SF1901: Sannolikhetslära och statistik
Föreläsning 15.
Enkel linjär regression

Jan Grandell & Timo Koski

07.03.2016

KTH Matematik

Innehåll

- Enkel linjär regression, repetition
- Enkel linjär regression: linjär algebra & numerisk analys
  - Modellens antaganden, teoretisk regressionslinje $y = \alpha + \beta x$
  - Minstakvadratskattningarna α^*_{obs} och β^*_{obs}
  - Skattade regressionslinjen, residualkvadratsumman
- Analys av enkel linjär regression
  - Extra formler för minstakvadratskattningarna α^*_{obs} och β^*_{obs}
  - Normalfördelning, väntevärde och varians för α^* och β^*
- Konfidensintervall
  - Konfidensintervall för α och β
  - Konfidensintervall för $\alpha + \beta x$
  - Prediktionsintervall för en ny observation
- Multipel regression: Statistisk prediktion av vinets kvalitet
- Bilaga i summaformler för regressionsanalys

KTH Matematik

Enkel linjär regression

Teorin och praktiken handlar om följande: Det föreligger n par av värden

$$(x_1, y_1), \ldots, (x_n, y_n)$$

där $x_1, \ldots, x_n$ är givna storheter och $y_1, \ldots, y_n$ är observationer av oberoende s.v. $Y_1, \ldots, Y_n$, där $Y_i \in N(\mu_i, \sigma)$.

Observera att σ förutsätts att *ej* bero av x, vilket ofta är det kritiska antagandet. Varje väntevärde μ_i är linjärt beroende av x_i, d.v.s.

$$\mu_i = \alpha + \beta x_i, \quad i = 1, \ldots, n.$$

Linjen

$$y = \alpha + \beta x \tag{1}$$

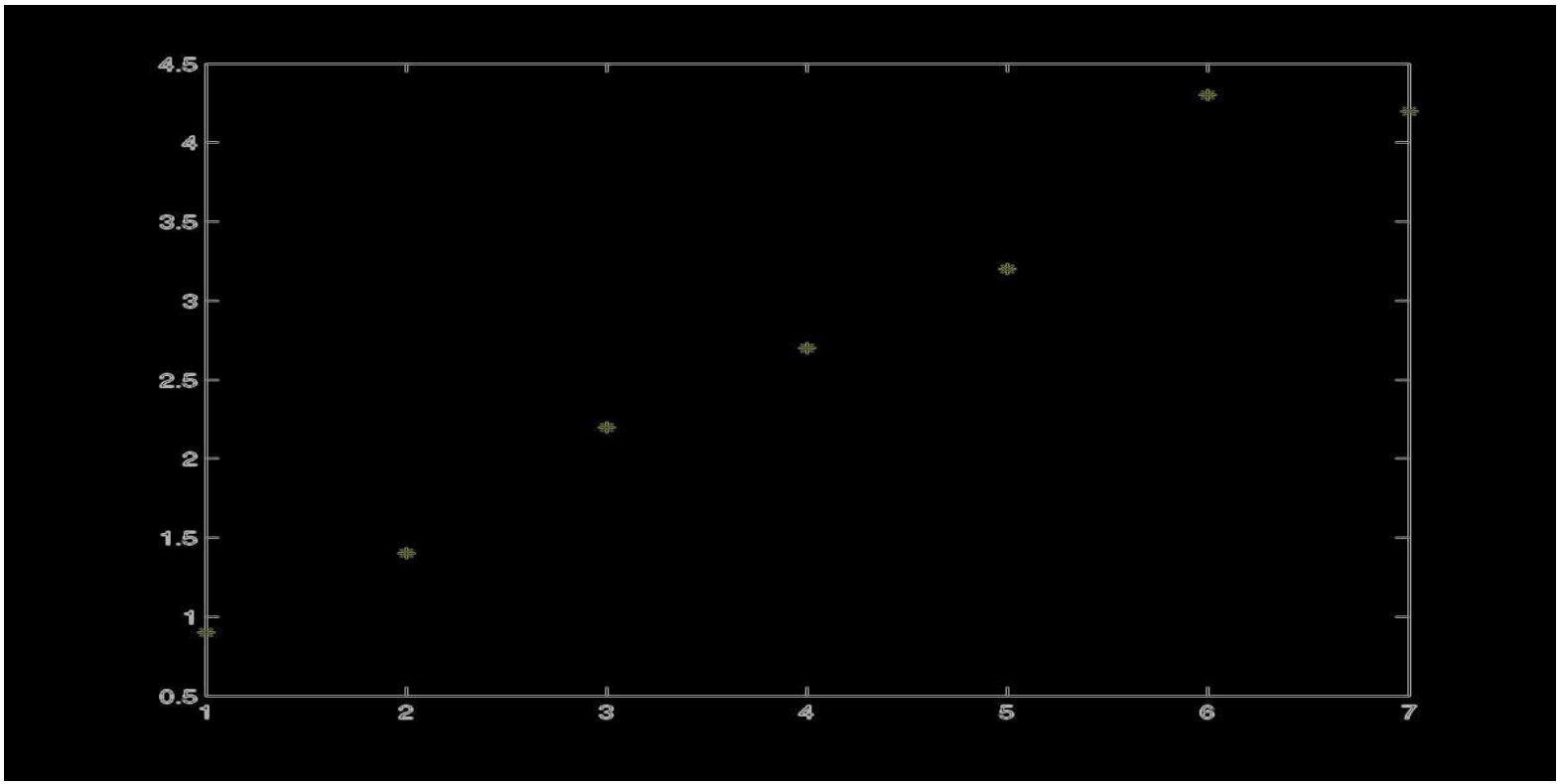
kallas den *teoretiska regressionslinjen*.

