



Johan Koskinen, Statistiska institutionen, Stockholms universitet

## Finansiell statistik, vt-05

F17 regressionsanalys

## Residualanalys

För modellen

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

antog vi att  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  är oberoende likafördelade  $N(0, \sigma^2)$

Då var skattningarna BLUE (bästa lineära vvr-skattningen)

och våra hypotestest har önskade signifikansnivåer

att undersöka

$$e_i = y_i - \hat{y}_i$$

ger oss indikationer på om antagandena gäller



Johan Koskinen, Department of Statistics

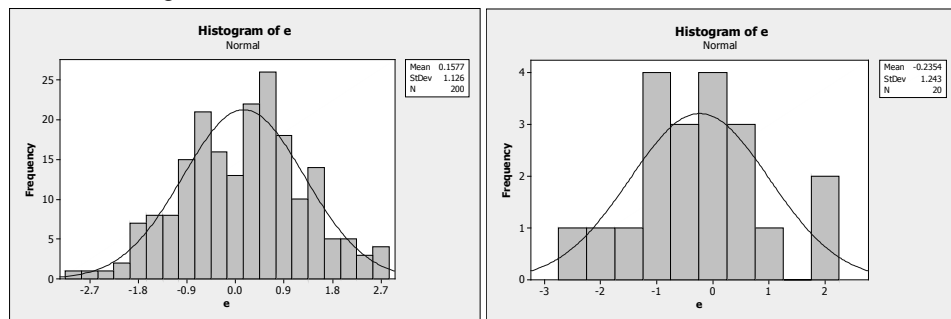
2005-05-09

2

## Normalfördelade?

Avvikelser från normalfördelningsantagandet

Histogram



200 tal från  $N(0,1)$

20 tal från  $N(0,1)$



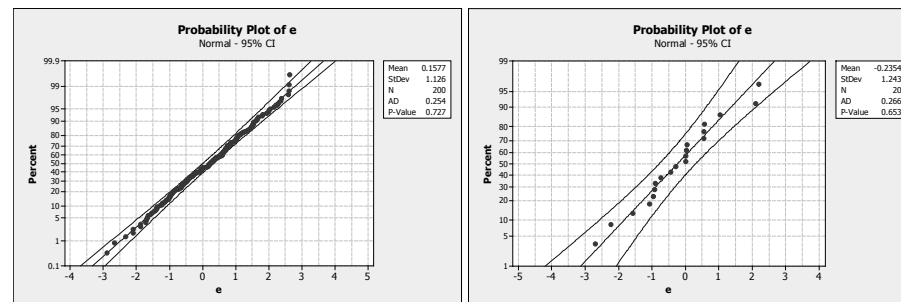
Johan Koskinen, Department of Statistics

2005-05-09

3

## Normalfördelade?

Normalfördelningsdiagram



200 tal från  $N(0,1)$

20 tal från  $N(0,1)$



I praktiken svårt att avgöra

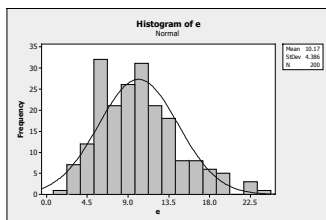
Johan Koskinen, Department of Statistics

2005-05-09

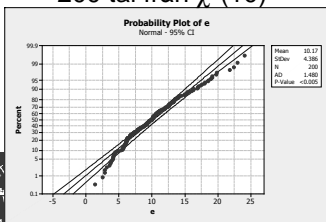
4

## Normalfördelade?

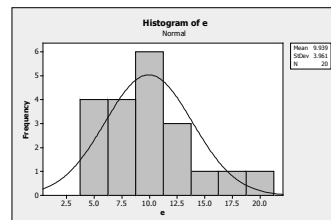
Normalfördelningsdiagram



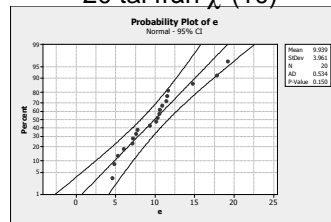
200 tal från  $\chi^2(10)$



Johan Koskinen, Department of Statistics



20 tal från  $\chi^2(10)$



2005-05-09

5

## Normalfördelade?

Om vi har starka skäl att tvivla på normalfördelningsantagandet  
tvivlar vi på att t.ex. estimatorernas samplingfördelning normal  
m.a.o. behöver inte hypotestesten gälla...



