Nomenclature:

Note that the nomenclature used here is slightly different from that used by Hastie et al. It has been modified for the purposes of this workshop.

We will define a dataset consisting of N observations to be given by $(\mathbf{y}, \mathbf{X})$**.**

- The set of response variables are given by the column vector $\mathbf{y}$.
  - Each row in $\mathbf{y}$ contains a single "observation" y_i, with $i = 1,2,3 \dots N$
- The set of feature variables are given by the design matrix $\mathbf{X}$.
  - Each row in $\mathbf{X}$ corresponds to a single "observation" $\mathbf{x}_i$, with $i = 1,2,3 \dots N$
  - Each observation $\mathbf{x}_i$ consists of p scalar features, $x_{i,j}$, with $i = 1,2,3 \dots p$
  - All the observations corresponding to a single feature can be represented by the column vector $\mathbf{X}_j$, such that $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 & \dots & \mathbf{X}_p\end{bmatrix}$
- We define our model coefficients using the column vector $\boldsymbol{\beta} = [\beta_1 \ \ \beta_2 \dots]^T$, as well as an intercept term β_0

In a linear regression model:

- We learn the $p + 1$ coefficients such that $\hat{y}(\mathbf{x}) = \sum_{j=0}^{p} x_j \beta_j = \mathbf{x} \cdot \boldsymbol{\beta}$.
- There are $p + 1$ coefficients as we have included the "dummy" predictor $x_0 = 1$, which corresponds to the intercept term β_0.
- Since $\mathbf{x}$ is row vector, we have $\mathbf{x} \cdot \boldsymbol{\beta}$ simply equal to $\mathbf{x}\boldsymbol{\beta}$ (since $\mathbf{x}$ is $1 \times (p + 1)$ and $\boldsymbol{\beta}$ is $(p + 1) \times 1$). We will however retain the dot-product sign for clarity.
- For the full set of observations, we have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$.
- It is commonly known that the solution to the optimization problem…

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{N} (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2 \right)$$

…is given by the solution of the system of linear equations, called the normal equations:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

…with $\mathbf{X}$ being the augmented design matrix, which contains a column of ones corresponding to the dummy predictor $x_0 = 1$.

Note: The handling of the intercept term often causes confusion. Careful attention should be paid to the way in which the intercept is handled when using different software packages. This will be pointed out clearly throughout the workshop.