Modelling with Indicator variables (Categorical variables)

Suppose we have

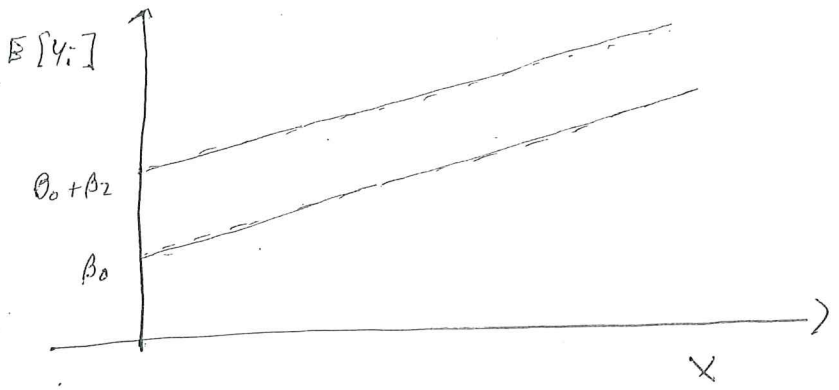
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_i + \varepsilon_i$$

with the usual assumptions.

Z is an indicator variable i.e

$$Z = \begin{cases} 0 & \text{if catalyst 1 is used} \\ 1 & \text{— " — 2 is used.} \end{cases}$$

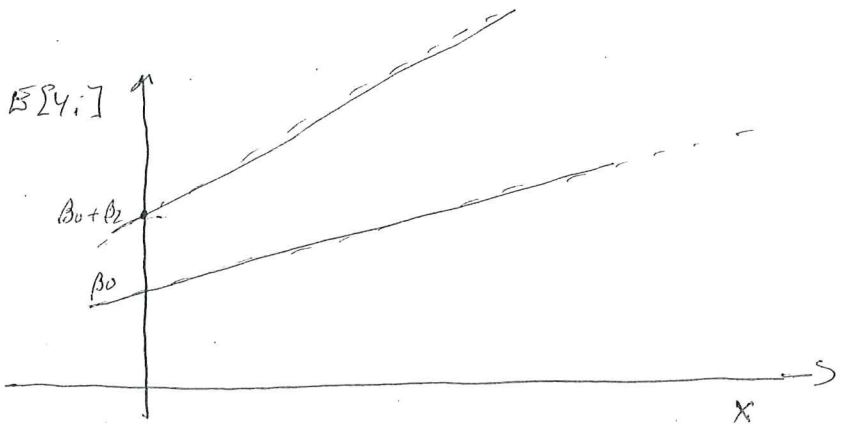
Then $$E[Y_i] = \begin{cases} \beta_0 + \beta_1 x_{1i} & \text{if catalyst 1 is used} \\ \beta_0 + \beta_2 + \beta_1 x_{1i} & \text{— " — 2 is used.} \end{cases}$$



Suppose,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_i + \beta_3 x_{1i} z_i + \varepsilon$$ with usual assumptions and Z an indicator variable.

Then $$E[Y_i] = \begin{cases} \beta_0 + \beta_1 x_{1i} & \text{if } Z = 0 \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_{1i}, & Z = 1 \end{cases}$$



15.3 Orthogonal columns in the Design matrix

$$\underset{\sim}{Y} = \underset{\uparrow}{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$$

Design matrix

Let $\underline{X} = [\underline{1}, \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_k]$

If $\underline{X}_p' \underline{X}_q = 0$ i.e $\sum_{i=1}^{n} x_{pi} x_{qi} = 0$, $p \neq q$

the columns $\underline{X}_p$ and $\underline{X}_q$ are orthogonal and

$$\underline{X}'\underline{X} = \begin{bmatrix} n & & & \\ & \underline{X}_1'\underline{X}_1 & & 0 \\ & & \ddots & \\ & 0 & & \underline{X}_k'\underline{X}_k \end{bmatrix}.$$

For $\hat{\beta}$ we obtain $\underline{\hat{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} = \begin{bmatrix} \frac{1}{n} & & & \\ & (X_1'X_1)^{-1} & & \\ & & \ddots & \\ & & & (X_k'X_k)^{-1} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \underline{X}_1'\underline{Y} \\ \vdots \\ \underline{X}_k'\underline{Y} \end{bmatrix}$

We observe that the estimator for β_j, $\hat{\beta}_j$ only depends on $\underline{X}_j$ and $\underline{Y}$

In addition we get that $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

$$= \sum_{i=1}^{n} (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - b_0)^2 = b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2$$

or $SS_R(\beta_0, \beta_1, \ldots, \beta_k) = R(\beta_1) + R(\beta_2) + \ldots + R(\beta_k)$

Chap 15. Two level designs

In industrial design of experiments, two-level designs play an important role. Such designs have two levels for each factor and are denoted 2^k designs where k is the number of factors.

