Matematisk statistik för B, K, N, BME och Kemister

Föreläsning 10

Johan Lindström

27 september 2017

Linjär regression

Modell (Kap. 10.2)

Vi har n st par av mätvärden (x_i, y_i), $i = 1, \ldots, n$ där y_i är observationer av

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

där ε_i är oberoende av varandra, och $\varepsilon_i \in \mathrm{N}(0, \sigma^2)$.

Parameterskattningarna (Kap. 10.4–10.5)

Skattningarna av α^*, β^*

$$\beta^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad \in \mathrm{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$\alpha^* = \bar{y} - \beta^* \bar{x} \quad \in \mathrm{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

och $s^2 = (\sigma^2)^*$ är

$$s^2 = \frac{Q_0}{n-2} \text{ där } Q_0 = \sum_{i=1}^n (y_i - \alpha^* - \beta^* x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\frac{Q_0}{\sigma^2} \in \chi^2(n-2)$$

Skattningarna α^* och β^* är dock **inte oberoende** av varandra.

Konfidens- & Prediktionsintervall (Kap. 10.6–10.7)

Konfidensintervall för linjen, μ_0, vid x_0:

$$I_{\mu_0} = \alpha^* + \beta^* x_0 \pm t_{a/2}(n-2) \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Prediktionsintervall för en *ny mätning*, $Y(x_0)$, vid x_0:

$$I_{Y(x_0)} = \alpha^* + \beta^* x_0 \pm t_{a/2}(n-2) \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Kalibreringsintervall (Kap. 10.8)

Kalibreringsintervall för $x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$ givet en mätning y_0,

$$I_{x_0} = x_0^* \pm t_{a/2}(n-2) \cdot \frac{s}{|\beta^*|} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0^* - \bar{x})^2}{S_{xx}}}$$

Linjärisering av exponentiella samband

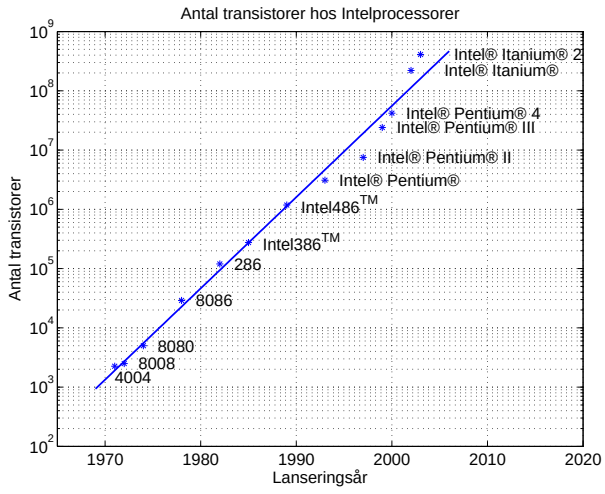
För att få ett linjärt samband

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

kan vissa exponent- och potenssamband logaritmeras.

$$z_i = a \cdot e^{\beta x_i} \cdot \varepsilon_i' \quad \xrightarrow{\ln} \quad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \cdot x_i + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

$$z_i = a \cdot t_i^{\beta} \cdot \varepsilon_i' \quad \xrightarrow{\ln} \quad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \underbrace{\ln t_i}_{x_i} + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

