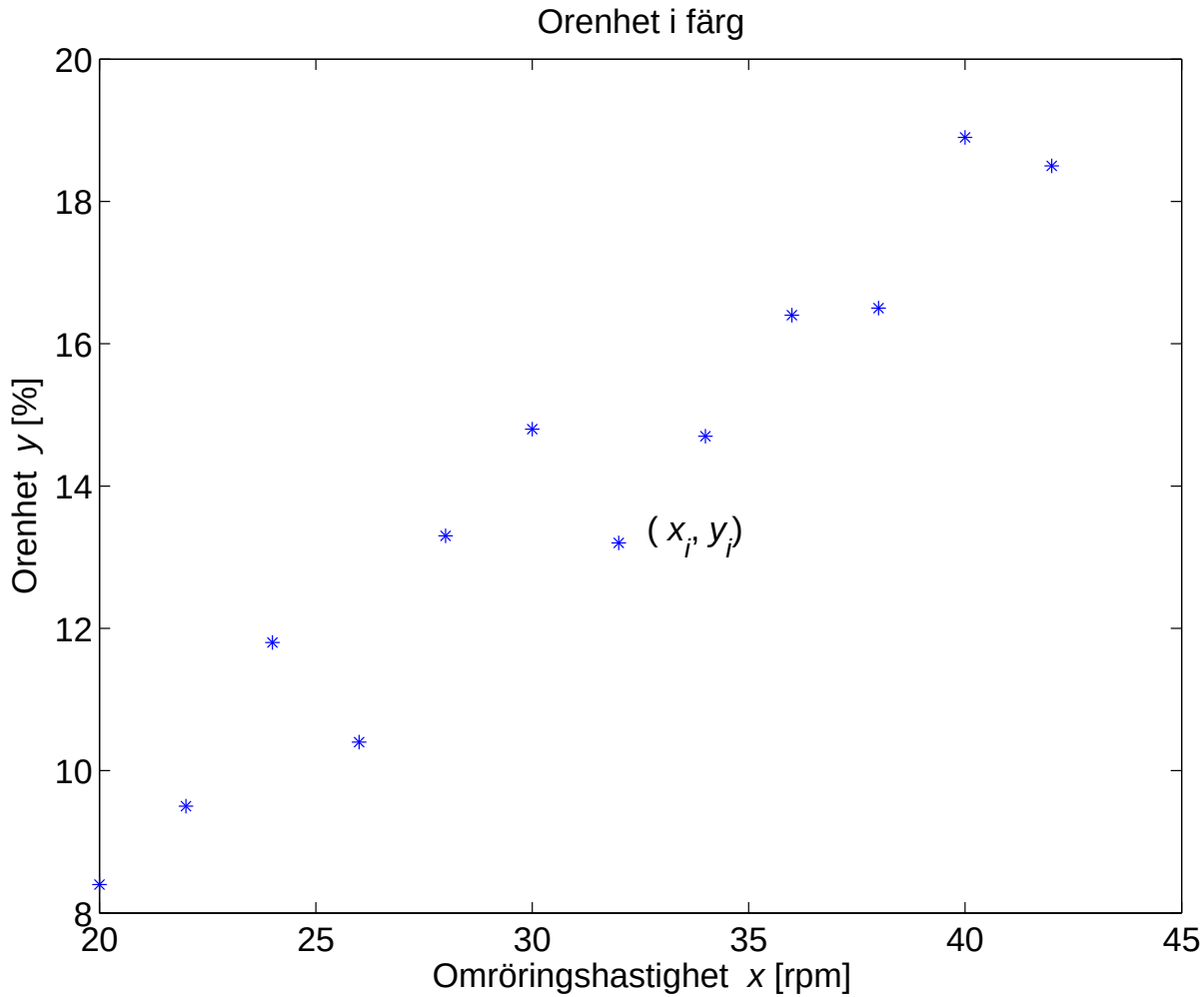
Föreläsning 12, Matematisk statistik 7.5 hp för E
Enkel linjär regression

Stas Volkov

Enkel linjär regression

Finns det ett linjärt samband mellan x och y? Hur ser det i så fall ut?



