

Simple Linear Regression

Relationships among Variables

Regression Analysis

1. Dependent variable (y) is continuous
2. Independent variables (x_i)—continuous or categorical
3. Direction and strength of relationship

Example:

1. Relate systolic blood pressure levels to a measure of obesity.
2. Relate cost of a house to the square footage of the house.

Applications

1. Characterize relationship
2. Obtain quantitative formula for y as a function of x_1, x_2, \dots, x_k
3. Control for other variables
4. Select subset of variables
5. Is relationship linear

y = dependent variable

x = independent variable

n = number of subjects

(x_1, y_1)

...

(x_n, y_n)

Example:

Age	Height
9	48
10	51
11	53
12	55
13	55

Strategies:

- 1. Try straight line first**
- 2. Does it explain a significant amount of variability**

Let $x = \text{age}$ and $y = \text{height}$. We can propose a relationship of the following form:

$$E(y|x) = \alpha + \beta x$$

For a given age, the average height $E(y|x)$ is $\alpha + \beta x$

The line

$y = \alpha + \beta x$ is called the regression line

α is called the intercept

β is called the slope

Generally, the relationship is not exact for every child. We introduce an error term, e , which represents the variance of height for children of the same age.

$$y = \alpha + \beta x + e$$

where e is normally distributed with mean 0 and variance σ^2 . Therefore, for a child of age x , the corresponding height will be normally distributed with mean $\alpha + \beta x$ and variance σ^2 . The latter is a measure of spread for a particular age. Note: If σ^2 were 0, every point would fall exactly on the regression line.

1. If $\beta > 0$, then as x increases, the expected value of y increases.
2. If $\beta < 0$, then as x increases, the expected value of y decreases.
3. If $\beta = 0$, there is no linear relationship between x and y .