F15

Korrelation och regression

Korrelation

- Om vi har konstaterat ett linjärt samband mellan variabler mätta på minst intervall-nivå kan vi uppskatta en regressionslinje
- Lutningen talar om ifall vi har ett positivt eller negativt linjärt samband
- En korrelationskoefficient är ett mått på styrkan av det linjära sambandet
  - Korrelation:samvariation
  - Hur starkt samvarierar två variabler?

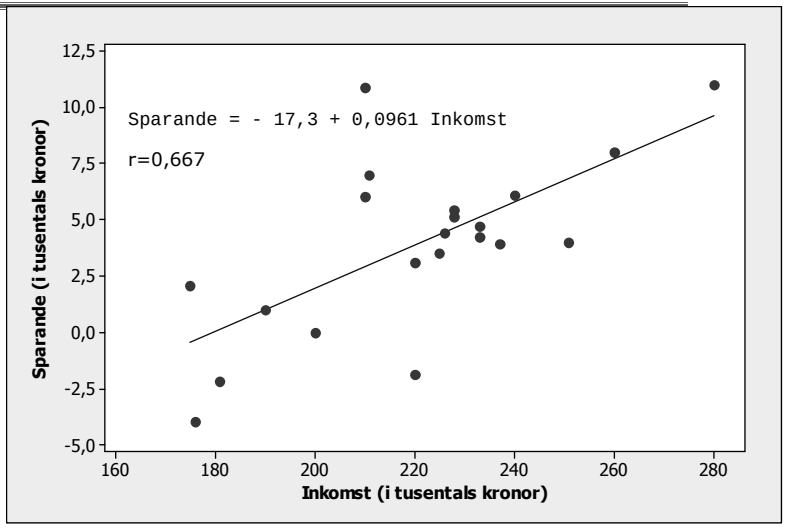
Korrelation

- Korrelationskoefficienten r_{XY}
  - Mäter styrkan på det linjära sambandet
  - Är oberoende av de sorter variablerna X och Y mäts i (till skillnad mot b)

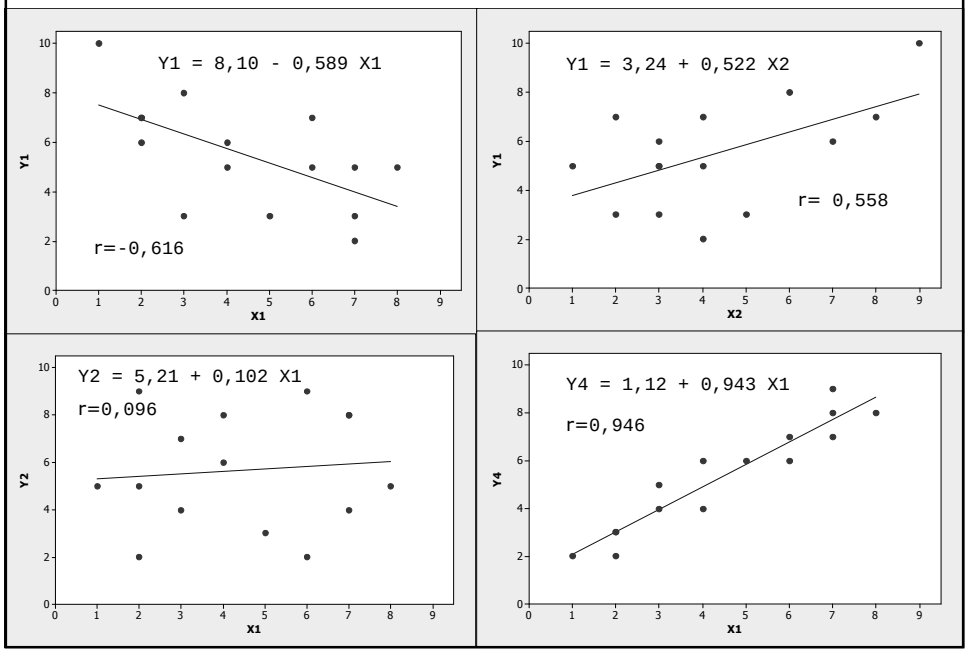
$$-1 \leq r_{XY} \leq 1$$

  - -1: perfekt negativt linjärt samband
  - 1: perfekt positivt linjärt samband
  - Se föreläsningsanteckningar (KD) F14-F16

Exempel, r_{XY}=0,667



Exempel på positiva och negativa korrelationskoefficienter av olika styrka



Korrelationskoefficienten, r_{XY} och determinationskoefficienten R^2

- Korrelationskoefficienten är ett mått på hur nära punkterna ligger den skattade linjen.
  - Om punkterna ligger på linjen – perfekt positivt (+1) eller negativt (-1) linjärt samband
    - Undantag: om linjen är parallell
- r_{XY} och b har alltid samma tecken
- Ett relaterat mått är determinationskoefficienten R^2
  - Korrelationen i kvadrat (om vi har *en* oberoende variabel)