Enkel linjär regression

Ett fiktivt exempel

$$x = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$$

$$y = 0.9 \quad 1.4 \quad 2.2 \quad 2.7 \quad 3.2 \quad 4.3 \quad 4.2$$



KTH Matematik

Enkel linjär regression

Vi skattar parametrarna α och β med Minsta-Kvadratmetoden, dvs. vi minimerar summan av de kvadrerade **lodräta** avstånden mellan den teoretiska linjen och y-värdena, d.v.s vi minimerar

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - E(Y_i))^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

m.a.p. α och β. De värden α^*_{obs} och β^*_{obs} som ger minimum kallas *MK-skattningarna* av α och β.

KTH Matematik

Vi får

$$\alpha^*_{\text{obs}} = \overline{y} - \beta^*_{\text{obs}}\overline{x} \quad \text{och} \quad \beta^*_{\text{obs}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Minsta kvadrat via linjär algebra & numerisk analys

Låt oss skriva det ovanstående med en matrisformalism.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$n \times 1$ -vektor, $n \times 2$ - matris, 2×1 -vektor. Den teoretiska regressionmodellen är då

$$\mathbf{y} = \mathbf{X}\theta + \epsilon$$

där $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ är en $n \times 1$ vektor av normalfördelade variabler.

KTH Matematik

Minsta kvadrat via linjär algebra & numerisk analys

För $\mathbf{z}$ en $n \times 1$ -vektor, definieras den euklidiska normen (i kvadrat) av $\mathbf{z}$ som

$$\| \mathbf{z} \|^2 = \sum_{i=1}^{n} z_i^2.$$

Då gäller att

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \| \mathbf{y} - \mathbf{X}\theta \|^2$$

och Q minimeras av ($\mathbf{X}^T$ är den transponerade matrisen, $^{-1}$ är matrisinversen)

$$\widehat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

I Matlab^{TM} ges detta av

$$\widehat{\theta} = \mathbf{X} \setminus \mathbf{y}$$

KTH Matematik

Minsta kvadrat via linjär algebra & numerisk analys

Genom att multiplicera matriserna och utveckla inversen kan man verifiera att

$$\begin{pmatrix} \alpha^*_{\text{obs}} \\ \beta^*_{\text{obs}} \end{pmatrix} = \widehat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

d.v.s vi har samma lösning som ovan. Vi har att residualerna

$$\mathbf{y} - \mathbf{X}\widehat{\theta}$$

är **ortogonala** mot kolonnrummet av matrisen $\mathbf{X}$

Enkel linjär regression

Linjen

$$\widehat{y} = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x.$$

kallas den *skattade regressionslinjen*. De lodräta avstånden ε_i från y_i till den skattade regressionslinjen i x_i,

$$\varepsilon_i = y_i - \alpha^*_{\text{obs}} - \beta^*_{\text{obs}} x_i$$

kallas *observerade residualer*. Q_0 definieras som

$$Q_0 = Q(\alpha^*_{\text{obs}}, \beta^*_{\text{obs}}) = \sum_{i=1}^{n} \varepsilon_i^2.$$

och kallas residualkvadratsumman.
σ^2 skattas med

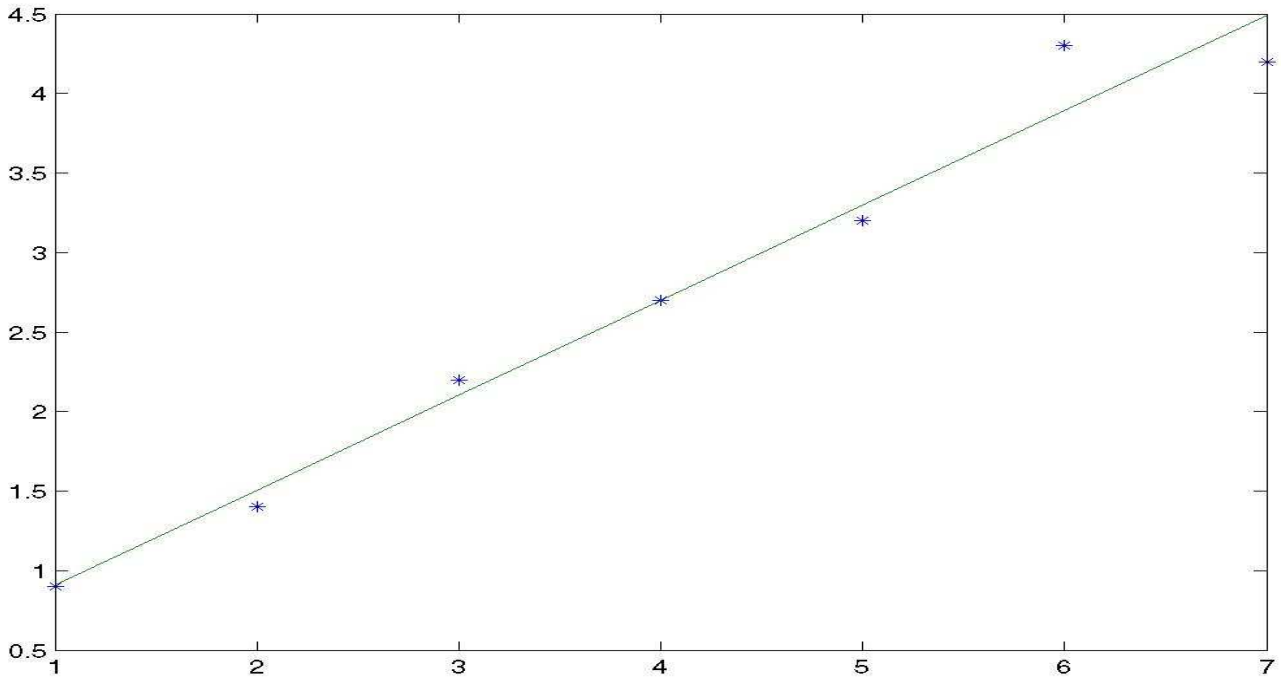
$$s^2 = \frac{Q_0}{n-2}.$$

KTH Matematik

Enkel linjär regression

I exemplet ovan är skattade regressionslinjen, residualkvadratsumman och s^2

$$y = 0.3143 + 0.5964x, \quad Q_0 = 0.2796, \quad s^2 = 0.0559$$



Formlerna för α^*_{obs} och β^*_{obs} omskrivna

$$\alpha^*_{\text{obs}} = \overline{y} - \beta^*_{\text{obs}}\overline{x} \quad \text{och} \quad \beta^*_{\text{obs}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