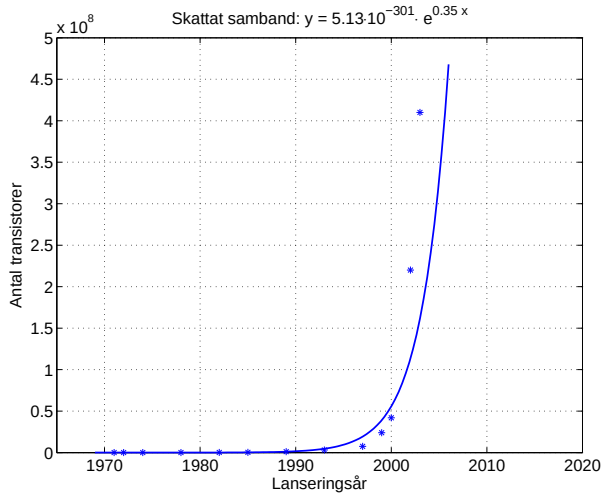


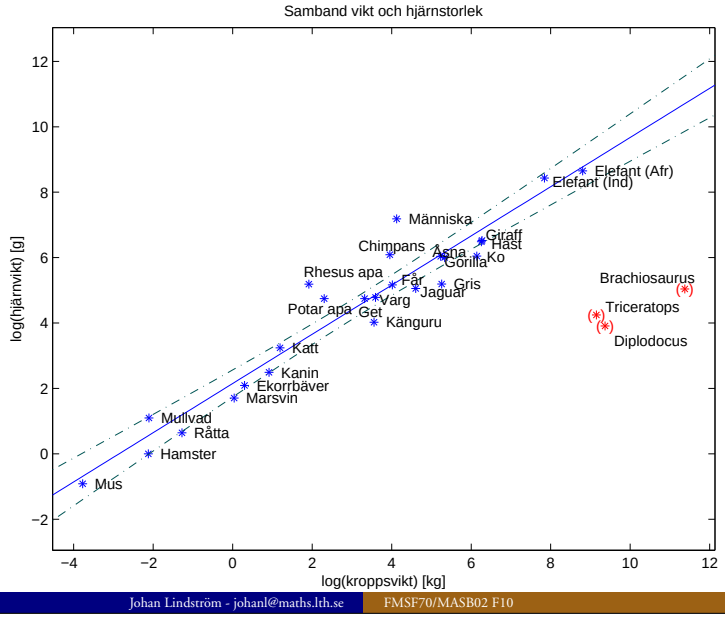
Exempel: Moores lag

Figuren på föregående slide är baserad på Moores Lag.
1965 framförde Gordon Moore (en av Intels grundare) tesen att antalet transistorer på ett chip fördubblas vartannat år (`www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf`).
Genom att anpassa en exponential funktion till data fås följande

$$\ln z_i = -691 + 0.35 x_i \qquad z_i = 5.13 \cdot 10^{-301} \cdot \exp(0.35 x_i)$$

där z_i är antalet transistorer och x_i är lanseringsår.





Multipel regression (Kap. 11.2)

Modellen

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \in \mathrm{N}\left(0, \sigma^2\right) \text{ oberoende}$$

kan skrivas på matrisform som

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{E}$$

där $\boldsymbol{Y}$ och $\boldsymbol{E}$ är $n \times 1$-vektorer, β en $(p+1) \times 1$-vektor och $\boldsymbol{X}$ en $n \times (p+1)$-matris

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{E} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Skattning av β och σ^2 (Kap. 11.3)

MK-skattningar av $\beta_0, \ldots, \beta_p$ (elementen i β) blir

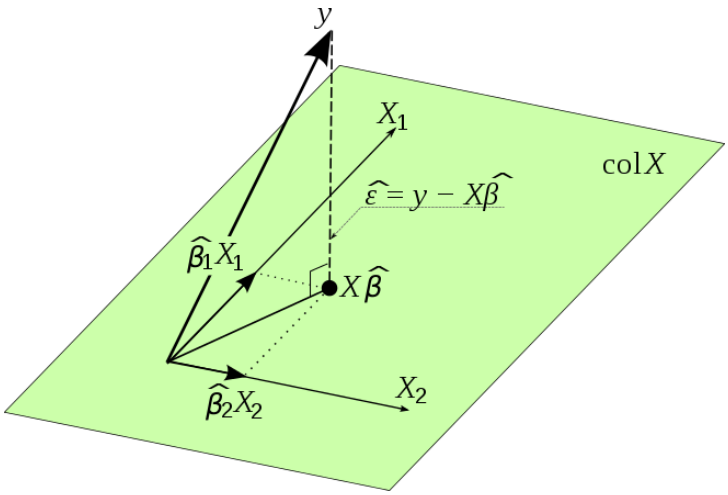
$$\beta^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \qquad \mathrm{V}(\beta^*) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

och skattning av σ^2 är

$$s^2 = \frac{Q_0}{n - (p+1)}$$

där **residualkvadratsumman** ges av

$$\begin{aligned} Q_0 &= \sum_{i=1}^{n} \left(y_i - \beta_0^* - \beta_1^* x_{1i} - \ldots - \beta_p^* x_{pi}\right)^2 \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - \beta^{*\top} \boldsymbol{X}^\top \boldsymbol{Y} \end{aligned}$$



en.wikipedia.org/wiki/Ordinary_least_squares#/media/File:
OLS_geometric_interpretation.svg

