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

Föreläsning 11 & 12

Johan Lindström

2 & 9 oktober 2017

Repetition
- Multipel linjär regression
- Skattningar
- Konfidensintervall

Försöksplanering
- Tolkning av hypotesprövning
- Statistiska undersökningar
- Observationsstudier
- Kontrollerade experiment

Faktorförsök
- 2 2-försök
- Exempel
- Modell för 2 2-försök
- Skattningar
- Exempel

2 3-försök
- Modell
- Exempel

Multipel linjär regression (Kap. 11.2)

Modellen

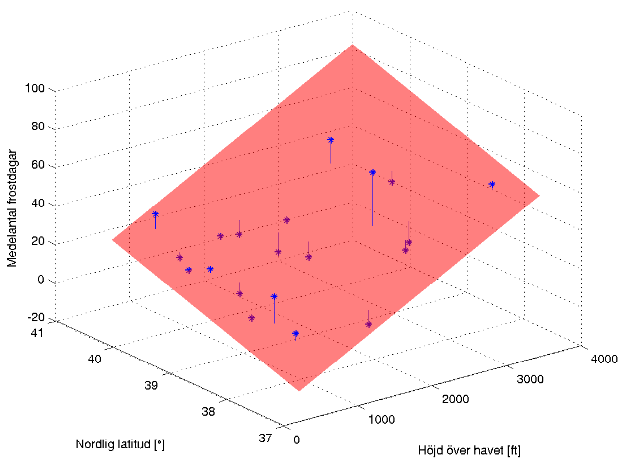
$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \in \mathsf{N}\left(0, \sigma^2\right) \text{ oberoende}$$

kan skrivas på matrisform som

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

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

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{21} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{E} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$



Skattning av β och σ^2 (Kap. 11.3–11.4)

MK-skattningar av $\beta_0, \ldots, \beta_p$ (elementen i $\boldsymbol{\beta}$) blir

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

och skattning av σ^2 är

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

där **residualkvadratsumman** ges av

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

Konfidensintervall för β_i och $\mu^*(\boldsymbol{x}_0)$ (Kap. 11.4–11.5)

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

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

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

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

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

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)?

Hypotestest — Tolkning

Årsmedelvärdet av kvävedioxid i utomhusluft bör inte överstiga $40\,\mu g/m^3$ (Miljökvalitetsnorm, 2006–01–01).

I en större svensk stad vill man undersöka om gränsvärdet överskrids och ett antal mätningar av NO_2 görs.

Gatukontoret (som måste vidta åtgärder om normen överskrids) anser att man bör testa

$$H_0 : \mu = 40\,\mu g/m^3 \qquad \text{mot} \qquad H_1 : \mu > 40\,\mu g/m^3$$

Miljöförvaltningen anser däremot att ett lämpligt test är

$$H_0 : \mu = 40\,\mu g/m^3 \qquad \text{mot} \qquad H_1 : \mu < 40\,\mu g/m^3$$

Statistiska undersökningar (Kap. 12.1)

Vi skiljer på två typer av statistiska undersökningar.

Deskriptiv undersökning

Syftar till att beskriva **egenskaper** hos en population. Exempel:

- Skattningar av medelvärde, varians
- Konfidensintervall för medelvärde, sannolikheter, etc.

Analytisk undersökning

Syftar till att undersöka **effekter** av olika **förklarande variabler** eller **faktorer** på en population. Exempel:

- Stickprov i par
- Regression

Problem med observationsstudier

1. Vid en undersökning av sjösäkerhet upptäcks ett positivt samband mellan försäljning av rosevin och antalet drunkningstillbud.
2. Man finner ett (positivt) samband mellan familjers utbildningsnivå och risken för Downs syndrom hos barnen.
3. Suppose you're trying to help the military decide how best to armor their planes for future bombing runs. They let you look over the planes that made it back, and you note that some areas get shot heavily, while other areas hardly get shot at all. So, you decide to increase the armor on the areas that get shot.

Kontrollerat experiment (Kap. 12.1.5)

Det finns flera olika metoder för att uppnå ett så bra experiment som möjligt:

Randomiseras för att förhindra **systematiska fel**

Homogen population Mindre varians för det lättare att upptäcka effekter, men kan **hindra generella slutsatser**.

