

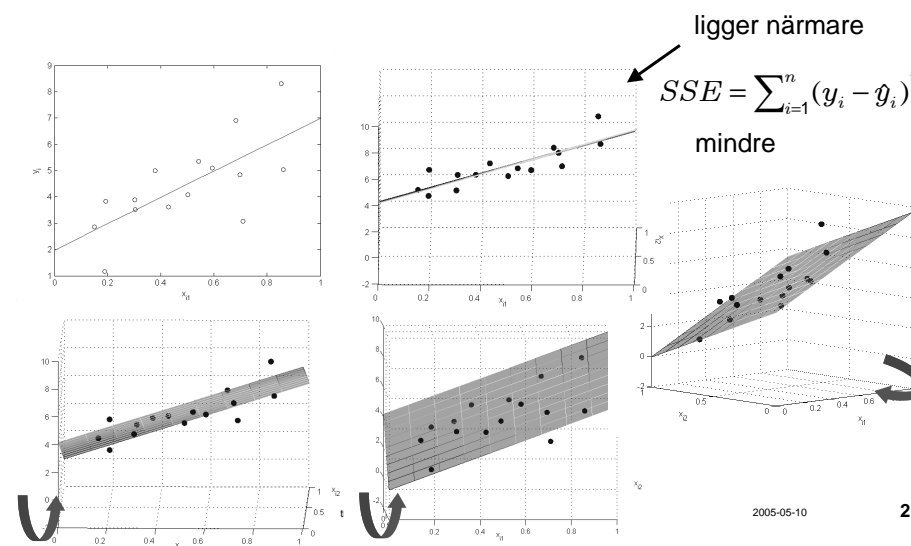


Johan Koskinen, Statistiska institutionen, Stockholms universitet

Finansiell statistik, vt-05

F18 regressionsanalys & logistisk regression

Förklaringsgrad och antalet prediktorer



Förklaringsgrad och antalet prediktorer

När man lägger till prediktorer kan

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

bara minska

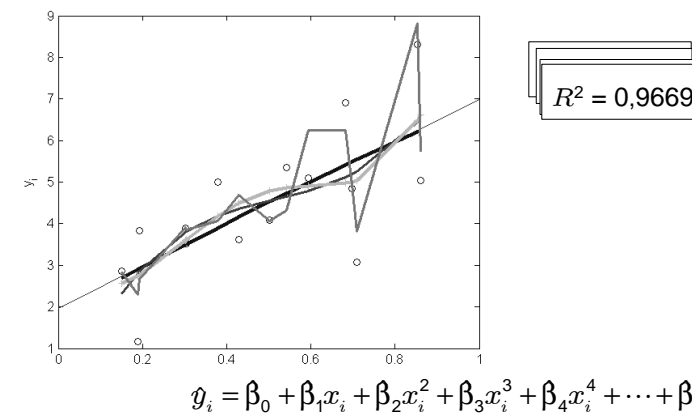
Eftersom andelen förklarad variation i Y

$$R^2 = \frac{\text{förklarad variation}}{\text{total variation}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

ökar R^2 alltid när man lägger till förklarande variabler

Förklaringsgrad och antalet prediktorer

Exempel på att R^2 ökar när man lägger till förklarande variabler



Förklaringsgrad och antalet prediktorer

P.g.a. detta är ett bättre mått på en models anpassning
den justerade determinationskoefficienten:
för en modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

definieras den justerade determinationskoefficienten som

$$\bar{R}^2 = 1 - \frac{\text{ej förklarad variaton/f.g.}}{\text{total variation/f.g.}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

i ex med 9 prediktorer

$$R^2 = 0,9669$$

$$\bar{R}^2 = 1 - \frac{1,344/(15-9-1)}{40,59/(15-1)} = 0,9073$$

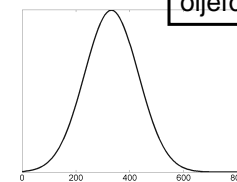
Johan Koskinen, Department of Statistics

2005-05-10

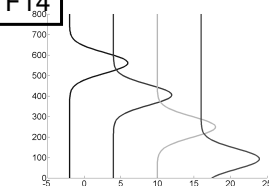
5

Prediktioner

Eftersom vi nu har en modell för Y_i för olika värden på $x_{i1}, x_{i2}, \dots, x_{ik}$ borde vi kunna säga var ett värde på borde ligga med större säkerhet när vi har $x_{i1}, x_{i2}, \dots, x_{ik}$ än utan



oljaförbrukningen från F14



I stället för t.ex. $Y_i \in N(\mu, \sigma^2)$

En fördelning för varje värde på x_i :