Johan Koskinen, Department of Statistics

2005-05-09

6

## Oberoende

Om  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  är oberoende borde vi ej se ngt mönster

- relativt tiden
- relativt ordning
- relativt beroende variabel
- relativt oberoende variabel
- relativt ngn annan variabel



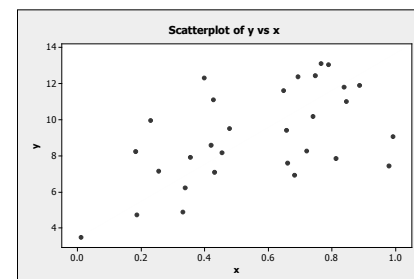
Johan Koskinen, Department of Statistics

2005-05-09

7

## Beroende över tid

Antag att vi har följande data



### Regression Analysis: y versus x

The regression equation is  
 $y = 6.05 + 5.42 x$

Predictor	Coef	SE Coef	T	P
Constant	6.0535	0.9772	6.19	0.000
x	5.419	1.575	3.44	0.002

S = 2.20771 R-Sq = 29.7% R-Sq(adj) = 27.2%



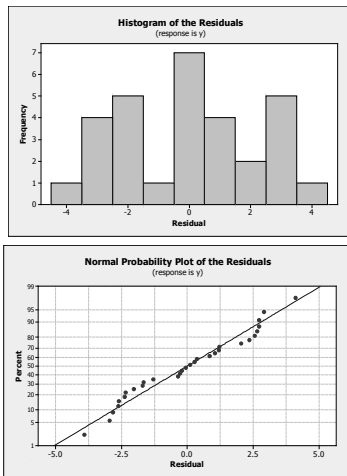
Johan Koskinen, Department of Statistics

2005-05-09

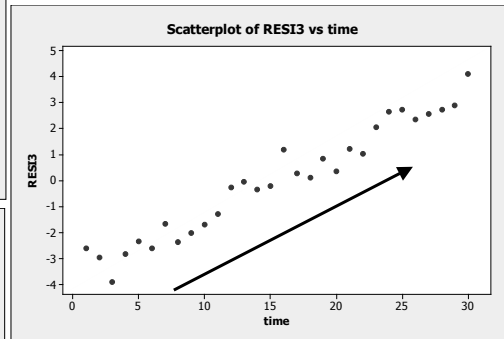
8

## Beroende över tid

Vilket ger residualerna



Inget märkligt tills

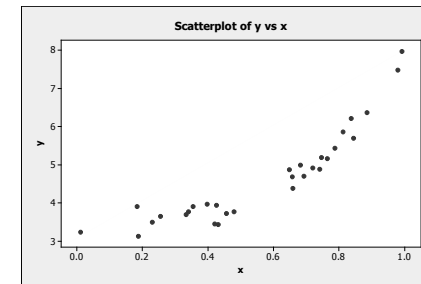


trend över tid

2005-05-09

9

## Beroende av förklarande var.

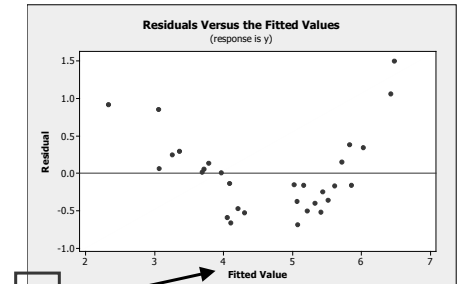


Regression Analysis: y versus x

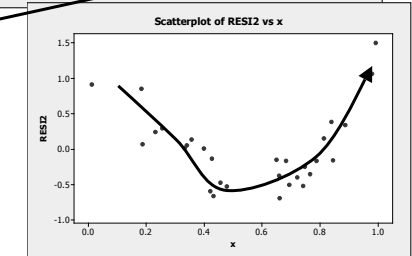
The regression equation is  
 $y = 2.28 + 4.23 x$

Predictor	Coef	SE Coef	T	P
Constant	2.2774	0.2412	9.44	0.000
x	4.2300	0.3888	10.88	0.000

