Statistical Models from Data

In any given problem where linear regression is applicable, multiple analysis steps are often necessary to obtain robust model that yields accurate and effective results [1]. The objective of this paper is to introduce and describe some of the basic steps that are required when a robust model is sought after.

In a regression problem the first step is to identify and understand a set of variables that contain the observed information or data that form each variable [1]. Suppose the variable names are $L_1, L_2, ..., L_n$ and the data are $X_1, X_2, ..., X_n$ which are columns of numerical data, for now, each containing n records. Of these variables, our first goal is to identify which X_k we wish to explore where $k \in i = 1, 2, ..., n$. In a linear regression problem, after X_k has been identified, the set $X_1, X_2, ..., X_n$ becomes $X_1, X_2, ..., X_{n-1}$ for $j = 1, 2, ..., n-1$ and $k \notin j$ which are often called *predictors* or *independent/explanatory variables* and the *response/dependent variable* of the data set where it is common to write $Y = X_k$ [1, 3, 4]. In other words, linear regression analysis can show how $X_1, X_2, ..., X_{n-1}$ influences the behavior or causes Y to change or vary as the set of explanatory variables take on different values [3]. In the case where we choose to"explain" the value of Y with only one independent variable X_j a linear model is

$$Y = \beta_0 + \beta_1 X_j + \epsilon \tag{1}$$

. Or if we want to explain Y with multiple independent variables, the linear model is

$$Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + ... + \beta_t X_t + \epsilon \tag{2}$$

where (1) is a *population simple linear regression model*, (2) is a *population multiple linear regression model*, ϵ is a *statistical error*, and $t \leq n-1$ given the response variable appears to depend linearly on the independent variable(s) [1, 4, 5].

It is easier to understand the statistical error in terms of the population simple linear regression model. The statistical error term in (1) is required because these lines generally do not pass through every pair of bivariate data which is the set of ordered pairs contained in (X_j, Y) for $k = 1, 2, ..., m$ records in the data set [1]. Furthermore, all of the βs and ϵ in (1) and (2) are unknown *parameters* and must be estimated from the data contained in the independent and dependent variables with $k = 1, 2, ..., m$ records in the data set [1]. In (1) the statistical error for record k is

$$\epsilon_k = y_k - \beta_0 - \beta_1 x_{k,j} \tag{3}$$

and the statistical error in (2) is

$$\epsilon_k = y_k - \beta_0 - \beta_1 x_{k,1} - \beta_2 x_{k,2} - ... - \beta_t x_{k,t} \tag{4}$$

[1, 4, 5].

In either case of (1) and (2), estimations of the model parameters are often obtained by minimizing a quantity called the *sum of squared residuals* or *residual sum of squares* which writes as

$$\sum_{k=1}^{m} \epsilon_k^2 = \sum_{k=1}^{m} (y_k - \beta_0 - \beta_1 x_{k,j})^2 \tag{5}$$

and

$$\sum_{k=1}^{m} \epsilon_k^2 = \sum_{k=1}^{m} (y_k - \beta_0 - \beta_1 x_{k,1} - \beta_2 x_{k,2} - ... - \beta_t x_{k,t})^2 \tag{6}$$

[2, 4, 5]. It turns out that the function (5) is minimized when

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{m} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{m} (x_k - \bar{x})^2} \tag{7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{8}$$

where $\bar{x}$ and $\bar{y}$ are the *sample means* of X and Y repectively [1, 2, 5]. On the other hand, the estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_t$ are found by utilizing matrix algebra