Johan Koskinen, Department of Statistics

2005-05-10

6

Prediktioner

Den i någon mening bästa gissningen vi kan göra av utfallet på Y_{n+1} när vi endast har $x_{n+1,1}, x_{n+1,2}, \dots, x_{n+1,k}$ är

$$\hat{y}_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \beta_2 x_{n+1,2} + \dots + \beta_k x_{n+1,k}$$

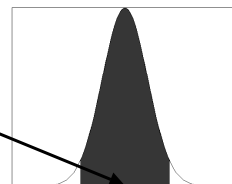
alltså

$$\hat{y}_{n+1} = \mu_{n+1} = E(Y_{n+1})$$

Eftersom Y_{n+1} är normalfördelad kan vi ge ett intervall som det är 95% chans att utfallet hamnar i

Vi behöver dock standardavvikelsen för Y_{n+1}

$$\hat{y}_{n+1} = \mu_{n+1}$$



Johan Koskinen, Department of Statistics

7

Prediktioner

För enkel linjär regression

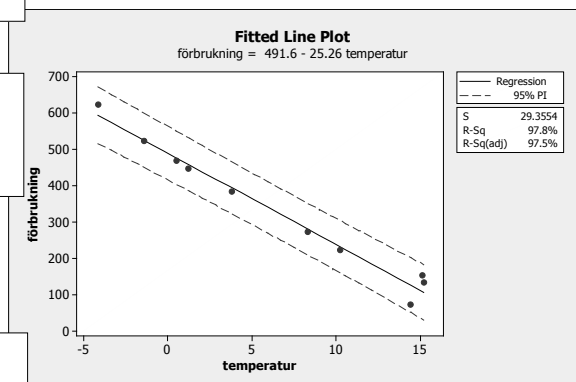
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

ges en skattning av standardavvikelsen för Y_{n+1} av

$$s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

och ett $(1 - \alpha) \times 100\%$ -igt prediktionsintervall ges av

$$\hat{y}_{n+1} = \beta_0 + \beta_1 x_{n+1} \pm t_{n-k-1, \alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$



2005-05-10

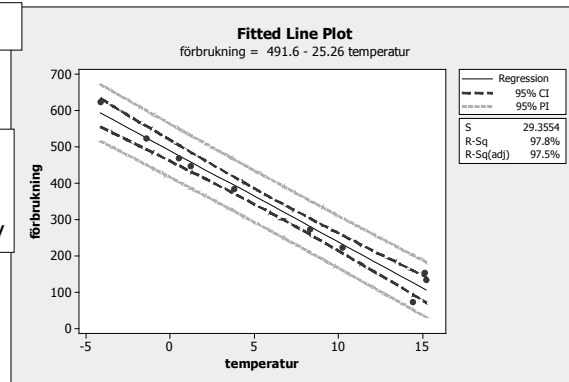
8

Prediktioner

För enkel linjär regression

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

ges ett $(1 - \alpha) \times 100\%$ -igt konfidenstintervall för det förväntade värdet för Y_{n+1} av



$$\hat{\mu}_{n+1} = \beta_0 + \beta_1 x_{n+1} \pm t_{n-k-1, \alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

motsvarande formler för multipel regression finns i Lee 15.6

9

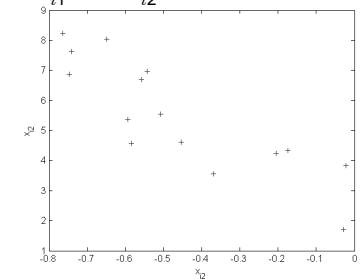
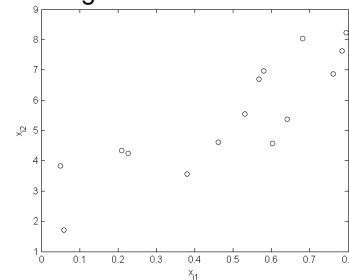
Multikolaritet

Om vi har en regressionsmodell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$$

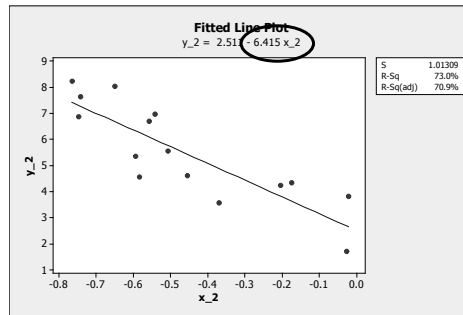
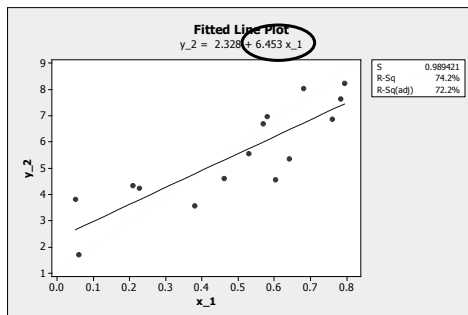
säger vi att det finns multikolaritet om två eller flera av de oberoende variablerna är högt korrelerade

