Johan Koskinen, Statistiska institutionen, Stockholms universitet

Finansiell statistik, vt-05

F15 regressionsanalys

Regressionsmodellen

Modellen kan skrivas:

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

där ε_1 , ε_2, ..., ε_n är oberoende likafördelade N(0,σ^2)

För alla i =1, 2,… , n.

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$$
$$= \beta_0 + \beta_1 x_i + E(\varepsilon_i)$$
$$= \beta_0 + \beta_1 x_i$$

konstanter

$\varepsilon_i \in$ N(0,σ^2)

Regressionsmodellen

För alla i =1, 2,… , n.

$$\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i)$$
$$= 0 + 0 + \text{Var}(\varepsilon_i)$$
$$= \sigma^2$$

konstanter

$\varepsilon_i \in$ N(0,σ^2)

För alla $i \neq j$

$$\text{Cov}(Y_i, Y_j) = E\big((\beta_0 + \beta_1 x_i + \varepsilon_i)(\beta_0 + \beta_1 x_j + \varepsilon_j)\big)$$
$$- E(\beta_0 + \beta_1 x_i + \varepsilon_i)E(\beta_0 + \beta_1 x_j + \varepsilon_j)$$

Regressionsmodellen

$$E\big((\beta_0 + \beta_1 x_i + \varepsilon_i)(\beta_0 + \beta_1 x_j + \varepsilon_j)\big)$$
$$= E\big((\beta_0 + \beta_1 x_i)(\beta_0 + \beta_1 x_j)\big) + (\beta_0 + \beta_1 x_i)E(\varepsilon_j) + (\beta_0 + \beta_1 x_j)E(\varepsilon_i) + E(\varepsilon_i \varepsilon_j)$$

konstanter

ε_i, $\varepsilon_j \in$ N(0,σ^2)

$$\boxed{= (\beta_0 + \beta_1 x_i)(\beta_0 + \beta_1 x_j)}$$

för normalfördelade variabler oberoende medför kovarians 0

$$E(\beta_0 + \beta_1 x_i + \varepsilon_i)E(\beta_0 + \beta_1 x_j + \varepsilon_j)$$
$$= (\beta_0 + \beta_1 x_i + E(\varepsilon_i))(\beta_0 + \beta_1 x_j + E(\varepsilon_j))$$
$$= (\beta_0 + \beta_1 x_i)(\beta_0 + \beta_1 x_j)$$

Regressionsmodellen

M.a.o., för alla $i \neq j$

$$\text{Cov}(Y_i, Y_j) = E\big((\beta_0 + \beta_1 x_i + \varepsilon_i)(\beta_0 + \beta_1 x_j + \varepsilon_j)\big)$$
$$- E(\beta_0 + \beta_1 x_i + \varepsilon_i)E(\beta_0 + \beta_1 x_j + \varepsilon_j)$$
$$= (\beta_0 + \beta_1 x_i)(\beta_0 + \beta_1 x_j) - (\beta_0 + \beta_1 x_i)(\beta_0 + \beta_1 x_j) = 0$$

Om vi sammanfattar (Lee)

- den oberoende variabelns x_i värden är fixa (A)
- variablen ε_i är normalfördelad (B)
- variablen ε_i har väntevärde 0 (C)
- ε_i och ε_j är statistiskt oberoende för alla $i \neq j$ (D)
- variablerna ε_i har konstant varians σ^2 (E)

Regressionsanalys

Vi har en modell för Y givet värdet på β_0 , β_1 och x är

$$\boxed{Y \in N(\beta_0 + \beta_1 x, \sigma^2)}$$

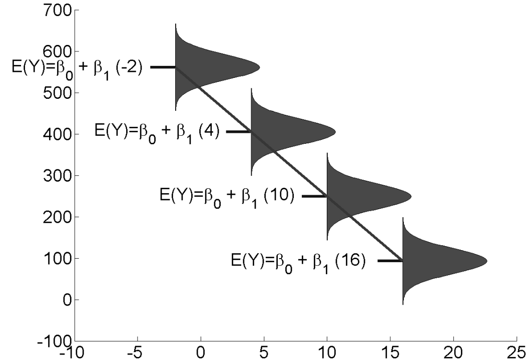
där alltså väntevärdet för Y beskrivs av en rät linje

Vi utgår från att x ej varierar utan fix

Okänt i modell

β_0 , β_1 och σ^2

Hur skatta dessa parametrar?