Linjär regression

Modell

Vi har n st par av mätvärden $(x_k, y_k),\ k = 1, \ldots, n$ där y_k är observationer av

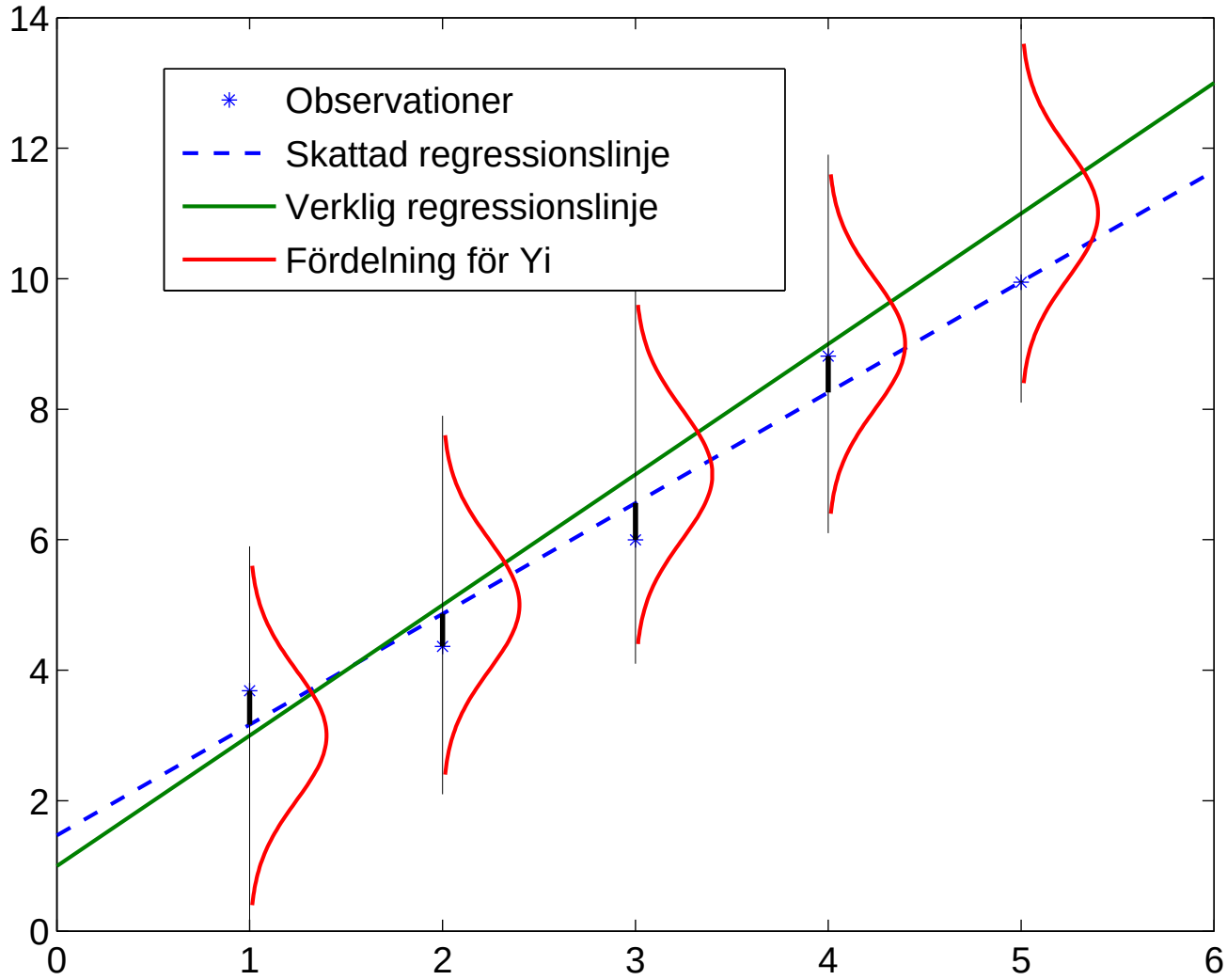
$$Y_k = \alpha + \beta x_k + \varepsilon_k$$

där ε_k är oberoende av varandra, och $\varepsilon_k \in \mathsf{N}(0, \sigma)$.

α och β är okända tal, x_k är kända **tal** (”oberoende varibel”). Vi får då

$$Y_k \in \mathsf{N}(\alpha + \beta x_k, \sigma) = \mathsf{N}(\mu_k, \sigma)$$

Y’s väntevärde ligger på en rät linje, $\mu(x) = \alpha + \beta x$, där β är linjens lutning och α dess skärning med y-axeln.