Antag vi har två oberoende variabler x_{i1} och x_{i2}



10

Multikolaritet



Om x_{i1} och x_{i2} nästan är lika kan

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

nästan skrivs

$$\text{Jof } Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \epsilon_i$$

11

Multikolaritet

Om vi nästan har

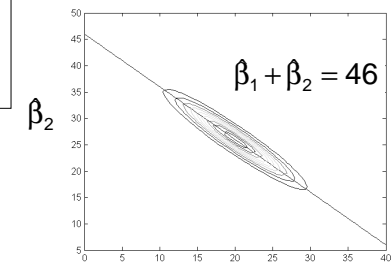
$$Y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \epsilon_i$$

alltså

$$Y_i \approx \beta_0 + (\beta_1 + \beta_2) x_{i1} + \epsilon_i$$

men när vi skall skatta β_1 och β_2 ger en given summa $\beta^* = \beta_1 + \beta_2$ oändligt många lösningar

i praktiken kanske endast stort samband mellan estimatorerna i samplingfördelningen

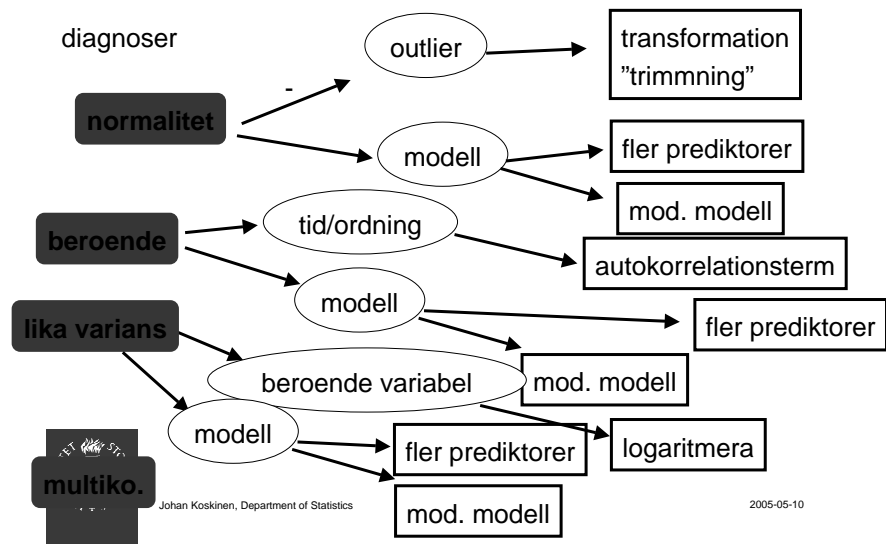


β_1

2005-05-10

12

Sammanfattning

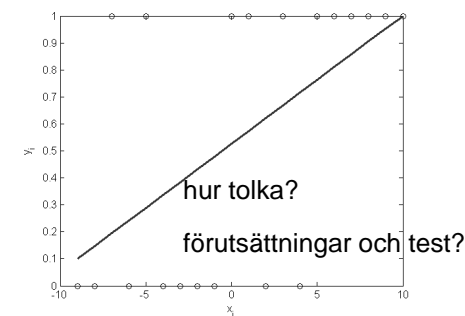
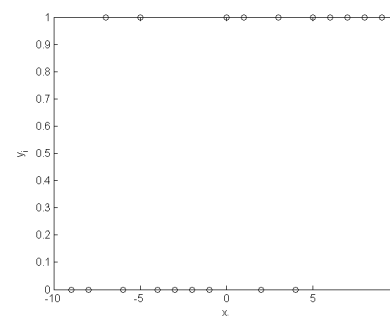


Logistisk regression

Antag att vi gör observationer på Y_i för olika värden på $x_{i1}, x_{i2}, \dots, x_{ik}$ där Y_i antar värden 1 eller 0

Ex en oberoende variabel x_i :

