Regression och tidsserieanalys

0010-LOW

① $y_i = \alpha + \beta x_i + \varepsilon_i$

a) $$b = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

siffror givna i tentan

$\Sigma x^2 = 1791{,}25$ $\quad \Sigma y^2 = 29923$

$\Sigma xy = 7316{,}60$

$\Sigma x = 162{,}5$ $\quad \Sigma y = 665$ $\quad n = 15$

$$b = \frac{15 \cdot 7316{,}50 - 162{,}5 \cdot 665}{15 \cdot 1791{,}25 - 162{,}5^2}$$

$$= \frac{1685}{462{,}5} = 3{,}643243243 \simeq 3{,}6432$$

$$a = \bar{y} - b\bar{x} = \frac{665}{15} - 3{,}6432 \cdot \frac{162{,}5}{15}$$

$= 4{,}864864865 \simeq 4{,}8649$

$\bar{y} = 44{,}33$

$\bar{x} = 10{,}83$

$$\hat{y}_i = 4{,}8649 + 3{,}6432 x_i$$

Regression och tidsserieanalys

0010 - LOW

b) givet: $SS_E = 32,076$

Upg 1. Fort: Bestäm R^2 och ge en tolkning:

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

$$SST = \Sigma (y_i - \bar{y})^2 = (n-1) s_y^2$$

$$s_y^2 = \frac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n(n-1)} = \frac{15 \cdot 29923 - 665^2}{15(15-1)}$$

$$= \frac{6620}{15 \cdot 14} = 31,5238$$

$$31,5238 \cdot (15-1) = 441,333$$

$$SSR = 441,333 - 32,076 = 409,257$$

$$\frac{SSR}{SST} = 0,9273$$

92,73% av variationen i y kan förklaras av (med hjälp av modellen) genom variationen i x

Fort. → sid 3

c) 95% KI för β → $\beta \pm t_{\frac{\alpha}{2}}(n-k-1) \cdot s_{bj}$

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

$$s_x^2 = (n\Sigma x_i^2 - (\Sigma x_i)^2) / (n(n-1))$$

$$= \frac{15 \cdot 1791,25 - 162,5^2}{15 \cdot (15-1)}$$

$$= 462,5 / (15 \cdot 14) = 2,2024$$

$$s_e^2 = MSE = \frac{SSE}{n-k-1}$$

$$= \frac{32,076}{15-1-1} = 2,4674$$

Regression och tidsserieanalys

0010 - LOW

c) Fort

upg. 1

$s_x^2 = 2,2024$

$s_e^2 = 2,4674$

$s_b^2 = \frac{2,4674}{(15-1)\cdot 2,2024} = 0,080023284 \simeq 0,0800$

$\sqrt{s_b^2} = s_b = 0,282888387 \simeq 0,2829$

$t_{\frac{\alpha}{2}}(n-k-1) = t_{0,025}(13) = 2,160$ (Från tabell)

$b \pm t_{\frac{\alpha}{2}} \cdot s_{bj} \rightarrow 3,6432 \pm 2,16 \cdot 0,2829$

$3,6432 \pm 0,6110$

(3,0322 ; 4,2542) 95% KI för β

(vi ser när också att $\beta \neq 0$ då intervallet inte täcker 0)

d) 95% KI för $\hat{y} \mid X = 12$

$\hat{y} = 4,8649 + 3,6432 \cdot 12 = 48,5833$

$\bar{x} = \frac{162,5}{15} = 10,8333$

$\hat{y} \pm t_{\frac{\alpha}{2}} \cdot \sqrt{s_e^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$

$48,5833 \pm 2,16 \cdot \sqrt{2,4674 \left(\frac{1}{15} + \frac{(12-10,8333)^2}{(15-1)\cdot 2,2024}\right)}$

$\rightarrow 48,5833 \pm 1,1294$

(47,4539 ; 49,7127)

95% KI för $\hat{y} \mid X = 12$

siffror redan uträknade innan.

