In matrix form, this can be formulated as

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \dots & \dots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = A \cdot \beta - y = f(A) - y.$$

The above error vector r is called residual. The L^2-error

$$E = \left(\sum_{i=1}^{n} r_i^2 \right)^{\frac{1}{2}} = \|r\|_2 = \|A \cdot \beta - Y\|_2.$$

Goal of linear regression: find the coefficient $\beta = [\beta_1, \beta_2, \dots, \beta_m]$ so that the error E is minimized. The explicit formula is available. However, the computation is timely expensive for large m, because it involves the numerical computation of the inverse of an m-by-m matrix.

The following is a method from Numerical Analysis (MAT 2680). The least square solution β can be found by solving the so-called normal equation

$$(A^T \cdot A) \cdot \beta = A^T \cdot y.$$

Example 1.2. Get familiar with the R function lm( ). See the shared program codes.

Link for R document

Project 6 - part 2 (2 points, due May 4)

Download data file "project6.csv" from Blackboard.

1. Find the linear regression function for model

$$X_3 = \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5.$$

2. Compute the root mean square error (RMSE) for your model.

1.3 Variation of linear regressions

1.3.1 How do we include/exclude the y-intercept?

For example, the line model $y = kx + b$ is in fact

$$Y = \beta_1 X_1 + \beta_2 \cdot 1 = \begin{bmatrix} X_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

For an observation a, the predicted value

$$f(a) = \beta_1 a + \beta_2 = \begin{bmatrix} a & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$