Regressionsanalys

Vi gör parvisa observationer. För fixa $x_1, x_2, \ldots, x_n$ gör vi observationer $Y_1, Y_2, \ldots, Y_n$, där vi antar

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

Vi skall hitta estimatorer av

β_0 , β_1 och σ^2

Som helst är

- väntevärdesriktiga
- konsistenta
- effektiva

Regressionsanalys - räta linjens ekv.

Vi har två observationspar

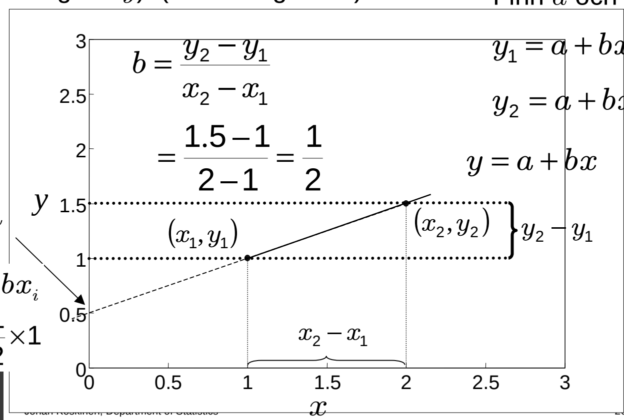
b = (förändringen i y)/ (förändringen i x)

Finn a och b så att

$y_1 = a + bx_1$

$y_2 = a + bx_2$ alltså

$y = a + bx$

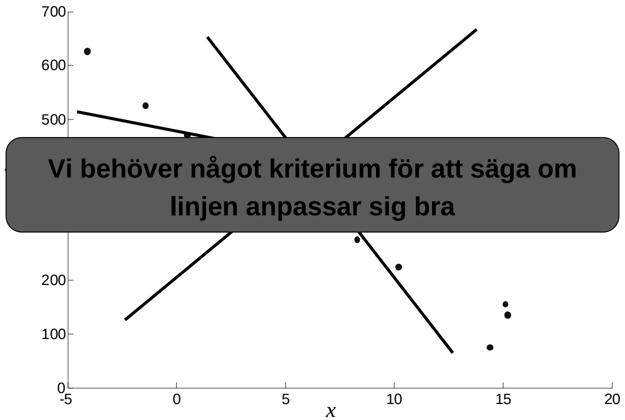


$a = y_i - bx_i$

$= 1 - \frac{1}{2} \times 1$

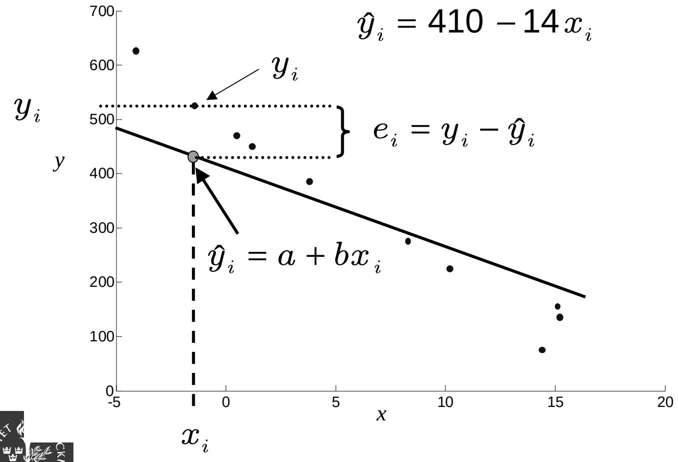
$= \frac{1}{2}$

Regr. analys - minsta kvadrat skattningen

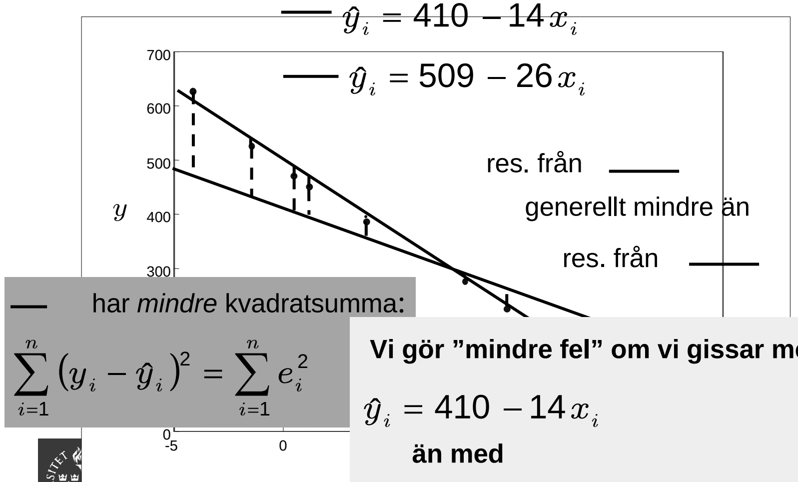


Vi behöver något kriterium för att säga om linjen anpassar sig bra

Minstakvadrat-skattningen - residual



Minstakvadrat-skattningen - residual



har *mindre* kvadratsumma:

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

Vi gör "mindre fel" om vi gissar med

$$\hat{y}_i = 410 - 14x_i$$

än med

$$\hat{y}_i = 509 - 26x_i$$

Minstakvadrat-skattningen (MKS)

Minstakvadratskattningen ges av att minimera kvadratsumman map a och b

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

$$= 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1)$$

$$= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} a + 2b\sum_{i=1}^{n} x_i$$

$$\boxed{a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n} b \sum_{i=1}^{n} x_i}$$

Minstakvadrat-skattningen (MKS)

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

$$= 2\sum_{i=1}^{n}(y_i-a-bx_i)(-x_i)$$

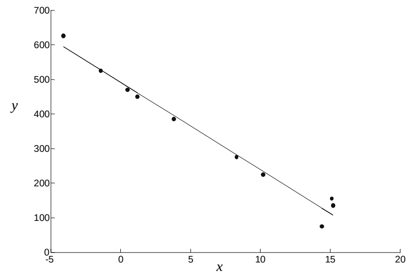
$$= -2\sum_{i=1}^{n}y_ix_i + 2\sum_{i=1}^{n}ax_i + 2b\sum_{i=1}^{n}x_i^{\;2}$$

$$\boxed{a=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{n}b\sum_{i=1}^{n}x_i}$$

$$= -2\sum_{i=1}^{n}y_ix_i + 2\left\{\frac{\sum y_i - b\sum x_i}{n}\right\}\sum_{i=1}^{n}x_i + 2b\sum_{i=1}^{n}x_i^{\;2}$$

$$\boxed{b=\frac{\sum_{i=1}^{n}y_ix_i-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}x_i\right)}{\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2}}$$

Minstakvadrat-skattningen (MKS)



x	x^2	y	xy
15.1	228.01	155	2340.5
14.4	207.36	75	1080
15.2	231.04	135	2052
10.2	104.04	225	2295
8.3	68.89	275	2282.5
3.8	14.44	385	1463
0.5	0.25	470	235
-1.4	1.96	525	-735
-4.1	16.81	625	-2562.5
1.2	1.44	450	540
63.2	874.24	3320	8990.5

$$b=\frac{\sum yx-\frac{1}{n}\sum y\sum x}{\sum x^2-\frac{1}{n}\left(\sum x\right)^2}=\frac{8990.5-\frac{1}{10}(63.2)(3320)}{874.24-\frac{1}{10}(63.2)^2}=-25.2559$$

$$a=\bar{y}-25.2559\bar{x}=332-(-25.2559)(6.32)=491.62$$

Kvadratsummor

Definiera kvadratsumman

$$SS_x=\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

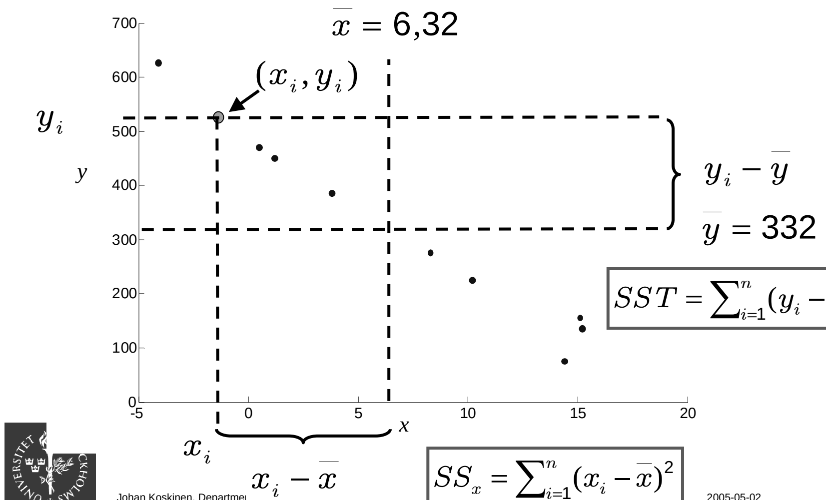
och

$$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2$$

så att

$$s_x^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{SS_x}{n-1} \qquad s_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}=\frac{SST}{n-1}$$