0010-LOW

Uppg 1 Fort.

e) 95% PI $\hat{y} \mid X = 12$

$\hat{y} \mid X = x$ samma som d) → 48,5833

$$\hat{y} \pm t_{\frac{\alpha}{2}} \cdot \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$$

$$\rightarrow 48{,}5833 \pm 2{,}16 \cdot \sqrt{2{,}4674 \left(1 + \frac{1}{15} + \frac{(12-10{,}833)^2}{(15-1)\cdot 2{,}2024}\right)}$$

uträknade sedan innan

$48{,}6833 \pm 3{,}575967$

$\simeq 3{,}6760$

$(45{,}0073 ; 52{,}1593)$ PI för $\hat{y} \mid X = 12$

Jämfört med d) går det att se att intervallet är längre.

Formeln för PI innehåller ju en extra +1 under $\sqrt{\ }$.

KI är för det "sanna" okända värdet, baserat på hela det observerade datamaterialet.

Medan PI undersöker vad framtida observation kan få för värde.

Det är mer osäkert på oobserverat data. Därmed är intervallet längre. Vilket skapar "större utrymme att hamna inom".

0010 - LOW

Upg (2)

a) Antaganden:

1. Sambandet mellan x och y ska vara linjärt!

   Dvs det ska gå att skriva sambandet enligt:

   $$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon_i$$

2. Felltermerna ska vara oberoende varandra. Dvs. oberoende observationer

3. Lika varians, dvs homoskedasticitet

   variansen för y ska inte bero på x.

4. Förklaringsvariablerna ska inte vara linjär kombinationer av varandra.

   Dvs. det ska inte finnas multikollinearitet mellan Förklaringsvariablerna

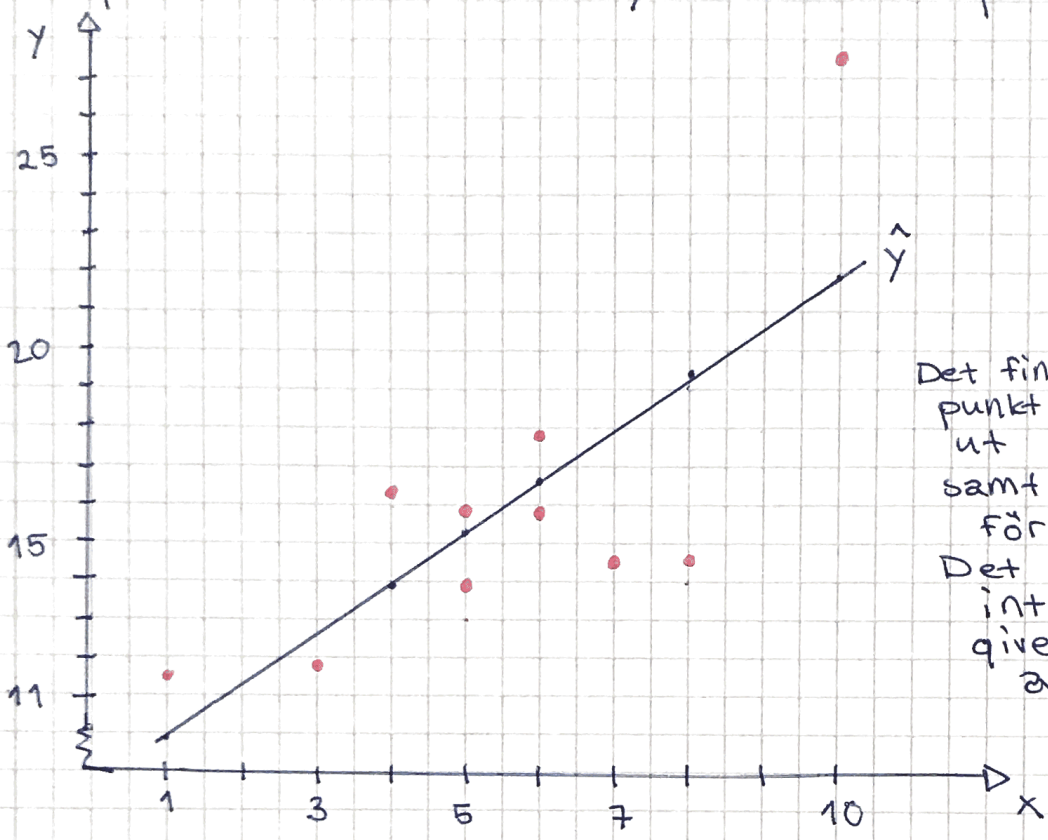
5. Existensvilkoret: parametrarna i modellen ska vara ändliga.
   dvs. de inte gå mot plus/minus oändligheten.