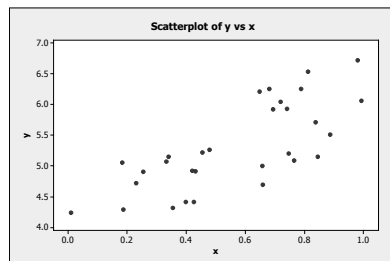
S = 0.544899 R-Sq = 80.9% R-Sq(adj) = 80.2%



$\hat{y}_i$



## Beroende av ytterligare variabler

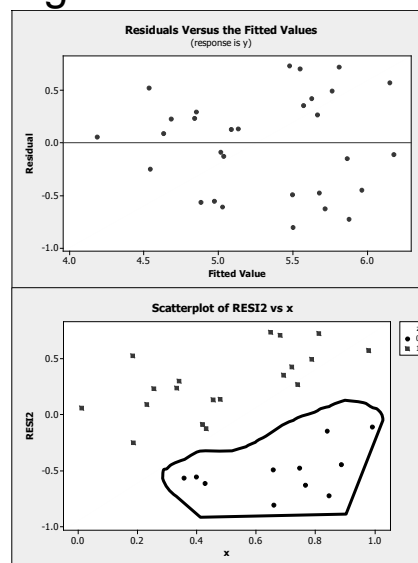


Regression Analysis: y versus x

The regression equation is  
 $y = 4.16 + 2.03 x$

Predictor	Coef	SE Coef	T	P
Constant	4.1619	0.2117	19.66	0.000
x	2.0307	0.3412	5.95	0.000

S = 0.478270 R-Sq = 55.8% R-Sq(adj) = 54.3%

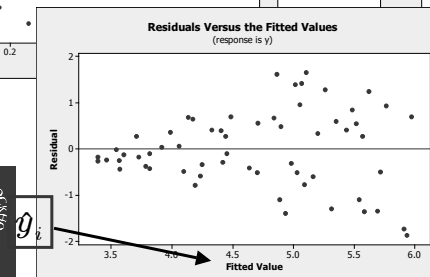
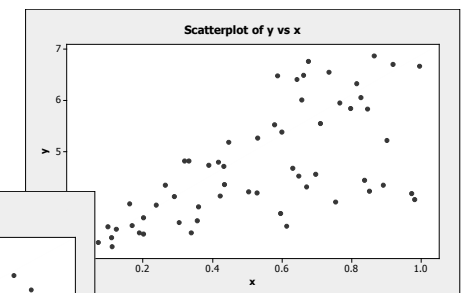
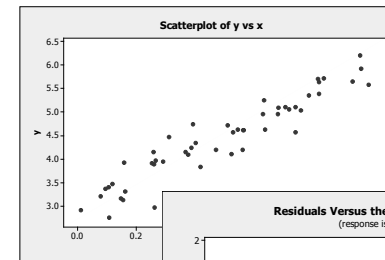


## Lika varians

Om  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  har samma varians  $\sigma^2$  ej se ngt mönster