An example

In production of carbon-steel springs there were severe problems with cracks.
Basic knowledge suggested two important factors that might cause the problem.

1. The temperature of the steel before quenching
2. The amount of carbon in the formulation.

4 experiments were conducted with two-levels for each of the factors. For each level combination it was registred the percentage of springs without cracks.

Factors, levels and data

Factor \ level	Low	High
Temperature	785C 1450 F	870 C 1600 F
Amount carbon	0.5 %	0.7 %

Experiment number	Temperature	Amount carbon	Persentage springs without cracks.
1	1450	0.5	67
2	1600	0.5	79
3	1450	0.7	61
4	1600	0.7	73

NB! Experiments shall always be performed in randomized order. In the table they are written up in standard form

How shall we analyse the experiment.

Let Y_i be the number of cracks in experiments number i,

$i = 1, 2, 3, 4$.

A natural model may be.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon$$

Y_i: response. x_{1i}: temperature. x_{2i}: carbon. $x_{1i} x_{2i}$: interaction term.

Our design matrix becomes.

$$\underline{X} = \begin{bmatrix} 1 & 1450 & 0.5 & 725 \\ 1 & 1600 & 0.5 & 800 \\ 1 & 1450 & 0.7 & 1015 \\ 1 & 1600 & 0.7 & 1120 \end{bmatrix} \quad \text{and} \quad \underline{y} = \begin{bmatrix} 67 \\ 79 \\ 61 \\ 75 \end{bmatrix}$$

The estimated coefficients are.

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_{12} \end{bmatrix} = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{y} = \begin{bmatrix} 14.3 \\ 0.047 \\ -126.7 \\ 0.067 \end{bmatrix}$$

Let us now introduce the factors

A = temperature in the steel before quenching

B = percentage amount of carbon in the steel.

and let:

$$X_A = \frac{A - 1525}{75} \quad \text{and} \quad X_B = \frac{B - 0.6}{0.1}$$

i.e the factors are first centered and thereafter divided down on half the distance between high and low level.

With the coded levels the design matrix becomes (where also an extra column $X_{AB} = X_A \cdot X_B$ is added:

$$\underline{X} = \begin{matrix} & X_A & X_B & X_{AB} \\ \end{matrix} \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \underline{X}'\underline{X} = \begin{bmatrix} 1/4 & & 0 \\ & 1/4 & \\ 0 & & 1/4 \end{bmatrix} \quad \text{and}$$

$$\underline{X}'\underline{y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ X_A' y \\ X_B' y \\ X_{AB}' y \end{bmatrix} \quad \text{and} \quad \underline{b}_C = \begin{bmatrix} \bar{y} \\ (-y_1 + y_2 - y_3 + y_4)/4 \\ (-y_1 - y_2 + y_3 + y_4)/4 \\ (y_1 - y_2 - y_3 + y_4)/4 \end{bmatrix} = \begin{bmatrix} 208.5 \\ 6.5 \\ -2.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} \bar{y} \\ b_A \\ b_B \\ b_{AB} \end{bmatrix}$$

If we substitute for X_A and X_B we get.

$$\hat{y} = 208.5 + 6.5\left(\frac{X_1 - 1525}{75}\right) - 2.5\left(\frac{X_2 - 0.6}{0.1}\right) + 0.5\left(\frac{X_1 - 1525}{75}\right)\left(\frac{X_2 - 0.6}{0.1}\right)$$

$$= 208.5 + \frac{6.5 \cdot 1525}{75} + \frac{2.5 \cdot 0.6}{0.1} + \frac{0.5 \cdot 1525 \cdot 0.6}{75 \cdot 0.1} + \frac{6.5}{75} X_1 - \frac{0.5 \cdot 0.6}{75 \cdot 0.1} X_1$$

$$+ \frac{2.5}{0.1} X_2 - \frac{0.5 \cdot 1525}{75 \cdot 0.1} X_2 + \frac{0.5 \cdot X_1 X_2}{75 \cdot 0.1}$$

$$= 14.33 + 0.047 X_1 - 126.67 X_2 + 0.067 X_1 X_2$$

Definition of main effects

For two-level design we define the main effect of a factor as: Expected average response when the factor is on the high level - expected average response when the factor is at the low level.

A natural estimate is $\bar{y}_H - \bar{y}_L$ which for temperature, A, and amount of carbon, B, becomes.

$$A_e = \frac{79+75}{2} - \frac{61+67}{2} = 13 = 2b_A$$

$$B_e = \frac{61+75}{2} - \frac{79+67}{2} = -5 = 2b_B$$

Definition of two-factor interaction

The interaction between two factors is defined as: Half the main effect of a factor when the other is on the high level - half the main effect when the other factor is on its low level.

An interaction between two factors tells us ~~that a~~ if a main effect of a factor depends on the level of the other factor.

$$AB_e = \frac{75-61}{2} - \frac{79-67}{2} = 1 = 2b_{AB}$$