Anonym kod: 0070-DCY (1)

Regressionsanalys och tidsserieanalys

1 a)

y = Slutpris (miljoner kronor)

x = begärt pris (miljoner kronor)

- $\hat{y} = \beta_0 + \beta_1 x_i + \varepsilon$
- $\beta_0 = \bar{y} - \beta_1 \bar{x} = 1{,}9975 - 1{,}0762 \cdot 1{,}8555 = 0{,}0006109$
- $\beta_1 = \dfrac{S_{xy}}{S_x^2} = \dfrac{0{,}24852922}{0{,}2309358889} = 1{,}0762$

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{2{,}236763}{10-1} =$$

$$= 0{,}24852922$$

$$S_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{2{,}078423}{10-1} = 0{,}2309358889$$

$$\bar{y} = \frac{1}{10} \cdot 19{,}975 = 1{,}9975$$

$$\bar{x} = \frac{1}{10} \cdot 18{,}555 = 1{,}8555$$

$$\hat{Y} = 0{,}0006109 + 1{,}0762\, x_1$$

Anonymkod : 0070-DCY ②

1b) För varje ökning i det begärda priset stiger slutpriset med 1,0762 miljoner vilket kan stämma för alla x-värden, däremot är slutpriset för x-värdet för x = 0 väldigt litet, med ett slutpris på 0,0006109 miljoner kronor.

1c) $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

SSE = 0,1827700964

SST = 2,590013

SSR = $b^2 \cdot 2{,}078423$ = $1{,}15820644 \cdot 2{,}078423$

= 2,40724

$b^2 = 1{,}0762^2$ (taget från 1a))

$R^2 = \frac{2{,}40724}{2{,}590013} = 0{,}9294$

Modellen förklarar att 92,94% av variationen i Y förklaras av x (och dess ändring).

Anonymkod : 0070-DCY (3)

1d) F-test

Hypotes: $H_0: \beta_1 = 0$ $H_a: \beta_j \neq 0$ $j=1$

Testvariabel: $F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{(1,8;0,05)}$

Beslutsregel: Förkasta H_0 om $F_{obs} > F_{(1,8;0,05)}$

$F_{(1,8;0,05)} = 5,32$ (enligt tabell 5)

Beräkning:

$SSR = 2,40724$ (från uppgift 1c)

$SSE = 0,18277$ (från uppgift 1c)

$$F = \frac{2,40724/1}{0,18277/8} = 105,3669$$

$F_{obs} = 105,3669 > 5,32$

Slutsats: H_0 förkastas på 95% signifikansnivå, modellen är signifikant då β_1 är skiljd från noll. X förklarar Y.

Anonymkod: 0070-DCY (4)

1e) t - test

Hypotes: $H_0: \beta_1 = 0$ $H_1: \beta_1 > 0$

Testvariabel: $t = \frac{b_j}{s_{b_j}} \sim t_{(8;0{,}05)}$

Beslutsregel: Förkasta H_0 om $t_{obs} > t_{(8;0{,}05)}$

H_0 kommer förkastas. Ett F-test (som i uppgift 1d) testar hela modellen, men då denna modell bara har ett β så blir det i stort sett som att göra ett t-test på endast en variabel.

Anonymkod : 0070 - DCY

1 f)

Söker : $\mu_{Y|x=3mkr}$

$$\hat{\mu}_{Y|x=3mkr} \pm t_{8;0{,}025} \cdot \sqrt{s_e^2\left(\frac{1}{10}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$$

$$s_e^2 = \frac{SSE}{8} = 0{,}02284625$$

$$(x-\bar{x})^2 = (3-1{,}8555)^2 = 1{,}30988$$

$$s_x^2 = 0{,}2309358889$$

$$\mu_{Y|x=x} = \beta_0 + \beta_1 x = 0{,}0006109 + 1{,}0762 \cdot 3$$
$$= 3{,}2292109$$

$$3{,}2292109 \pm 2{,}306 \cdot \sqrt{0{,}02284625\left(\frac{1}{10}+\frac{1{,}30988}{9 \cdot 0{,}2309358889}\right)}$$

$$3{,}2292109 \pm 2{,}306\sqrt{0{,}0166829674}$$

$$3{,}2292109 \pm 0{,}2978488675$$

$$(2{,}93136 \ ; \ 3{,}52706)$$

Med det begärda priset på 3 miljoner kommer slutpriset hamna i ett intervall mellan 2,9314 miljoner och 3,52706 miljoner. Allt detta gäller med en signifikansnivå på 95% ($\alpha = 0.05$)

Anonymkod: 0070-DCY (6)

