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

Föreläsning 9

Joakim Lübeck

(Johan Lindström)

25 september 2017

Konfidensintervall (Kap. 9.1.2 & 9.2.2 & 9.3.2)

Ett **konfidensintervall** för en parameter ϑ täcker rätt värde på ϑ med sannolikheten $1-\alpha$.

$1-\alpha$ kallas **konfidensgrad**. Vanliga värden är 0.95, 0.99 och 0.999.

Normalfördelad skattning, $\vartheta^* \in \mathsf{N}(\vartheta, \mathsf{V}(\vartheta^*))$

$\mathsf{D}(\vartheta^*)$ **känd:** $I_\vartheta = \vartheta^* \pm \lambda_{\alpha/2}\mathsf{D}(\vartheta^*)$
$\mathsf{D}(\vartheta^*)$ **okänd:** $I_\vartheta = \vartheta^* \pm t_{\alpha/2}(f)d(\vartheta^*)$

Normalapproximation, $\vartheta^* \underset{\sim}{\in} \mathsf{N}(\vartheta, \mathsf{V}(\vartheta^*))$ (Ex: CGS)

$\mathsf{D}(\vartheta^*)$ **känd:** $I_\vartheta = \vartheta^* \pm \lambda_{\alpha/2}\mathsf{D}(\vartheta^*)$
$\mathsf{D}(\vartheta^*)$ **okänd:** $I_\vartheta = \vartheta^* \pm \lambda_{\alpha/2}d(\vartheta^*)$ (alltid λ-kvantil)

Konfidensintervall för σ^2 i $\mathsf{N}(\mu, \sigma^2)$ (Kap. 8.1.2)

$x_1, \dots, x_n$ observationer av $X_i \in \mathsf{N}(\mu, \sigma^2)$
Ett $1-\alpha$ konfidensintervall för σ^2 ges av

$$I_\sigma^2 = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \right)$$

Hypotestest – Vilken metod?

- Normalfördelad skattning.
  - σ känd: Vilken som helst.
  - σ okänd: Direktmetoden kräver t-fördelningens fördelningsfunktion.
- Fördelning där $\mu^* = \bar{X} \underset{\sim}{\in} \mathsf{N}(\mu, \mathsf{V}(\mu^*))$ …enl. CGS.
  - Vilken som helst
- Bin, Po, …där $\mathsf{D}(\vartheta^*)$ innehåller ϑ.
  - Direktmetoden **Går alltid** att använda, ibland med normalapproximation.
  - Testkvantitet Kräver normalt **normalapproximation**.

Vid **styrkefunktion** är det naturligt att **utgå från testkvantitet**.

Testkvantiter

Antag att vi vill testa $H_0: \vartheta = \vartheta_0$.

Model		Skattning	$T(X)$	$D(\vartheta^*)/d(\vartheta^*)$	kvantil
$X_i \in \mathrm{N}(\mu, \sigma^2)$	σ känd	$\mu^* = \bar{X}$	$\frac{\mu^* - \mu_0}{\mathrm{D}(\mu^*)}$	$\frac{\sigma}{\sqrt{n}}$	λ
	σ okänd		$\frac{\mu^* - \mu_0}{d(\mu^*)}$	$\frac{s}{\sqrt{n}}$	$t(f)$
$X \in \mathrm{Bin}(n, p)$		$p^* = \frac{X}{n}$	$\frac{p^* - p_0}{\mathrm{D}_0(p^*)}$	$\sqrt{\frac{p_0(1-p_0)}{n}}$	λ
$X_i \in \mathrm{Po}(\mu)$		$\mu^* = \bar{X}$	$\frac{\mu^* - \mu_0}{\mathrm{D}_0(\mu^*)}$	$\sqrt{\frac{\mu_0}{n}}$	λ

Notera:

1. Standardavvikelse/medelfel **räknas under H_0**.
2. Bin och Po fallet kräver **normalapproximation**.
3. α-kvantil om **ensidigt**, $\alpha/2$-kvantil om **tvåsidigt**.

Statistikteori – översikt

Punktskattning
: Hur gör man en bra gissning av en okänd storhet? Hur vet man att den är bra?

Intervallskattning
: Hitta istället ett intervall som täcker den okända storheten med en given (stor) sannolikhet.

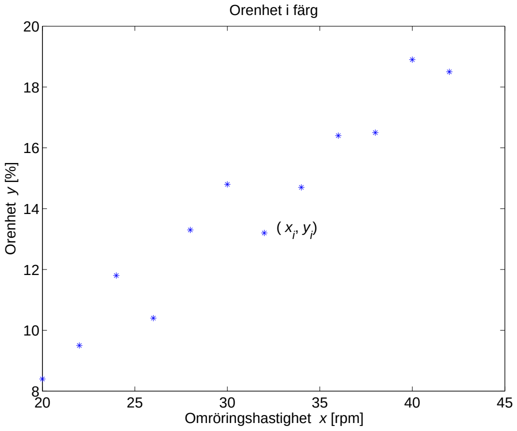
Hypotestest
: Om gissningen blev 0.013, kan rätt värde på den okända storheten ändå vara 0.01?

Regression
: Hur vet vi om två variabler påverkar varandra?

Försöksplanering & Faktorförsök
: Hur konstruerar man studier som på bäst sätt (minst antal mätningar) undersöker effekten av olika faktorer (behandlingar)?

Enkel linjär regression (Kap. 10)

Det verkar finnas ett linjärt samband mellan x och y.



