# Linear Models

It is of interest to see if a nice relationship exists between two random variables, $X$ and $Y$. Eventual objective may be either prediction of a future value or utilization of the relaionship for understanding the structure.

**Ex.** $X$ = height, $Y$ = weight of indviduals. One may ask: is there an optimal weight for a given height?
Data: $(x_i, y_i)$, observations from $n$ randomly chosen individuals, $i = 1, 2, \ldots, n$.

**Ex.** $X$ = temperature, $Y$ = pressure of a certain volume of gas.
Data: $(x_i, y_i)$, $i = 1, 2, \ldots, n$ from a *controlled experiment* where a certain volume of gas is *subjected to different temperatures* and the resulting pressure is measured.

**Ex.** In a biological assay, $Y$ = response corresponding to a dosage level of $X = x$. Again, $(x_i, y_i)$, $i = 1, 2, \ldots, n$ from $n$ laboratory subjects.

**Ex.** In an agricultural experiment, $y$ is the yield of a crop. A piece of land is divided into $I$ plots according to soil fertility; $J$ different fertilizer levels are also used. Then, if $y_{ij}$ is the yield from the $i$th plot receiving $j$th level of fertilizer, we might like to try the model:
$y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}$. Why do we need $\epsilon_{ij}$? It is a random error (measurement error, noise or uncontrolled variability) needed to explain the variation in the model, which is needed in each of the other cases as well.

In general,

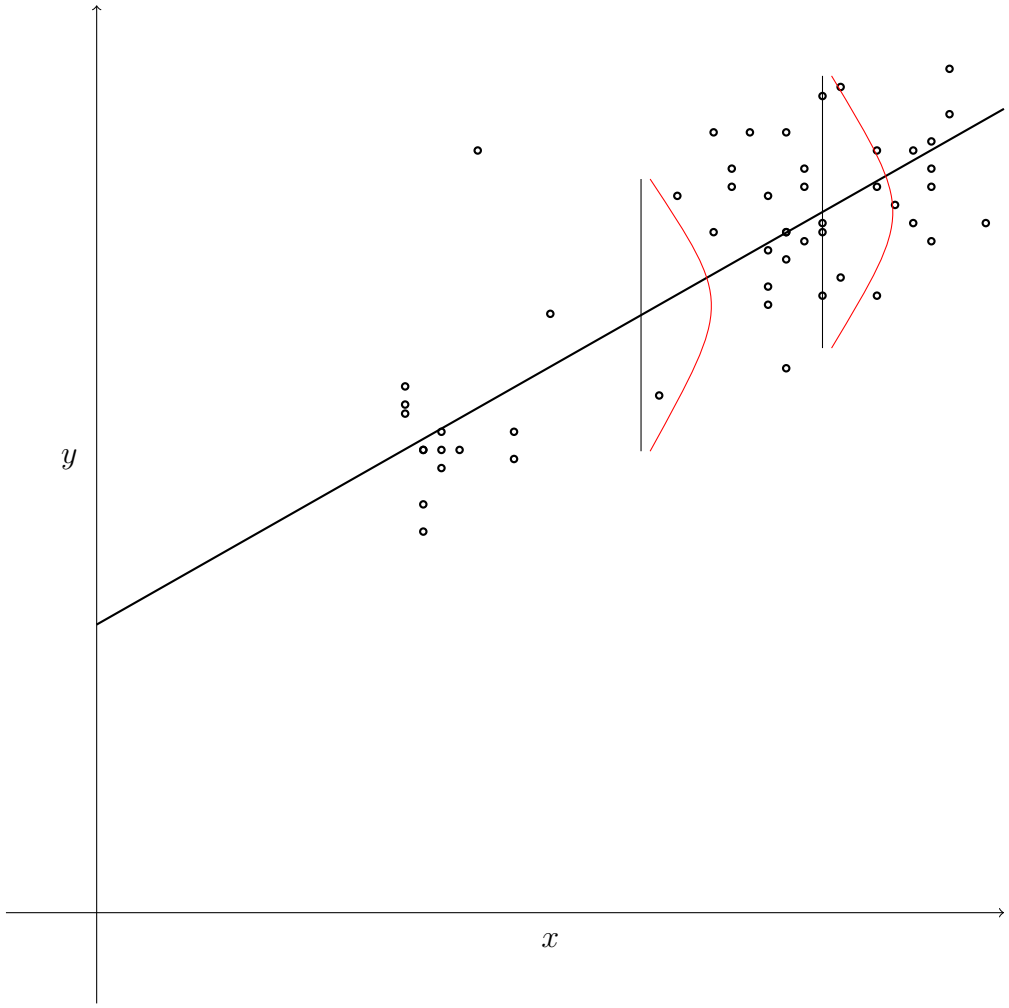
$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

where $y$ is the response variable and $x$ is the predictor variable, and $\alpha$ and $\beta$ are unknown coefficients is called a linear model. Here 'linear' stands for linear space, linear or additive in the coefficients and not for linear in $x$, as will be seen later. Equation (1) expresses the linear or additive relationship between $E(Y|X = x)$ and the influencing factors.