Observationer
Skattad regressionslinje
Verklig regressionslinje
Fördelning för Yi
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6

Skattning av parametrarna α^* och β^*

Parametrarna kan skattas t.ex. med ML-metoden, dvs det $(\alpha, \beta, \sigma^2)$ som maximerar (med avseende på (α, β))

$$L(\alpha, \beta, \sigma^2) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_k-(\alpha+\beta x_k))^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-(\alpha+\beta x_k))^2}$$

Parameterskattningarna

ML-skattningarna av α^* och β^* är (OBS! MK-skattningarna också!)

$$\beta^* = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2} =: \frac{S_{xy}}{S_{xx}}, \qquad \alpha^* = \bar{y} - \beta^* \cdot \bar{x}$$

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

Skattning av σ^2

Definera **residualerna** som avstånden från y_k-värden till den skattade linjen:

$$e_k = y_k - (\alpha^* + \beta^* x_k) =: y_k - \mu^*(x_k)$$

och **residualkvadratsumma**

$$Q_0 = \sum_{k=1}^{n} e_k^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

För σ^2, dvs. variansen för ε_k samt för Y_k, blir då skattningen

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

Skattningen är korrigerad med ”-2” för att bli väntevärdesriktig.

Räkna ut kvadratsummorna

För att räkna ut kvadratsummorna S_{xx}, S_{yy} och S_{xy} kan man ha användning av sambanden

$$S_{xx} = \sum_{k=1}^{n}(x_k - \bar{x})^2 = \left(\sum_{k=1}^{n} x_k^2\right) - n\bar{x}^2$$

$$S_{yy} = \sum_{k=1}^{n}(y_k - \bar{y})^2 = \left(\sum_{k=1}^{n} y_k^2\right) - n\bar{y}^2$$

$$S_{xy} = \sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) = \left(\sum_{k=1}^{n} x_k y_k\right) - n\bar{x}\bar{y}$$

Exempel, x-Cu-konc. y-absorption

Vi har följande 10 par av mätvärden:

k:	1	2	3	4	5	6	7	8	9	10
x_k:	0	0	50	50	100	100	150	150	200	200
y_k:	−0.005	−0.003	0.092	0.084	0.185	0.185	0.329	0.302	0.372	0.336

För att skatta parametrarna i regressionsmodellen behövs

$$n = 10, \quad \sum x_k = 1000, \quad \sum y_k = 1.8770,$$
$$\sum x_k y_k = 282.05, \quad \sum x_k^2 = 150000, \quad \sum y_k^2 = 0.5347$$

$$\Longrightarrow \quad S_{xx} = \sum_{k=1}^{n} x_k^2 - n\bar{x}^2 = 150000 - 10 \cdot 100^2 = 50000$$

