The general linear model (also known as the multiple regression model):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad i = 1, 2, \cdots N \ (population\ model)$$

$$Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \quad i = 1, 2, \cdots n \ (sample\ model)$$

In matrix notation:

$$\underset{(n\times 1)}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{Design\ matrix\ \ (n\times k)}{\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}} \underset{(k\times 1)}{\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}} + \underset{(n\times 1)}{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}}$$

$$Y = Xb + e \ (estimated\ sample\ model)$$

$$Y = X\beta + \varepsilon \ (assumed\ population\ model)$$

These matrix equations are used by R and work for any number of X variables.

For the special case of simple linear regression (where there is only 1 X variable), the design matrix is just:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ \vdots \\ b_0 + b_1 x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_1 + e_1 \\ b_0 + b_1 x_2 + e_2 \\ \vdots \\ b_0 + b_1 x_n + e_n \end{bmatrix}$$

Similarly, in matrix notation, the least squares estimates become (see page 6 of the notes):

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

This matrix equation again holds for any number of X variables.

For simple linear regression (SLR):

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad \text{and} \quad X^T Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix}$$

Finding the inverse of a $2 \times 2$ matrix is relatively easy (note that finding the inverse of larger matrices is not as easy, and in this course, we will leave that work to R and trust that it has good reliable numerical routines for doing such calculations):

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow M^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad \text{so} \quad \left( X^T X \right)^{-1} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

Putting these pieces together, we find that for SLR, the least squares estimates have the same formulae you would have used in an introductory statistics course (and are the same formulae we used earlier in the revision topic):

$$\hat{\beta} = b = \left(X^T X\right)^{-1}\left(X^T Y\right) = \frac{1}{nS_{xx}}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}\begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix} = \begin{bmatrix} \overline{Y} - \frac{S_{xy}}{S_{xx}}\overline{x} \\ \frac{S_{xy}}{S_{xx}} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

(see page 7 of the notes for the missing details).

Note that even though we will leave the calculations to R, some of the matrices formed in this process have interesting properties and occasionally we will want R to show us the contents of key matrices, so that we can examine these properties.

A type of matrix that often has interesting properties is a variance-covariance matrix.

For example the vector of errors has variance-covariance matrix:

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_N) \\ Cov(\varepsilon_1, \varepsilon_2) & Var(\varepsilon_2) & & Cov(\varepsilon_2, \varepsilon_N) \\ \vdots & & \ddots & \\ Cov(\varepsilon_1, \varepsilon_N) & & & Var(\varepsilon_N) \end{bmatrix}$$

If two errors are independent of each other, then their covariance will be 0, so the model-specific assumptions that the errors are independent and have constant variance, can be summarised as:

$$Var(\varepsilon) = \sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix}$$

Similarly, the variance-covariance matrix of the least squares estimates can be shown (see pages 8 and 9 of the notes) to be:

$$Var(b) = \sigma^2 \left( X^T X \right)^{-1}$$

For SLR, the diagonal elements of this matrix are:

$$Var(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right) \quad \text{and} \quad Var(b_1) = \frac{\sigma^2}{S_{xx}}$$

We could take the square root of these variances to find standard errors for the least squares estimates of the coefficients, but first we will need to estimate the error variance.

Another matrix with interesting properties is the so-called hat matrix:

$$\hat{Y} = Xb = X\left(X^T X\right)^{-1} X^T Y = HY, \quad \text{where} \quad H = X\left(X^T X\right)^{-1} X^T$$

The matrix $H$ is called the hat matrix because when it is multiplied by the observed $Y$ values, it produces the fitted values (the $Y$ "hats"). The hat matrix is an $n \times n$ matrix and the diagonal elements of this matrix are called the leverage values:

$$h_{ii} = \frac{1}{nS_{xx}} \sum_{j=1}^{n} \left(x_j - x_i\right)^2$$

There is one leverage value for each of the original $n$ observations and the leverage value can be used as a measure of how influential the corresponding observation was in the fitting of the regression model.

Note that the vector of residuals (observed errors) can be written as:

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

And the variance-covariance matrix of the residuals is then:

$$Var(e) = \sigma^2 (I - H)$$

Which will be useful later in the course when we want to standardise the residuals.

Finally, we can also use the hat matrix to estimate the error variance (so we can find standard errors for the estimated coefficients). For SLR, where the error degrees of freedom are $n - 2$:

$$\hat{\sigma}^2 = s^2 = \frac{SS_{Errors}}{(n-2)} = \frac{e^T e}{(n-2)} = \frac{(Y - \hat{Y})^T (Y - \hat{Y})}{(n-2)} = \frac{Y^T (1 - H) Y}{(n-2)}$$

(again, see page 9 of the notes for the missing details).