LMA201/LMA521: Faktorförsök

Föreläsning 1

Anders Hildeman

Innehåll

- Försöksplanering
- Faktorförsök med två nivåer
- Skattning av effekterna.
- Diagram för huvudeffekter
- Diagram för samspelseffekter
- Paretodiagram

- Den här veckan kommer tillägnas faktorförsök.
- En annan bok: “Försöksplanering - Faktorförsök” av Ulla Dahlbom. Den lilla lila boken.

Statistik kan användas för att undersöka hur olika faktorer (x) påverkar någon given storhet (y).

Om faktorerna är det enda som påverkar y och man kan mäta dessa exakt så är det inte någon slump inblandad. Problemet är då enbart av matematisk natur och statistik behövs inte.
I verkligheten är dock detta antagande sällan sant. Mätningar är inexakta och det finns oftast ytterligare faktorer som man inte känner till eller kan mäta. Utan kontroll över mätfelen och de "gömda" faktorerna så beter sig den uppmätta storheten slumpmässigt.

Eftersom slump är inblandat så modelleras storheten som en slumpvariabel (Y). Man kan då observera slumpvariabel betingat på olika värden hos de kända faktorerna, $(Y|\mathbf{x})$.

Om man bara har diskreta värden hos faktorerna (nivåer) så kan man dela upp observerad data i grupper. En grupp för varje unik uppsättning av nivåer hos faktorerna.

Exempel

- $Y \in (0, \infty)$: Tiden det tar för 4:ans spårvagn att åka sin rutt.
- x_1: Tiden på dagen ({ 'morgon', 'dag', 'kväll' })
- x_2: Väderlek ({ 'sol', 'regn' })

Antal grupper: $3 \cdot 2 = 6$.

I exemplet med spårvagnen så kan man fråga sig om sannolikhetsfördelningen för dessa 6 olika grupper är olika och i så fall på vilket sätt?
Ofta nöjer man sig med att ta reda på om väntevärdena är olika och vilken grupp som ger det bästa eller värsta värdet (den kortaste eller längsta tiden i exemplet med spårvagnen).

Exempel

Är $\mathbb{E}\left[Y|x_1 = morgon, x_2 = sol\right] < \mathbb{E}\left[Y|x_1 = dag, x_2 = regn\right]$?

Planerade försök

För att få tillgång till observationer av storheten man är intresserad av så kan man göra på två sätt.

- Planerat försök: Man kontrollerar själv nivåerna på faktorerna för varje mätning av storheten.
- Observationsstudie: Man observerar storheten och antecknar vilka nivåer faktorerna hade under observationen.

Kan man välja så är data från ett planerat försök att föredra eftersom man kan balansera antalet observationer i varje grupp samt hindra att någon okänd faktor, som också påverkar Y, systematiskt stör slutsatsen. Nackdelen är att ett planerat försök kräver möjlighet att kontrollera nivåerna samt oftast innebär en större arbetsinsats/kostnad.

Faktorförsök

Den här veckan kommer vi lära oss planera faktorförsök med två nivåer. D.v.s. hur vi skall utföra ett planerat försök då faktorerna bara kan anta två möjliga nivåer.
Då vi endast tänker oss två nivåer per faktor så kan man koda dessa som 'låg nivå'(-1) och 'hög nivå' ($+1$). Man tänker sig sedan att väntevärdet, $\mu|\mathbf{x}$, för $Y|\mathbf{x}$ kan modelleras som funktion av faktornivåerna med parametrar $\boldsymbol{\beta} = [\beta_a, \beta_b, \beta_{ab}, \beta_c, ...]$.

Exempel: Två faktorer

$$\mathbb{E}[Y|\mathbf{x}] = \mu|\mathbf{x} = \beta_0 + \beta_a x_a + \beta_b x_b + \beta_{ab} x_a x_b$$

Termen β_a motsvarar hur Y påverkas då faktor A byter nivå om man medelvärdesbildar mellan båda nivåerna på faktor B. Alltså om man tittar på slumpvariabeln $\frac{Y|(x_a, x_b=-1)+Y|(x_a, x_b=+1)}{2}$.