$$S_{yy} = 0.5347 - 10\left(\frac{1.8770}{10}\right)^2 = 0.1824$$

$$S_{xy} = 282.06 - \frac{1}{10} \cdot 1000 \cdot 1.8770 = 94.35$$

Exempel, x-Cu-konc. y-absorption

Skattningarna blir

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{94.35}{50000} = 0.0019$$

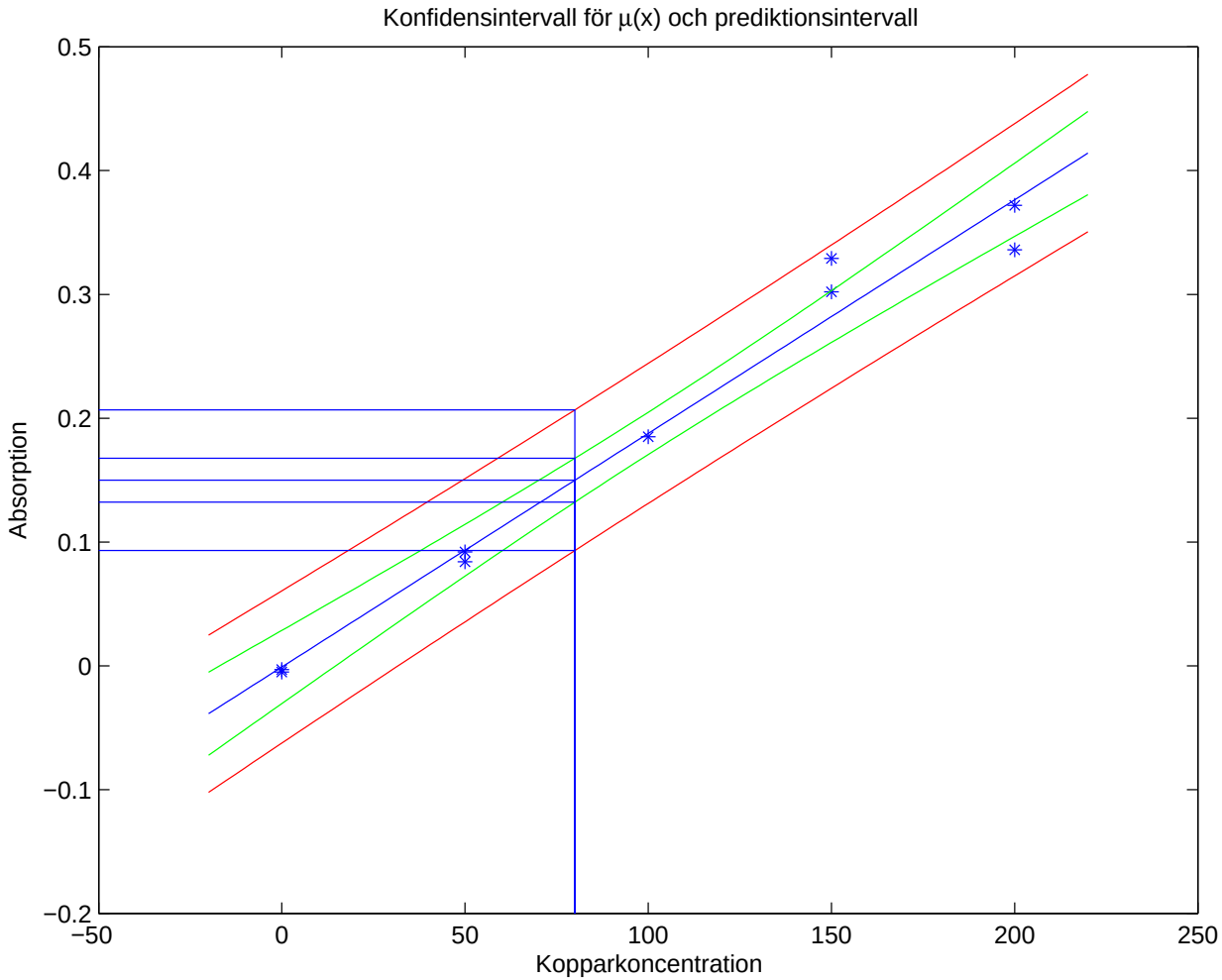
$$\alpha^* = \bar{y} - \beta^* \cdot \bar{x} = \frac{1.8770}{10} - 0.0019 \cdot \frac{1000}{10} = -0.0023$$

Samt

$$Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 0.1824 - \frac{94.35^2}{50000} = 0.0044$$

$$\sigma^* = s = \sqrt{\frac{Q_0}{n-2}} = \sqrt{\frac{0.0044}{8}} = 0.0234$$

Exempel, Regressionslinje



Är skattningarna bra? Nära de rätta värdena?

Skattningarnas fördelning

Eftersom

$$\alpha^* = \bar{Y} - \beta^* \cdot \bar{x} \qquad \beta^* = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(Y_k - \bar{Y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

båda är **linjära funktioner av Y_k** (men inte av talen x_k), som är normalfördelade, måste även α^* och β^* vara **normalfördelade**.

Skattningarnas fördelning

Fördelningarna blir

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

... men de är **inte dock oberoende** av varandra!

Skattningarnas fördelning (forts)

Men man kan visa att β^* och $\bar{Y}$ är oberoende av varandra.
Kovariansen mellan α^* och β^* blir då

$$C(\alpha^*, \beta^*) = C(\bar{Y} - \beta^* \cdot \bar{x}, \beta^*) = C(\bar{Y}, \beta^*) - \bar{x} \cdot C(\beta^*, \beta^*) =$$
$$= 0 - \bar{x} \cdot V(\beta^*) = -\bar{x} \cdot \frac{\sigma^2}{S_{xx}}.$$

Vi ser att kovariansen är negativ då $\bar{x} > 0$ och positiv då $\bar{x} < 0$.