Observe the following data and the scatter plot of $y$ versus $x$, where $x$ = duration and $y$ = interval (both in minutes) for eruptions of Old Faithful Geyser.

| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.4 | 78 | 3.9 | 74 | 4.0 | 68 | 4.0 | 76 | 3.5 | 80 | 4.1 | 84 |
| 2.3 | 50 | 4.7 | 93 | 1.7 | 55 | 4.9 | 76 | 1.7 | 58 | 4.6 | 74 |
| 3.4 | 75 | 4.3 | 80 | 1.7 | 56 | 3.9 | 80 | 3.7 | 69 | 3.1 | 57 |
| 4.0 | 90 | 1.8 | 42 | 4.1 | 91 | 1.8 | 51 | 3.2 | 79 | 1.9 | 53 |
| 4.6 | 82 | 2.0 | 51 | 4.5 | 76 | 3.9 | 82 | 4.3 | 84 | 2.3 | 53 |
| 3.8 | 86 | 1.9 | 51 | 4.6 | 85 | 1.8 | 45 | 4.7 | 88 | 1.8 | 51 |
| 4.6 | 80 | 1.9 | 49 | 3.5 | 82 | 4.0 | 75 | 3.7 | 73 | 3.7 | 67 |
| 4.3 | 68 | 3.6 | 86 | 3.8 | 72 | 3.8 | 75 | 3.8 | 75 | 2.5 | 66 |
| 4.5 | 84 | 4.1 | 70 | 3.7 | 79 | 3.8 | 60 | 3.4 | 86 | | |

Table 1: Eruptions of Old Faithful Geyser, August 1 – 4, 1978




(1) is a linear model for $E(y|x)$, so $\epsilon$ denotes the spread or dispersion around this line. i.e., $y = E(y|x) + \epsilon$. If we let $g(x) = E(y|x)$, assuming $g$ to be smooth, we could consider the approximation:

$$\begin{aligned} g(x) &= g(0) + g'(0)x + \frac{g''(0)}{2!}x^2 + \ldots + \frac{g^{(k)}(0)}{k!}x^k \\ &= \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k. \end{aligned}$$

This is linear in the coefficients $\beta_0, \beta_1, \ldots$ but not in $x$. Also, recall Weirstrass theorem on being able to uniformly approximate by polynomials any contiuous function on a closed interval. Thus, on a reasonable range of $x$ values, such a 'linear' approximation may be quite acceptable. More importantly, special tools and techniques from linear spaces and linear algebra are available for studying linear models.

## MULTIPLE LINEAR REGRESSION MODEL

The response $y$ is often influenced by more than one predictor variable. For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used. These variables are controlled by the experimenter, but the yield may also depend on uncontrollable variables such as those associated with weather. A linear model relating the response y to several predictors has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \epsilon. \tag{2}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are called regression coefficients. The presence of $\epsilon$ provides for random variation in $y$ not explained by the $x$ variables. This random variation may be due partly to other variables that affect $y$ but are not known or not observed. The model in (2) is linear in the $\beta$ parameters; it is not necessarily linear in the $x$ variables. Thus models such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3 + \beta_4 \sin(x_2) + \epsilon$$

are included in the designation linear model. A model provides a theoretical framework for better understanding of a phenomenon of interest. Thus a model is a mathematical construct that we believe may represent the mechanism that generated the observations at hand. The postulated model may be an idealized oversimplification of the complex real-world situation, but in many such cases, empirical models provide useful approximations of the relationships among variables. These relationships may be either associative or causative.

Regression models such as (2) are used for various purposes, including the following:

**Prediction.** Estimates of the individual parameters $\beta_0, \beta_1, \ldots$ are of less importance for prediction than the overall influence of the x variables on y. However, good estimates are needed to achieve good prediction performance.

**Data Description or Explanation.** The scientist or engineer uses the estimated model to summarize or describe the observed data.

**Parameter Estimation.** The values of the estimated parameters may have theoretical implications for a postulated model.

**Variable Selection or Screening.** The emphasis is on determining the importance of each predictor variable in modeling the variation in y. The predictors that are associated with an important amount of variation in y are retained; hose that contribute little are deleted.

**Control of Output.** A cause-and-effect relationship between y and the x variables is assumed. The estimated model might then be used to control the output of a process by varying the inputs. By systematic experimentation, it may be possible to achieve the optimal output.

There is a fundamental difference between purposes 1 and 5. For prediction, we need only assume that the same correlations that prevailed when the data were collected also continue in place when the predictions are to be made. Showing that there is a significant relationship between $y$ and the $x$ variables in (2) does not necessarily prove that the relationship is causal. To establish causality in order to control output, the researcher must choose the values of the $x$ variables in the model and use randomization to avoid the effects of other possible variables unaccounted for. In other words, to ascertain the effect of the $x$ variables on $y$ when the $x$ variables are changed, it is necessary to change them.

**Vector-matrix form of linear model.**

Data is of the form: $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, n$, $\mathbf{x}_i = (x_{i0} = 1, x_{i1}, \ldots, x_{i(p-1)})'$. The linear model is:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \\ &= \sum_{j=0}^{p-1} \beta_j x_{ij} + \epsilon_i, i = 1, 2, \ldots, n; x_{i0} = 1. \end{aligned}$$

Equivalently,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1(p-1)} \\ 1 & x_{21} & \ldots & x_{2(p-1)} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{ or}$$

$$\mathbf{y} = X\beta + \epsilon.$$

$\mathbf{y}_{n\times 1}$ is the response vector, $X_{n\times p}$ is the matrix of predictors or covariates, $\beta_{p\times 1}$ is the vector of regression coefficients, and $\epsilon$ is random noise. $\mathbf{y}$ is random since $\epsilon$ is random. $X$ is treated as a fixed matrix and $\beta$ is a fixed but unknown vector of parameters. Note that the model involves random vectors and matrices, so some preliminaries on these are needed before we can proceed further.