$$\Leftrightarrow$$

$$\beta^*_{\text{obs}} = \frac{S_{xy}}{S_{xx}} \quad \text{och} \quad \alpha^* = \bar{y} - \beta^*_{\text{obs}}\bar{x}, \tag{2}$$

där vi sätter

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \text{ och } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Covariance and Correlation Coefficient

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

with S_x and S_y,

$$S_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}, S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

and get the *correlation coefficient* as

Definition

$$r_{xy} \stackrel{\text{def}}{=} \frac{c_{xy}}{S_x S_y}$$

$$| r_{xy} | \leq 1.$$

KTH Matematik

Return to the mean (1)

$$\alpha^*_{\text{obs}} = \bar{y} - \beta^*_{\text{obs}}\bar{x}$$

and

$$\widehat{y} = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}}x = \bar{y} + \beta^*_{\text{obs}}(x - \bar{x})$$

i.e.,

$$\widehat{y} - \bar{y} = \beta^*_{\text{obs}}(x - \bar{x}).$$

KTH Matematik

Return to the mean (1)

$$\widehat{y} - \bar{y} = \beta^*_{\text{obs}}(x - \bar{x}).$$

But by the above

$$\beta^*_{\text{obs}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Hence (factor in $1/n$)

$$\beta^*_{\text{obs}} = \frac{c_{xy}}{s_x s_y}\frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x}$$

KTH Matematik

Return to the mean (2)

with

$$\beta^*_{\text{obs}} = r_{xy}\frac{S_y}{S_x}$$

we get

$$\widehat{y} - \bar{y} = \beta^*_{\text{obs}}(x - \bar{x}) = r_{xy}\frac{S_y}{S_x}(x - \bar{x})$$

or

$$\frac{\widehat{y} - \bar{y}}{S_y} = r_{xy}\frac{(x - \bar{x})}{S_x}$$

KTH Matematik

Return to the mean (3)

If $-1 < r_{xy} < 1$, then we see by

$$\frac{\widehat{y} - \bar{y}}{S_y} = r_{xy} \frac{(x - \bar{x})}{S_x}$$

that the predicted standardized value $\widehat{y}$ of y is closer to its mean $\bar{y}$ than the standardized value of x is to its mean $\bar{x}$. Thus the data points $(x_i, y_i)_{i=1}^n$ display regression toward the mean (as found and formulated by Francis Galton).

KTH Matematik

Return to the mean (4)

The idea of regression toward the mean resolved an important difficulty of Darwinian selection:

> *if offspring were always identical to parents, then evolution by natural selection was not possible. But, on the other hand, there was also intergenerational stability, as all experience under fairly constant environmental conditions showed that the range of variability on short time scales, as between two generations, was essentially constant. (Stephen Stigler)*

KTH Matematik

Omskrivning av formlerna

Vi kommer nu att ägna oss åt en rätt så minutiös exercis i omskrivning av formlerna för α^*_{obs} och β^*_{obs}. De (i sig enkla) räkneregler som kommer till användning här ges i Bilagan till detta dokument (se nedan).
Slutresultaten ingår avsnitt 13.3 i Formelsamlingen.

KTH Matematik

Omskrivning av formlerna

Vinsten med detta är att vi överför skattningarna på en form som gör det tydligt, hur vi får normalfördelningarna för de mot svarande α^*_{obs} och β^*_{obs} stickprovsvariablerna α^* och β^* (se avsnitt 13.1 i Formelsamlingen). Detta ger i sin tur s.g.s automatiskt ett antal konfidensintervall (se avsnitt 13.2 i Formelsamlingen) för enkel linjär regression.

KTH Matematik

Formlerna omskrivna

Korssumman S_{xy} kan skrivas på flera sätt och vi har (se (7) & (8) i Bilagan)

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$
$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

Vi har här utnyttjat att summorna $\sum_{i=1}^{n}(x_i - \bar{x})$ (och $\sum_{i=1}^{n}(y_i - \bar{y})$ båda) är 0. Se formel (6) i Bilagan .

KTH Matematik

Formlerna omskrivna

Motsvarande likhet gäller för kvadratsumman S_{xx}, formel (9) i Bilagan och/eller avsnitt 13.3 i Formelsamlingen:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

KTH Matematik

Formlerna omskrivna

Det är viktigt att notera att båda skattningarna är linjära uttryck i observationerna y_i. Man har ju (använd formel (4) i Bilagan upprepade ggr)

- $$\beta^*_{\text{obs}} = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) =$$

$$= \frac{1}{S_{xx}}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right) y_i = \sum_{i=1}^{n} c_i y_i$$

- $$\alpha^* = \bar{y} - \beta^*_{\text{obs}}\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} c_i y_i =$$

$$= \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}c_i\right) y_i = \sum_{i=1}^{n} d_i y_i.$$

Formlerna omskrivna

D.v.s.