2a) Modell 1

Source	DF	SumSq	MeanSq	F-value	Pr > F
Model	1	3 861,630	3 861,630	45,1769	
Error	30	2 564,338	85,4779		
Corrtot	31	6425,968			

	DF	b_j	s_{b_j}	t-value
Intercept	1	59,092	12,816	4,61079
X_1 (age)	1	1,6045	0,2387	6,7218

$$F\text{-value} = \frac{MSR}{MSE} = 45,1769 = F_{obs}$$

$$s_e^2 = \frac{2564,338}{30} = 85,4779$$

$$R^2 = \frac{SSR}{SST} = 0,60094 \approx 60,1\,\%$$

F_{krit} med frihetsgrader 1 och 30 och $\alpha = 0,05$

$F_{krit} = 4,17$

$\Longrightarrow$

Nästa blad för modell 2 och 3.

Anonymkod: 0070-DCY (7)

2a) Modell 2

Source	DF	Sum Sq	Mean Sq	F_{obs}	Pr>F
Model	2	4689,684	2344,842	39,1643	
Error	29	1736,285	59,87189		
Corr tot	31	6425,969			

Variabel	DF	b_j	S_{b_j}	t-value	VIF
Intercept	1	48,050	11,12956	4,31733	
X1 (age)	1	1,7092	0,201759	8,47144	
X2 (smk)	1	10,294	2,768107	3,71878	

$$S_e^2 = \frac{SSE}{n-k-1} = \frac{1736,285}{29} = 59.87189$$

$$R^2 = \frac{SSR}{SST} = \frac{4689,684}{6425,969} = 0,72980 \approx 72,9\%$$

$$F_{obs} = \frac{MSR}{MSE} = \frac{2344,842}{59,87189} = 39,1643$$

F_{krit} med frihetsgrader 2, 29 samt $\alpha = 0.05$

Då dessa frihetsgrader ej finns med i tabellsamlingen (tabell 5) får man antingen välja med fg. 2,25 eller fg. 2,30. fg. 2,30 är närmre modellens frihetsgrader, än 2,25, därför väljer jag 30. Ibland är det önskat att leta upp ett medelvärde av dessa alt. ta 25 för att ta det säkra förr det osäkra. $F_{krit\,2,30} = 3.32$

Anonymkod: 0070-DCY

8

2a) Modell 3

	DF	Sum Sq	Mean Sq	F-value
Model	3	4889,826	1629,942	29,7097
Error	28	1536,143	54,86225	
Corr.	31	6425,969		

	DF	bj	Sbj	t-value
Intercept	1	45,103	10,76468	4,189828
x1	1	1,2127	0,323819	3,74499
x2	1	9,9456	2,656057	3,74448
x3	1	0,0085924	0,0044987	1,90997

$$S_e^2 = \frac{SSE}{n-k-1} = \frac{1536,143}{28} = 54,86225$$

$$R^2 = \frac{SSR}{SST} = \frac{4889,826}{6425,969} = 0,76094 \approx 76,1\%$$

$$F_{obs} = \frac{MSR}{MSE} = 29,7097$$

F_{krit} för fg. 3, 28 samt $\alpha = 0,05$ enl tabell 5. Av samma anledning som i Modell 2 väljer jag att ta F-värdet för frihetsgraderna för 3, 30 = $F_{3,30;0.05} = 2,92$.

Anonymkod: 0070 - DCY

2b) KI för X1, X2, X3 med 95% signifikansnivå

$b_j \pm t_{28;0.025} \cdot S_{b_j}$

X_1: $1{,}12127 \pm 2{,}048 \cdot 0{,}323819$

$(0{,}5495 \ ; \ 1{,}8759)$

X2: $9{,}9456 \pm 2{,}048 \cdot 2{,}656057$

$(4{,}5059 \ ; \ 15{,}3852)$

X3: $0{,}0085924 \pm 2{,}048 \cdot 0{,}0044987$

$(-0{,}0006209 \ ; \ 0{,}017805)$

Resultaten för X1 och X2 är bra, men då X3 har ett intervall som innefattar noll och/eller negativa värden kan man dra slutsatsen att X3 inte är signifikant skiljd från noll, och därmed inte bidrar till modellen.

(vilket även är det som man tittar efter).

(I denna modell är det bara fastställt att detta gäller på ett 95% signifikansnivå via mitt test, kan dock hända att X3 är bra med en annan signifikansnivå).

Anonymkod : 0070 - DCY

2c)

Jag hade valt Modell 2,

detta då förklaringsvariabeln X3
inte bidrar till att förklara modellen
(enligt svaret i 2c)) mer utvecklat än
vad X1 och X2 redan gjort.
Hade inte valt modell 1 då 2 är mer
utvecklad, och hade inte valt modell pga
det jag skrev ovan.
Ja, detta är säkerställt då testet för modell 3
är okej för X1 och X2, test för modell
1 och 2 hade kommit till samma slutsats
då en ytterligare X (som X3) endast
försvårar för X1 och X2 och deras
signifikans.