Lika varians: homoscedasticitet

olika varians: heteroscedasticitet

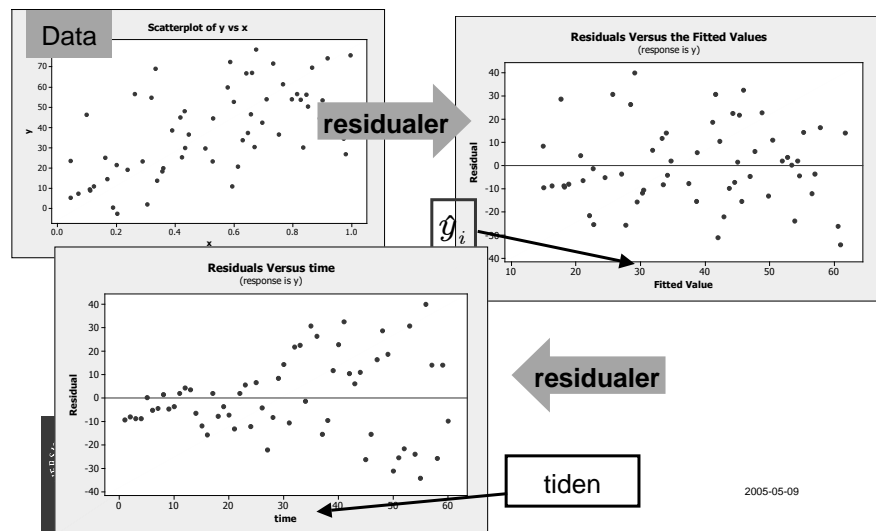


residualer

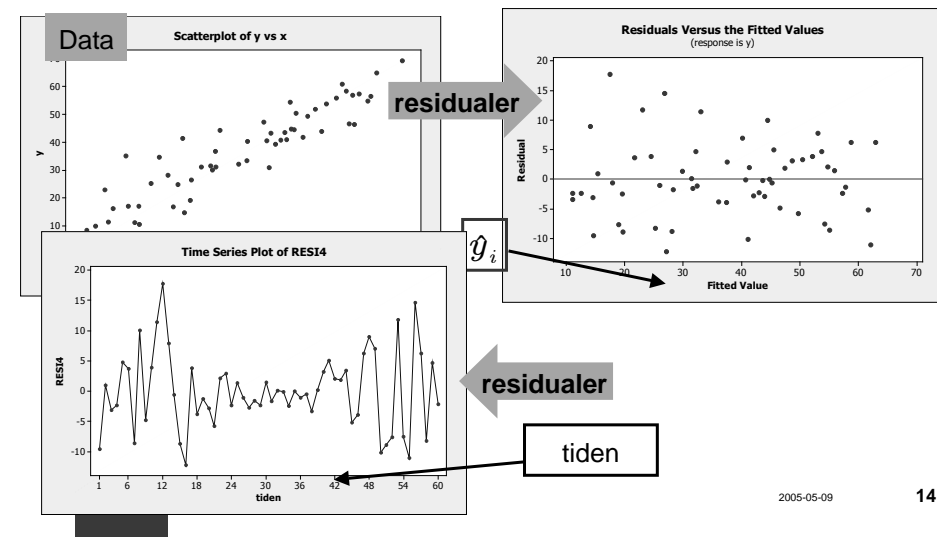
2005-05-09

12

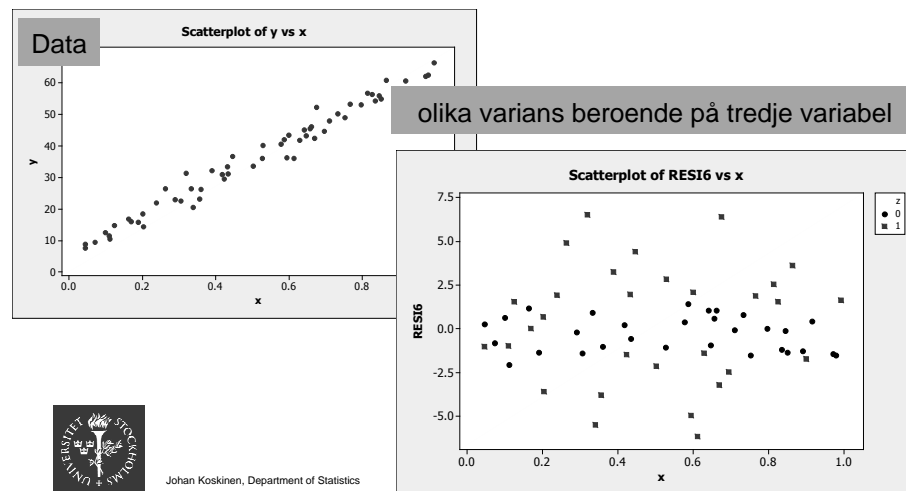
## Heteroscedasticitet



## Heteroscedasticitet



## Heteroscedasticitet

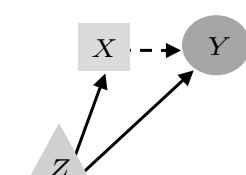
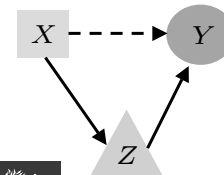
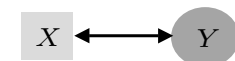
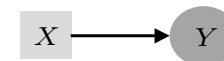


## Andra exempel på fel modellantaganden

Utöver ovan nämnda tekniska bitar

"Vad påverkar vad"?