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

  - Anger den del av variationen i Y som förklaras av variationen i X (kan anges i %)

Determinationskoefficienten

- Om r_{XY} antar värdet –1 eller 1 så antar R^2 värdet 1, dvs all variation i Y kan förklaras av variationen i X.
  - Alla residualer är noll – alla punkter ligger på linjen
  - Obs! Betyder ej att det det är ett *kausalt* samband
- Se formeln i KD F14-F16
- Determinationskoefficienten kan också ses som 1-andelen oförklarad variation

Prediktioner

- Korrelationskoefficienten och determinationskoefficienten kan användas för att bedöma hur väl vi kan förutse (predicera) värdet på Y om vi vet X.
- Om vi har ett perfekt samband kan Y-värdet förutses exakt i vårat datamaterial

Outliers/Unusual observations

- Residualer
  - Residualer är observationer som avviker kraftigt från den skattade linjen
  - Anges med ”R” i Minitab
- Andra extremobservationer
  - Observationer som avviker från X-medelvärdet
  - Bidrar mycket till lutningen av linjen
  - Anges med ”X” i Minitab

Regression Analysis: Sparande versus Inkomst

The regression equation is
Sparande = - 17,3 + 0,0961 Inkomst

Predictor	Coef	SE Coef	T	P
Constant	-17,280	5,492	-3,15	0,005
Inkomst	0,09607	0,02472	3,89	0,001

S = 2,96905 R-Sq = 44,3% R-Sq(adj) = 41,4%

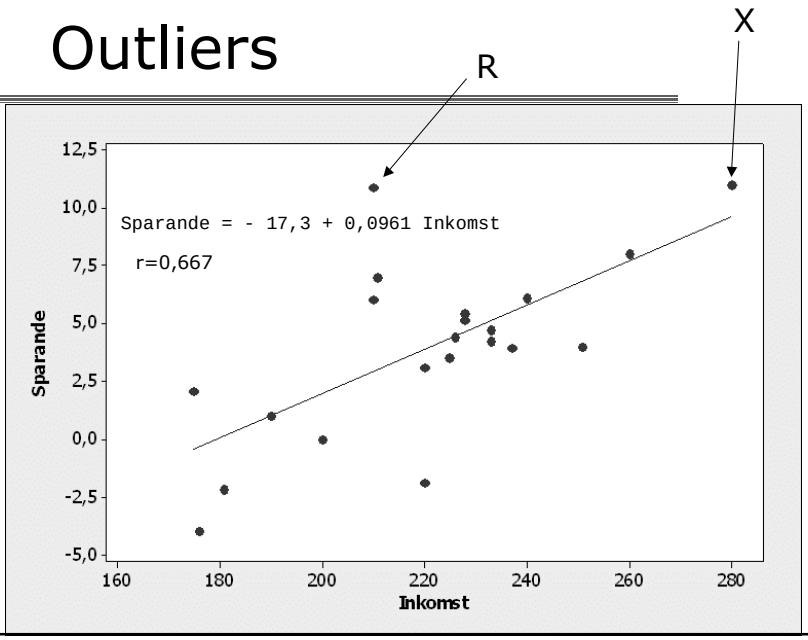
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	133,18	133,18	15,11	0,001
Residual Error	19	167,49	8,82		
Total	20	300,67			

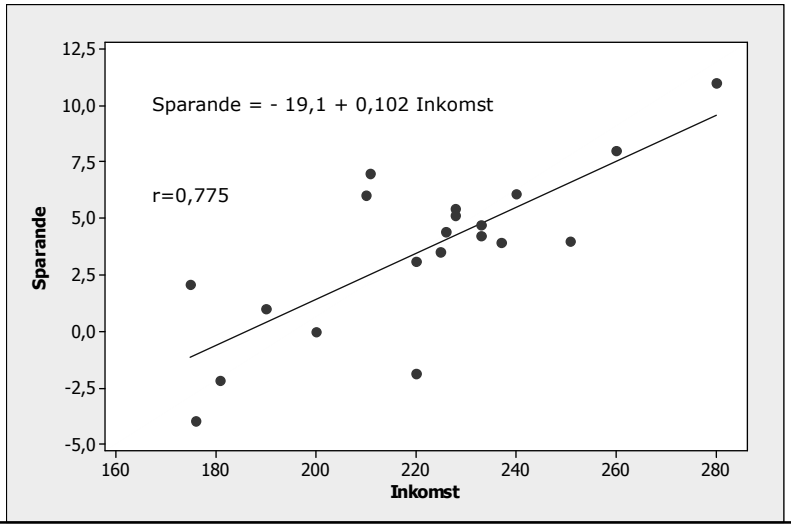
Unusual Observations

Obs	Inkomst	Sparande	Fit	SE Fit	Residual	St Resid
7	210	10,900	2,894	0,699	8,006	2,77R
21	280	11,000	9,619	1,603	1,381	0,55 X

Outliers



Spridningsdiagram utan residualen



Spridningsdiagram utan extrema x-värdet