$$\beta^*_{\text{obs}} = \sum_{i=1}^{n} c_i y_i \quad \text{och} \quad \alpha^*_{\text{obs}} = \sum_{i=1}^{n} d_i y_i$$

där

$$c_i = (x_i - \bar{x}) / S_{xx} \quad \text{och} \quad d_i = \frac{1}{n} - c_i \bar{x}. \tag{3}$$

De motsvarande stickprovsvariablerna är

$$\beta^* = \sum_{i=1}^{n} c_i Y_i \quad \text{och} \quad \alpha^* = \sum_{i=1}^{n} d_i Y_i$$

De statistiska egenskaperna

Därmed är stickprovsvariablerna α^* och β^* linjära funktioner i Y-variablerna, kom ihåg att $x$en är givna tal.

$$\beta^* = \sum_{i=1}^{n} c_i Y_i \quad \text{och} \quad \alpha^* = \sum_{i=1}^{n} d_i Y_i$$

Därmed är β^* och α^* normalfördelade ! Det gäller att beräkna de respektive väntevärdena och varianserna.

KTH VETENSKAP OCH KONST
KTH Matematik

Väntevärdet på α^*

Vi inleder med

$$E(\alpha^*) = \sum_{i=1}^{n} d_i E(Y_i).$$

Men vi vet att $Y_i \in N(\alpha + \beta x_i, \sigma)$, så att

$$\sum_{i=1}^{n} d_i E(Y_i) = \sum_{i=1}^{n} d_i (\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} d_i + \beta \sum_{i=1}^{n} d_i x_i$$

Vi kommer att visa att

(1) $\sum_{i=1}^{n} d_i = 1$

(2) $\sum_{i=1}^{n} d_i x_i = 0$

KTH Matematik

Väntevärdet på α^*: (1) $\sum_{i=1}^n d_i = 1$

Håll i minnet att $c_i = (x_i - \bar{x})/S_{xx}$.

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (\frac{1}{n} - c_i\bar{x}) = \sum_{i=1}^n \frac{1}{n} - \sum_{i=1}^n c_i\bar{x}$$

$$= 1 - \bar{x}\sum_{i=1}^n c_i = 1 - \bar{x}\frac{1}{S_{xx}}\sum_{i=1}^n (x_i - \bar{x})$$

$$= 1 - \bar{x}\frac{1}{S_{xx}} \cdot 0 = 1.$$

ty $\sum_{i=1}^n (x_i - \bar{x}) = 0$ enligt (6) i Bilagan. Observera att vi även fick

$$\sum_{i=1}^n c_i = 0.$$

KTH Matematik

Väntevärdet på α^*: (2) $\sum_{i=1}^n d_i x_i = 0$

$$\sum_{i=1}^n d_i x_i = \sum_{i=1}^n (\frac{1}{n} - c_i \bar{x}) x_i$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n c_i x_i .$$

Men

$$\sum_{i=1}^n c_i x_i = (1/S_{xx}) \sum_{i=1}^n (x_i - \bar{x}) x_i = (1/S_{xx}) \sum_{i=1}^n (x_i^2 - \bar{x} x_i)$$

$$= (1/S_{xx}) \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) =$$

och från ovan

$$= (1/S_{xx}) \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = (1/S_{xx}) \cdot S_{xx} = 1.$$

KTH Matematik

VÄNTEVÄRDET PÅ α^*: (2) $\sum_{i=1}^n d_i x_i = 0$

Så vi fick att $\sum_{i=1}^n c_i x_i = 1$ och detta ger

$$\sum_{i=1}^n d_i x_i = \sum_{i=1}^n (\frac{1}{n} - c_i \bar{x}) x_i$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \underbrace{\sum_{i=1}^n c_i x_i}_{=1}.$$

$$= \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} = 0.$$

Väntevärdet på α^*:

Således har vi visat att

$$E\left(\alpha^*\right) = \alpha \sum_{i=1}^{n} d_i + \beta \sum_{i=1}^{n} d_i x_i = \alpha.$$

dvs. α^* är väntevärdesriktig.

KTH Matematik

Hjälpresultat

Vissa hjälpresultat har vi dessutom visats:

(A)

$$\sum_{i=1}^{n} c_i x_i = 1$$

(B)

$$\sum_{i=1}^{n} c_i = 0,$$

(C)

$$\sum_{i=1}^{n} d_i x_i = 0$$

KTH Matematik

Väntevärdet på β^*

Vi har $\beta^* = \sum_{i=1}^{n} c_i Y_i$ och detta ger

$$E(\beta^*) = \sum_{i=1}^{n} c_i E(Y_i).$$

Men $Y_i \in N(\alpha + \beta x_i, \sigma)$, så att

$$\sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i (\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} c_i + \beta \sum_{i=1}^{n} c_i x_i$$

Hjälpresultaten (A) och (B) ovan ger

$$E(\beta^*) = \beta$$

och β^* är väntevärdesriktig.

KTH Matematik

VARIANS FÖR α^*

Vi ha p.g.a oberoendet mellan $Y_i$na att

$$V(\alpha^*) = \sum_{i=1}^{n} d_i^2 V(Y_i).$$

Men $Y_i \in N(\alpha + \beta x_i, \sigma)$, så att

$$\sum_{i=1}^{n} d_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^{n} d_i^2$$

Näst visar vi att

- $\sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{x}}{S_{xx}}$

KTH VETENSKAP OCH KONST
KTH Matematik

VARIANS FÖR α^*

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

$$= \sum_{i=1}^{n} \frac{1}{n^2} - \frac{2\bar{x}}{n} \sum_{i=1}^{n} c_i + \bar{x}^2 \sum_{i=1}^{n} c_i^2$$

Men hjälpresultatet (B) ovan ger $\sum_{i=1}^{n} c_i = 0$. Av definition på c_i och S_{xx} fås