Blockindela Dela upp experimentet i grupper och randomisera inom grupperna.

Efterjustering Tar hänsyn till **kovariater**.

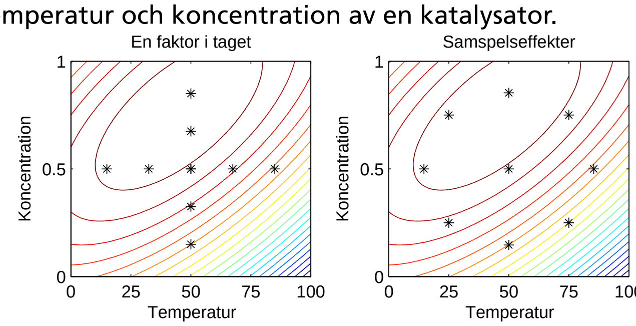
Replikat Minskar osäkerheten.

Flerfaktorförsök För att upptäcka **samspelseffekter**.

Faktorförsök (Kap. 12.2–12.4)

Undersöker hur en **responsvariabel** påverkas av olika **faktorer** när de varieras på **olika nivåer**.

Ex: Effekten av en kemiskreaktion som funktion av temperatur och koncentration av en katalysator.



2^k-försök (Kap. 12.2)

I ett 2^k-försök har man k **faktorer** som alla kan varieras på 2 **nivåer**.

Det enklaste fallet är ett

2^2-försök (Kap. 12.3)

I ett 2^2-försök har man 2 **faktorer** som alla kan varieras på 2 **nivåer**.

Exempel 2^2-försök:

Koncentration	Hög	μ_{12}	μ_{22}
	Låg	μ_{11}	μ_{21}
		Låg	Hög
		Temperatur	

2^2-försök

B	Hög	μ_{12}	μ_{22}
	Låg	μ_{11}	μ_{21}
		Låg	Hög
		A	

Enkel effekt: Effekten av **en faktor** om den andra faktorn är **fix**.

Huvudeffekt: Effekten av **en faktor** för **alla värden** på den andra faktorn.

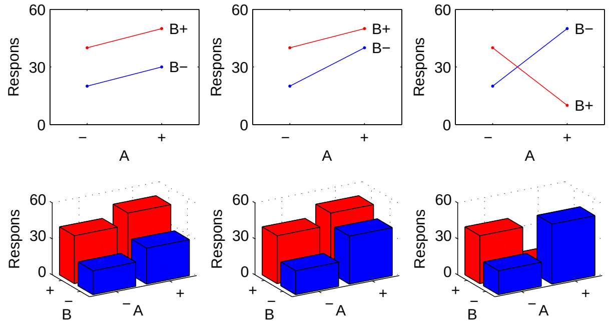
Samspelseffekten: Skillnaden mellan de **enkla effekterna**.

Teckenschema för 2^2-försök

	Försök	Respons	μ	A	B	AB
A och B låg	(1)	μ_{11}	+	-	-	+
A hög	a	μ_{21}	+	+	-	-
B hög	b	μ_{12}	+	-	+	-
A och B hög	ab	μ_{22}	+	+	+	+

Teckenschema för 2^3-försök i formelsamlingen. För 2^2-försök använd endast de rader och kolumer med (1), a, b, och ab.

2^2-försök – Effekter



Om responsändringen för en faktor **inte** beror på nivån av andra faktorn är faktorerna **additiva**, och det finns **inget samspel**.

Exempel: 2^2-försök

Man vill undersöka hur olika typer av konstgödsel påverkar avkastningen från en vete-odling. Två olika gödsel, ett kväve och ett fosfor baserat, testas och avkastningen (ton/ha) mäts.