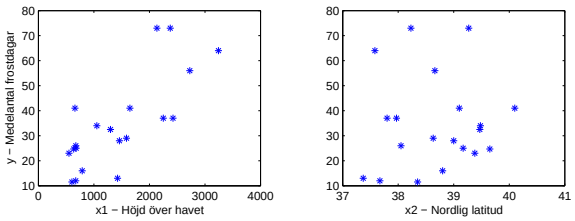
Exempel: Antal frostdagar

I West Virginia har man under ett antal år mätt antalet frostdagar på olika orter. Följande data har registrerats

- Y: Medelantalet frostdagar per år.
- x_1: Ortens höjd över havet (ft).
- x_2 Ortens nordlig breddgrad (°).

Skatta parametrarna i modellen

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$



Exempel: Antal frostdagar

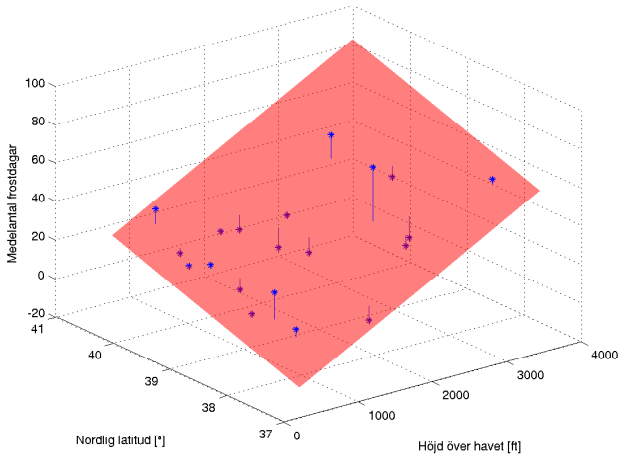
För data fås följande värden:

$$\boldsymbol{X}^\top \boldsymbol{Y} = \begin{bmatrix} -27.0 \\ 1.89 \cdot 10^5 \\ -1.07 \cdot 10^3 \end{bmatrix} \qquad Q_0 = 1.7798 \cdot 10^3$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{bmatrix} 1.59 \cdot 10^2 & -1.64 \cdot 10^{-3} & -4.06 \\ -1.64 \cdot 10^{-3} & 9.14 \cdot 10^{-8} & 3.91 \cdot 10^{-5} \\ -4.06 & 3.91 \cdot 10^{-5} & 1.03 \cdot 10^{-1} \end{bmatrix}$$

Bestäm:

1. Skattningar av β.
2. Konfidensintervall för β_1.



Det anpassade regressionplanet mellan antalet frostdagar och h.ö.h. samt latitud.

Konfidensintervall för β_i (Kap. 11.5)

Konfidensintervall för β_i blir alltså

$$I_{\beta_i} = \beta_i^* \pm t_{a/2}(n-p-1) \cdot d(\beta_i^*)$$

Där $d(\beta_i^*)$ är

$$d(\beta_i^*) = s \cdot \sqrt{\text{element}(ii) \text{ i } (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}$$

Skattning av punkt på ”planet” (Kap. 11.4–11.5)

Y-s väntevärde i en punkt $\begin{bmatrix} x_1^0 & x_2^0 & \cdots & x_p^0 \end{bmatrix}$ ges nu av

$$\mu_Y^*(\boldsymbol{x}^0) = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i^0.$$

$$\mathrm{V}(\mu_Y^*(\boldsymbol{x}_0)) = \sigma^2 \cdot \boldsymbol{x}^{0\top} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}^0.$$