Kvadratsummor



$$\boxed{SST=\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\boxed{SS_x=\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Korsprodukter

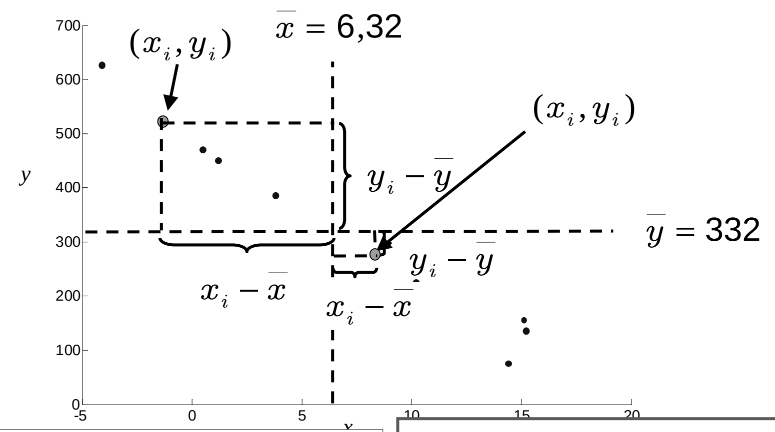
Definiera korsproduktsumman

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \;=\sum_{i=1}^{n} y_i x_i - \tfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)$$

så att stickprovskovariansen

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

Korsprodukter



m.a.o. linjens lutning:
tecknen på produkten av avvikelserna:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} = \frac{SS_{xy}}{SS_x}$$

Väntevärdesriktiga skattningar

Antag modellen med antagandena A-E uppfyllda

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

För fixa $x_1, x_2, \ldots, x_n$ gör vi observationer $Y_1, Y_2, \ldots, Y_n$

Låt estimatorerna av β_0 , β_1 vara

$$\boxed{\hat{\beta}_0 = a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n} b \sum_{i=1}^{n} x_i}$$

$$\boxed{\hat{\beta}_1 = b = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{SS_{xy}}{SS_x}}$$

Väntevärdesriktiga skattningar

Är estimatorn av β_1 vvr?

$$E(\hat{\beta}_1) = \frac{1}{SS_x} E\left(\sum_{i=1}^{n} Y_i x_i - \tfrac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} x_i\right)\right)$$

$$= \frac{1}{SS_x} E\left(\sum_{i=1}^{n} Y_i x_i\right) - \tfrac{1}{n}\frac{1}{SS_x} E\left(\sum_{i=1}^{n} Y_i\right)\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{SS_x} \sum_{i=1}^{n} \underbrace{E(Y_i x_i)} - \tfrac{1}{n}\frac{\left(\sum_{i=1}^{n} x_i\right)}{SS_x} \sum_{i=1}^{n} \underbrace{E(Y_i)}$$

$$= \frac{1}{SS_x} \sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i) \;-\; \frac{\sum_{i=1}^{n} x_i}{nSS_x} \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)$$

Väntevärdesriktiga skattningar

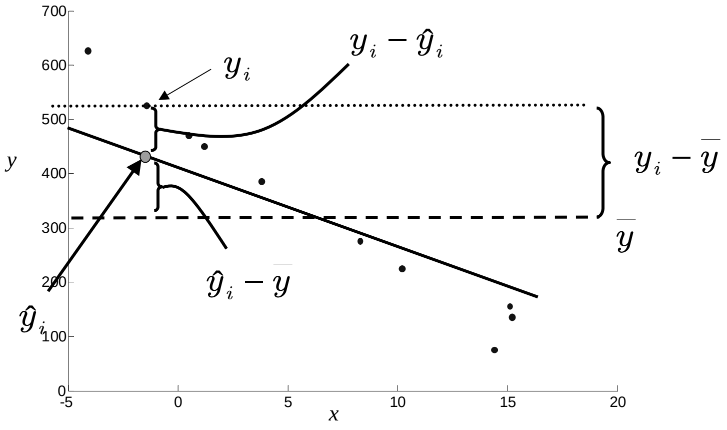
$$= \frac{1}{SS_x}\sum_{i=1}^{n} x_i(\beta_0+\beta_1 x_i) - \frac{\sum_{i=1}^{n} x_i}{nSS_x}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i)$$

$$= \frac{n}{nSS_x}\sum_{i=1}^{n} x_i(\beta_0+\beta_1 x_i) \quad - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i(\beta_0+\beta_1 x_j)}{nSS_x}$$

$$= \frac{n\beta_0\sum_{i=1}^{n} x_i + n\beta_1\sum_{i=1}^{n} x_i^2}{nSS_x} \quad - \frac{\beta_0 n\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}{nSS_x}$$

$$= \frac{\beta_1 n\sum_{i=1}^{n} x_i^2 - \beta_1\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}{nSS_x} = \frac{\beta_1 n\sum_{i=1}^{n} x_i^2 - \beta_1\left(\sum_{i=1}^{n} x_i\right)^2}{nSS_x} \boxed{= \frac{\beta_1 nSS_x}{nSS_x}}$$

