.14331}{4.00662/4} = 1.14142$$

$$\Rightarrow \sum_{j=1}^{4} S_j^+ = 4$$

/4

säsongrensning = y / S_j^+

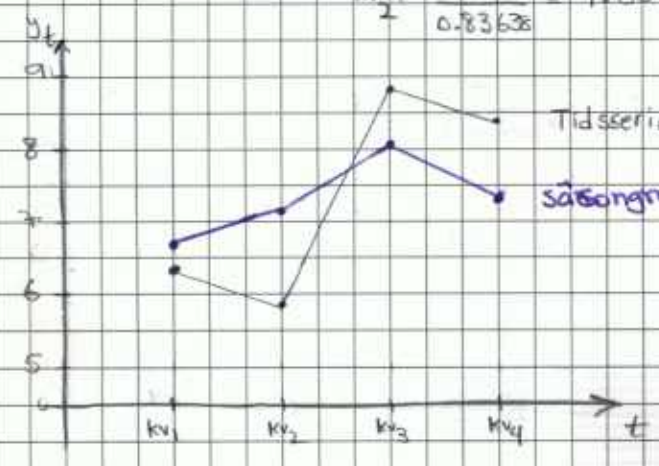
säsongrensade serie 2017:kv1: $\frac{6.3}{0.93066} = 6.7694$

kv_2: $\frac{5.9}{0.83638} = 7.0542$

$kv_3 = \frac{8.8}{1.09154} = 8.062$

$kv_4 = \frac{8.4}{1.14142} = 7.3592$

4-c)



/4

(10)

(5)

5-a)

$$\widehat{\text{Log odds}}(Y=1|x_i) = -1.1499 - 0.0283\,X_1 + 2.6052\,X_2 + 1.896\,X_3$$

$$\widehat{\text{Log odds}}(Y=1|x_1=20, x_2=1, x_3=1) = -1.1499 - 0.0283(20) + 2.6052 + 1.896 = 2.7853$$

$$P(Y=1|X_i) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)} = \frac{1}{1+\exp(-\beta_0-\beta_1X_1-\beta_2X_2-\beta_3X_3)}$$

$$P(Y=1|x_1=20, x_2=1, x_3=1) = \frac{1}{1+e^{-2.7853}} = \underline{0.9419}$$

Överlevde en 20-årig kvinna som reste i första klass med 94.2 % sannolikhet. R

$$\widehat{\text{log odds}}(Y=1|X_1=20, X_2=0, X_3=0) = -1.1499 - 0.0283(20) = -1.7159$$

$$P(Y=1|X_1=20, X_2=0, X_3=0) = \frac{1}{1+\exp(+1.7159)} = \underline{0.152}$$

Med 15% sannolikhet överlevde en man som inte reste med första klass. R

/6

5-b) Skattade parameter för age (b_1) visar att hur påverkas sannolikheten för överlevnade av åldet.

Skattade parameter har tecknet minus (-) vilket betyder att hur äldre du är, mindre sannolikt att överlevn. dvs ett negativt samband mellan ålder och överlevnads sannolikhet. ~~ERA~~

/4

(10)