För Q_0 gäller dessutom

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

(”2” är fortfarande **antalet skattade parametrar** i Q_0 – kom ihåg tumregeln!)

Intervallskattningar

Skattningarna av α^* och β^* är båda på formen

$$\vartheta^* \in \mathsf{N}\left(\vartheta, \mathsf{D}(\vartheta^*)\right)$$

där $\mathsf{D}(\vartheta^*)$ innehåller ett σ som skattas med ett s med $f = n - 2$ frihetsgradet. Så vi får konfidensintervall med konfidensgrad $1 - a$ ("a" används eftersom α är upptagen) som vanligt:

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

$$I_\beta = \beta^* \pm t_{a/2}(f) \cdot d(\beta^*) = \beta^* \pm t_{a/2}(n-2) \cdot \frac{s}{\sqrt{S_{xx}}}$$

Konfidensintervall för μ_0 (eller för linjen)

För ett givet x-värde, $x = x_0$, kan vi skatta Y-s väntevärde med $\mu_0^* = \mu^*(x_0) = \alpha^* + \beta^* \cdot x_0$ dvs. en punkt på den skattade linjen.

$$
\begin{aligned}
\mathsf{V}(\mu_0^*) &= \mathsf{V}(\alpha^* + \beta^* \cdot x_0) \\
\langle \alpha^* = \bar{Y} - \beta^* \cdot \bar{x} \rangle &= \mathsf{V}\left(\bar{Y} + \beta^* \cdot (x_0 - \bar{x})\right) \\
\langle \beta^*, \bar{Y} \text{ är ju oberoende} \rangle &= \mathsf{V}(\bar{Y}) + (x_0 - \bar{x})^2 \cdot \mathsf{V}(\beta^*) \\
&= \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \cdot \frac{\sigma^2}{S_{xx}} \\
\Longrightarrow \quad \mu_0^* &\in \mathsf{N}\left(\mu_0, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right).
\end{aligned}
$$

Vi får således direkt ett **konfidensintervall** för μ_0 som

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

Prediktionsintervall

Intervallet ovan gäller **väntevärdet** för Y då $x = x_0$. Om man vill uttala sig om **en** framtida **observation** av Y för $x = x_0$ blir ovanstående intervall för smalt.

Vi kan få ett **prediktionsintervall** för en framtida observation för ett givet x_0

