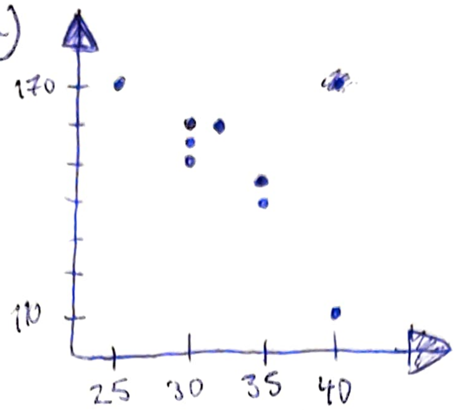
0065-YJZ

Tentamen regressionsanalys 210111

uppgift 1

a)



$n = 8$

$\bar{x} = 32{,}125$

$\bar{y} = 148{,}75$

x_i	y_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
40	110	62,015663	1507,563
35	145	8,265625	14,0625
35	140	8,265625	76,5625
32	160	0,015625	126,5625
30	160	4,515625	126,5625
30	150	4,515625	1,5625
25	170	50,765663	457,5625
30	155	4,515625	39,0625
257	1190	142,875	2337,5

$x_i \cdot y_i$	x_i^2
4400	1600
5075	1225
4900	1225
5120	1024
4800	900
4500	900
4250	625
4650	900
37695	8399

$$s_x^2 = \frac{1}{n-1}\sum_{n=1}^{n}(x_i-\bar{x})^2 = 142{,}875/7 \approx 20{,}411$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\cdot 37695 =$$

$$= \frac{37695}{7} = 5385$$

$$s_y^2 = \frac{1}{n-1}(y_i-\bar{y})^2 = \frac{1}{7}\cdot 2337{,}5 \approx 333{,}93$$

$$\boxed{\begin{aligned} s_x^2 &= 20{,}411 \\ s_y^2 &= 333{,}93 \\ s_{xy} &= 5385 \end{aligned}}$$

b) $Y_i = \alpha + \beta x + \varepsilon$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{8\cdot 37695 - 257\cdot 1190}{8\cdot 8399 - 257^2} = \frac{-4270}{1143} \approx -3{,}7358$$

$$a = \bar{y} - b\bar{x} = 148{,}75 - (-3{,}7358)\cdot 32{,}125 \approx 268{,}76$$

$$\hat{Y} = 268{,}76 - 3{,}74x$$

Antaganden:
- Det råder ett linjärt samband mellan beroendevariabeln (y) och förklaringsvariabeln (x)
- Feltermerna är normalfördelade med väntevärde 0 och konstant varians σ_ε^2.

Forts. på sida 2

- Feltermerna är oberoende av varandra
- X är oberoende av feltermen ε_i

▫ Det går inte att dra några slutsatser om kausalitet. Samvariation betyder inte att det måste finnas ett kausalt samband.

c) Jag skattade visst redan parametrarna.

$$\hat{y} = \underbrace{269}_{\alpha/\beta_0} - \underbrace{3{,}74}_{\beta_1}x$$

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

jag använder exakta värden från b) för α och β och erhåller följande:

$\hat{y}_i$	$(\hat{y}_i - \bar{y})^2$
119,3307	865,4947
138,0096	115,3557
138,0096	115,3557
149,217	0,218064
156,6885	63,0204
156,6885	63,0204
175,3675	708,4889
156,6885	63,0204
	1993,974 → $\Sigma(\hat{y}_i - \bar{y})^2$

Vi räknade ut SST i b) och kan sätta in värdena:

$$\frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y_i - \bar{y})} = \frac{1993{,}974}{2337{,}5} \approx 0{,}853$$

Tolkningar: • Modellen förklarar ca 85% av variationen i Y, dvs kvoten mellan modellens variation och den totala variationen är ca 0,85.

• Det råder ett ganska starkt negativt samband mellan pris i kr och säld kvantitet av jordgubbar (åtminstone under obs. perioden).

Forts. på sida 3

Man ska vara mycket försiktig med att tolka α, men vi gör det för skojs skull: om jordgubbarna är gratis förväntar man sig att sälja 269 liter under en dag. För varje krona som priset höjs minskar den förväntade försäljningen med 3,74 liter.

