Chapter 1 Simple Linear Regression (Part 1)

1 Simple linear regression model

Suppose for each subject, we observe/have two variables X and Y. We want to make inference (e.g. prediction) of Y based on X. Because of random effect, we cannot predict Y accurately . Instead, we can only predict its "expected/mean" value, i.e. $E(Y) = f(X)$ or $E(Y|X) = f(X)$. A simple functional form is $f(X) = \beta_0 + \beta_1 X$ with β_0 **and** β_1 **being unknown**, i.e.

$$E(Y|X) = \beta_0 + \beta_1 X$$

In statistics, people like to write it as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

which is called the "linear regression model". In the model,

- X is called: independent variable(s); covariate or predictor(s).
- Y is called: dependent variable; response.
- ε is called random error with $E\varepsilon = 0$, which is not observable and not estimable. Thus

  $$E(Y) = \beta_0 + \beta_1 X$$

  or

  $$E(Y|X) = \beta_0 + \beta_1 X$$

- β_0 and β_1 are unknown, called regression coefficients. β_0 is also called intercept (value of EY when $X = 0$); β_1 is called slope indicating the change of Y on average when X increases one unit.

Suppose we have observed n subjects. We have n observations

$$(X_1, Y_1), \quad (X_2, Y_2), \quad ..., \quad (X_n, Y_n)$$

The linear regression model also means

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1, \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n. \end{aligned}$$

In the model,

- X_i is known, observable, and non-random,
- ε_i is called random error (unobservable).
- Thus Y_i is random.

Assumptions and features of the model

- X_i is non-random, but ε_i is random. Thus the first part of Y_i: $\beta_0 + \beta_1 X_i$ is due to regression on (X); the second part: ε_i is due to the random effect.
- [Mean of random errors] $E\varepsilon_i = 0$, thus

$$\begin{aligned} E\{Y_i\} &= E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} \\ &= \beta_0 + \beta_1 X_i + E\varepsilon_i \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

- [Homogeneity of Variance] $Var(\varepsilon_i) = \sigma^2$
- [independence (no serial correlation)] $Cov(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$.
- Thus (please prove it based on the previous point), $Var(Y_i) = \sigma^2$ and $Cov(Y_i, Y_j) = 0$ for any $i \neq j$. Thus Y_i and Y_j are uncorrelated. (Do you think it is reasonable?)

Parameters in the model: β_0, β_1 and σ^2. They need to be estimated.

2 The estimation of the parameters and the model

2.1 Least Squares Estimation (LSE)

The deviation of Y_i from its expected value is

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i).$$

Since β_0, β_1 are unknown, "Good" estimators of β_0, β_1, denoted by b_0 and b_1, should minimize the overall deviations, e.g.

$$L = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |Y_i - b_0 - b_1 X_i|,$$

leading to the Least Absolute Deviation (LAD) Estimator. This is not our interest in this module (due to its complexity). Another approach to find "good" b_0 and b_1 is to minimize

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{Y_i - b_0 - b_1 X_i\}^2,$$

called the (ordinary) least squares estimation (LSE, or OLS).

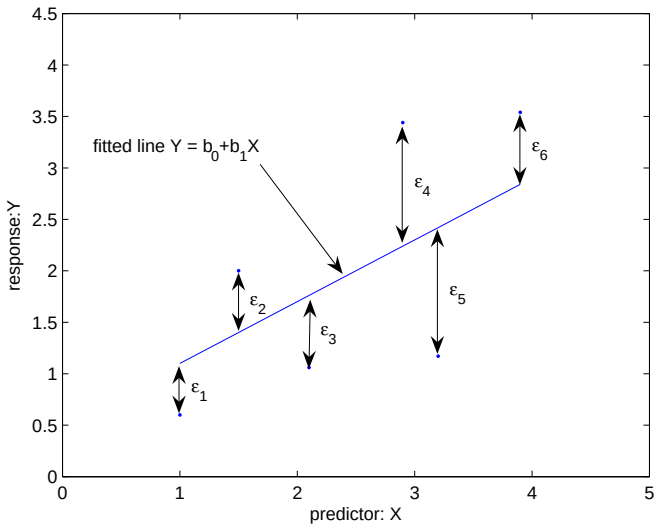


Figure 1: *An example of linear regression model (see the example below)*

By calculus, we have

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} \{Y_i - b_0 - b_1 X_i\},$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^{n} X_i \{Y_i - b_0 - b_1 X_i\}.$$

The solution of (b_0, b_1) that minimize Q is such that

$$-2\sum_{i=1}^{n}\{Y_i - b_0 - b_1 X_i\} = 0,$$
$$-2\sum_{i=1}^{n} X_i\{Y_i - b_0 - b_1 X_i\} = 0.$$

The least squares estimators b_0, b_1 are calculated by solving **normal equations**:

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$-2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

Finally, we have the LSE (least squares estimators)

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$
$$b_0 = \frac{1}{n}\{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\} = \bar{Y} - b_1\bar{X}.$$

The estimators are sometimes written as $\hat{\beta}_1$ and $\hat{\beta}_0$ respectively.

Terminology for the estimation

- The estimated/fitted model is

$$\hat{Y} = b_0 + b_1 X$$

(Note that we use $\hat{Y}$, to denote the predicted/fitted value of Y for a given X)

- The fitted values for the n observations are

$$\hat{Y}_i = b_0 + b_1 X_i, \qquad i = 1, ..., n$$

(for a new subject with $X = X'$, we also call the fitted value $\hat{Y}' = b_0 + b_1 X'$ predicted value)

- The fitted residuals for the n subjects are respectively

$$e_i = Y_i - \hat{Y}_i, \ i = 1, 2, ..., n.$$

Example 2.1 Data

Obs.	X	Y
1	1.0	0.60
2	1.5	2.00
3	2.1	1.06
4	2.9	3.44
5	3.2	1.17
6	3.9	3.54

By simple calculation,

$$\bar{X} = 2.4333, \quad \bar{Y} = 1.9683, \quad \sum_{i=1}^{6}(X_i - \bar{X})(Y_i - \bar{Y}) = 4.6143, \quad \sum_{i=1}^{6}(X_i - \bar{X})^2 = 5.9933$$

Thus,

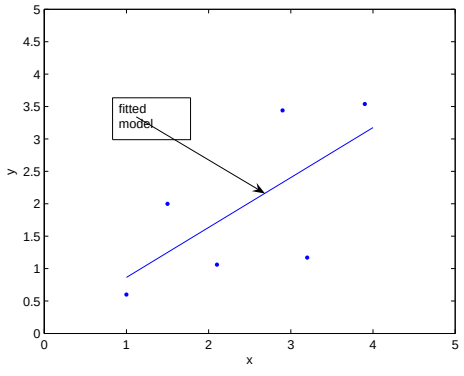
$$b_1 = 0.7699, \quad b_0 = 0.0949.$$

The estimated model is

$$\hat{Y} = 0.0949 + 0.7699X$$

Obs.	X	Y	fitted Y: $\hat{Y}$	residuals
1	1.0	0.60	0.8648	-0.2648
2	1.5	2.00	1.2497	0.7503
3	2.1	1.06	1.7117	-0.6517
4	2.9	3.44	2.3276	1.1124
5	3.2	1.17	2.5586	-1.3886
6	3.9	3.54	3.0975	0.4425

The model indicates that Y increase with X. As X increases one unit, Y increases 0.7699 unit.



Suppose we have a new subject with $X = 3$, then our prediction of Y is $\hat{Y} = 0.0949 + 0.7699 * 3 = 2.4046$.