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / S_{xx}^2 = \frac{1}{S_{xx}}$$

och $\sum_{i=1}^{n} \frac{1}{n^2} = 1/n$. Sammanfattningsvis

$$\sum_{i=1}^{n} d_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

VARIANS FÖR α^*

$$V(\alpha^*) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

I härledningen fick vi även hjälpresultatet

(D)

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{S_{xx}}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Varians för β^*

Vi har $\beta^* = \sum_{i=1}^{n} c_i Y_i$ och p.g.a oberoendet mellan $Y_i$na att

$$V(\beta^*) = \sum_{i=1}^{n} c_i^2 V(Y_i).$$

Men $Y_i \in N(\alpha + \beta x_i, \sigma)$, så att

$$\sum_{i=1}^{n} c_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}}$$

enligt det senaste hjälpresultatet (D).

KTH VETENSKAP OCH KONST
KTH Matematik

De statistiska egenskaperna

Sats

$$\alpha^* \in N\left(\alpha, \sigma\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}\right)$$

$$\beta^* \in N\left(\beta, \sigma\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right).$$

KTH Matematik

De statistiska egenskaperna

Sats

Vidare gäller att, här betyder Q_0 stickprovsvariabeln $Q(\alpha^*, \beta^*)$,

$$\frac{Q_0}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2}$$

är $\chi^2(n-2)$-fördelad och att S^2 är oberoende av α^* och β^*.

KTH VETENSKAP OCH KONST
KTH Matematik

En algebraisk exercis i ovanstående stil visar även att

$$Q_0 = Q(\alpha^*_{\text{obs}}, \beta^*_{\text{obs}}) = \sum_{i=1}^{n} \varepsilon_i^2 =$$

$$= \sum_{i=1}^{n} (y_i - \alpha^*_{\text{obs}} - \beta^*_{\text{obs}} x_i)^2 = S_{yy} - \frac{S_{xy}}{S_{xx}}.$$

KTH Matematik

Konfidensintervall

Det ovanstående gör att vi kan konstruera konfidensintervall och test som förut, både då σ är känt och okänt. För att inte behöva skriva alla intervall två gånger så betraktar vi fallet då σ är okänt.

KTH VETENSKAP OCH KONST
KTH Matematik

Konfidensintervall

Eftersom α nu förekommer som en parameter, så ger vi konfidensintervall med konfidensgrad 95%. Metoden är definiera ett θ och sedan bilda

$$I_\theta = \theta^*_{\text{obs}} \pm t_{0.025}(n-2)d(\theta^*),$$

där $d(\theta^*)$ är skattningen av $D(\theta^*)$.

KTH Matematik

Medelfelet $d(\theta^*)$

Vi använder som skattning av σ i konfidensintervallen

$$s = \sqrt{\frac{Q_0}{n-2}}.$$

(dvs.kvadratroten av residualkvadratsumman$/n-2$)

KTH Matematik

Konfidensintervall för α

$$I_\theta = \theta^*_{\text{obs}} \pm t_{0.025}(n-2)d(\theta^*),$$

(1) $\underline{\theta = \alpha}$

Vi har $V(\theta^*) = V(\alpha^*) = \sigma^2 \cdot \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$, vilket ger

$$I_\alpha = \alpha^* \pm t_{0.025}(n-2)s\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}.$$

KTH Matematik

Konfidensintervall för β

(2) $\underline{\theta = \beta}$

Vi har $V(\beta^*) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$, vilket ger

$$I_\beta = \beta^*_{\text{obs}} \pm t_{0.025}(n-2) \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} .$$

KTH VETENSKAP OCH KONST
KTH Matematik

Hypotesprövning för β

Det kan hända att inget samband i form av en teoretisk regressionslinje finns, dvs. $\beta = 0$. Vi kan testa detta genom

$$H_0: \ \beta = 0$$

mot

$$H_1: \ \beta \neq 0$$

Förkasta H_0 på nivån 0.05 om

$$0 \notin I_\beta = \beta^*_{\text{obs}} \pm t_{0.025}(n-2)\frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\,.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Konfidensintervall för regressionslinjen

(3) $\underline{\theta = \alpha + \beta x_0}$

Detta är alltså ett konfidensintervall för den teoretiska regressionslinjen i punkten $x = x_0$.

Vi observerar först

$$E(\alpha^* + \beta^* x_0) = E(\alpha^*) + E(\beta^* x_0) = \alpha + \beta x_0.$$

$$V(\alpha^* + \beta^* x_0) = V\left(\sum_{i=1}^{n} d_i Y_i + \sum_{i=1}^{n} c_i Y_i x_0\right)$$

$$= V\left(\sum_{i=1}^{n} (d_i + c_i x_0)\, Y_i\right)$$

och p.g.a. oberoendet

$$= \sum_{i=1}^{n} (d_i + c_i x_0)^2\, V(Y_i) = \sigma^2 \sum_{i=1}^{n} (d_i + c_i x_0)^2 .$$

KTH Matematik

$d_i = \frac{1}{n} - c_i\bar{x}$ så att

$$\sum_{i=1}^{n}(d_i + c_i x_0)^2 = \sum_{i=1}^{n}\left(\frac{1}{n} + c_i(x_0 - \bar{x})\right)^2 =$$

$$= \sum_{i=1}^{n}\frac{1}{n^2} + 2\frac{(x_0 - \bar{x})}{n}\sum_{i=1}^{n}c_i + (x_0 - \bar{x})^2\sum_{i=1}^{n}c_i^2 = \frac{1}{n} + (x_0 - \bar{x})^2/S_{xx}.$$

ty hjälpresultatet (B) ovan ger $\sum_{i=1}^{n} c_i = 0$ och hjälpresultatet (D) ger $\sum_{i=1}^{n} c_i^2 = 1/S_{xx}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Vi slår ihop från ovan

$$V(\alpha^* + \beta^* x_0) = \sigma^2 \cdot \left(\frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx} \right).$$