Obs! Farligt att extrapolera utanför observationsområdet. Det går t.ex. inte att få negativ försäljning, hur högt priset än är.

d) Ett 95% konfidensintervall erhålls enligt

$$\beta \pm t(n-2) \cdot s_\beta$$

$$s_\beta = \frac{s_e}{\sqrt{\Sigma x_i^2 - n\bar{x}^2}} \qquad s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} \qquad \Sigma(y_i-\hat{y}_i)^2 = SSE$$

$$SSE = SST - SSR$$

- $SST - SSR = 2337{,}5 - 1993{,}974 = 343{,}562$

$$s_e = \sqrt{\frac{343{,}562}{6}} = 7{,}5667$$

$$s_\beta = \frac{7{,}5667}{\sqrt{8399 - 8 \cdot 32{,}125^2}} = \frac{7{,}5667}{11{,}95303} \approx 0{,}633$$

$$\beta \pm t(n-2) \cdot s_\beta = -3{,}7358 \pm t_{0{,}025}(6) \cdot 0{,}63304 =$$

$$= -3{,}7358 \pm 2{,}447 \cdot 0{,}63304 \approx -3{,}74 \pm 1{,}55$$

Alternativt från $-5{,}29$ till $-2{,}19$

e) Om priset <u>inte</u> hade inverkan på såld kvantitet skulle $\beta = 0$. Vårt 95% intervall täcker emellertid inte 0, så svaret är ja. Priset <u>har</u> signifikant påverkan på såld kvantitet.

Uppgift 2

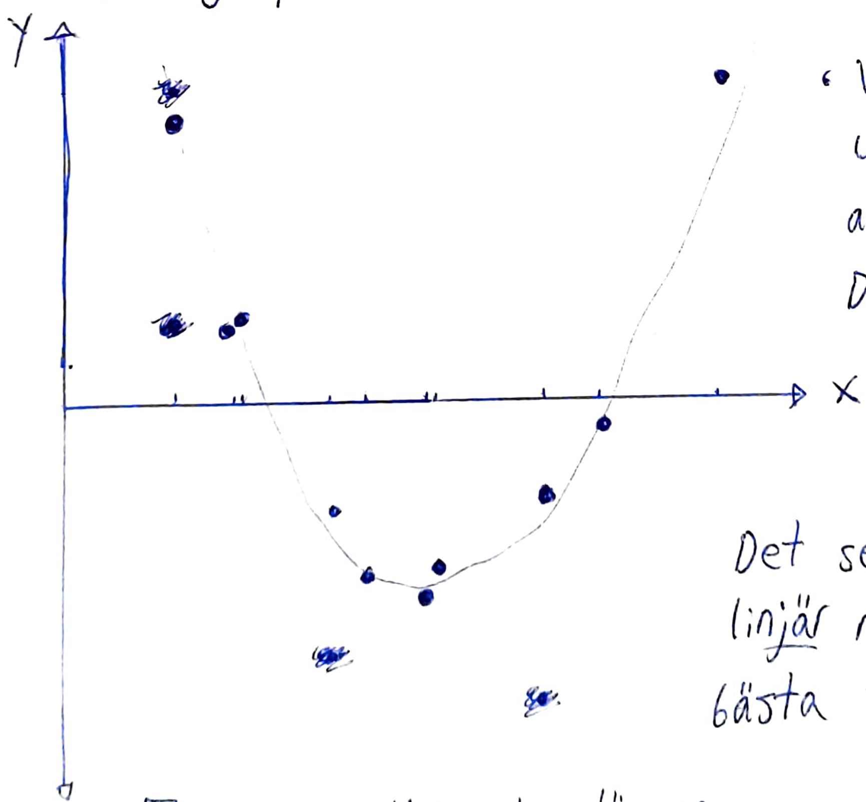
a) $\boxed{1}$ • Beroendevariabeln är en linjär funktion av de oberoende variablerna (förklaringsvariablerna)

$\boxed{2}$ • samtliga förklaringsvariabler är oberoende av feltermerna ε_i

$\boxed{3}$ • Feltermerna är oberoende av varandra

$\boxed{4}$ • Feltermerna är normalfördelade med väntevärde 0 och konstant varians σ_ε^2 (ingen heteroskedasticitet)

b) Jag plottar residualerna:



• Variansen ser inte ut att vara oberoende av förklaringsvariabeln. Den ser heller inte ut att vara konstant.

Det ser ut som att en linjär modell inte är den bästa i detta fall.