vi antar linjär model och  $x$  fix



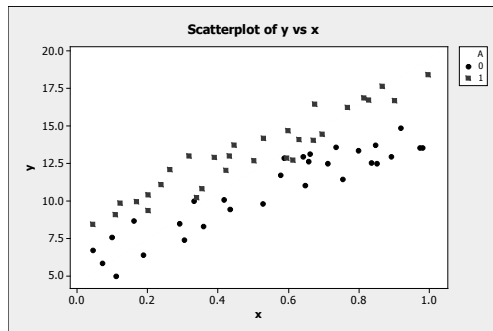
## Multipel regression - dummyvariabler

Antag att vi har typer av enheter i populationen A & B

och vi tror på modellen

$$Y_i = \begin{cases} \beta_0^* + \beta_1 x_{i1} + \varepsilon_i & \text{om } i \text{ är A} \\ \beta_0 + \beta_1 x_{i1} + \varepsilon_i & \text{om } i \text{ är B} \end{cases}$$

ska vi då skatta en modell  
för A och en för B?



Johan Koskinen, Department of Statistics

2005-05-09

17

## Multipel regression - dummyvariabler

Låt

$$x_{i2} = \begin{cases} 1 & \text{om } i \text{ är A} \\ 0 & \text{om } i \text{ är B} \end{cases}$$

och skriv om modellen

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

så får vi

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{i1} + \beta_2 + \varepsilon_i & \text{om } i \text{ är A} \\ \beta_0 + \beta_1 x_{i1} + \varepsilon_i & \text{om } i \text{ är B} \end{cases}$$



Johan Koskinen, Department of Statistics

2005-05-09

18

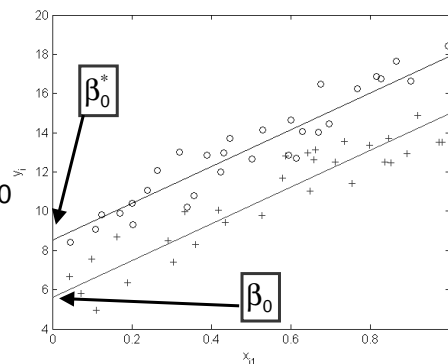
## Multipel regression - dummyvariabler

Vilket vi kan tolka som

$$\beta_2 = \beta_0^* - \beta_0$$

alltså är de förväntade värdet på  
den oberoende variabeln när  $x_{i1} = 0$

$$Y_i = \begin{cases} \beta_0 + \beta_2 = \beta_0^* & \text{om } i \text{ är A} \\ \beta_0 & \text{om } i \text{ är B} \end{cases}$$



m.a.o. sambandet är "likadant" mellan  $x_{i1}$  och  $Y_i$  men linjerna  
ligger på olika nivåer



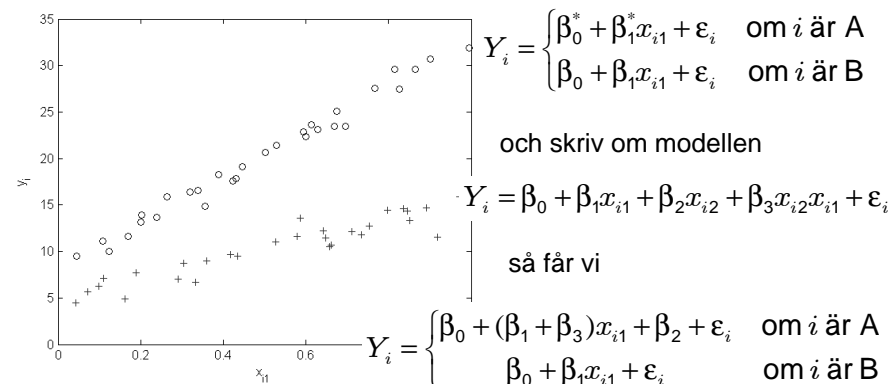
Johan Koskinen, Department of Statistics

2005-05-09

19

## Multipel regression - dummyvariabler

Om vi tror på en modell med olika lutning för A & B



$$Y_i = \begin{cases} \beta_0^* + \beta_1^* x_{i1} + \varepsilon_i & \text{om } i \text{ är A} \\ \beta_0 + \beta_1 x_{i1} + \varepsilon_i & \text{om } i \text{ är B} \end{cases}$$

och skriv om modellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2} x_{i1} + \varepsilon_i$$

så får vi

$$Y_i = \begin{cases} \beta_0 + (\beta_1 + \beta_3) x_{i1} + \beta_2 + \varepsilon_i & \text{om } i \text{ är A} \\ \beta_0 + \beta_1 x_{i1} + \varepsilon_i & \text{om } i \text{ är B} \end{cases}$$