2d)

På Nästa sida →

Anonymkod : 0070 - DCY

11

2d) I en multipel regressionsmodell måste följande vara uppfyllt:

1 Den oberoende variabeln (x) och den beroende modellen (Y) måste ha ett samband, ökar x måste även Y öka.

2. Y (den beroende) variabeln måste vara av intervalldata (mätvar), som exempelvis tempentur, som då blir påverkad av olika faktorer (vind, väder, årstider).

3. Summan av samtliga differenser mellan det verkliga Y-värdet och det skattade Y-värdet måste alltid vara lika med noll.

4. Medelvärdet av samma Y- och $\hat{Y}$-värden som i "3" ska alltid vara lika med noll.

5 Observationer för den oberoende variabeln (x) är okorrelerad, annat fall är detta autokorrelerat, vilket ofta är en dum modell.

De oberoende variablerna (x) kan vara korrelerade med varandra, men det får inte bli ett tillfälle där summan av samtliga $(C_0 + C_1 X_{1i} + C_2 X_{2i} + C_k X_{ki} = 0)$, detta kallas multikollinearitet, och går att (via en Anova) tabell utläsa med ett VIF (varians inflation) värde, som bör vara mindre än 5 eller 10.

Anonymkod: 0070 - DCY

2e)

	Y	x1	x2
Y	1		
x1	0,775	1	
x2	0,247	-0,139	1

$r_{x_1x_2} = -0,139$

$r^2_{x_1x_2} = 0,019321 = R^2_{x_1x_2}$

$VIF = \frac{1}{1-0,019321} = 1,0197$

Denna modell utstår ingen risk för multikollinearitet då VIF-värdet är långt under 5, vilket tumregeln (inom VIF) visar på.

Anonymkod: 0070-DLY (13)

3a) Med ett spuriöst samband menar man att det ej finns någon orsaksmässig (kausal) effekt mellan den beroende (Y) variabeln och den oberoende variabeln (X):
Exempelvis kan ett samband mellan längd och hastighet vara spuriöst, det kan dock finnas ett indirekt samband mellan variablerna.

3b) Överanpassning och modellanpassning är nära besläktade, dessa anpassningar är problemet huruvida man ska dra sin regressionslinje baserat på hur datan ser ut.
Hur ska man skatta en linje på en modell som har väldigt "känsliga" data.
Det går vid många fall att använda både en rät linje, exponentiell ökning (andragrad) och tredjegradsekvationer, problemet ligger i vilken av dessa man bör använda sig av samt varför.

Anonymkod: 0070-DCY (14)

5a) $\text{Log Odds}(Y=1 | x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$\text{Log Odds}(Y=1 | 40000, 1, 1) =$

$-11{,}3828 + 0{,}000163 \cdot 40000 + 1{,}3420 \cdot 1 + 1{,}0606 \cdot 1$

$= -2{,}4602$

$\text{Odds} = \exp(-2{,}4602) = 0{,}0854178657$

$$P(Y=1 | 40000,1,1) = \frac{0{,}0854178657}{1 + 0{,}0854178657} = 0{,}0786958$$

$P(Y=1 | 40000,1,1) = 0{,}0786958 \approx 7{,}86\%$

5b) 95% KI for $OR(X_1)$

$(\exp(b_j - z_{\alpha/2} \cdot s_{bj}) ; (\exp(b_j + z_{\alpha/2} \cdot s_{bj})$

$(\exp(0{,}000163 - 1{,}96 \cdot 0{,}000016) ;$

$(\exp(0{,}000163 + 1{,}96 \cdot 0{,}000016)$

$(1{,}000131649 ; 1{,}000194379)$

Ja, jag anser att inkomsten spelar en viktig roll. Koefficienten är såpass liten då den i de flesta fall kommer multipliceras med stora tusental, och inte 1 och 0 som de andra koefficienterna multipliceras med.
Om personen är en singelman med en inkomst är det endast inkomsten som separerar, mao är den viktig

4a) En tidsserie är ofta karaktäriserad
på ~~s~~ sätt att x-axeln är en tidsförändring
och Y är den beroende variabeln som undersöks.
Observationerna är beroende av varandra
och det finns samband mellan dessa.
Oftast går observationerna med ett stigande
värde över tid och har en periodisk
uppdelning (eller variation) över en viss tidsperiod,
årsdata, kvartalsdata, månadsdata osv.

Just denna modell ser ut som att passa
med en multiplikativ modell. Detta då
svängningarna ökar när trendens nivå går
upp.