Kväve (N)	Ja	4.5	6
	Nej	4	5
		Nej	Ja
		Fosfor (P)	

Bestäm

1. De enkla effekterna av fosfor
2. Huvudeffekterna
3. Samspelseffekterna

Modell för 2^2-försök (Kap. 12.3.1)

Vid n mätningar (**replikat**) av varje faktorkombination ges varje observation av

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1,2;\ j = 1,2;\ k = 1,\ldots,n$$

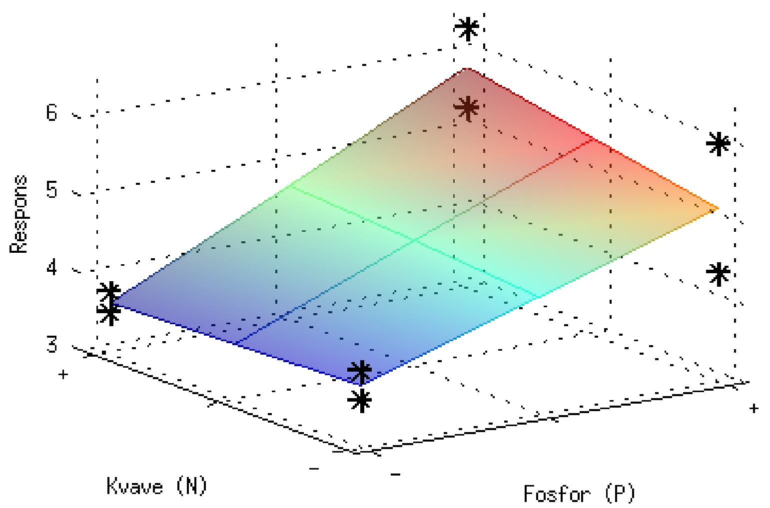
och felen antas vara **oberoende** $\varepsilon_{ijk} \in \mathsf{N}\left(0, \sigma^2\right)$.

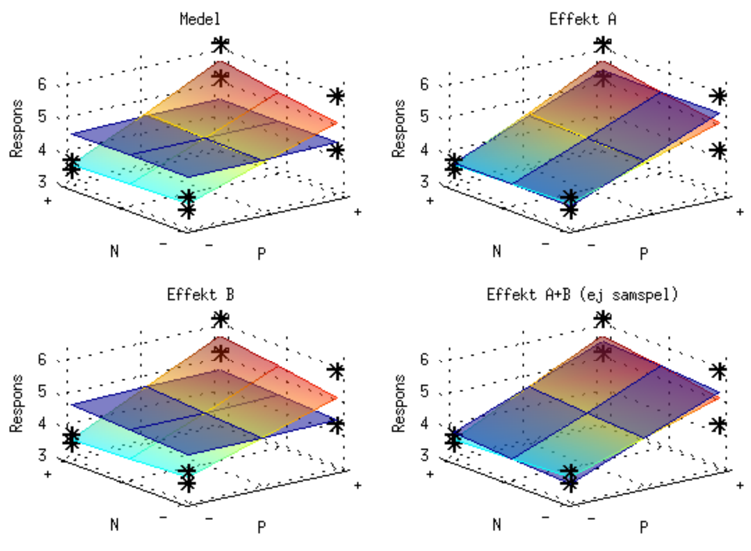
Exempel: 2^2-försök

I odlings exemplet ovan görs två replikat för varje faktor

		Fosfor (P): Nej	Fosfor (P): Ja
Kväve (N)	Ja	$\begin{pmatrix}3.50\\3.76\end{pmatrix}$	$\begin{pmatrix}5.27\\6.30\end{pmatrix}$
	Nej	$\begin{pmatrix}3.59\\3.97\end{pmatrix}$	$\begin{pmatrix}4.37\\6.03\end{pmatrix}$

1. Skatta huvud- och samspelseffekter.
2. Skatta variansen.
3. Gör 95%-konfidensintervall för huvud- och samspelseffekter.
4. Gör 95%-konfidensintervall för variansen σ^2.





Skattningar för 2^2-försök (Kap. 12.3.2)

Givet observationer y_{ijk}

$$\mu^*_{ij} = \bar{y}_{ij\cdot} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk},$$

$$\widehat{A} = \frac{-\bar{y}_{11\cdot} + \bar{y}_{21\cdot} - \bar{y}_{12\cdot} + \bar{y}_{22\cdot}}{2^2},$$

$$s^2_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2,$$

$$s^2 = \frac{s^2_{11} + s^2_{21} + s^2_{12} + s^2_{22}}{2^2}.$$

Övriga effekter ($\widehat{\mu}$, $\widehat{B}$ och $\widehat{AB}$) fås från skattningarna μ^*_{ij}

Konfidensintervall för 2^2-försök

Givet n replikat blir konfidensintervallen för effekterna

$$I_\bullet = \widehat{\bullet} \pm t_{\alpha/2}(2^2 \cdot (n-1))\frac{s}{\sqrt{2^2 \cdot n}}$$

Där $\bullet$ är någon av A, B eller AB.

