THE MULTIPLE LINEAR REGRESSION MODEL

Notation:
- \( p \) predictors \( x_1, x_2, \ldots, x_p \)
- (Some might be indicator variables for categorical variables.)
- \( k-1 \) non-constant terms \( u_1, u_2, \ldots, u_{k-1} \)
  - Each \( u_j \) is a function of \( x_1, x_2, \ldots, x_p \): \( u_j(x_1, x_2, \ldots, x_p) \)
  - For convenience, we often set \( u_0 = 1 \) (constant function/term)

The Basic Multiple Linear Regression Model: Two assumptions:
1. \( E(Y|\mathbf{x}) \) (or \( E(Y|\mathbf{u}) = \eta_0 + \eta_1 u_1 + \ldots + \eta_{k-1} u_{k-1} \)) (Linear Mean Function)
2. \( \text{Var}(Y|\mathbf{x}) \) (or \( \text{Var}(Y|\mathbf{u}) = \sigma^2 \)) (Constant Variance)

Assumption (1) in vector notation:
- \( \mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_{k-1} \end{bmatrix} = \begin{bmatrix} 1 \\ u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_{k-1} \end{bmatrix} \) \( \eta = \begin{bmatrix} \eta_0 \\ \eta_1 \\ \cdot \\ \cdot \\ \cdot \\ \eta_{k-1} \end{bmatrix} \)

Then \( \eta^T = [\eta_0 \ \eta_1 \ \ldots \ \eta_{k-1}] \) and
- \( \eta^T \mathbf{u} = \eta_0 + \eta_1 u_1 + \ldots + \eta_{k-1} u_{k-1} \)

so (1) becomes:
- \( (1') E(Y|\mathbf{x}) = \eta^T \mathbf{u} \)

If we have data with \( i^{th} \) observation \( x_{i1}, x_{i2}, \ldots, x_{ip}, y_i \), recall
- \( \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{ip} \end{bmatrix} = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T \)

Define similarly
- \( u_{ij} = u_j(x_{i1}, x_{i2}, \ldots, x_{ip}) \) = the value of the \( j^{th} \) term for the \( i^{th} \) observation, and
So in particular, the model says
\[ E(Y|x_i) = \eta^T u_i \]

**Estimation of Parameters:** Analogously to the case of simple linear regression, consider functions of the form
\[ y = h_0 + h_1 u_1 + \ldots + h_{k-1} u_{k-1} = h^T u. \]

(The graph of such an equation is called a "hyperplane.")

The *least squares estimate* of \( \eta \) is the vector
\[ \hat{\eta} = \begin{bmatrix} \hat{\eta}_0 \\ \hat{\eta}_1 \\ \vdots \\ \hat{\eta}_{k-1} \end{bmatrix} \]

that minimizes the "objective function"
\[ \text{RSS}(h) = \sum_{i=1}^{n} (y_i - h^T u_i)^2 \]

The solution (when it exists; see below) can be found by setting all partial derivatives of \( \text{RSS}(h) \) equal to zero and solving the resulting simultaneous equations.

**Recall:** In simple linear regression, the solution for \( \hat{\eta}_1 \) had \( SXX = \sum_{i=1}^{n} (x_i - \overline{x})^2 \) in the denominator. So the formula for \( \hat{\eta}_1 \) won't work if all \( x_i \)'s = \( \overline{x} \). In that case, there is not a unique solution to the least squares problem. (Draw a picture in the case \( n = 2 \! \! \! \! \).)

**In multiple regression:** There is a unique solution \( \hat{\eta} \) *provided* both:

i) \( k < n \) (the number of terms is less than the number of observations)

ii) no \( u_j \) is (as a function) a linear combination of the other \( u_j \)'s

If there is a unique solution, it is called the *ordinary least squares (OLS) estimate* of the (vector of) coefficients.
Examples where there is not a unique solution:

1. When \( k = 2 \) (simple linear regression) and there is only one data point.

2. \( k = 2 \) and both data points have the same \( x \) value.

3. Similar examples for larger \( k \).

4. Two predictors, three terms with
   \[ u_1 = x_1, \ u_2 = x_2, \ u_3 = x_1 + x_2 \]
   
   e.g., Scholastic Aptitude Test Scores (SAT) with terms SATM, SATM, SATM + SATV

Multicollinearity:

When condition (ii) is violated, we say there is (strict) multicollinearity. (e.g., example 4 above.)

A situation close to strict multicollinearity is typically called \textit{multicollinearity}.
Technically, there is a solution, but
a. The solutions involved small denominators, which can make calculation virtually impossible. (e.g., if \( p = 1 \) and if \( x \) is close to constant, then \( SXX \) is very small.)
   b. The variances will be large, making inference virtually useless.