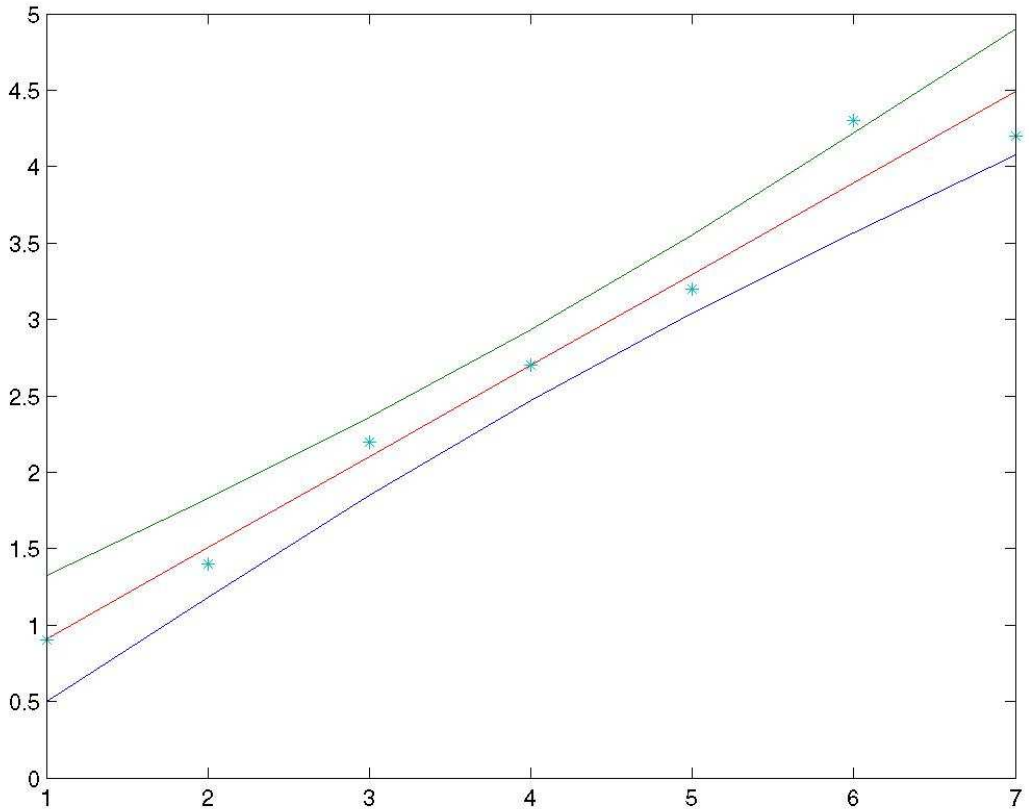
Detta ger det slutliga konfidensintervallet som

$$I_{\alpha+\beta x_0} = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0 \pm t_{0.025}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}.$$

KTH Matematik

Konfidensintervall för linjen i exemplet ovan

Data, regressionslinje och $\alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0 \pm t_{0.025}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}$



KTH Matematik

Prediktionsintervall

Låt Y_0 vara en ny observation som svarar mot x_0 dvs. $Y_0 \in N(\alpha + \beta x_0, \sigma)$. Vi predikterar Y_0 genom $\widehat{Y}_0$ som ges av den skattade regressionslinjen

$$\widehat{Y}_0 = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0.$$

Prediktionsfelet

$$Y_0 - \widehat{Y}_0$$

är normalfördelat med väntevärdet noll. Eftersom Y_0 och $\widehat{Y}_0$ är oberoende

$$V\left(Y_0 - \widehat{Y}_0\right) = V\left(Y_0\right) + V\left(\widehat{Y}_0\right)$$

Ovan har vi i fallet (3) visat att

$$V\left(\widehat{Y}_0\right) = V(\alpha^* + \beta^* x_0) = \sigma^2 \cdot \left(\frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}\right).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Prediktionsintervall

Alltså

$$V\left(Y_0 - \widehat{Y}_0\right) = \sigma^2\left(1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}\right)$$

Vi skattar σ^2 med s^2.

KTH Matematik

Prediktionsintervall

Man kan visa att kvoten

$$\frac{Y_0 - \widehat{Y}_0}{s \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}}.$$

är t-fördelad med $n - 2$ frihetsgrader. Då kan vi på bekant manér visa att intervallet