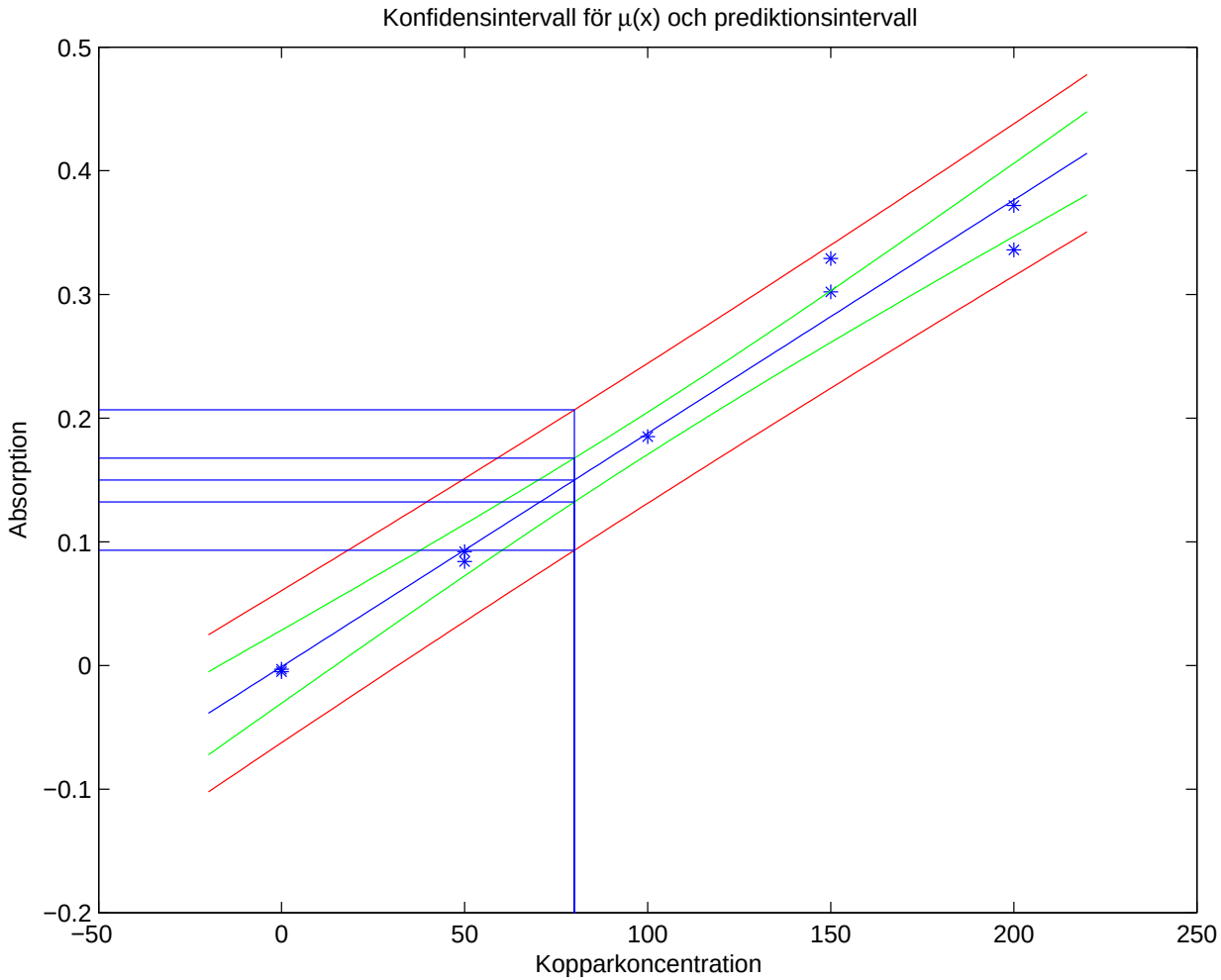
$$Y(x_0) = \alpha^* + \beta^* \cdot x_0 + \varepsilon_0 = \mu_0^* + \varepsilon_0,$$
$$\mathsf{E}(Y(x_0)) = \mu_0 + 0, \quad \mathsf{V}(Y(x_0)) = \mathsf{V}(\mu_0^*) + \mathsf{V}(\varepsilon_0)$$

Prediktionsintervallet blir

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

Observera att det bara är **ettan i kvadratroten** som skiljer mellan prediktionsintervallet och konfidensintervall I_{μ_0}.

Konfidens- och prediktionsintervall



Kom ihåg...

Gauss approximationsformler i en variabel

$Y = g(X)$. Taylorutveckla funktionen g kring $\mu = E(X)$

$$g(X) \approx g(\mu) + (X - \mu)g'(\mu) \quad \Longrightarrow$$

- $E(Y) \approx g(E(X))$
- $V(Y) \approx (g'[E(X)])^2 \cdot V(X)$

Kalibreringsintervall

Om man observerat ett värde y_0 av y, vad var då x_0?????
Man löser ut x_0 ur $y_0 = y(x_0^*) = \alpha^* + \beta^* \cdot x_0^*$ och får

