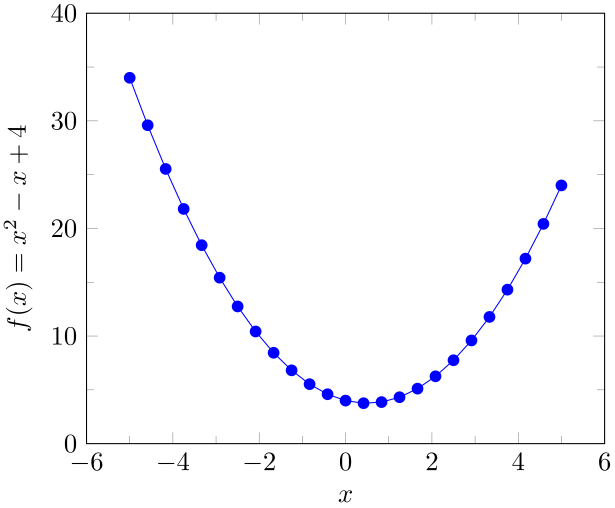
Linear Regression

TDung

1 Introduction



Linear regression is used to predict the value of a variable based on the value of another variable. The variable which one to predict is called the dependent variable (or sometimes, it can be called *the outcome variable*). The variable using to predict the other variable's value is called the independent variable (or sometimes, it can be called *the predictor variable*).

For example, you can use linear regression to predict population in any country based on population of previous years, predict a person's height depends on weight...

In case, when we have more than one independent variables, you need to use multiple regression.

2 Ordinary Least Squares Estimation

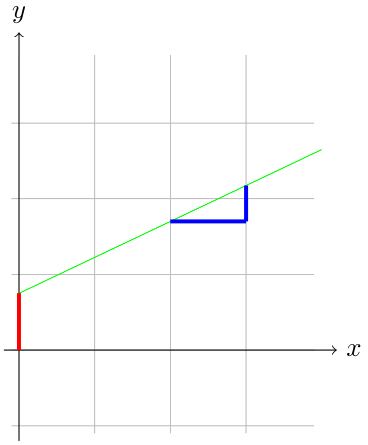
The *simple linear regression* model consist of the *mean function* and the *variance function.*

$$E = (Y|X = x) = \beta_0 + \beta_1 x \tag{1}$$

$$Var(Y|X = x) = \sigma^2 \tag{2}$$

The parameters in the mean function are the intercept β_0 and the slope β_1. The intercept is the value of $E(Y|X = x)$ when x equals zero and the slope is rate of change in $E(Y|X = x)$ for a unit change in X. The parameters are usually unknown and must be estimated by using data.

The *variance function* in (2) is assumed to be constant, with a positive value σ^2 that is usually unknown. When $\sigma^2 > 0$, the observed value of the i^{th} response y_i will typically not equal its expected value $E(Y|X = x_i)$.



Equation of a straight line $E(Y|X = x) = \beta_0 + \beta_1 x$
Which the red line is β_0 and the blue line is β_1.

We have many methods have been suggested for obtaining estimates of parameters in a model. And the method we discuss here is called *ordinary least squares* (or ols), in which parameter estimates are chosen to minimize a quantity called the residual sum of squares. A formal development of the least squares estimates is given in Appendix A.3.

Parameters are unknown quantities that characterize a model. Estimates of parameters are computable functions of data and are therefore statistics. To keep this distinction clear, parameters are denoted by Greek letters like α, β, γ and σ, and estimates of parameters are denoted by putting a "hat" over the corresponding Greek letter. For example with $\hat{\beta}_i$, read "beta i^{th} hat" is the estimator of β_i, and $\hat{\sigma}^2$ is the estimator of σ^2. The *fitted value* for case i is given by $\hat{E}(Y|X = x_i)$, for which we use the shorthand notation y,

$$\hat{y_i} = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3}$$

All least squares computations for simple regression depend only on averages, sums of squares and sums of cross-products. Definitions of the quantities used are given in Table 2.1. Sums of squares and cross-products have been centered by subtracting the average from each of the values before squaring or taking cross-products. Appropriate alternative formulas for computing the corrected sums of squares and cross products from uncorrected sums of squares and crossproducts that are often given in elementary textbooks are useful for mathematical proofs, but they can be highly inaccurate when used on a computer and should be avoided.

Table 2.1: Definitions of Symbols

Quantity	Definition	Description
$\overline{x}$	$\sum \left(\frac{x_i}{n}\right)$	Average of x
$\overline{y}$	$\sum \left(\frac{y_i}{n}\right)$	Average of y
SXX	$\sum (x_i - \overline{x})^2$	Sum of squares for the x
SYY	$\sum (y_i - \overline{y})^2$	Sum of squares for the y

The symbol $\sum$ means to add over all the values or pairs of values in the data.

This inconsistency is deliberate since in many regression situations, these statistics are not estimates of population parameters.

To illustrate computations, we will use Forbes' data, page 4, for which $n = 17$. The data are given in Table 2.2. In our analysis of these data, the response will be taken to be $Lpres = 100xlog10(Pressure)$, and the predictor is Temp. We have used the values for these variables shown in Table 2.2 to do the computations.

Table 2.2: Forbes' 1857 Data on Boiling Point and Barometric Pressure for 17 Locations in the Alps and Scotland

Case Number	Temp (F)	Pressure (Inches Hg)	Lpres = 100 x log(Pressure)

The symbol $\sum$ means to add over all the values or pairs of values in the data.