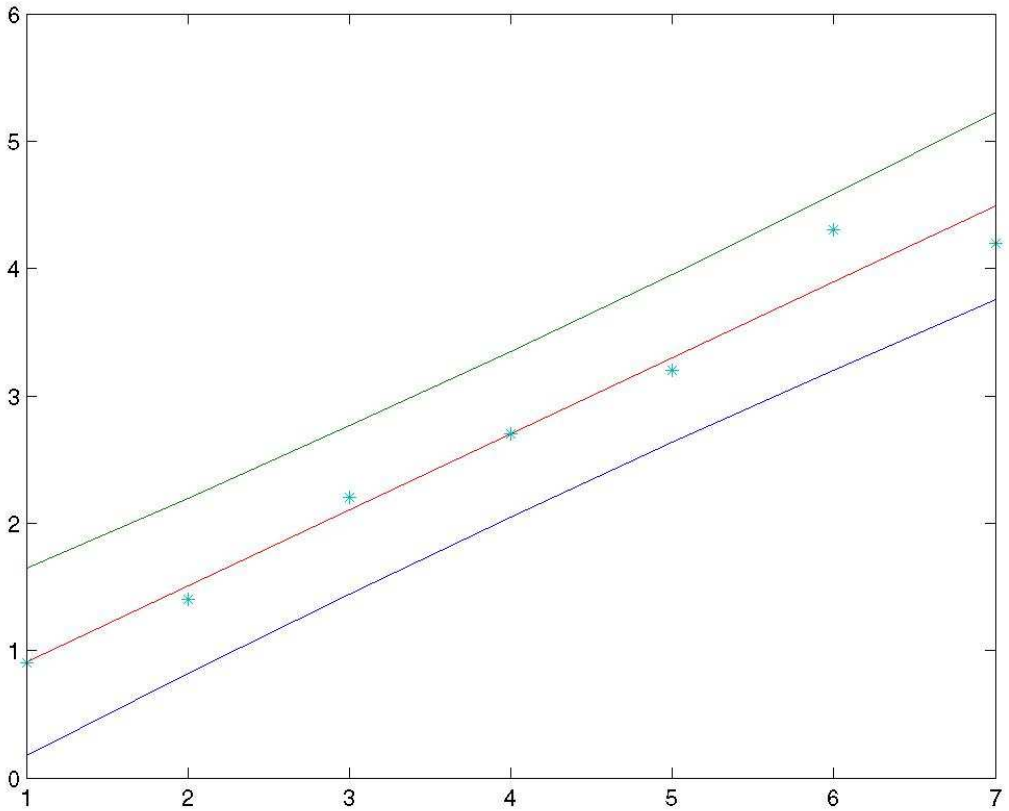
$$\widehat{Y}_0 \pm t_{0.025}(n-2)s \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}$$

täcker Y_0 med sannolikheten 0.95. Intervallet kallas ett 95%-igt prediktionsintervall för Y_0

KTH VETENSKAP OCH KONST
KTH Matematik

Prediktionsintervall i exemplet ovan

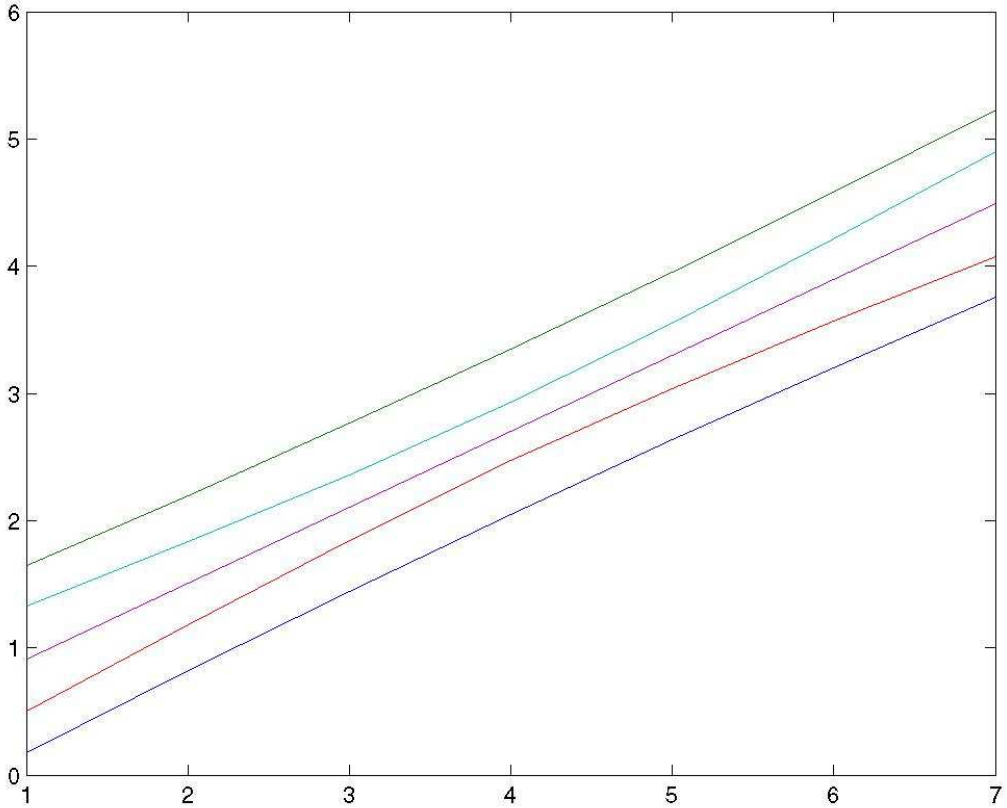
Data, regressionslinje och $\widehat{Y}_0 \pm t_{0.025}(n-2)s \cdot \sqrt{1+\frac{1}{n}+(x_0-\bar{x})^2/S_{xx}}$



Prediktionsintervall och konfidensintervall i exemplet ovan

Regressionslinje, $\alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0 \pm t_{0.025}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}$ och

$\widehat{Y}_0 \pm t_{0.025}(n-2)s \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}$



KTH Matematik

Multipel linjär regressionsanalys: vinets kvalitet

Multipel regression med tre förklarande variabler $x1, x2, x3$ för statistisk prediktion (SPR) av (Bordeaux-)vinernas kvalitet:

$$\text{vinets kvalitet} = 12.145 + 0.00117x1 + 0.0614x2 - 0.00386x3$$

där

$$x1 = \text{ nederbörd under vintern}$$

$$x2 = \text{ genomsnittstemperaturen under växtsäsongen}$$

$$x3 = \text{ nederbörd under skördeperioden}$$

av Orley Ashenfelter, se `http://www.liquidasset.com/` .