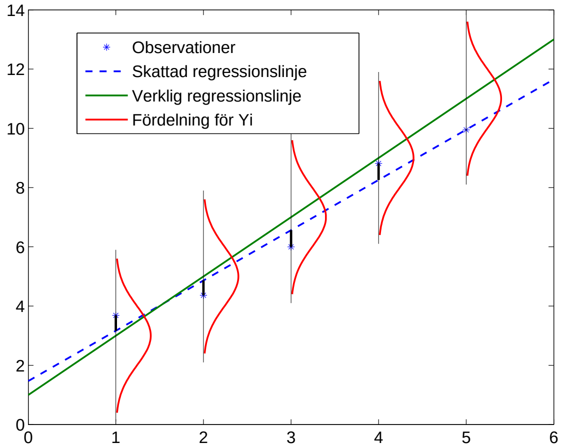
Linjär regression

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 $\varepsilon_i \in \mathrm{N}\left(0, \sigma^2\right)$.



Parameterskattningarna (Kap. 10.4–10.5)

Skattningarna av α^*, β^*

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

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

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.

Räkna ut kvadratsummorna

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

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

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

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

Exempel, x-Cu-konc. y-absorption

Vi har följande 10 par av mätvärden

i:	1	2	3	4	5	6	7	8	9	10
x_i:	0	0	50	50	100	100	150	150	200	200
y_i:	−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

$$n = 10, \qquad \sum x_i = 1000, \qquad \sum y_i = 1.8770,$$

$$\sum x_i y_i = 282.05, \qquad \sum x_i^2 = 150000, \qquad \sum y_i^2 = 0.5347$$

Konfidens- och prediktionsintervall

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

Konfidensintervall (Kap. 10.6)

Konfidensintervall för linjen, μ_0, vid x_0:

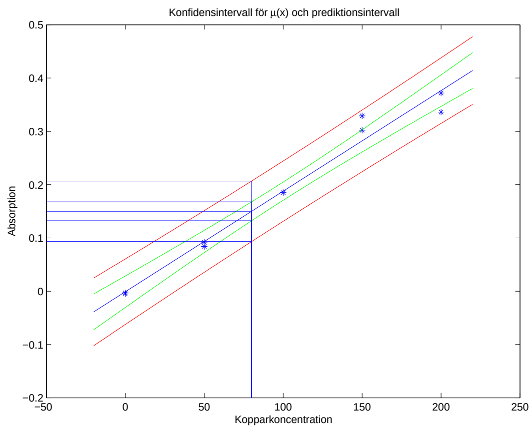
$$I_{\mu_0} = \alpha^* + \beta^* x_0 \pm t_{a/2}(n-2) \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Prediktionsintervall (Kap. 10.7)

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

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

Konfidens- och prediktionsintervall



Kalibreringsintervall

Om man observerat ett värde y_0 av y, vad blir då x_0? Man kan lösa ut x_0 ur $y_0 = \alpha^* + \beta^* x_0$ och får

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

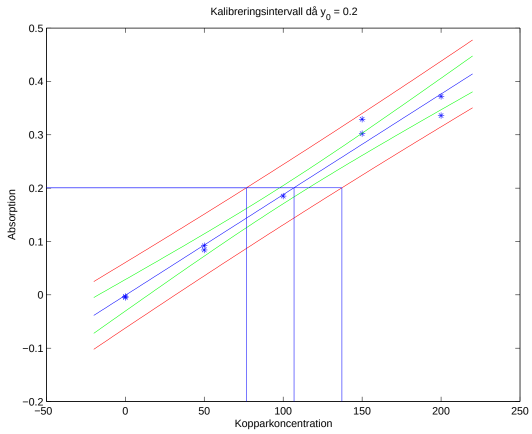
Denna skattning är **inte normalfördelad**, men **Delta metoden** ger ett approximativt värde på $D(x_0^*)$.

Kalibreringsintervall (Kap. 10.8)

Kalibreringsintervall för x_0 givet en mätning y_0, vid x_0:

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

Kalibreringsintervall



Residualanalys/Modellvalidering (Kap. 10.10)

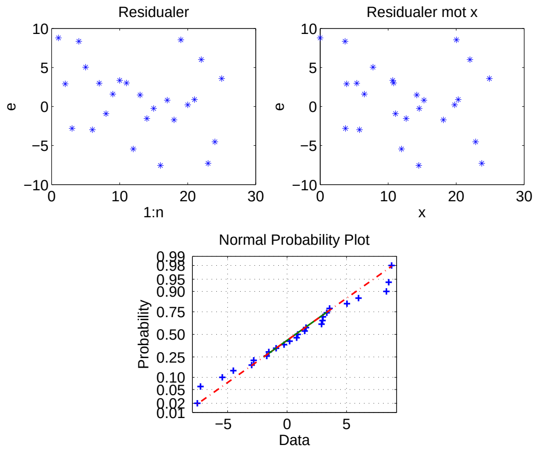
För att undersöka hur bra modellen stämmer kan vi kan studera **residualerna**, dvs avvikelserna mellan observerade y-värden och den skattade linjen.

$$e_i = y_i - \alpha^* - \beta^* x_i, \quad i = 1, \dots, n$$

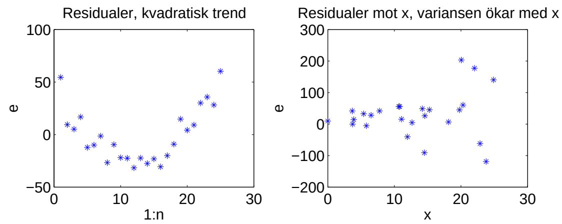
Dessa är observationer av ε_i, och residualerna bör alltså:

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

Residualplottar



Mindre bra residualplottar



I en modellvalidering bör man även testa

$$H_0: \quad \beta = 0 \qquad\qquad H_1: \quad \beta \neq 0$$