



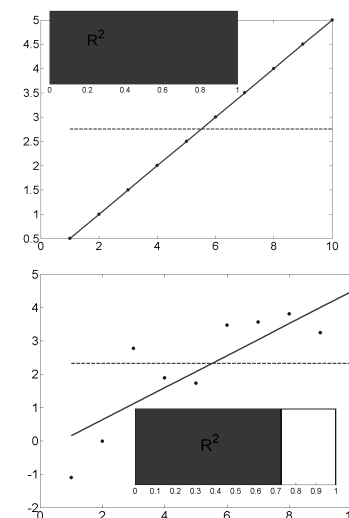
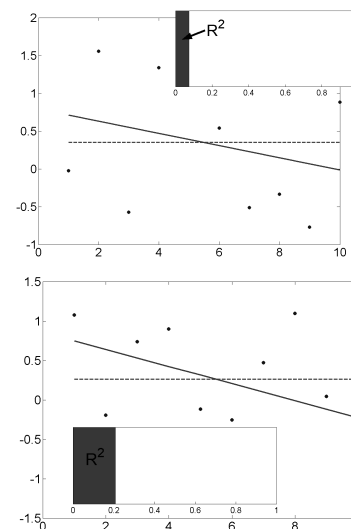
Johan Koskinen, Statistiska institutionen, Stockholms universitet

Finansiell statistik, vt-05

F16 regressionsanalys

Andelen förklarad variation

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



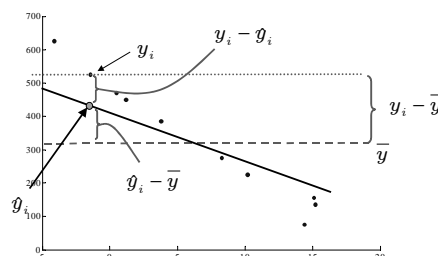
2

ANOVA-tabell

Kom ihåg

$$SST = SSE + SSR$$

Vi ställer upp hur variationen fördelar sig i en ANOVA-tabell



variation i	kvadratsumma	frihetsgrader	medelkvadratsumma
regression	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	$MSR = SSR / 1$
residualer	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - 2$	$MSE = SSE / (n - 2) = s_e^2$
totalt	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	$MST = SST / (n - 1) = s_y^2$

Johan Koskinen, Department of Statistics

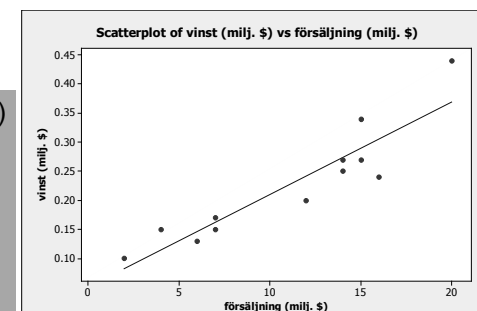
2005-05-08

3

Exempel - hamburgare

Nyttiga hamburgare i Illinois

försäljning (milj. \$)	vinst (milj. \$)
7	0.15
2	0.10
6	0.13
4	0.15
14	0.25
15	0.27
12	0.24
14	0.20
20	0.27
15	0.44
15	0.34
7	0.17



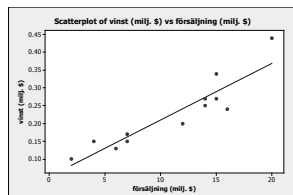
Regression Analysis: vinst (milj. \$) versus försäljning (milj. \$)

The regression equation is
vinst (milj. \$) = 0.0506 + 0.0159 försäljning (milj. \$)

Predictor	Coef	SE Coef	T	P
Constant	0.05060	0.02687	1.88	0.089
försäljning (milj. \$)	0.015930	0.002196	7.25	0.000

S = 0.0407357 R-Sq = 84.0% R-Sq(adj) = 82.4%

Exempel - hamburgare



variation i kvadratsumma frihetsgrader medelkvadratsumma

regression $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 1 $MSR = SSR/1$

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

totalt $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ $n-1$ $MST = SST/(n-1) = s_y^2$

Analysis of Variance

Source	DF	SS	MS	Predictor
Regression	1	0.087298	0.087298	Constant
Residual Error	10	0.016594	0.001659	försäljning (milj. \$)
Total	11	0.103892		

Regression Analysis:

The regression equation
vinst (milj. \$) = 0.0

Predictor
Constant
försäljning (milj. \$)

S = 0.0407357 R-Sq

2005-05-08 5

Johan Koskinen, Department of Statistics

Kvadratsummor - varför

För just det observerade datamaterialet har vi fått en observerad förklaringsgrad, t.ex.