KTH Matematik

Multipel linjär regressionsanalys: vinets kvalitet

Statistisk prediktion (SPR) av (Bordeaux-)vinernas kvalitet:

$$\text{vinets kvalitet} = 12.145 + 0.00117x1 + 0.0614x2 - 0.00386x3$$

Poängen här alltså att vi vill uttala oss om kvaliten hos en årgång av (Bordeaux-)vin INNAN NÅGON smakat på vinet. Detta tycks fungera, men upplevs trots detta (eller kanske just p.g.a detta) som stötande av många goda vinexperter.

KTH Matematik

Statistisk prediktionsregel (SPR)

For a very wide range of prediction problems, statistical prediction rules (SPRs), often rules that are very easy to implement, make predictions than are as reliable as, and typically more reliable than, human experts. The success of SPRs forces us to reconsider our views about what is involved in understanding, explanation, and good reasoning.

KTH Matematik

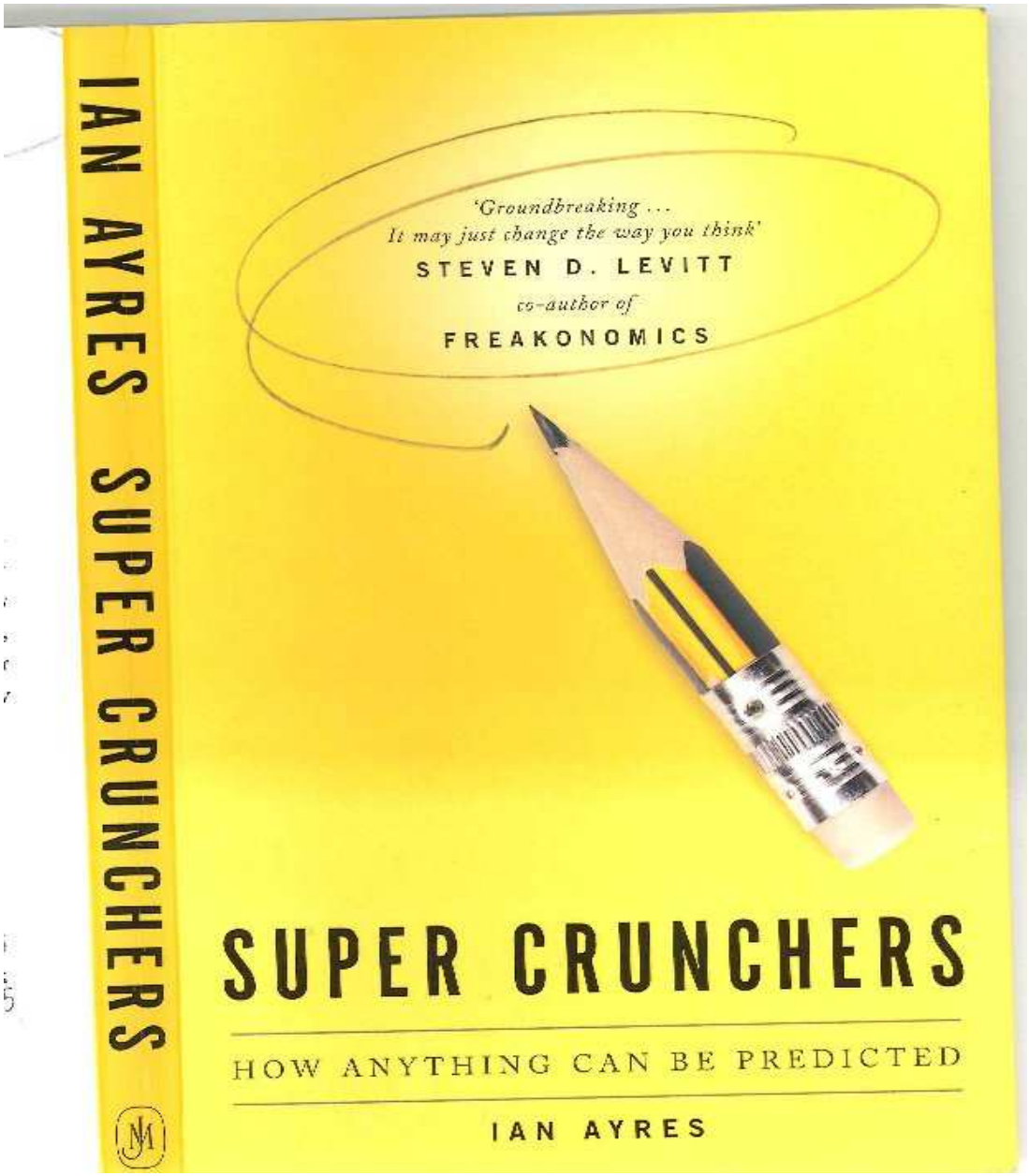
En bok om SPR

Orley Ashenfelters SPR för vinets kvalitet är hämtad ur boken

I.Ayres: Super Crunchers. How anything can be predicted. John Murray (Publishers), Paperback edition 2008, London.

KTH Matematik

En bok om SPR



KTH VETENSKAP OCH KONST
KTH Matematik

Bilaga : räkneoperationer med summor

Definition

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$

Sats

(2) $\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.$

Sats

(3) $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

KTH Matematik

Bilaga : räkneoperationer med summor

Sats

(4) $\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$

Sats

(5) $\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$

KTH Matematik

Bilaga : räkneoperationer med summor

Låt $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Då gäller

Sats

(6) $\underline{\sum_{i=1}^{n}(x_i - \overline{x}) = 0.}$

Sats

(7) $\underline{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})\, y_i = \sum_{i=1}^{n} x_i\,(y_i - \overline{y})\,.}$

KTH Matematik

Bilaga : räkneoperationer med summor

Sats

(8) $\underline{\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-n\overline{xy}.}$

Sats

(9) $\underline{\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2.}$

KTH Matematik