Eftersom variansen för (t.ex.) skattningen $\widehat{A}$ ges av

$$\mathrm{V}\left(\widehat{A}\right) = \mathrm{V}\left(\frac{-\bar{y}_{11\cdot} + \bar{y}_{21\cdot} - \bar{y}_{12\cdot} + \bar{y}_{22\cdot}}{4}\right)$$

$$= \frac{1}{4^2}\Big(\mathrm{V}(\bar{y}_{11\cdot}) + \mathrm{V}(\bar{y}_{21\cdot}) + \mathrm{V}(\bar{y}_{12\cdot}) + \mathrm{V}(\bar{y}_{22\cdot})\Big) = \frac{\sigma^2}{4n} = \frac{\sigma^2}{2^2 \cdot n}$$

Upgf. 1, 2004-10-19

I ett 2^2-faktorförsök vill man studera hur utbytet påverkas av tryck och temperatur. Man har gjort **tre replikat**. Försöksresultatet blev

Tryck	Temp. Låg	Temp. Hög
Högt	$\mu^*_{12} = 45.8,\ s^2_{12} = 1.235$	$\mu^*_{22} = 49.2,\ s^2_{22} = 0.103$
Lågt	$\mu^*_{11} = 48.2,\ s^2_{11} = 0.335$	$\mu^*_{21} = 48.8,\ s^2_{21} = 1.431$

1. Verkar det finnas något samspel? Motivera med hjälp av tabellen.
2. Skatta huvud- och samspelseffekter och avgör vilka som är signifikanta på 5% nivå.
3. Gör ett 95% konfidensintervall för variansen.

2^3-försök

I ett 2^3-försök har man en **tredje faktor** C som kan varieras mellan **två nivåer** (låg/hög).

Modell för 2^3-försök (Kap. 12.4)

Vid n mätningar (**replikat**) av varje faktorkombination ges varje observation av

$$y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}, \quad i = 1, 2;\ j = 1, 2;\ k = 1, 2;\ l = 1, \ldots, n$$

och felen antas vara **oberoende** $\varepsilon_{ijkl} \in \mathrm{N}\left(0, \sigma^2\right)$.

Respons för varje försök kan delas upp i huvud- och samspelseffekter.

$$\mu_{ijk} = \mu \pm A \pm B \pm C (\pm)(\pm) AB (\pm)(\pm) AC (\pm)(\pm) BC (\pm)(\pm)(\pm) ABC$$

Exempel

Vid ett försök studerades hur utbytet påverkades av

- A Temperatur på nivåerna 160, 180 ($^\circ C$)
- B Katalysatorkoncentration 20, 40 (%)
- C Typ av katalysator I, II

	Faktorer			Obs.		Skattningar	
Försök	A	B	C	y_{ijk1}	y_{ijk2}	$\bar{y}_{ijk\cdot}$	s^2_{ijk}
(1)	160	20	I	59	61	60	2
a	180	20	I	74	70	72	8
b	160	40	I	53	55	54	2
ab	180	40	I	69	67	68	2
c	160	20	II	50	54	52	8
ac	180	20	II	82	84	83	2
bc	160	40	II	46	44	45	2
abc	180	40	II	79	81	80	2

Vad gör man sen? – Mer statistik!

För B, K & N:

KLGN10 Kemometri (ges av Livsmedelsteknologi)

FMSF65 Försöksplanering LP4

FMSN30 Linjär och logistisk regression LP4

För BME:

FMSF10 Stationära stokastiska processer LP1

FMSF15 Markovprocesser LP1

FMSN45 Tidsserieanalys LP2

FMSN30 Linjär och logistisk regression LP4