0010-LOW

Fort ②

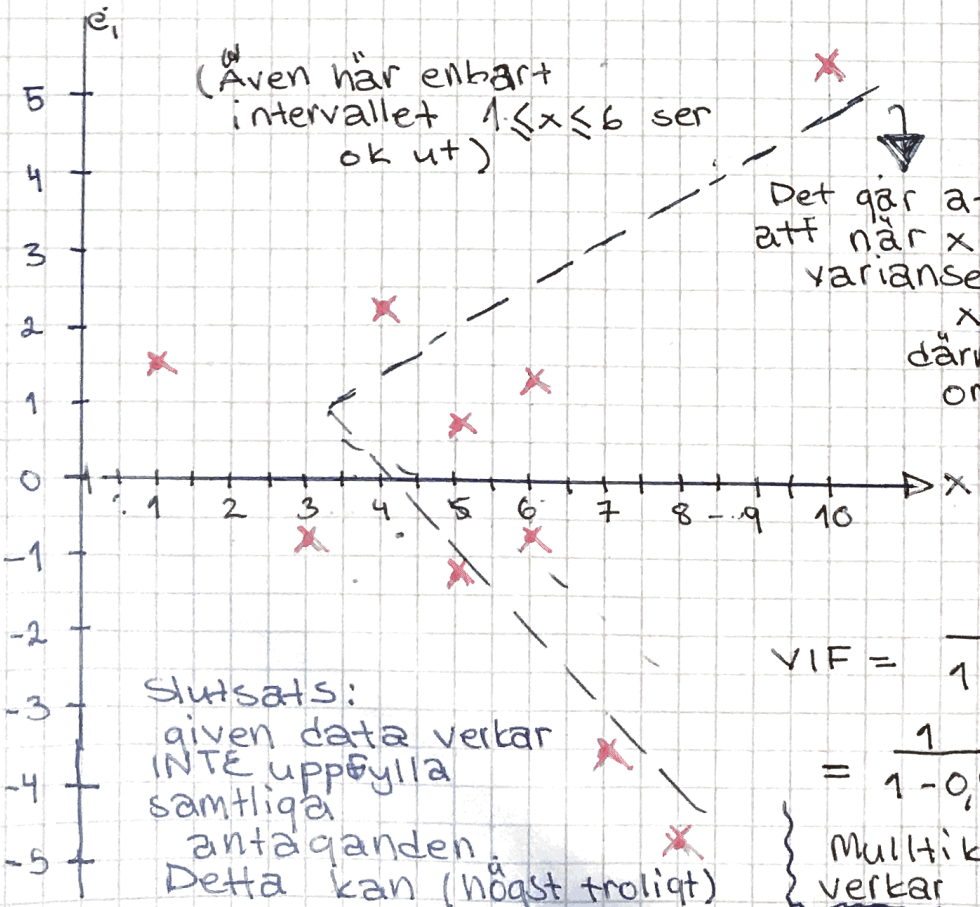
b) Ändligt? JA vi har fasta värden som inte går $+-\infty$

Linjärt? Plottar x mot y: Sätter in givna värden.



Det finns en tydlig punkt som "sticker" ut $x=10$. samt verka y minsta för $x=7$ $x=8$. Det är därmed inte säkert givet given data att modellen är linjär. (Bortser man får $x=7$ $x=8$ $x=10$ kan dock ett linjärt mönster antydas.)

Residualerna? Plottar dem (värderna: $y_i - \hat{y}_i = e_i$)



(Även här enbart intervallet $1 \leq x \leq 6$ ser ok ut)

Det går att se att när x ökar ökar variansen för värderna $x \geq 7$. Det finns därmed misstanke om heteroskedasticitet. (Det syns ett "tratt" mönster)

Slutsats: given data verkar INTE uppfylla samtliga antaganden. Detta kan (högst troligt) bero på $n=10$. Krävs minst $n=30$ för ok data analys

$$VIF = \frac{1}{1 - R^2_{yx}}$$

$$= \frac{1}{1 - 0{,}58} = 2{,}381 < 10$$

Multikollaritet verkar INTE föreligga.