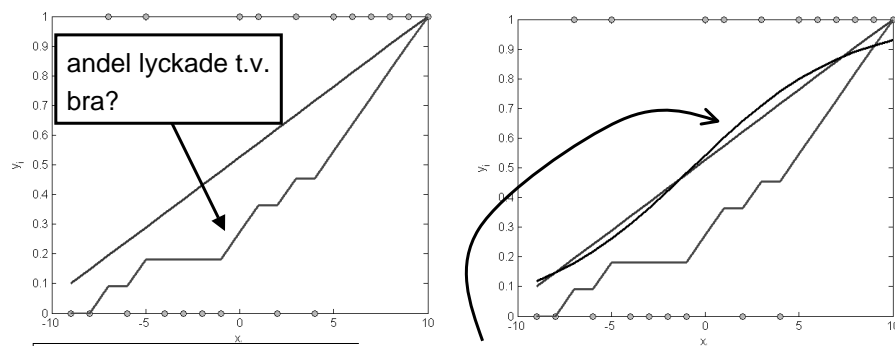
vi kan anpassa regressionslinje



Logistisk regression

Mer relevant: sannolikheten att vi skall lyckas givet visst värde på x_i

$P(Y_i = 1 | x_i)$ om Bernoulli $p_{x_i} = P(Y_i = 1 | x_i) = E(Y_i)$



en annan anpassad linje: $\hat{p}_{x_i} = \hat{P}(Y_i = 1 | x_i) = \hat{E}(Y_i)$

Logistisk regression

Antag att vi gör observationer på Y_i för olika värden på $x_{i1}, x_{i2}, \dots, x_{ik}$ där Y_i antar värden 1 eller 0

vi kan då anpassa en funktion

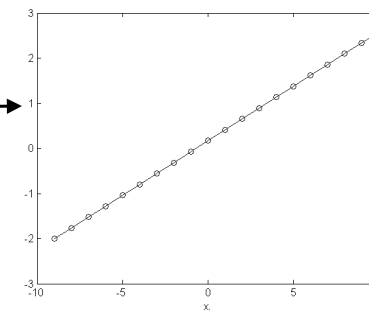
$$\hat{p}_{x_{i1}, \dots, x_{ik}} = \hat{P}(Y_i = 1 | x_{i1}, \dots, x_{ik}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}$$

Detta betyder att ju högre

$$\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

Desto större sannolikhet att Y_i antar värdet 1

linjär funktion...



Logistisk regression

Man definierar ofta

$$\text{logit}(\hat{p}_{x_i}) = \ln\left(\frac{\hat{p}_{x_i}}{1 - \hat{p}_{x_i}}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

så att man kan studera den linjära funktionen som i vanlig regression

Som i vanlig regression skattar

$$\hat{p}_{x_{i1}, \dots, x_{ik}} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}} \quad \text{en sann modell}$$



Johan Koskinen, Department of Statistics

$$p_{x_{i1}, \dots, x_{ik}} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}}$$

2005-05-10

17

Logistisk regression

Hur man tar fram skattningar av koefficienterna är lite mer invecklat än i vanlig regression

Test av modellanpassning och signifikant förklarande variabler ungefär samma

EX: för 40 olika bolag låt

Y_i vara 1 om haft uppgång och 0 om nedgång

förklarande variabel x_i

$$x_i = \frac{\text{verklig vinst per aktie} - \text{prognosticerad vinst per aktie}}{\text{aktiens värde}}$$



Johan Koskinen, Department of Statistics

diffvinst	uppgång
-1.2	1
-0.7	1
-0.5	1
-0.2	1
-0.2	1
-0.1	1
-0.1	1
0.0	1
0.0	1
0.1	1
0.1	1
0.2	1
0.3	1
0.5	1
0.5	1
0.6	1
0.8	1
0.8	1
0.9	1
1.1	1
1.2	1
1.4	1
1.4	1
1.7	1
2.3	1
3.7	1
-2.2	0
-1.7	0
-1.0	0
-0.6	0
-0.2	0
-0.1	0
0.0	0
0.0	0
0.3	0
0.4	0
0.8	0
1.4	0
1.4	0

2005-05-10

18

Logistisk regression

Binary Logistic Regression: uppgång versus diffvinst

Link Function: Logit

Response Information

Variable	Value	Count
uppgång	1	27 (Event)
	0	13
Total		40

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	0.563839	0.359835	1.57	0.117			
diffvinst	0.759181	0.412498	1.84	0.066	2.14	0.95	4.80

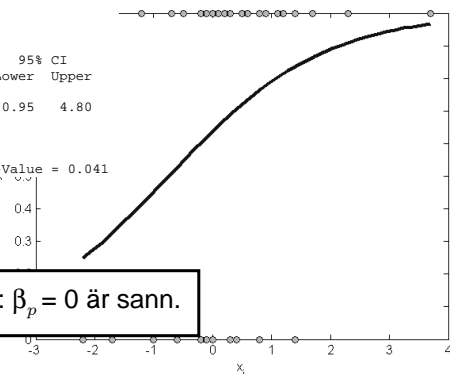
Log-Likelihood = -23.127

Test that all slopes are zero: G = 4.193, DF = 1, P-Value = 0.041

β_1

$$\frac{\beta_p - 0}{s_{\beta_p}} \in \text{approx. } N(0,1)$$

då $H_0: \beta_p = 0$ är sann.



diffvinst	uppgång
-1.2	1
-0.7	1
-0.5	1
-0.2	1
-0.2	1
-0.1	1
-0.1	1
0.0	1
0.0	1
0.1	1
0.1	1
0.2	1
0.3	1
0.5	1
0.5	1
0.6	1
0.8	1