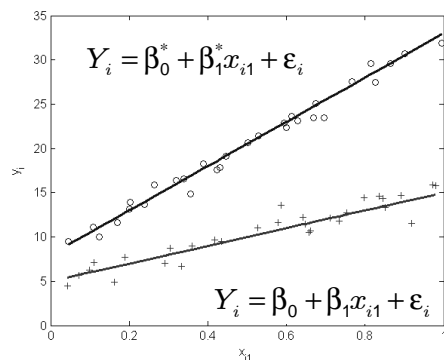


Johan Koskinen, Department of Statistics

2005-05-09

20

## Multipel regression - dummyvariabler



Vilket vi kan tolka som

$$\beta_2 = \beta_0^* - \beta_0$$

$$\beta_3 = \beta_1^* - \beta_1$$



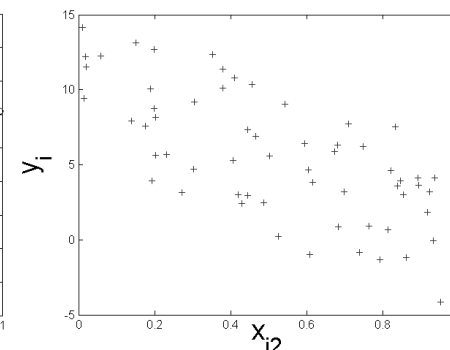
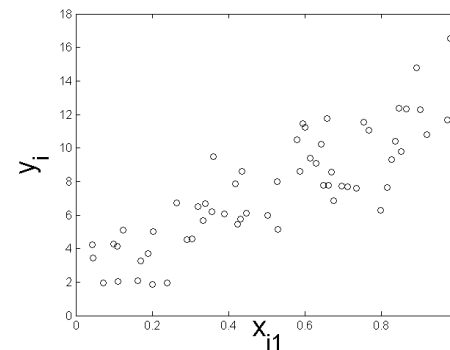
Johan Koskinen, Department of Statistics

2005-05-09

21

## Multipel regression - mer generellt

Antag att det t.ex. finns 2 variabler  $x_{i1}$  och  $x_{i2}$  som "förklarar" den beroende variabeln  $Y_i$



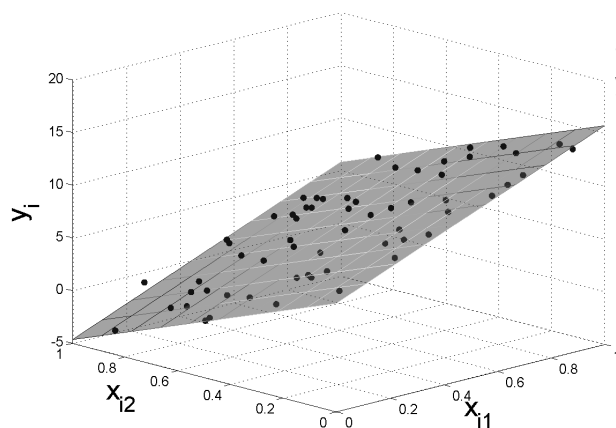
Johan Koskinen, Department of Statistics

2005-05-09

22

## Multipel regression - mer generellt

Att anpassa en modell  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$   $\epsilon_i \in N(0, \sigma^2)$