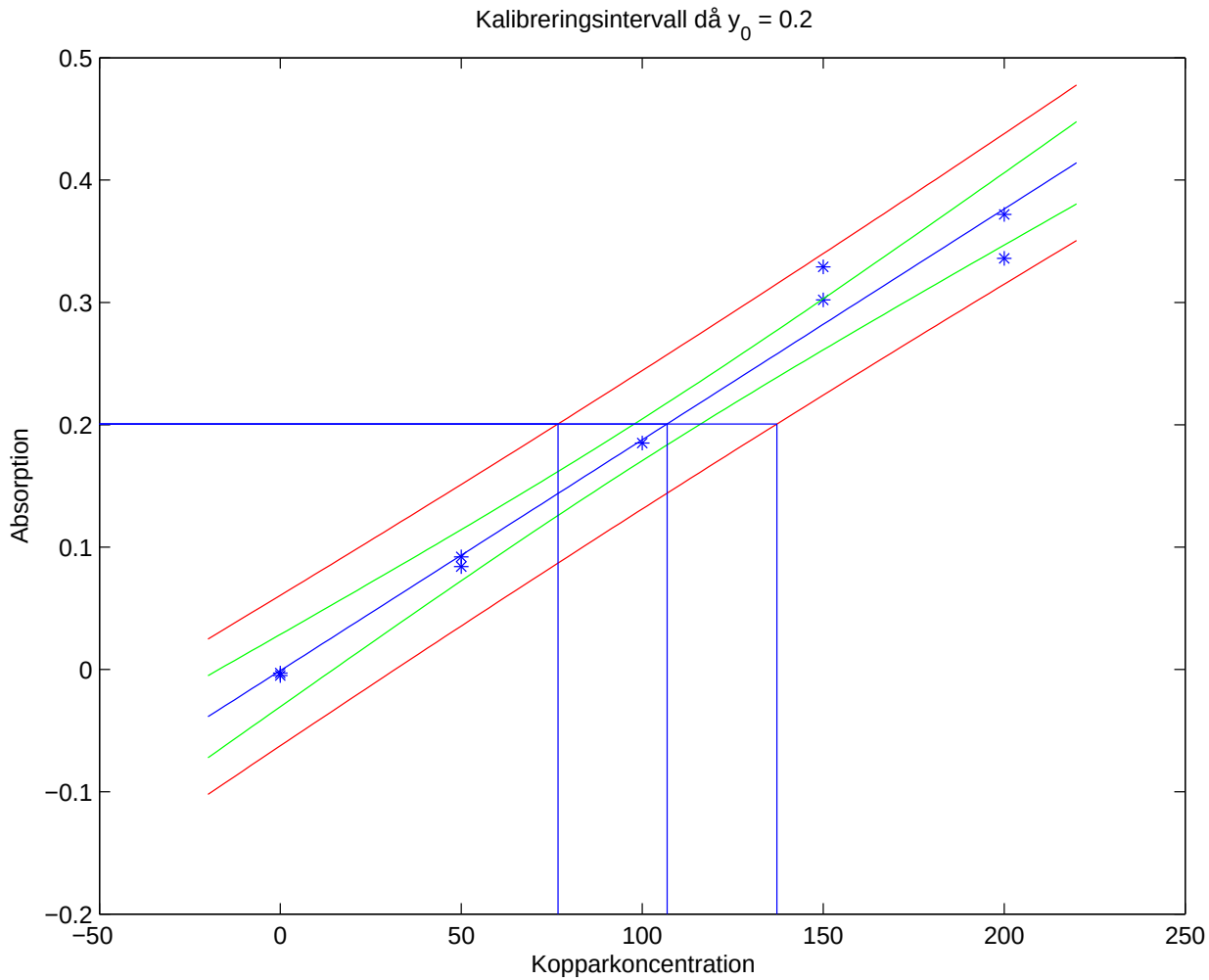
$$x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$$

Denna skattning är tyvärr **inte normalfördelad**, men vi kan t.ex. använda Gauss approximationsformler för att få fram ett approximativt värde på $\mathsf{D}(x_0^*)$.

Kalibreringsintervallet blir

$$I_{x_0} = x_0^* \pm t_{a/2}(n-2) \cdot \frac{s}{|\beta^*|}\sqrt{1 + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{(\beta^*)^2 S_{xx}}}.$$

Kalibreringsintervall



Modellvalidering

I modellen antar vi att variationen kring linjen är

$$\varepsilon_k \in \mathsf{N}(0, \sigma), \qquad \text{där fel } \varepsilon_k \text{ är oberoende av varandra}$$