07357 R-Sq = 84.0% R-Sq

m.a.o. kan vi förklara en del av variationen i Y utifrån värdena på x

m.h.a. en regressionslinje

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{där } \beta_0, \beta_1 \quad \text{skattar } \beta_0, \beta_1$$

men eftersom vi alltid har

$$0 \leq R^2 \leq 1$$

kommer vi alltid få en

observerad förklaringsgrad som är större än 0

p.g.a. slumpen



Johan Koskinen, Department of Statistics

2005-05-08

6

Kvadratsummor - varför

Om vi inte kan förklara en del av variationen i Y utifrån värdena på x

m.h.a. en regressionslinje

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

hur troligt är det då att vi av "ren slump" får t.ex.

07357 R-Sq = 84.0% R-Sq

hur troligt är det då att vi av "ren slump" får t.ex.



Johan Koskinen, Department of Statistics

2005-05-08

7

Test av regressionssamband

Trots inget samband: vi kan alltid anpassa linje

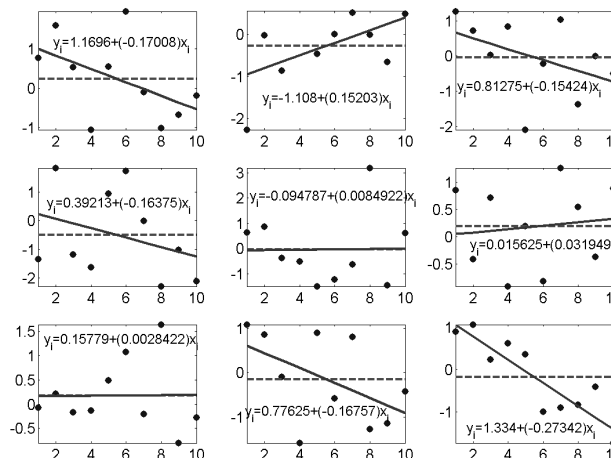
Oberoende $N(0,1)$

variabler

$$Y_i \in N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$= N(\beta_0, \sigma^2)$$

$$= N(\mu = 0, \sigma^2 = 1)$$



● (x_i, y_i)

--- \bar{y}

— \hat{y}_i

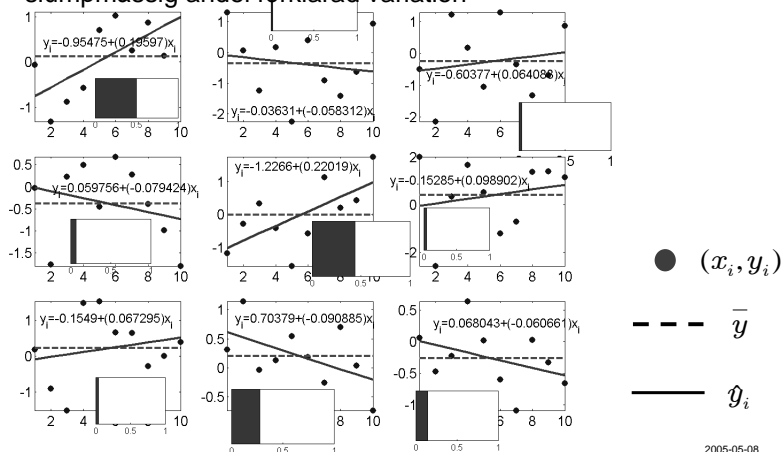
2005-05-08

8

Test av regressionssamband

10 nya datamaterial med oberoende $N(0,1)$ variabler:

slumpmässig andel förklarad variation



9

Test av regressionssamband

Under nollhypotesen: de oberoende variablerna x_1, x_2, \dots, x_n förklarar inget av variationen i den beroende variabeln Y_1, Y_2, \dots, Y_n

variation i	kvadratsumma	frihetsgrader	medelkvadratsumma	F-kvoten
regression	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	$MSR = SSR/1$	MSR/MSE
residualer	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n-2$	$MSE = SSE/(n-2) = s_e^2$	
totalt	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	$n-1$	$MST = SST/(n-1) = s_y^2$	

$$\frac{\text{regressionskvadratsumman}/1}{\text{residualkvadratsumman}/(n-2)} = \frac{MSR}{MSE} \in F(1, n-2)$$