är då att anpassa en yta

MK-skattningen är

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

som gör kvadratsumman

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

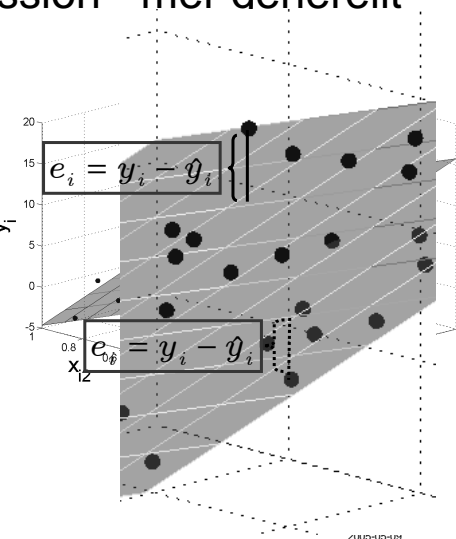
så liten som möjligt

## Multipel regression - mer generellt

Precis som för enkel linjär regression

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

$$= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$



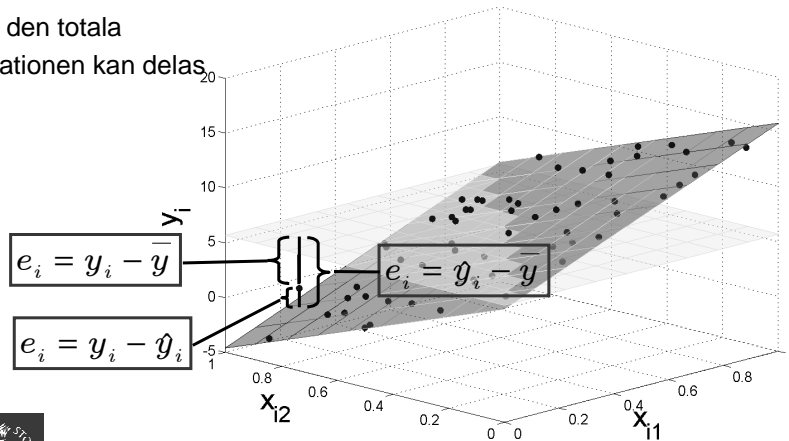
Johan Koskinen, Department of Statistics

2005-05-09

24

## Multipel regression - mer generellt

och den totala variationen kan delas upp



Johan Koskinen, Department of Statistics

2005-05-09

25

## Multipel regression - mer generellt

Vi kan ha "många",  $k$  stycken förklarande variabler  $x_{i1}, x_{i2}, \dots, x_{ik}$  som tillsammans förklarar den beroende variabeln

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \in N(0, \sigma^2)$$

Det förväntade värdet på  $Y_i$  givet  $x_{i1}, x_{i2}, \dots, x_{ik}$  är alltså en linjär funktion av de  $k$  oberoende variablerna

Dock svårt att rita ytan i fler dimensioner...

fortfarande har vi däremot

$$SST = SSR + SSE$$

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



Johan Koskinen, Department of Statistics

2005-05-09

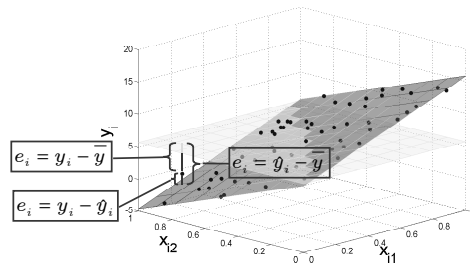
26

## Multipel regression - ANOVA

För anpassad modell

$$Y_i = \beta_0 + \sum_{p=1}^k \beta_p x_{ip}$$

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



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

2005-05-09

27

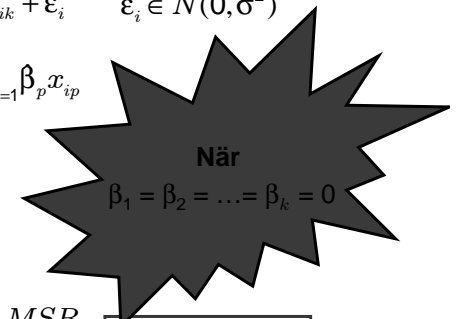
## Multipel regression - signifikant förklarande?