På samma sätt så är β_b påverkan av faktor B om man medelvärdesbildar bort faktor A.

Termen β_{ab} är en samspelseffekt. β_{ab} förklarar hur de två faktorerna tillsammans förstärker eller motverkar den påverkan som redan givits av β_a- och β_b-effekterna.

Det kan kanske vara lättare att förstå vad β-termerna egentligen symboliserar ifall man beskriver dem på omvänt sätt istället. Låt $\mu_{(k,l)} = \mu|(x_a = k, x_b = l)$.

$$\beta_0 = \frac{\mu_{(-1,-1)} + \mu_{(+1,-1)} + \mu_{(-1,+1)} + \mu_{(+1,+1)}}{4}$$
$$\beta_a = \frac{\mu_{(+1,-1)} + \mu_{(+1,+1)}}{2} - \frac{\mu_{(-1,-1)} + \mu_{(-1,+1)}}{2}$$
$$\beta_b = \frac{\mu_{(-1,+1)} + \mu_{(+1,+1)}}{2} - \frac{\mu_{(-1,-1)} + \mu_{(+1,-1)}}{2}$$
$$\beta_{ab} = \frac{\mu_{(+1,+1)} + \mu_{(-1,-1)}}{2} - \frac{\mu_{(-1,+1)} + \mu_{(+1,-1)}}{2}$$

Målet med ett planerat försök är ofta att upptäcka bästa eller värsta kombinationen av faktorer för ett önskat resultat. Det kan t.ex. vara hur man skall ställa in produktionsprocessen i en fabrik, hur man skall äta och träna för att göra ett bra resultat i en tävling eller vad som är de mest optimala parametrarna i ett inbyggt elektriskt system.

Om $\beta_a > 0$ så innebär det t.ex. att om man låter faktor A vara hög så kommer väntevärdet av Y i genomsnitt vara större än om faktor A är låg. På samma sätt för faktor B. Det kan dock vara så att sampselseffekten gör det mer effektivt att låta faktor B vara låg om det är så att samspelseffekten är negativ och $\beta_a + \beta_b + \beta_{ab} < \beta_a - \beta_b - \beta_{ab}$.

Mer än två faktorer?

Fungerar på exakt samma sätt. Man behöver nu en parameter för varje faktor (K st om vi antar K faktorer) samt en parameter för varje samspel. Samspel kan vara allt från två-faktorsamspel till K-faktorsamspel. Alltså finns det $\sum_{k=2}^{K} \binom{K}{k}$ möjliga samspel.

$$
\begin{aligned}
\mu|\boldsymbol{x} =& \beta_0 + \beta_a x_a + \beta_b x_b + \beta_c x_c + \dots \\
&+ \beta_{ab} x_a x_b + \beta_{ac} x_a x_c + \beta_{bc} x_b x_c + \dots \\
&+ \beta_{abc} x_a x_b x_c + \beta_{bcd} x_b x_c x_d + \dots
\end{aligned}
$$

När man utför de planerade försöken kommer man nu få ett antal olika mätningar. Antag att vi gör minst en mätning för varje grupp av samlade nivåer. Vi har då redan kunskapen för att göra punktskattningar av β-parametrarna eftersom vi vet hur man gör punktskattningar av väntevärden. T.ex. eftersom

$$\beta_a = \frac{\mathbb{E}\left[Y|(+1,+1)\right] + \mathbb{E}\left[Y|(+1,-1)\right]}{2} - \frac{\mathbb{E}\left[Y|(-1,+1)\right] + \mathbb{E}\left[Y|(-1,-1)\right]}{2}$$

så är vår punktskattning av β_a givet datan (låt oss kalla den l_a)

$$l_a = l_{a^+} - l_{a^-} = \frac{\hat{\mu}_{(+1,+1)} + \hat{\mu}_{(+1,-1)}}{2} - \frac{\hat{\mu}_{(-1,+1)} + \hat{\mu}_{(-1,-1)}}{2},$$

där $\hat{\mu}_{(+1,+1)} = \frac{1}{n_{(+1,+1)}} \sum_{i=1}^{n_{(+1,+1)}} y_{i,(+1,+1)}$, $n_{(+1,+1)}$ är antalet mätningar i grupp $(+1,+1)$ och $y_{i,(+1,+1)}$ är den i:te mätningen i grupp $(+1,+1)$.