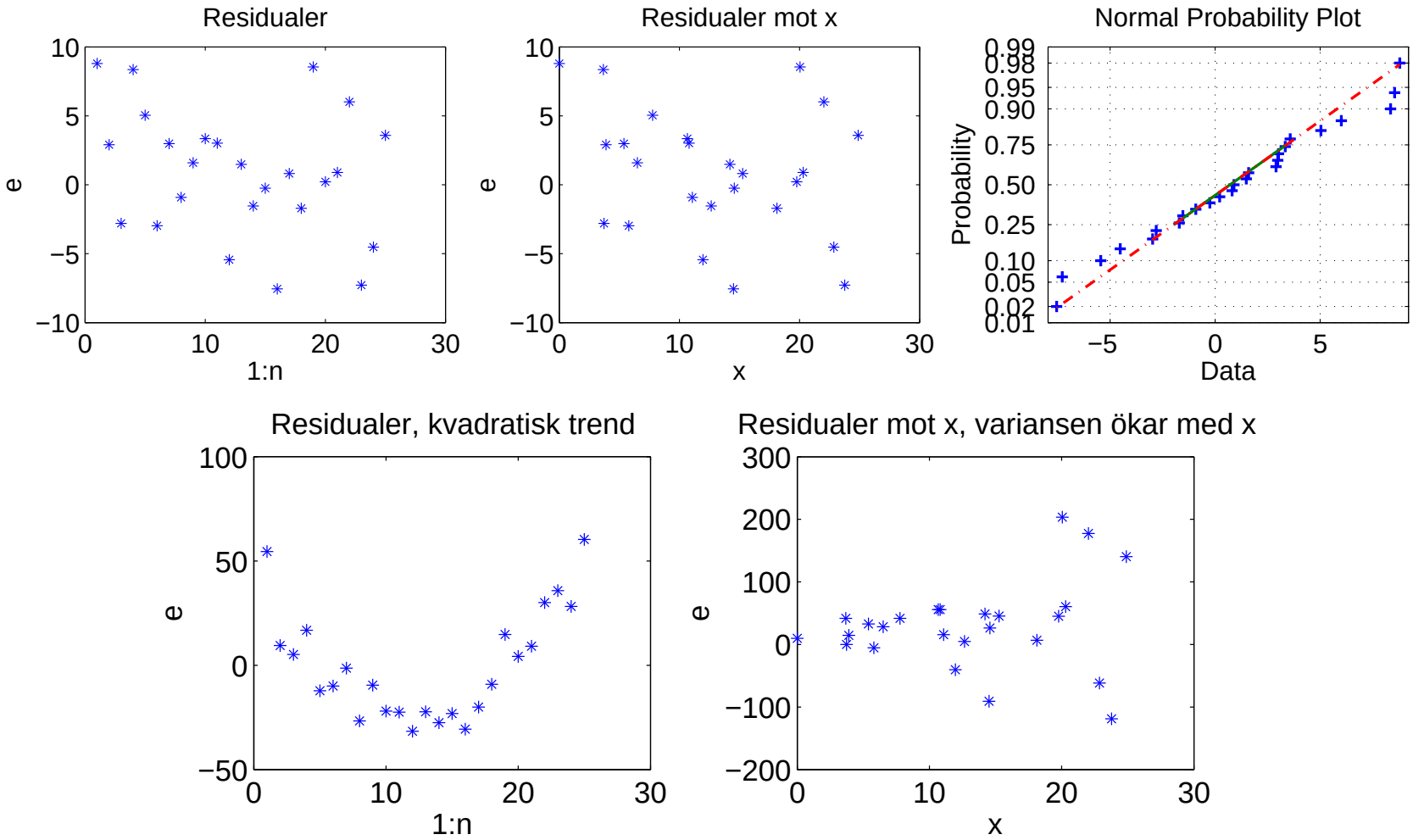
Eftersom skattningarnas fördelning och konfidensintervall osv. baseras på normal-antagandet är det **viktigt att undersöka** om antagandet är rimligt.
Vi kan studera **residualerna**, dvs avvikelserna mellan observerade y-värden och den skattade linjen $\mu^*(x) = \alpha^* + \beta^* x$.

$$e_k = y_k - \mu^*(x_k) = y_k - (\alpha^* + \beta^* \cdot x_k), \quad k = 1, \ldots, n$$

Dessa är observationer av ε_k, och residualerna bör alltså:

- se ut att komma från en och samma **normalfördelning**
- vara **oberoende** av varandra
- vara **oberoende** av alla x_k.

Mindre bra residualplottar



Residualerna bör:

- se ut att komma från en och samma **normalfördelning**

Hypotesprövning

Man vill testa

$$\begin{aligned} H_0: &\quad \beta = \beta_0 \qquad \text{mot} \\ H_1: &\quad \beta \neq 0. \end{aligned}$$

I en modellvalidering bör man ha $\beta_0 = 0$, dvs. man vill veta om x verkligen påverkar y eller inte (åtminstone på ett linjärt sätt...).

Det görs t.ex. genom att förkasta H_0 med felrisken[1] a om

- punkten β_0 ej täcks av I_β, eller
- $|T| > t_{a/2}(n-2)$ där $T = \dfrac{\beta^* - \beta_0}{d(\beta^*)} = \dfrac{\beta^*}{d(\beta^*)}$.

[1] α upptagen

Residualplottar för Cu-kalibreringen