0010 - LOW

Uppg ③ Givna siffror:

$n = 22$ $\quad SS_E = 140$ $\quad SS_R = 260$

$a = 5,2$

$b = -4,4$

F-test då vi enbart har 1 beta.

$F_{obs} = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{260/1}{140/(22-1-1)}$

$= \frac{260}{7} = 37,1429$

Bestäm R^2 och testa

$H_0: \beta = 0$ mot $H_a: \beta \neq 0$

$R^2 = \frac{SSR}{\underbrace{SSR + SSE}_{SST}}$

$= \frac{260}{260+140} = \frac{260}{400} = \frac{0,65}{\approx}$

$F_{krit} = 8,10$

$v_1 = 1$ (från tabell)

$v_2 = 20$

$\alpha = 0,01$

$F_{obs} > F_{krit}$

därmed kan vi förkasta H_0

β_0 är $\neq 0$

Hypotestest: $\frac{b-\beta}{s_b}$ under H_0 sann $\rightarrow$ $\frac{b}{s_b}$

Alternativt F-test

går inte räkna då vi ej kan räkna ut s_x^2.

Använder istället:

$\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

då vid enkel linjär regression är $R^2 = r_{xy}^2$

$r_{xy} = \sqrt{R_2} = 0,80622257...$

$\rightarrow \frac{\sqrt{0,65}\cdot\sqrt{22-2}}{\sqrt{1-0,65}} = 6,0945 = t_{obs}$

$\rho = 0$ mot $\rho \neq 0$

undersöker också om $\beta = 0$ mot $\beta \neq 0$ och därav kan vi använda korrelationskoefficient-testet istället!

Testen är ekvivalenta.

$t_{krit\frac{\alpha}{2}(n-k-1)}$

$t_{0,005}(20) = 2,845$

$t_{obs} > t_{krit}$ vi kan därmed förkasta H_0.

på 1% är β skilt från noll!

0010-LOW

Uppg. (4)

a) Kan tänkas bestå av komponenterna

Tr = Trend

[K = cyklisk variation (konjuktur)
Detta är dock svårt särskilja från trenden och därav brukar Tr få innefatta K]

S = Säsong

[E = slumpkomponent. Går inte heller veta exakt.]

$Y = Tr + S$ ← Därmed slutmodell.

S → beroende på ex olika årstider förbrukas elen olika. Antagande ex mer under vintern och mindre under sommare

Tr → Urskiljning av en långtidsutveckling. det går att urskilja om elförbrukningen exempel ökar / minskar genom tiden.

Forts. ny sida

Uppg (4)

Fort.

b)

Additativ.

År	kvartal	Y_t	$\hat{T}_t$	$Y_t - \hat{T}_t = \hat{S}_t$	
2016	1	7,5	*	*	
	2	1,5	*	*	
	3	2,5	4	-1,5	(2,5-4)
	4	5	3,9375	1,0625	
2017	1	6,5	4,0625	2,4375	(6,5-4,0625)
	2	2	4,3125	-2,3125	
	3	3	4,75	-1,75	(3-4,75)
	4	6,5	5,25	1,25	
2018	1	8,5	5,625	2,875	(8,5-5,625)
	2	4	5,9375	-1,9375	
	3	4	6,1875	-2,1875	(4-6,1875)
	4	8	6,3125	1,6875	
2019	1	9	6,5625	2,4375	(9-6,5625)
	2	4,5	6,8125	-2,3125	
	3	5,5	*	*	
	4	8,5	*	*	

Kvartalsdata:

$$\hat{T}_t = \frac{Y_{t-2} + 2Y_{t-1} + 2Y_t + 2Y_{t+1} + Y_{t+2}}{8}$$

* $\hat{T}_t$ för kv 1 och 2 2016 samt för kv 3 och 4 2019 försvinner då vi inte har data för $t-2$ / $t-1$ resp. $t+2$ / $t+1$ för dessa.