På samma sätt så kan vi punktskatta alla andra β-parametrar.

Försök nr	A	B	AB	Resultat
1	-	-	+	$\bar{y}_1 = 4$
2	+	-	-	$\bar{y}_2 = 2$
3	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	$\bar{y}_4 = 4$

Se på tabellen vilka y-värden som skall vara positiva och negativa för respektive effekt.

På samma sätt så kan vi punktskatta alla andra β-parametrar.

Försök nr	A	B	AB	Resultat
1	-	-	+	$\bar{y}_1 = 4$
2	+	-	-	$\bar{y}_2 = 2$
3	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	$\bar{y}_4 = 4$

Se på tabellen vilka y-värden som skall vara positiva och negativa för respektive effekt.

$$
\begin{aligned}
l_0 &= \frac{4+2+3+4}{4} = 3.25 \\
l_a &= \frac{2+4}{2} - \frac{4+3}{2} = 3 - 3.5 = -0.5 \\
l_b &= \frac{3+4}{2} - \frac{4+2}{2} = 3.5 - 3 = +0.5 \\
l_{ab} &= \frac{4+4}{2} - \frac{2+3}{2} = 4 - 2.5 = +1.5
\end{aligned}
$$

Som ni ser så kommer skattningar av effekterna att bero på flera av mätvärdena $\bar{y}_i$.
Detta stabiliserar skattningen (precis som när vi räknar ut ett stickprovsmedelvärde) fastän om man kanske bara gjort en mätning för varje grupp. Detta skulle ju inte varit fallet om man skattade $\mu|(-1,-1)$ direkt som $\bar{y}_1$.

Försök nr	A	B	AB	Resultat
1	-	-	+	$\bar{y}_1 = 4$
2	+	-	-	$\bar{y}_2 = 2$
3	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	$\bar{y}_4 = 4$

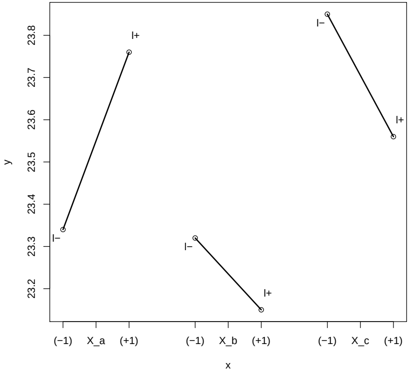
$\bar{y}_1$ är här stickprovsmedelvärdet av alla mätningar gjorda med uppsätningen av faktornivåer (-1, -1) osv.

Om det är mer än en mätning för varje uppsättning faktornivåer, kom ihåg att välja ut när mätningarna sker i experimentet slumpmässigt. Alltså, gör inte alla mätningar för (-1,-1) efter varandra utan slumpa ut ordningen.

Detta minskar risken att något systematiskt fel av misstag inkluderas i analysen.

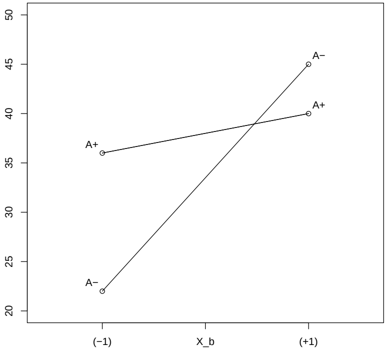
Diagram för huvudeffekter

Man kan rita ut huvudeffekterna i ett diagram för att lättare få känsla för vad effekterna är när faktorn är låg resp. hög. För faktor A ritar man då ut l_{a-}- och l_{a+}-värdena på y-axeln och motsvarande -1 och $+1$ på x-axeln. Man fortsätter därefter med faktor B, C, etc på samma sätt.