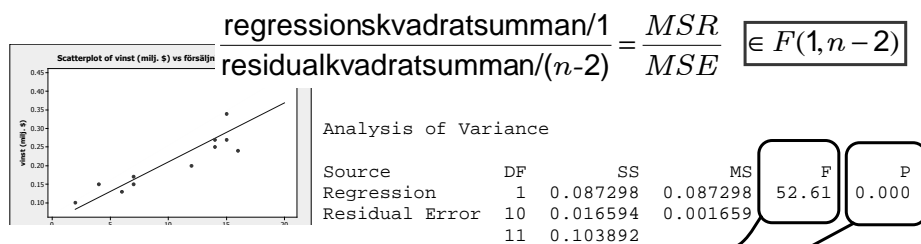
Förutsatt: antaganden A-E uppfyllda & nollhypotesen sann

Johan Koskinen, Department of Statistics

2005-05-08

10

Test av regressionssamband - hamburgare



11

Test av regressionssamband

För s.v. Y_1, Y_2, \dots, Y_n , $Y_i \in N(\mu_i = \beta_0 + \beta_1 x_i, \sigma^2)$

testa $H_0: \mu_i = \mu$ för alla i , alltså ingen regression, alltså $\beta_1 = 0$

mot $H_1: \beta_1 \neq 0$ på signifikansnivån α

Förutsatt A-E är teststatistikan

$$\frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \in F(1, n-2) \quad \text{då } H_0 \text{ är sann.}$$

Vi förkastar H_0 : (ingen regression) på signifikansnivån α då det observerade värdet på teststatistikan är större än $F^{1, n-2}_{\alpha}$ det kritiska värdet i F-fördelningen med frihetsgraderna 1 och $n-2$, på signifikansnivån α ,



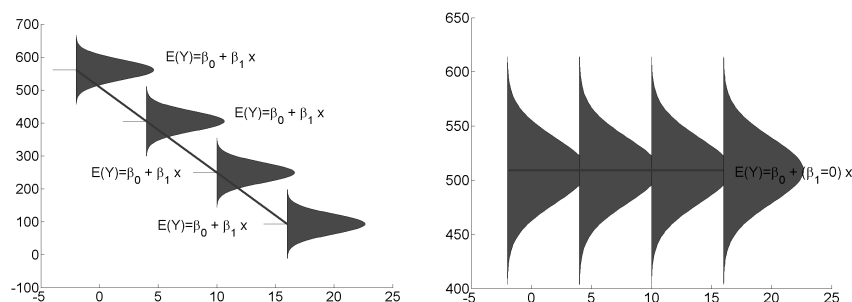
Johan Koskinen, Department of Statistics

2005-05-08

12

Test av regressionssamband

M.a.o. är "test av regression" ekvivalent med test av lutningen



Vi vill testat $H_0: \beta_1 = 0$ mot $H_1: \beta_1 \neq 0$

Johan Koskinen, Department of Statistics

2005-05-08

13

Test av lutningskoefficient

Vi har redan sett

$$E(\beta_1) = E\left(\frac{SS_{xy}}{SS_x}\right) = \beta_1$$

Man kan visa att givet att A-E uppfyllt

$$\beta_1 \in N\left(\beta_1, \frac{\sigma^2}{SS_x}\right)$$

Alltså måste

$$\frac{\beta_1 - E(\beta_1)}{SD(\beta_1)} = \frac{\beta_1 - \beta_1}{\sqrt{\sigma^2 / SS_x}} \in N(0,1)$$

Johan Koskinen, Department of Statistics

2005-05-08

14

Test av lutningskoefficient

Eftersom

$$\frac{\beta_1 - E(\beta_1)}{SD(\beta_1)} = \frac{\beta_1 - \beta_1}{\sqrt{\sigma^2 / SS_x}} \in N(0,1)$$

och om vi kan skatta

$$SD(\beta_1) \text{ med } s_b = \sqrt{\frac{s_e^2}{SS_x}}$$

borde

$$\frac{\beta_1 - \beta_1}{s_b} = \frac{\beta_1 - \beta_1}{\sqrt{s_e^2 / SS_x}} \in t(n-2)$$

frihetsgraderna = # obs - # skattade regresionskoeff.

