Part 4: Design of Experiments (DOE) with 2^k factorial designs

TMA4267 L17
21.03.2017

Regression $\quad Y = X\beta + \varepsilon \;, \quad \varepsilon \sim N_n(0, \sigma^2 I)$

$n \times p$, intercept and k covariates

$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$ + more!

And we used observational data.
Now: we design the experiment = choose X!

How should we choose X? Achieve some kind of optimality.

- minimize "$\mathrm{Var}(\hat{\beta})$" $= \mathrm{tr}(\sigma^2 (X^T X)^{-1})$
- minimize $\det(\mathrm{Cov}(\hat{\beta}))$

Our focus:

- maximize interpretability; e.g. by choosing X so that $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = 0 \iff (X^T X)$ is diagonal which we may achieve by choosing the columns of X to be orthogonal to each other.

We focus on 2^k factorial design

look at k factors each at 2 levels

Ex: Pilot plant

$Y =$ yield

x_1 : A Temperature: 160 $\xrightarrow{\text{recode}}$ -1 ; 180 $\rightarrow$ 1 z_1

x_2 : B Concentration: 20 $\rightarrow$ -1 ; 40 $\rightarrow$ 1 z_2

	A	B	AB	Y
1	-1	-1	1	y_1
2	1	-1	-1	y_2
3	-1	1	-1	y_3
4	1	1	1	y_4

A, B: standard order

AB: multiplying A and B

Observe that each factor column has $\sum_{i=1}^{n} x_{ij} = 0$,
and we also include an
intercept term with $\sum_{i=1}^{n} x_{i1} = n \Rightarrow \underset{4\times 4}{X}$

$\hat{\beta}$ and SSR will have simple formulas for this
full 2^2 design

do all combinations

2^k full factorial designs

Ex: Lima beans, $k=3$ factor at two levels.

All possible $2 \cdot 2 \cdot 2 = 2^3 = 8$ experiments performed

$$Y = \mathbb{X}\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I)$$

with $\mathbb{X}$ given as (8×8)

Intercept	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
1	1	-1	-1	-1	-1	1	1
1	-1	1	-1	-1	1	-1	1
1	1	1	-1	1	-1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	-1	1	-1	1	-1	-1
1	-1	1	1	-1	-1	1	-1
1	1	1	1	1	1	1	1

A, B, C: standard order

AB, AC, BC, ABC: by multiplying the relevant columns

Intercept, A, B, C: perform experiment

Hands-on:

1) Show that any two columns of $\mathbb{X}$ are orthogonal, $\sum_{i=1}^{n} x_{ij} x_{ik} = 0$

2) Show that $\sum_{i=1}^{n} x_{ij} = 0$ for all j except $\underset{\text{intercept column}}{j=0}$

3) And that $\sum_{i=1}^{n} x_{ij}^2 = n$.

Now: How does this (1+2+3) influence our formulas for

i) $\hat{\beta}$ ii) $\mathrm{Cov}(\hat{\beta})$ iii) $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

i) $\hat{\beta}$

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

$$(\mathbb{X}^T\mathbb{X}) = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \ddots & & \\ & & & n \end{bmatrix}$$

$$\{\mathbb{X}^T\mathbb{X}\}_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} = \begin{cases} 0 & j \neq k \\ n & j = k \end{cases}$$

$$\hat{\beta} = \begin{bmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ 0 & \frac{1}{n} & & \\ \vdots & & \ddots & \\ 0 & & & \frac{1}{n} \end{bmatrix} \begin{bmatrix} \\ \mathbb{X} \\ \\ \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \underline{\underline{\left\{ \frac{1}{n} \sum_{i=1}^{n} \overset{-1,1}{x_{ij}} y_i \right\}_{\text{element } j}}}$$

$$\hat\beta_0 = \frac{1}{n}\sum_{i=1}^{n} 1 \cdot y_i = \underline{\underline{\bar y}}$$

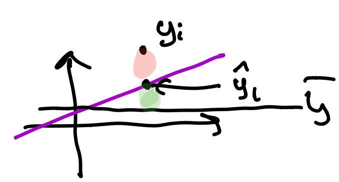
Observe that $\hat\beta_j$ is only dependent on x_{ij}, and not on x_{ik} $k \neq j$, so $\hat\beta_j$ will not change if we change the model. ← NEW now!

ii) $\mathrm{Cov}(\hat\beta) = (X^T X)^{-1}\sigma^2$

$$= \begin{bmatrix} \frac{1}{n} & 0 \cdots & 0 \\ 0 & \frac{1}{n} & 0 \cdots \\ & 0 & \frac{1}{n} \end{bmatrix} \sigma^2$$ so $\mathrm{Var}(\hat\beta_j) = \frac{1}{n}\sigma^2$ for all $j = 0, \dots, p$

and $\mathrm{Cov}(\hat\beta_j, \hat\beta_k) = 0$ for all $j \neq k$

↑ uncorrelated



iii) SST = SSE + SSR

$$\sum_{i=1}^{n}(y_i - \bar y)^2 = \sum_{i=1}^{n}(y_i - \hat y_i)^2 + \boxed{\sum_{i=1}^{n}(\hat y_i - \bar y)^2}$$

$\bar y = \hat\beta_0$ $\qquad$ $\hat y_i = \sum_{j=0}^{p-1} \hat\beta_j \cdot x_{ij}$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\sum_{j=0}^{p-1} \hat{\beta}_j \cdot x_{ij} - \hat{\beta}_0 \right)^2$$

$$= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} - \hat{\beta}_0 \right)^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} \right)^2$$

$$(\hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 x_{i2} + \cdots)(\hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \cdots + \hat{\beta}_{p-1} x_{i,p-1})$$

$$= \sum_{i=1}^{n} \left(\hat{\beta}_1^2 x_{i1}^2 + \hat{\beta}_1 \hat{\beta}_2 x_{i1} \cdot x_{i2} + \cdots + \hat{\beta}_{p-1}^2 x_{i,p-1}^2 \right)$$

remember $\sum_{i=1}^{n} x_{ij} x_{ih} = 0$ for $j \neq h$

$\sum_{i=1}^{n} x_{ij}^2 = n$

so $\hat{\beta}_1 \cdot \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} = 0$

etc.

$$= \sum_{i=1}^{n} \left(\hat{\beta}_1^2 x_{i1}^2 + \cdots + \hat{\beta}_{p-1}^2 x_{i,p-1}^2 \right) = n \cdot \sum_{j=1}^{p-1} \hat{\beta}_j^2$$

$$= \underbrace{n \cdot \hat{\beta}_1^2}_{\substack{SSR(X_1) \\ A}} + \underbrace{n \hat{\beta}_2^2}_{\substack{SSR(X_2) \\ B}} + \cdots + n \cdot \hat{\beta}_{p-1}^2$$

amount of variability due to each of the different covariates