I den här figuren ser man att faktor A:s huvudeffekt har störst påverkan eftersom det är störst skillnad mellan l_+ och l_- för faktor A. Dessutom ser vi att det är en positiv påverkan då $l_+ > l_-$.
(B och C:s huvudeffekter påverkar negativt.)

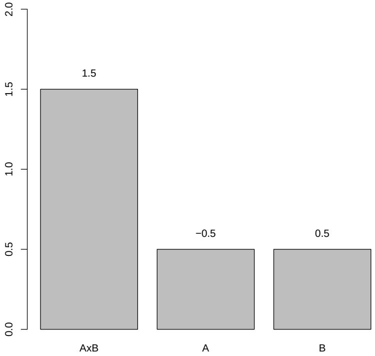
Man kan även rita ut diagram för samspelseffekterna. Är man t.ex. intresserad av samspelseffekten β_{ab} så kan man rita ett streck mellan medelvärdet inom grupperna där faktor B är låg respektive hög då faktor A är låg respektive hög.



I den här figuren ser man att det existerar ett samspel eftersom linjerna inte är parallella mot varandra. $\mu|x = \beta_0 + \beta_a x_a + \beta_b x_b + \beta_{ab} x_a x_b$

Paretodiagram

För att tydligt kunna se storleken på de uppskattade effekterna så kan man rita upp ett Paretodiagram.

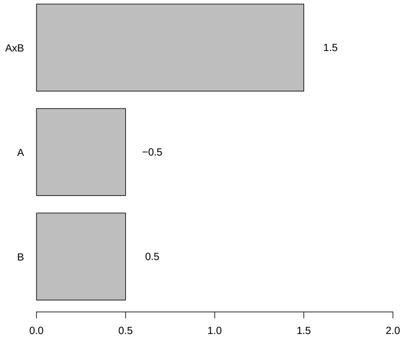


- Rangordna effekterna i storleksordning utan hänsyn till tecken.
- Rita in staplarna (störst till vänster och sedan i avtagande ordning).

Figur: Paretodiagram över de uppskattade effekterna från exemplet ovan.

Paretodiagram

I boken ritar de dem liggande.



Figur: Paretodiagram över de uppskattade effekterna från exemplet ovan.

Sammanfattning

- Vi vill utföra planerade faktorförsök för att ta reda på hur en viss storhet påverkas av förändringar i bakomliggande faktorer.
- Vi bryr oss nu bara om hur väntevärdet förändras då faktorerna skiftas mellan “höga” och “låga” nivåer.
- Om vi har K olika faktorer så kommer vi behöva göra minst 2^K olika mätningar för att uppskatta alla effekter i vår modell.
- Vid mer än en mätning för varje nivåuppsättning så välj ut ordningen på mätföljden slumpmässigt. Då undviker man att få med eventuella systematiska fel.
- Visualisera de skattade effekterna med diagram.

Exempel

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$y_1 = [3.7, 2.8]$
2	+	-	-	-	+	-	+	$y_2 = [4.8, 4.8]$
3	-	+	-	-	-	+	+	$y_3 = [18.7, 17.1]$
4	+	+	-	+	-	-	-	$y_4 = [13.5, 14.1]$
5	-	-	+	+	-	-	+	$y_5 = [10.1, 11.7]$
6	+	-	+	-	-	+	-	$y_6 = [8.8, 9.3]$
7	-	+	+	-	+	-	-	$y_7 = [17.7, 16.9]$
8	+	+	+	+	+	+	+	$y_8 = [0.4, -0.2]$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$l_0 = \frac{3.25 + 4.80 + 17.9 + 13.8 + 10.9 + 9.05 + 17.30 + 0.10}{8} = 9.64$$

$$l_a = \frac{4.80 + 13.8 + 9.05 + 0.10}{4} - \frac{3.25 + 17.9 + 10.9 + 17.30}{4} = -5.4$$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$l_b = \frac{17.9 + 13.8 + 17.3 + 0.1}{4} - \frac{3.25 + 4.8 + 10.9 + 9.05}{4} = 5.28$$

$$l_c = \frac{10.9 + 9.05 + 17.3 + 0.1}{4} - \frac{3.25 + 4.8 + 17.9 + 13.8}{4} = -0.60$$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

