Part II

Linear Regression Modelling

2. Linear regression Modelling

In the previous section, we attempted to explain the variation in an observed response variable by fitting models with one or more factors.

Factors are **discrete** variables taking different levels; in this section we will now utilize **continuous** variables that can similarly explain variation in an observed response.

2.1 Simple Linear Regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight-line).

Note: variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Example: Pharmacokinetic Model.

If a dose of drug is taken at time $x = 0$, the amount (concentration) of drug still in the bloodstream at time x is often well-modelled by a simple equation. Let

- D denote the amount of drug taken at $x = 0$
- x time
- $y^\star$ is the amount (concentration per unit volume) in the bloodstream.

Then

$$y^\star = \frac{D}{V} \exp\{-\lambda x\}$$

where

- λ is the elimination rate
- V is the volume of bloodstream.

Example: Pharmacokinetic Model (continued).

Taking logs of both sides, setting $y = \log y^\star$, then

$$y = -\lambda x + \log(D/V) = -\lambda x + (\log D - \log V)$$

that is, $y = ax + b$ where

- $a = -\lambda$
- $b = (\log D - \log V)$

Such models are **DETERMINISTIC**, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

2.1.1 Probabilistic Models

In a **probabilistic** model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where $ERROR$ is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

Note that we do not treat x and y exchangeably; x is a fixed observed variable that is measured *without error*, whereas y is an observed variable that is measured *with random error*.

We model the variation in y as a function of x. We observe pairs $(x_i, y_i), i = 1, \ldots, n$.

A Basic Probabilistic Model

Terminology:

- y - *Dependent variable* or *independent variable*
- x - *Independent variable*, or *predictor*, or *covariate*

The model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random error term, a random variable with mean zero and finite variance ($E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of y.

- β_0 - *Intercept* parameter
- β_1 - *Slope* parameter

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1. We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.