För modell med  $k$  stycken förklarande variabler  $x_{i1}, x_{i2}, \dots, x_{ik}$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad \varepsilon_i \in N(0, \sigma^2)$$

För anpassad modell  $Y_i = \beta_0 + \sum_{p=1}^k \beta_p x_{ip}$

gäller att givet att  
förutsättningarna A-E  
är uppfyllda



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



Johan Koskinen, Department of Statistics

2005-05-09

28

## Multipel regression - signifikant förklarande?

För s.v.  $Y_1, Y_2, \dots, Y_n$ ,  $Y_i \in N(\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}, \sigma^2)$

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

alltså  $\beta_1 = \beta_2 = \dots = \beta_k = 0$

mot  $H_1$ : för minst ett  $p$ ,  $\beta_p \neq 0$  på signifikansnivån  $\alpha$

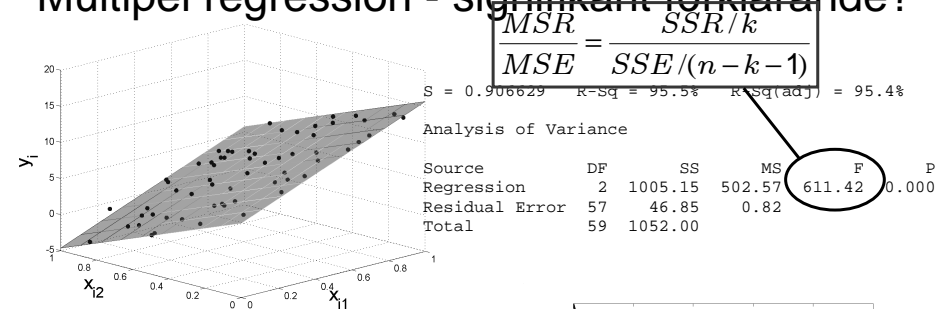
Förutsatt A-E är teststatistikan

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

Vi förkastar  $H_0$ : (ingen regression) på signifikansnivån  $\alpha$  då det observerade värdet på teststatistikan är större än  $F^{k, n-k-1}_{\alpha}$  vi säger att regressionen är signifikant förklarande

29

## Multipel regression - signifikant förklarande?

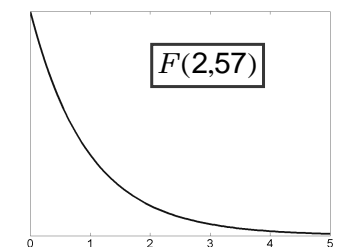


### Regression Analysis: förs versus inköp; nederbörd

The regression equation is  
förs = 5.28 + 10.5 inköp - 9.91 nederbörd

Predictor	Coef	SE Coef	T	P
Constant	5.2803	0.3323	15.89	0.000
inköp	10.4732	0.4271	24.52	0.000
nederbörd	-9.9142	0.4117	-24.08	0.000

Johan Koskinen, Department of Statistics



## Multipel regression - test av individuella koefficienter

Att vi förkastar  $H_0$  betyder

alltså att vi tror på

$H_1$ : för minst ett  $p$ ,  $\beta_p \neq 0$

lite "trubbigt"

testa:

givet

förutsättningarna

A-E är uppfyllda:

(1) testa  $H_0: \beta_1 = 0$

mot  $H_1: \beta_1 \neq 0$

(2) testa  $H_0: \beta_2 = 0$

mot  $H_1: \beta_2 \neq 0$

...

(k) testa  $H_0: \beta_k = 0$

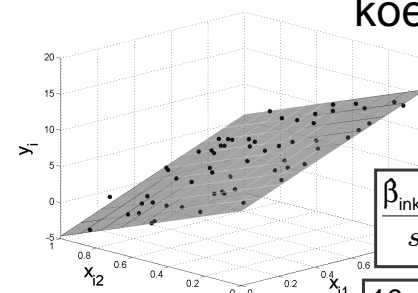
mot  $H_1: \beta_k \neq 0$

$$\frac{\beta_p - 0}{s_{\beta_p}} \in t(n-k-1) \quad \text{då } H_0: \beta_p = 0 \text{ är sann.}$$

2005-05-09

31

## Multipel regression - test av individuella koefficienter



(1) testa  $H_0: \beta_{\text{inköp}} = 0$

mot  $H_1: \beta_{\text{inköp}} \neq 0$

då  $H_0: \beta_p = 0$   
är sann.

$$\frac{10,47}{0,427} = 24,52$$

### Regression Analysis: förs versus in

The regression equation is  
förs = 5.28 + 10.5 inköp - 9.91 nederbörd

Predictor	Coef	SE Coef	T	P
Constant	5.2803	0.3323	15.89	0.000
inköp	10.4732	0.4271	24.52	0.000
nederbörd	-9.9142	0.4117	-24.08	0.000

Johan Koskinen, Department of Statistics