$\boxed{1}$ verkar alltså inte stämma

$\boxed{2}$ verkar inte heller uppfyllas

$\boxed{4}$ Variansen ej konstant för den linjära modellen

Uppgift 3

$Y = \alpha + \beta X + \varepsilon$

$SSR = 15,36$
$F = 16$

$F = \frac{MSR}{MSE}$ $MSR = \frac{SSR}{k}$ $MSE = \frac{SSE}{n-k-1}$

k = antal förklarande variabler = 1
n = antal observationer = 7

$MSR = \frac{SSR}{1} = SSR = 15,36$

$MSE = \frac{MSR}{F} = \frac{15,36}{16} = 0,96$

$0,96 = \frac{SSE}{n-k-1} = \frac{SSE}{7-1-1} = \frac{SSE}{5} \rightarrow SSE = 0,96 \cdot 5 = 4,8$

$SST = SSE + SSR = 4,8 + 15,36 = 20,16$

$R^2 = SSR/SST = 15,36/20,16 \approx 0,762$

$\boxed{R^2 = 0,762}$

Observationernas varians runt regressionslinjen, d.v.s. residualvariansen, är $\Sigma(Y_i - \hat{Y}_i)^2 = \frac{SSE}{n-2} = 4,8$

Vi testar om villornas storlek signifikant påverkar uppvärmningskostnaden på följande vis:

Testvariabel: F som är F-fördelad med k och n-k-1 frihetsgrader.

$H_0: \beta = 0$ $H_1: \beta_1 \neq 0$

H_0 förkastas om $F_{obs} > F_{1;5;0,05}$

$F_{obs} = 16$
$F_{1;5;0,05} = 6,61$
$16 > 6,61$, H_0 förkastas på 5%-nivån.

Svar: Villastorlek påverkar signifikant uppvärmningskostnaden

Uppgift 4

a) Man brukar skilja mellan fyra komponenter:

- Trend (T)
- Cyklisk variation (C)
- säsongsvariation (S)
- slumpkomponent (E)

Trenden är långtidsutvecklingen av Y, alltså den allmänna utvecklingen över lång tid.

Cyklisk variation kallas även konjunktur. Konjunkturen har lång omloppstid och varierar runt trenden.

Säsongsvariation är också cyklisk men har kortare omloppstid, i regel ett år.

Slumpkomponenten är, som namnet antyder, en oregelbunden komponent som alltså saknar cyklicitet/trend.

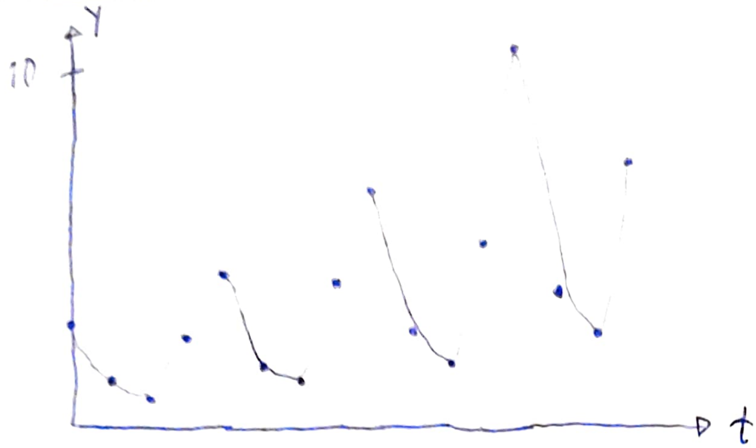
- Komponenterna kan komponeras till additiva eller multiplikativa modeller:

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

En skillnad mellan modellerna är att säsongsvariationen i en serie med positiv trend blir allt större, medan den är konstant i en additiv modell. Vid starka trender blir alltså skillnaderna mellan modellerna väldigt stora.

b) Grov plot av tidsserien:

(se nästa sida)



vi delar upp observationerna i kvartal:

År	Kvartal			
	1	2	3	4
2016	2,9	1,2	0,8	2,4
2017	4,1	1,7	1,2	4,0
2018	6,6	2,7	1,8	5,1
2019	10,7	3,9	2,7	7,5

Jag antar additiv modell.

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

Tiden rann ut

Uppgift 5

a) $\log odds(Y=1 \mid 25;28) = -3{,}388 - 0{,}102 \cdot 25 + 0{,}119 \cdot 28 = -2{,}606$

$odds(Y=1 \mid 25;28) = \exp(-2{,}606) = 0{,}07383$

$\therefore P(Y=1 \mid 25;28) = \dfrac{0{,}07383}{1+0{,}07383} \approx 0{,}069$

Svar: 6,9 %

b) Jag ser direkt att parameterskattningarna är mindre än en standardavvikelse från 0, så nollhypotesen att de inte påverkar sannolikheten för 0 kan inte förkastas.