Assumptions (A) for Multiple Linear Regression

**Notation** The \((k + 1)\)-variable population \(\{Y, X_1, \ldots, X_k\}\) is the study population.

**(Population) Assumption 1** The mean \(\mu_Y(x_1, \ldots, x_k)\) of the subpopulation of \(Y\) values with \(X_i = x_{i1}, \ldots, X_k = x_{ik}\) is

\[
\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]

where \(\beta_0, \beta_1, \ldots, \beta_k\) are unknown parameters and \(x_1, \ldots, x_k\) belong to the set of allowable values (sometimes called the *domain*) of the predictor variables.

**(Population) Assumption 2** The standard deviation of the \(Y\) values in the subpopulations with \(X_i = x_{i1}, \ldots, X_k = x_{ik}\) does not depend on the values \(x_1, \ldots, x_k\) (i.e., the standard deviations are the same for each subpopulation determined by specified values of the predictor variables \(X_1, \ldots, X_k\)). This common standard deviation of all the subpopulations is denoted by \(\sigma_{Y|X_1, \ldots, X_k}\). When there is no possibility of confusion, we use the simpler notation \(\sigma\) instead of the more complete notation \(\sigma_{Y|X_1, \ldots, X_k}\).

**(Population) Assumption 3** Each subpopulation of \(Y\) values, determined by specified values of \(X_1, \ldots, X_k\) is a Gaussian population.

**(Sample) Assumption 4** The sample data are obtained by simple random sampling or by sampling with preselected values of \(X_1, \ldots, X_k\), discussed in Section 2.3. The number of items in the sample is \(n\).

**(Sample) Assumption 5** All sample values \(y_i, x_{i1}, \ldots, x_{ik}\) for \(i = 1, \ldots, n\) are observed without error (but read Section 3.10).
Assumptions (B) for Multiple Linear Regression

(Population) Assumption 1 The study population \( \{(Y, X_1, \ldots, X_k)\} \) is a \((k + 1)\)-variable Gaussian population.

(Sample) Assumption 2 The sample data are obtained by simple random sampling described in Section 2.3; i.e., a simple random sample of \( n \) items is selected from the population and the values of the variables \( Y, X_1, \ldots, X_k \) are observed.

(Sample) Assumption 3 The sample values \( y_i, x_{i1}, \ldots, x_{ik} \) for \( i = 1, \ldots, n \) are measured without error.

Simple Random Sampling

Sample data are obtained by selecting a simple random sample of \( n \) items from the entire population of \( N \) items and recording the values for the response variable \( Y \) and the predictor variables \( X_1, \ldots, X_k \), for each item in the sample. Refer to Section 1.6.

Random Sampling with Preselected \( X \) values

Specific values of the predictor variables \( X_1, \ldots, X_k \) are preselected by the investigator, and each of these preselected sets of values determines a subpopulation of \( Y \) values. A simple random sample of one or more \( Y \) values is selected from each of these subpopulations. The number of observations to be sampled from each subpopulation is also predetermined by the investigator.
Regression Analysis: Concepts and Applications