Estadística industrial

Relació entre regressió i dissenys factorials

Qualsevol disseny factorial es pot analitzar fent servir models de regressió. De fet, són uns models de regressió amb unes característiques molt especials.

Utilitzarem l'exemple de la molla per il·lustrar aquest fet.
L'objectiu era maximitzar el nombre de compressions fins al trencament d'una molla. Els factors de disseny disponibles i els nivells on es va experimentar eren :

Factors:		**Nivells**
Longitud	(L)	10 i 15 cm
Gruix	(G)	5 i 7 mm
Tipus d'acer	(T)	A i B

Els experiments realitzats (recorda que es van fer dues rèpliques) van ser els següents:

Exp.	**Factors**			**N. compr. (milers)**		**Mitjana**	**s^2**
	L	**G**	**T**				
1	-1	-1	-1	$77^{(6)}$	$81^{(13)}$	79	8
2	1	-1	-1	$98^{(12)}$	$96^{(4)}$	97	2
3	-1	1	-1	$76^{(1)}$	$74^{(16)}$	*75*	*2*
4	1	1	-1	$90^{(15)}$	$94^{(10)}$	*92*	*8*
5	-1	-1	1	$63^{(8)}$	$65^{(2)}$	*64*	*2*
6	1	-1	1	$82^{(9)}$	$86^{(14)}$	84	8
7	-1	1	1	$72^{(3)}$	$74^{(11)}$	73	2
8	1	1	1	$92^{(7)}$	$88^{(5)}$	90	8

Per l'anàlisi dels dissenys factorials a dos nivells, estimem l'efecte de cada factor sobre la resposta així:

$\text{Efecte} = \bar{y}_+ - \bar{y}_-$
$\bar{y}_+$: mitjana de les $^{N}/_{2}$ observacions amb signe $+$
$\bar{y}_-$: mitjana de les $^{N}/_{2}$ observacions amb signe $-$

i si tenim el mateix nombre de rèpliques a cada condició experimental podem estimar la variància dels efectes així:

$$V(ef) = V(\bar{y}_+ - \bar{y}_-) = \frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2} = \frac{4\sigma^2}{N}$$

on σ^2 s'estima mitjançant s_R^2

En l'exemple de la molla, el valors dels efectes era el següents:

L	18.0
G	1.5
LG	-1.0
T	-8.0
LT	0.5
GT	6.0
LGT	-0.5

I la variància dels efectes era

$$s_R^2 = \frac{8+2+2+8+2+8+2+8}{8} = 5, \text{amb } \nu = 8 \text{ graus de llibertat}$$
$$s_{ef}^2 = \frac{4 \cdot (5)}{16} = 1{,}25$$
$$s_{ef} = 1{,}12$$

Noteu que la variància és exactament igual per tots els efectes ja siguin efectes principals com interaccions (tots són diferències de mitjanes amb el mateix nombre de dades cadascuna)

Aquest anàlisi és equivalent a ajustar aquest model de regressió:

$$y = \beta_0 + \beta_1 L + \beta_2 G + \beta_3 T + \beta_{12} LG + \beta_{13} LT + \beta_{23} GT + \beta_{123} LGT + \varepsilon$$

on les variables explicatives són les columnes del disseny factorial (efectes principals) i les columnes d'interaccions surten de multiplicar les columnes corresponents de la matriu de disseny.

En particular, $b_0 = \bar{y}$
$b_i = \frac{Efecte}{2}, i \neq 0$

Això és perquè cadascuna de les b_i del model de regressió mesura quant augmenta la resposta (y) quan incrementem en una unitat la variable explicativa (x_i). L'efecte del factor x_i mesura quant augmenta la resposta quan incrementen en dues unitats (passem de –1 a +1) la variable explicativa (el factor).

La desviació tipus dels coeficients b_i del model de regressió també són iguals per tots els factors i exactament la meitat de la desviació tipus dels efectes.

Pel nostre exemple l'ajust de la regressió és el següent:

Regression Analysis: Num. compr versus L; G; T; LG; LT; GT; LGT

```
The regression equation is
Num. compr = 81,8 + 9,00 L + 0,750 G - 4,00 T - 0,500 LG + 0,250 LT + 3,00 GT
             - 0,250 LGT


Predictor     Coef  SE Coef       T      P
Constant   81,7500   0,5590  146,24  0,000
L           9,0000   0,5590   16,10  0,000
G           0,7500   0,5590    1,34  0,217
T          -4,0000   0,5590   -7,16  0,000
LG         -0,5000   0,5590   -0,89  0,397
LT          0,2500   0,5590    0,45  0,667
GT          3,0000   0,5590    5,37  0,001
LGT        -0,2500   0,5590   -0,45  0,667

S = 2,23607   R-Sq = 97,7%   R-Sq(adj) = 95,7%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       7  1711,00  244,43  48,89  0,000
Residual Error   8    40,00    5,00
Total           15  1751,00

Source  DF   Seq SS
L        1  1296,00
G        1     9,00
T        1   256,00
LG       1     4,00
LT       1     1,00
GT       1   144,00
LGT      1     1,00
```

Els coeficients són la meitat dels efectes i la desviació tipus és igual per tots els coeficients

La particularitat de les dades d'un disseny factorial és que, per construcció, les columnes de cadascun dels efectes és ortogonal a les altres (pots comprovar que el producte escalar de dues columnes de la matriu del disseny factorial és zero). Això té dues conseqüències importants:

1) El càlcul i la interpretació del model es simplifica perquè la matriu $X'X$ és diagonal,
$$X'X = \begin{pmatrix} N & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & N \end{pmatrix}$$
cosa que provoca que els coeficients del model també siguin independents.

2) El fet de treure una variable del model no fa que els altres coeficients canviïn. Això fa que puguem saber quines variables són significatives només mirant els p-valors de la regressió completa (sense la necessitat de netejar el model)

Què passa quan no tenim rèpliques?

Que es impossible estimar la variància dels efectes (o dels coeficients) perquè no queden graus de llibertat per estimar-la. Quan no hi ha rèpliques estimem la variància dels efectes suposant que les interaccions d'ordre gran no són significatives. Això equival a ajustar el model sense les columnes que corresponen a les interaccions que suposem no significatives.

Els dissenys factorials fraccionals s'analitzen anàlogament.

Casos en que pot anar bé fer l'anàlisi via regressió

1. Quan volem analitzar dades que provenen de 2 o més dissenys (estratègia seqüencial) realitzats en zones diferents de l'espai de les X's
2. Quan no podem controlar de forma precisa els nivells baix i alt d'un disseny.
3. Quan falten algunes observacions del disseny complert o només es tenen rèpliques per algunes condicions experimentals.
4. Quan algunes de les variables de disseny estan a més de dos nivells
5. Quan a més de variables controlables, s'observen variables no controlades que afecten a la resposta