



## Why

We model a real-valued output as corrupted by small random errors. Thus, we can talk about a dataset which is “close” to being consistent with a linear predictor.

## Definition

Let  $(\Omega, \mathcal{A}, \mathbf{P})$  be a probability space. Let  $x \in \mathbf{R}^d$  and  $e : \Omega \rightarrow \mathbf{R}^n$ . For  $A \in \mathbf{R}^{n \times d}$ , define  $y : \Omega \rightarrow \mathbf{R}^n$  by  $y = Ax + e$ . We call  $(x, A, e)$  a *probabilistic errors linear model*. We call  $y$  the *response vector*,  $A$  the *model matrix* and  $e$  the *error vector*.

## Moment assumptions

The most basic distributional assumption for a probabilistic errors linear model pertain to the expectation and variance. Since  $\mathbf{E}(y) = Ax + \mathbf{E}(e)$  and  $\mathbf{var}(y) = \mathbf{var}(e)$ , these assumptions can be given for  $e$  or for  $y$ .

If  $\mathbf{E}(x) = 0$  and  $\mathbf{var}(y) = \sigma^2 I$  then we call  $(x, A, e)$  a *classical linear model with moment assumptions*. Notice that the components of  $e$  are assumed uncorrelated. We have  $d + 1$  unknowns (the  $d \times 1$  entires of  $\theta$  and scalar parameter  $\sigma^2$ ).

In this case  $\mathbf{E}(y_i) = a^i{}^\top \theta$  and so  $\theta$  is called the *mean parameter vector* and  $\sigma^2$  is called the *model variance*. The model variance indicates the variability inherent in the observations. Neither the mean nor variance of the error depends on the regression vector  $x$  nor on the parameter vector  $\theta$ .

## Examples

Consider the *two-sample problem* in which we have two populations with (unknown) mean responses  $\alpha_1, \alpha_2 \in \mathbf{R}$ . We observe these responses with (perhaps unknown) common variance  $\sigma^2$ , and assume that errors are uncorrelated.

We define  $y^1 = \alpha_1 \mathbf{1} + e^1$  and  $y^2 = \alpha_2 \mathbf{1} + e^2$  so that we can stack these and obtain

$$y = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} \alpha_1 \mathbf{1} \\ \alpha_2 \mathbf{1} \end{bmatrix} + \begin{bmatrix} e^1 \\ e^2 \end{bmatrix}.$$

To cast this in our standard form we define

$$A = \begin{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \cdots & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \cdots & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{bmatrix}^\top, \quad x = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

with regression vectors  $a_1 = (1, 0)$  and  $x_2 = (0, 1)$  repeated  $n_1$  and  $n_2$  times, respectively. An input design for this model involves specifying a sequence of these two vectors, which (with the uncorrelated assumption) reduces to dictating how many responses should be collected from each population. The inputs here is really the set  $\mathcal{X} = \{1, 2\}$ . The feature function is  $\phi : \mathcal{X} \rightarrow \mathbf{R}^2$  defined by  $\phi(1) = (1, 0)$  and  $\phi(2) = (0, 1)$ . And so the regression range is  $\{(1, 0), (0, 1)\}$ .