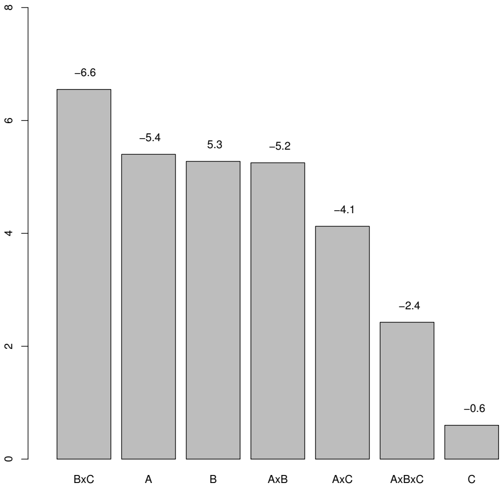
$$l_{ab} = \frac{3.25 + 13.8 + 10.9 + 0.1}{4} - \frac{4.8 + 17.9 + 9.05 + 17.3}{4} = -5.25$$

$$l_{bc} = \frac{3.25 + 4.8 + 17.3 + 0.1}{4} - \frac{17.9 + 13.8 + 10.9 + 9.05}{4} = -6.55$$

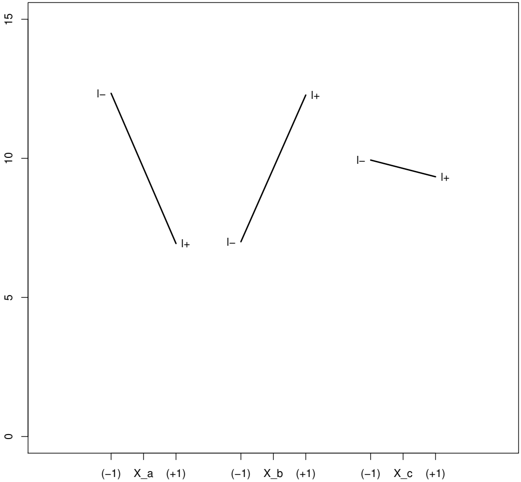
Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$l_{ac} = \frac{3.25 + 17.9 + 9.05 + 0.1}{4} - \frac{4.8 + 13.8 + 10.9 + 17.3}{4} = -4.13$$

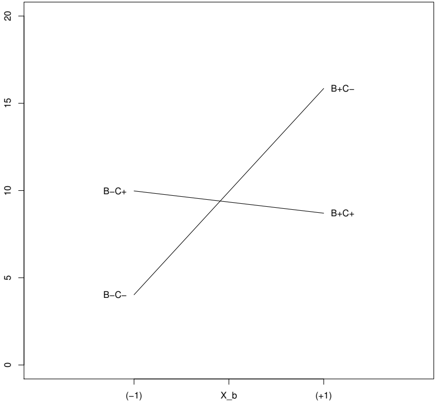
$$l_{abc} = \frac{4.8 + 17.9 + 10.9 + 0.1}{4} - \frac{3.25 + 13.8 + 9.05 + 17.3}{4} = -2.43$$



Figur: Paretodiagram.



Figur: Diagram för huvudeffekter.



(a) Samspelsgraf för BxC.

- B_+C_-: medelvärdet av alla $\bar{y}_i$ där faktor B är hög och faktor C är låg.
- Samspelseffekten är tydlig då linjerna inte är parallella. Att vinkeln är stor mellan linjerna visar att samspelseffekten är stor.