3. Fitting a linear model to a sample by least squares

- Recall the sample version of the linear model. Data are $y_1, y_2, \ldots, y_n$ and on each individual i we have p explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$.

$$\text{(LM1)} \qquad y_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i \quad \text{for } i = 1, 2, \ldots, n$$

- Using summation notation, we can equivalently write

$$\text{(LM2)} \qquad y_i = \sum_{j=1}^{p} x_{ij} b_j + e_i \qquad \text{for } i = 1, 2, \ldots, n$$

- We can also use matrix notation. Define column vectors $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_p)$. Define the matrix of explanatory variables, $\mathbb{X} = [x_{ij}]_{n \times p}$. In matrix notation, (LM1) and (LM2) is exactly

$$\boxed{\text{(LM3)} \qquad \mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e}}$$

- Matrices give a compact way to write the linear model, and also a good way to carry out the necessary computations.

The least squares formula

- We seek the **least squares** value of $\mathbf{b}$ that minimizes the sum of squared error, $\sum_{i=1}^{n} e_i^2$.
- Since n is usually much bigger than p, there is usually no value of $\mathbf{b}$ for which we can exactly explain the data using the explanatory matrix $\mathbb{X}$.
- In other words, there is no choice of $\mathbf{b}$ which solves $\mathbb{X}\,\mathbf{b} = \mathbf{y}$.
- The least squares choice of $\mathbf{b}$ turns out to be

$$\boxed{\text{(LM4)} \qquad \mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}}$$

- We will gain understanding of (LM4) by studying the **simple linear regression** model with $p = 2$.