Ett **konfidensintervall** för $\mu_Y(\boldsymbol{x}_0)$ blir

$$I_{\mu_Y(\boldsymbol{x}^0)} = \mu_Y^*(\boldsymbol{x}^0) \pm t_{a/2}(n-p-1) \cdot s \cdot \sqrt{\boldsymbol{x}^{0\top} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}^0}$$

För **prediktionsintervallet** fås, som tidigare, genom att lägga till en etta under kvadratroten

$$I_{Y(\boldsymbol{x}^0)} = \mu_Y^*(\boldsymbol{x}^0) \pm t_{a/2}(n-p-1) \cdot s \cdot \sqrt{1 + \boldsymbol{x}^{0\top} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{x}^0}$$

Exempel: Antal frostdagar

För data fås följande värden:

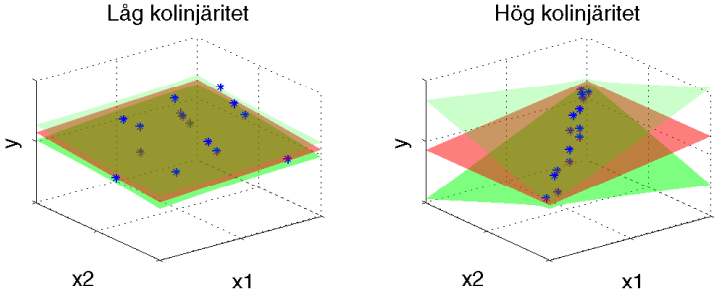
$$\boldsymbol{X}^\top \boldsymbol{Y} = \begin{bmatrix} -27.0 \\ 1.89 \cdot 10^5 \\ -1.07 \cdot 10^3 \end{bmatrix} \qquad Q_0 = 1.7798 \cdot 10^3$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{bmatrix} 1.5945 \cdot 10^2 & -1.6445 \cdot 10^{-3} & -4.0590 \\ -1.6445 \cdot 10^{-3} & 9.1434 \cdot 10^{-8} & 3.9094 \cdot 10^{-5} \\ -4.0590 & 3.9094 \cdot 10^{-5} & 1.0346 \cdot 10^{-1} \end{bmatrix}$$

Skatta medelantalet frostdagar och ett 95%-konfidensintervall då $x_1 = 3\,000$ och $x_2 = 39$.

Kolinjäritet (ex. två variabler) (Kap. 11.6)

Man bör om möjligt välja sina (x_{1i}, x_{2i})-värden så att de blir utspridda i (x_1, x_2)-planet och inte klumpar ihop sig längs en linje. Detta ger ”en mer stabil grund” åt regressionsplanet.



Polynomregression

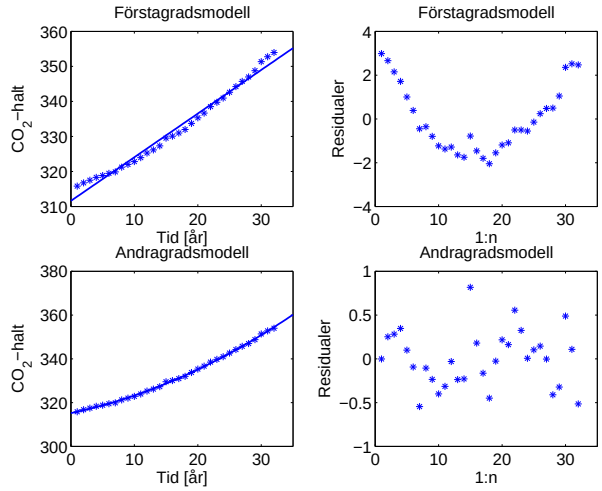
Om y är ett polynom av x, dvs vi har

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i$$

och funktionen är **linjär** i β_p.
Genom att samla polynomen av x i en matris

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

kan parametrar skattas på samma sätt som tidigare.



Linjär $y = \alpha + \beta x$, och kvadratisk, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, anpassning av årlig CO_2-halten vid Mauna Loa som funktion av året (sedan 1960).