Variansuppdelning



$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} \quad = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Variansuppdelning

Eftersom

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

kan vi visa

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

$$SST = SSE + SSR$$

Till att börja med

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

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

visa att försvinner

Variansuppdelning

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)(a + bx_i - (a + b\bar{x}))$$

$$= b\sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \bar{x})$$

$$= b\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i - b\bar{x}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

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

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\left(\bar{y} - b\bar{y}\right) - \sum_{i=1}^{n} bx_i$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} bx_i - \sum_{i=1}^{n} bx_i = 0$$

Variansuppdelning

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

$$=\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}ax_i-\sum_{i=1}^{n}bx_i^2$$

$$=\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}\left(\tfrac{1}{n}\sum_{j=1}^{n}y_j-\tfrac{1}{n}b\sum_{j=1}^{n}x_j\right)x_i-\sum_{i=1}^{n}bx_i^2$$

$$=\sum_{i=1}^{n}y_ix_i-\tfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}y_jx_i+\tfrac{1}{n}b\sum_{i=1}^{n}\sum_{j=1}^{n}x_jx_i-\sum_{i=1}^{n}bx_i^2$$

$$=SS_xb-SS_xb=0$$

Residualvarians

För fixa $x_1, x_2, \ldots, x_n$ gör vi observationer $Y_1, Y_2, \ldots, Y_n$, där vi antar

$$Y_i \in N(\beta_0+\beta_1x_i,\sigma^2)$$

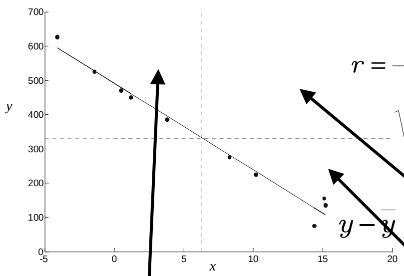
och eftersom vi vvr estimatorer av regressionskoefficienterna kan man visa att

$$s_e^2=\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}=\frac{\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{n-2}=\frac{SSE}{n-2}$$

är en vvr estimator av σ^2. Alltså

$$E(s_e^2)=E\left(\frac{\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{n-2}\right)=E\left(\frac{SSE}{n-2}\right)=\sigma^2$$

Hur starkt är sambandet - korrelationskoefficienten



$$r=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\frac{SS_{xy}}{\sqrt{SST\times SS_x}}$$

$y-\bar{y}$ stort positivt & $x-\bar{x}$ stort positivt

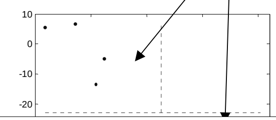
$y-\bar{y}$ stort negativt & $x-\bar{x}$ stort positivt

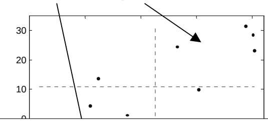
$y-\bar{y}$ stort positivt & $x-\bar{x}$ stort negativt

Korrelationskoefficienten

De parvisa avvikelserna har olika tecken - negativ korrelation

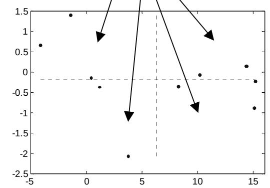
De parvisa avvikelserna har samma tecken - positiv korrelation

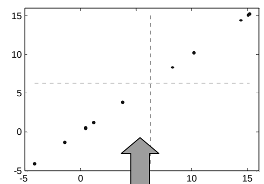




De parvisa avvikelserna har olika *och* samma tecken

- ingen korrelation





Stark positiv korrelation

Korrelationskoefficienten

Eftersom

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{SS_{xy}}{\sqrt{SST \times SS_x}}$$

och

$$b = \frac{SS_{xy}}{SS_x}$$

kan vi skriva

$$b = r\frac{\sqrt{SST}}{\sqrt{SS_x}}$$

Determinationskoefficienten

Determinationskoefficienten R^2 anger andelen variation i Y som förklaras av regressionslinjen

$$R^2 = \frac{SSR}{SST} \quad = 1 - \frac{SSE}{SST}$$

och kan också skrivas

$$R^2 = r^2$$