Johan Koskinen, Department of Statistics

2005-05-08

15

Test av lutningskoefficient

För s.v. Y_1, Y_2, \dots, Y_n , $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

testa $H_0: \beta_1 = 0$

mot $H_1: \beta_1 \neq 0$ på signifikansnivån α

Förutsatt A-E är teststatistikan

$$\frac{\beta_1 - \beta_1}{s_b} = \frac{\beta_1 - 0}{\sqrt{s_e^2 / SS_x}} \in t(n-2) \quad \text{då } H_0: \beta_1 = 0 \text{ är sann.}$$

Vi förkastar $H_0: \beta_1 = 0$ på signifikansnivån α då det observerade värdet på teststatistikan är större till beloppet än $t_{n-2, \alpha/2}$ det kritiska värdet i t-fördelningen med frihetsgraden $n - 2$, på signifikansnivån α .

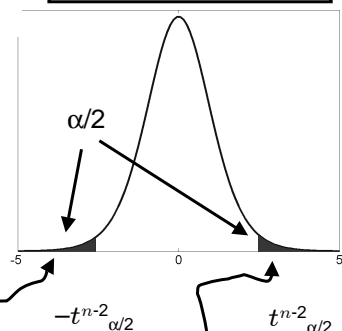
16

Test av lutningskoefficient

$$\frac{\beta_1 - \beta_1}{s_b} = \frac{\beta_1 - 0}{\sqrt{s_e^2 / SS_x}} \in t(n-2) \quad \text{då } H_0: \beta_1 = 0 \text{ är sann.}$$

Vi förkastar $H_0: \beta_1 = 0$ på signifikansnivån α då

$$\left| \frac{\beta_1 - 0}{\sqrt{s_e^2 / SS_x}} \right| > t_{\alpha/2}^{n-2}$$



Johan Koskinen, Department of Statistics

2005-05-08

17

Test av lutningskoefficient F och t

Eftersom F-testet och t-testet har samma hypoteser (i enkel linjär regression) hur hänger de ihop?

$$t = \frac{\beta_1 - 0}{\sqrt{s_e^2 / SS_x}} = \frac{SS_{xy} / SS_x}{\sqrt{s_e^2 / SS_x}} = \frac{SS_{xy} / \sqrt{SS_x}}{\sqrt{MSE}}$$

utnyttja att

$$r = \frac{SS_{xy}}{\sqrt{SST \times SS_x}}$$

$$t^2 = \frac{SS_{xy}^2}{MSE \times SS_x} = \frac{SST}{MSE} (r)^2 = \frac{SST}{MSE} R^2 = \frac{SSR/1}{MSE}$$



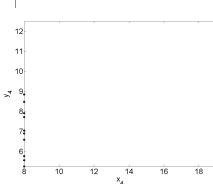
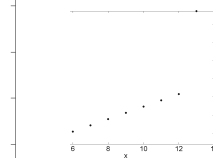
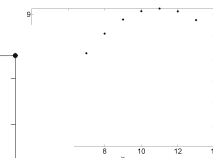
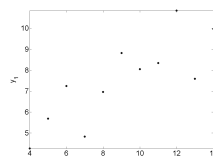
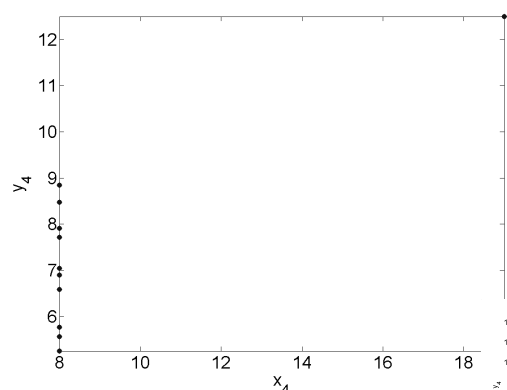
Johan Koskinen, Department of Statistics

2005-05-08

18

Förutsättningar (Anscombe, 1973)

x	y_1	y_2	y_3	x_4	y_4
10	8.04	9.14	7.46	8	6.58
8	6.95				
13	7.58				
9	8.81				
11	8.33				
14	9.96				
6	7.24				
4	4.26				
12	10.84				
7	4.82				
5	5.68	4.74	5.73	8	6.89



Förutsättningar

n 11

\bar{x} 9

\bar{y} 7,5

ekvation:

$$y = 3 + 0,5x$$

SS_x : 110

SSE : 13,75 (9 df)

s_b : 0,118

r : 0,816

R^2 : 0,667