2016 Kv3 : $(7,5 + 2\cdot 1,5 + 2\cdot 2,5 + 2\cdot 5 + 6,5) / 8 = 4$

kv4 : $(1,5 + 2\cdot 2,5 + 2\cdot 5 + 6,5\cdot 2 + 2) / 8 = 3,9375$

⋮

2019 kv2 : $(8 + 9\cdot 2 + 4,5\cdot 2 + 5,5\cdot 2 + 8,5) / 8 = 6,8125$

Upg 4 fort.

säsongskomponenter

c)

kvartal	1	2	3	4
2016	*	*	-1,5	1,0625
2017	2,4375	-2,3125	-1,75	1,25
2018	2,875	-1,9375	-2,1875	1,6875
2019	2,4375	-2,3125	*	*
$\bar{S}_i$: Medel värde:	2,58333	-2,1875	-1,8125	1,33333

$\Sigma = -0,08333$

kv 1 $(2,4375 + 2,875 + 2,4375) / 3$

kv 2 $(-2,3125 + (-1,9375) + (-2,3125)) / 3$

kv 3 $(-1,5 + (-1,75) + (-2,1875)) / 3$

kv 4 $(1,0625 + 1,25 + 1,6875) / 3$

$2,58333 - 2,1875 - 1,8125 + 1,33333$

Σ ska bli noll och måste därmed justeras. Detta görs genom att summan av medelvärdet subtraheras för varje kvartal.

$\frac{-0,08333}{4} = -0,0208333$

$2,5833 - (-0,0208333) = 2,6041666 = S_1^+$

$-2,1875 - (-0,0208333) = -2,16666 = S_2^+$

$-1,8125 - (-0,0208333) = -1,79166 = S_3^+$

$1,3333 - (-0,0208333) = 1,3541666 = S_4^+$

$\Sigma = 0$

skattade justerade säsongskomponenter

Gör på liknande vis men tar $\bar{Y}_t / \bar{T}_t = \bar{S}_t$ sen tar man $\bar{S}_i / (\bar{S}_1 + \bar{S}_2 + \bar{S}_3 + \bar{S}_4)$ $\times 400$ (4 kvartals data)

Hade vi istället använd multiplikativ hade

$S_1 = 149,93$ $S_2 = 60,14$ $S_3 = 63,65$ $S_4 = 126,28$

0010-LOW

Upg (5) $n = 20$

$y = \begin{cases} 1 = \text{ny hjärtattack inom 1år} \\ 0 = \text{annars} \end{cases}$

$x_1 = \begin{cases} 1 = \text{behandling mot oro} \\ 0 = \text{annars} \end{cases}$

x_2 = orosindex (högre värden - större oro)

Modell från utskrift:

$$\text{LogOdds} = -7{,}388 + 1{,}024x_1 + 0{,}119x_2$$

a) $P(Y=1 \mid x_1=1, x_2=60)$

$$= \exp(\alpha+\beta_1x_1+\beta_2x_2) \Big/ (1+\exp(\alpha+\beta_1x_1+\beta_2x_2))$$

$$\text{logodds} = -7{,}388 + 1{,}024 \cdot 1 + 0{,}119 \cdot 60 = 0{,}776$$

$$\exp(0{,}776) / [1+\exp(0{,}776)] = 0{,}684817738$$

sannolikheten $P(Y=1 \mid x_1=1, x_2=60)$ är 0,6848

b) 1. Vi ser i tabellen att variabeln inte är signifikant. Vilket gör att man där kan fundera att den bör tas ur modellen.

2. Vi vet också att för att det ska vara effektivt för att minska sannolikheten ska $\beta_1 < 0$ alt. $e^{\beta_1} < 1$. Vi ser i tabellen att $\beta = 1{,}024 > 0$ och kan räkna ut $e^{1{,}024} = 2{,}78 > 1$. Båda är större än 0 resp. 1. vilken antyder det ökar sannolikheten.

En check med $P(Y=1 \mid x_1=0, x_2=60)$ ger oss sannolikhet på 0,4383 vilket är lägre än 0,6848.

3. Men att sannolikheten ökar givet terapi låter högst otroligt. Och då vi ser att variabeln inte är signifikant bör vi inte dra några slutsatser om hur den påverkar. Vi bör testa med ett större urval alt. ta bort den variabeln.